

# The Pluralistic Nature of Emotion: Human and Machine Interpretations of Textual Emotional Content

Zayn Jameel Abbas  
School of Computer Science  
University of Guelph  
Guelph, Ontario, Canada  
abbasr@uoguelph.ca

Stacey D. Scott  
School of Computer Science  
University of Guelph  
Guelph, Ontario, Canada  
stacey.scott@uoguelph.ca

## Abstract

This paper explores the complex nature of emotion interpretation in text-based communication by comparing human and machine approaches to emotion detection. Human emotions, shaped by personal experiences and cultural backgrounds, reflect individuality, yet many detection systems overlook these nuances.

Through a three-part study involving human participants and advanced large language models (LLMs), the research shows that humans naturally embrace emotional ambiguity. Preliminary findings suggest that higher EI may correlate with recognising interpretive nuances rather than seeking consensus, though this relationship requires validation with larger samples.

This paper introduces innovative methodologies, such as circumplex-based scoring systems that acknowledge interpretive plurality. The findings suggest that emotion detection systems should complement human interpretation, enhancing human-AI collaboration in tasks requiring individual perspectives and contextual sensitivity.

## CCS Concepts

• **Human-centered computing** → **User studies**.

## Keywords

Emotion Detection, LLMs, HCI, AI

### ACM Reference Format:

Zayn Jameel Abbas and Stacey D. Scott. 2026. The Pluralistic Nature of Emotion: Human and Machine Interpretations of Textual Emotional Content. In *31st International Conference on Intelligent User Interfaces (IUI '26)*, March 23–26, 2026, Paphos, Cyprus. ACM, New York, NY, USA, 16 pages. <https://doi.org/10.1145/3742413.3789161>

## 1 Introduction

The digital revolution has fundamentally transformed how humans communicate, with text-based interactions now dominating our daily exchanges through social media platforms, messaging applications, and online forums [12, 20]. As billions of people worldwide increasingly rely on written communication for both personal and professional interactions, the ability to accurately interpret emotions in text has become critically important for maintaining meaningful human connections and preventing misunderstandings [6, 10].

Text lacks critical emotional cues such as tone, facial expression, and body language, all elements that humans typically rely on to interpret emotions [10]. This absence of non-verbal cues has profound implications for digital communication, often leading to misinterpretations that can escalate into conflicts and damaged relationships. The familiar refrain “That’s not how I meant it” has become quite common in digital exchanges [6, 10]. This challenge intensifies for individuals who already struggle with emotional interpretation, as text-based communication removes the very contextual cues that might otherwise aid their understanding [10, 11].

Recent advancements in artificial intelligence (AI), particularly large language models (LLMs), have transformed natural language processing (NLP) approaches [23]. This has resulted in systems that can detect and classify emotions in text into sophisticated single and multi-label taxonomies [13, 16]. Despite these capabilities, a critical gap persists between computational emotion recognition and genuine human emotional understanding [6, 7, 13]. While current methods can identify multiple emotions, they often lack the nuanced and context-aware interpretation processes that define human emotional intelligence (EI) [7, 13]. Current approaches treat emotional complexity as a technical problem rather than recognizing how individual differences in EI, cultural background, and personal experiences fundamentally shape interpretation [13].

## 1.1 Research Problem: The Pluralistic Challenge

This research reveals a fundamental paradox that challenges the assumptions underlying emotion detection systems. These systems assume consensus on emotional interpretation exists and should be achieved, yet human emotional interpretation is inherently pluralistic. Our study of 25 participants demonstrates that emotional understanding fundamentally operates through pluralistic rather than singular classification, with 78% of all emotional interpretations involving multiple emotions.

More strikingly, participants with higher EI showed a tendency toward lower agreement with single-label dataset annotations, though this pattern did not reach statistical significance ( $r=-0.176$ ,  $p=0.40$ ,  $n=25$ ). This suggestive pattern, if validated with larger samples, emerged not from poor emotional understanding but from recognizing complexity that simplified labels cannot capture.

We term this preliminary finding the “Emotional Intelligence Paradox,” which may expose a critical flaw in how we design and evaluate emotion detection systems. By embracing rather than resolving emotional ambiguity, emotionally sophisticated participants appeared to perform worse, not because they misunderstood



This work is licensed under a Creative Commons Attribution 4.0 International License. *IUI '26, Paphos, Cyprus*

© 2026 Copyright held by the owner/author(s).  
ACM ISBN 979-8-4007-1984-4/26/03  
<https://doi.org/10.1145/3742413.3789161>

emotions, but because evaluation metrics penalize complexity. Traditional accuracy measures, designed to reward consensus on single labels, systematically penalize the very sophistication they should capture.

## 1.2 Research Questions

This paper addresses two fundamental questions that challenge conventional approaches to emotion detection:

*1.2.1 RQ1: How do humans naturally interpret emotions in text-based communication?* This question examines not only which emotions participants identify, but the reasoning processes underlying their interpretations and how individual differences in EI, demographics, and social media experience influence interpretation patterns.

*1.2.2 RQ2: How do human interpretations align with state-of-the-art LLMs when applied in zero-shot scenarios?* Rather than assuming alignment should be achieved, this question investigates where humans and AI converge or diverge, what these patterns reveal about complementary strengths, and which emotion categories present universal challenges that transcend the human-AI divide.

To address these questions, we conducted a two-phase investigation: human interpretation study (n=25 participants), LLM replication study using identical prompts (12 curated texts), which included an extended LLM evaluation across diverse datasets (570 prompts). This systematic comparison reveals that both humans and AI naturally adopt pluralistic approaches, but with distinctly complementary strengths. While AI excels at comprehensive emotion recognition—with Gemini achieving 67.1% multi-label alignment—humans demonstrate superior confidence calibration and contextual prioritization, reaching 54.8% position-weighted performance. These divergent capabilities suggest that optimal emotion-aware systems should leverage both human and AI strengths rather than pursuing perfect alignment on oversimplified classification tasks.

## 1.3 Contributions

This work makes several key contributions that advance emotion-aware interface design:

- (1) **Empirical evidence for pluralistic interpretation:** We provide systematic evidence that pluralistic emotion interpretation is widespread among humans (78% of responses), with preliminary findings suggesting this tendency may be stronger among emotionally sophisticated individuals, though this latter claim requires validation with larger samples. Our findings reveal that recognizing emotional complexity represents sophisticated understanding rather than interpretive failure.
- (2) **Novel evaluation frameworks:** We introduce circumplex-based scoring and position-weighted metrics that honor interpretive sophistication rather than penalizing it. These methodologies provide more psychologically grounded approaches to assessing emotion detection that acknowledge degrees of emotional similarity rather than enforcing binary correct-incorrect classifications.

- (3) **Discovery of complementary human-AI strengths:** We reveal that humans and AI demonstrate distinct but complementary capabilities—AI for comprehensive emotion recognition, humans for confidence calibration and contextual judgment. This finding suggests new interaction paradigms where AI presents multiple plausible interpretations while humans provide the contextual understanding needed to prioritize them effectively.

These findings challenge the consensus-seeking paradigm that has dominated emotion detection research and point toward human-AI collaborative interfaces that leverage pluralistic understanding. Rather than pursuing ever-higher accuracy on oversimplified benchmarks, we advocate for developing emotion-aware systems that embrace the beautiful complexity of human emotional communication while leveraging AI's pattern recognition capabilities.

## 2 Background & Related Work

To understand emotion interpretation, we must first examine how emotions have been conceptualized and measured in both psychological theory and computational systems. Two dominant frameworks have shaped emotion detection research: categorical and dimensional models.

### 2.1 Categorical Emotion Models

Categorical approaches, exemplified by Ekman's basic emotions theory [9], classify emotions into discrete categories such as joy, sadness, anger, fear, disgust, and surprise. The clear, discrete labels make this approach particularly appealing for practical applications [16]. Early lexicon-based approaches created dictionaries that mapped specific words to emotion categories, while modern systems employ transformer architectures like BERT that demonstrate sophisticated pattern recognition capabilities [8].

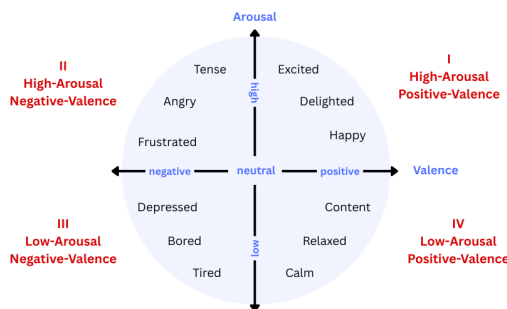
Despite their effectiveness, categorical models face fundamental limitations that reflect the inherent complexity of human emotional experience. The primary challenge stems from emotions frequently manifesting as blended states that resist discrete categorisation [10]. Research demonstrates that emotions expressed by writers are significantly more complex than those perceived by readers [1], suggesting categorical models systematically underestimate emotional nuance. This writer-reader disconnect reveals a critical gap in current emotion detection paradigms. Most systems are designed and evaluated based on reader interpretations of emotional content, yet the original emotional intent of writers may be fundamentally different and more complex than what readers perceive.

The training process for these systems reveals another significant limitation. When researchers collect datasets where multiple human annotators label texts with emotion categories, standard practice resolves disagreements between annotators by selecting majority votes or having experts make final decisions [7, 17]. This process effectively eliminates natural variation in human emotional interpretation, training models to converge on single "correct" emotional labels rather than recognising that multiple valid interpretations may coexist [3]. While these models achieve high accuracy against their training data, they fundamentally misrepresent the pluralistic nature of human emotional understanding by design.

## 2.2 Dimensional Emotion Models

The dimensional model of emotion represents an alternative theoretical framework that conceptualises emotions as points within a continuous multidimensional space rather than as discrete categories (often as 2D or 3D) [19]. This approach emerged from psychometric studies that identified underlying dimensions across diverse emotional experiences, revealing that emotions can be systematically organised along core affective dimensions (valence, arousal, and dominance).

Most often, the dimensional model characterises emotions using valence (pleasantness-unpleasantness) and arousal (level of activation) [18]. Russell’s circumplex model arranges emotions in a circular pattern along these two dimensions, where psychologically similar emotions appear close to each other on the circle, while opposing emotions appear on opposite sides [19]. This spatial arrangement reflects the underlying psychological reality that certain emotional transitions are more natural and likely than others. People more commonly shift between similar emotions like contentment to joy than between opposing emotions like joy directly to despair.



**Figure 1: Russell’s Circumplex Model of Affect showing emotions arranged by valence and arousal dimensions [14]**

The mathematical elegance of the circumplex model lies in its ability to quantify emotional relationships through geometric distance. Emotions located close to each other on the circle share similar valence and arousal characteristics, while emotions separated by greater distances represent more psychologically distinct states [18]. This quantitative framework enables precise measurement of emotional similarity using Euclidean distance calculations in the 2D valence-arousal space, providing an objective method for assessing the psychological plausibility of emotional predictions or transitions.

The circumplex model is particularly significant for this research because it provides the theoretical foundation for our alignment scoring methodology. Rather than treating emotion classifications as binary correct-incorrect judgments, we employ Russell’s circumplex framework to calculate psychological proximity between

predicted and ground truth emotions using valence and arousal coordinates. This approach recognises that emotional “errors” are not equivalent. Predicting sadness when the ground truth is grief represents a more psychologically plausible interpretation than predicting joy. This is because sadness and grief occupy nearby positions in the circumplex space while joy and grief are located on opposite sides.

Despite the complementary strengths of categorical and dimensional approaches, both share a critical limitation when implemented computationally. Their training processes typically require resolving annotator disagreements into single ground-truth values, whether categorical labels or dimensional coordinates [4, 22]. This practice fundamentally misrepresents the pluralistic nature of authentic human emotional understanding, as it systematically eliminates the natural variation that this research demonstrates as fundamental to emotional sophistication.

## 2.3 The Consensus Assumption in Emotion Detection

Current emotion detection systems, regardless of their theoretical foundation, operate under a problematic assumption: that consensus on emotional interpretation can and should be achieved. This manifests in three critical ways that collectively constrain how we conceptualise and evaluate emotional understanding.

Dataset annotation practices systematically eliminate interpretive diversity. Standard protocols resolve annotator disagreements through majority voting or expert adjudication [7, 17], effectively training models that emotions have single “correct” interpretations. While multi-label frameworks like GoEmotions [7] acknowledge that texts may express multiple emotions, they still produce fixed label sets that all annotators must converge upon. The result is that individual differences in emotional perception, differences that may reflect legitimate variation in how people experience and interpret emotions, are treated as noise to be averaged out rather than signal to be preserved.

Evaluation metrics compound this problem by rewarding consensus over sophistication. Traditional accuracy measures treat any divergence from ground truth as error, failing to distinguish between psychologically plausible alternatives and implausible confusions. For instance, these metrics penalise predicting “disappointment” when ground truth is “sadness” exactly as much as predicting “joy”, despite the former representing a subtle distinction within negative emotions and the latter representing a fundamental misunderstanding. This creates a fundamental tension: emotionally sophisticated interpreters who recognise complexity and ambiguity may perform worse on standard metrics despite demonstrating superior emotional understanding [5].

Individual differences remain largely unexamined despite their clear relevance to emotional interpretation. While EI research demonstrates that individuals vary substantially in their emotion perception and interpretation capabilities [21], emotion detection systems typically assume universal interpretive standards. The relationship between EI and emotion detection performance has received minimal attention, leaving a critical gap in understanding how interpretive sophistication manifests in text-based contexts. If emotionally intelligent individuals naturally recognise greater complexity in

emotional expression, current evaluation frameworks may systematically mischaracterise their sophistication as poor performance.

These limitations suggest that the field’s pursuit of ever-higher accuracy on consensus-based benchmarks may be optimising for the wrong objective. Rather than developing systems that capture the rich complexity of authentic human emotional understanding, we may be producing systems that excel at replicating simplified annotations whilst failing to recognise the legitimate diversity in how different individuals interpret emotional content.

### 2.4 Toward Pluralistic Approaches

Emerging research suggests that emotional interpretation may be inherently pluralistic rather than consensus-seeking. Alvarez-Gonzalez et al. [1] demonstrated that writer-expressed emotions are significantly more complex than reader-perceived emotions, revealing a fundamental gap between expression and interpretation that single-label systems cannot capture. Barrett’s work on emotion categorisation [2] challenges the notion of discrete emotional categories, arguing that emotion perception involves constructed interpretations shaped by individual experience and context rather than universal recognition of objective emotional states.

Recent advances in LLMs present an intriguing opportunity to examine this pluralistic nature. When applied in zero-shot scenarios, where models receive natural language instructions without emotion, specific training examples, LLMs demonstrate sophisticated language understanding that may reflect more natural approaches to emotional interpretation [15]. However, their emotional interpretation capabilities and alignment with human pluralistic understanding remain largely unexplored. Do these models, trained on vast corpora of human emotional expression, naturally adopt pluralistic approaches to emotion recognition? When they make errors, are those errors psychologically plausible or do they reveal fundamental gaps in emotional understanding?

This gap motivates our investigation. Rather than pursuing consensus-based accuracy, we examine how humans naturally interpret emotions in text, what individual differences influence these interpretations, and how human approaches compare to state-of-the-art AI systems. By employing evaluation frameworks that honour rather than penalise interpretive diversity, including circumplex-based scoring that measures psychological plausibility and position-weighted metrics that acknowledge confidence calibration, we aim to reveal patterns that inform more human-centred emotion-aware interface design. The question is not whether we can train systems to match human consensus on simplified labels, but whether we can develop approaches that capture the legitimate complexity that characterises authentic human emotional understanding.

## 3 Study Design & Methodology

We employed a mixed-methods approach combining quantitative emotion labelling with qualitative analysis of interpretive strategies. Our investigation consisted of two interconnected phases designed to examine human emotion interpretation patterns and compare them with state-of-the-art AI systems, as illustrated in Figure 2.

Phase 1 focused on understanding how humans naturally interpret emotions in text through a user study where 25 participants labelled emotions in curated Reddit comments from the GoEmotions dataset [7]. Participants selected from 27 predetermined emotion categories, completed the Schutte Self-Report Emotional Intelligence Test (SSEIT) [21], and participated in semi-structured interviews about their interpretive strategies.

Phase 2 evaluated three state-of-the-art LLMs (ChatGPT-4, Gemini 2.5 Flash, Cohere Command R+) in zero-shot scenarios to compare machine approaches with human interpretation. This phase proceeded in two steps: first, models were evaluated on the same 12 prompts used in the human study, enabling direct human-AI comparison on identical text samples. Second, models were tested on an extended set of 570 prompts from three datasets (GoEmotions, DailyDialog, ISEAR) to assess robustness and scalability beyond the initial replication.

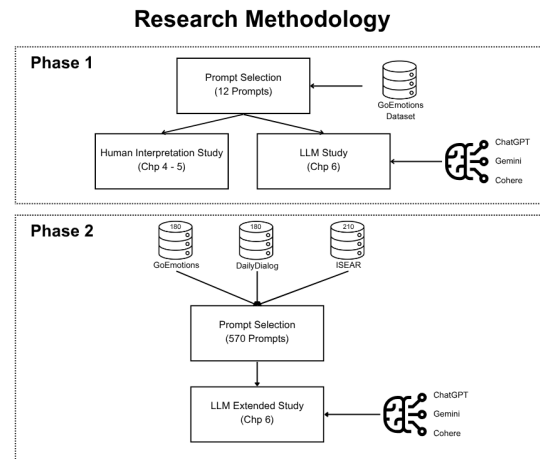


Figure 2: Research Design Flow Chart

### 3.1 Data Sources

The study utilized the GoEmotions dataset, containing 58,000 Reddit comments annotated with 27 fine-grained emotions organised hierarchically under six parent categories derived from Ekman’s basic emotions: anger, disgust, fear, joy, sadness, and surprise [7]. This hierarchical structure enables analysis at multiple levels of granularity, from broad emotional categories to specific emotions like admiration, embarrassment, or relief.

For the human study and the initial step of Phase 2, we selected 12 text samples through systematic review emphasizing emotional clarity, accessible language, and ethical suitability. The selection process involved several steps designed to minimize bias and maximize diversity of emotional expressions. Text samples were reviewed to ensure they conveyed clear emotional content whilst avoiding neutral or ambiguous statements. Ethical screening excluded explicit language, hate speech, and controversial topics to ensure participant comfort and meet research standards. The selection strategy specifically targeted two representative examples from each of Ekman’s six basic emotion categories as represented in the

**Table 1: Emotion coverage across datasets. GoEmotions provides the most granular taxonomy; overlap enables cross-dataset validation.**

Emotion Family	GoEmotions	DailyDialog	ISEAR
Anger	3 emotions	✓	✓
Disgust	1 emotion	✓	✓
Fear	2 emotions	✓	✓
Joy	12 emotions	✓	✓
Sadness	5 emotions	✓	✓
Surprise	4 emotions	✓	✓
Guilt/Shame	—	—	✓
Total	27	7	7

GoEmotions hierarchy, ensuring comprehensive coverage of the emotional spectrum whilst maintaining a manageable number of prompts for participants.

The second step of Phase 2 incorporated 570 prompts distributed across GoEmotions, DailyDialog, and ISEAR datasets. This extended evaluation deliberately balanced single-labelled and multi-labelled samples (15 of each per parent emotion category), enabling assessment of how models handle both focused emotional expressions and complex multi-emotion scenarios. The combination of three datasets ensures robust evaluation across different communication contexts: social media comments (GoEmotions), conversational dialogues (DailyDialog), and personal emotional narratives (ISEAR). Table 1 summarizes the emotion taxonomies across datasets.

### 3.2 Evaluation Methodology

Traditional emotion detection evaluation treats predictions as binary classifications (either correct or incorrect) failing to capture the psychological plausibility of different interpretive choices. We introduce two complementary metrics that honour rather than penalise interpretive sophistication.

The circumplex-based alignment scoring leverages Russell’s dimensional model [19] to measure psychological distance between predictions and ground truth. Rather than treating emotion classifications as binary judgements, we employ Russell’s circumplex framework to calculate psychological proximity between predicted and ground truth emotions using valence and arousal coordinates from validated norms [24]. This approach recognises that emotions exist within a continuous dimensional space, allowing for more nuanced assessment of interpretive alignment.

Alignment scores are calculated as follows:

$$S_{\text{circumplex}} = \begin{cases} 1.0 & \text{if exact match} \\ 0.8 \cdot e^{-d} + 0.2 & \text{if same parent category} \\ 0.6 \cdot e^{-d} & \text{if different parent} \end{cases} \quad (1)$$

where  $d$  represents Euclidean distance in valence-arousal space. This graduated scoring system acknowledges that emotional “errors” are not equivalent. Predicting disappointment when the ground truth is sadness ( $d = 0.11$ ) represents a more psychologically plausible interpretation than predicting joy when the ground truth is

sadness ( $d = 1.55$ ), even though both predictions are technically incorrect in traditional binary classification schemes.

The circumplex-based scoring system builds on Russell’s validated dimensional model of affect [19], which has demonstrated robust cross-cultural validity in representing emotions as points in a continuous valence-arousal space [18]. The emotion coordinates used in our calculations derive from Warriner et al.’s [24] norms for 13,915 English lemmas, collected from large-scale human ratings and widely validated in affective computing research.

Our scoring formula incorporates three key design decisions: Exponential Distance Decay, Categorical Bonuses, and Distance Weighting Coefficients.

- *Exponential Distance Decay* The  $e^{-d}$  term reflects psychological evidence that emotional similarity decays non-linearly with dimensional distance [18, 19]. Emotions that are close in circumplex space (small  $d$ ) are perceived as highly similar, while those separated by moderate distances show rapidly decreasing similarity.
- *Categorical Bonuses* The differential weighting for same-family ( $0.8 \times e^{-d} + 0.2$ ) versus cross-family ( $0.6 \times e^{-d}$ ) predictions acknowledges that emotions share both dimensional properties and categorical relationships [9, 10]. The baseline 0.2 bonus for same-family emotions reflects this categorical component.
- *Distance Weighting Coefficients* The choice of 0.8 and 0.6 as primary coefficients was determined through iterative testing to create appropriate score separation. These specific values have not been empirically validated against.

The multi-label score provides an unweighted assessment of overall prediction quality, whilst the position-weighted score incorporates confidence ranking to reflect that primary predictions carry more weight than secondary alternatives in practical applications. The position weights [1.0, 0.75, 0.5, 0.25] implement a linear decay in importance across ranked predictions. The top prediction reflects the interpreter’s highest-confidence judgment (weight = 1.0), the second represents a plausible alternative (weight = 0.75), while third and fourth predictions capture additional possibilities with decreasing confidence (weights = 0.5, 0.25). Alternative schemes (e.g., exponential decay) could be justified but would require empirical validation. We selected linear decay for interpretability and as a conservative choice that gives meaningful but declining weight to lower-ranked predictions.

Additionally, K-n accuracy measures whether ground truth appears within the top-n predictions, acknowledging legitimate interpretive uncertainty. K-1 represents exact top-prediction accuracy; K-4 accepts the ground truth appearing among the top four choices. The progression from K-1 to K-4 scores reveals whether systems capture psychologically reasonable alternatives even when missing the exact ground truth as their primary prediction.

This dual framework enables separate assessment of the breadth of emotional recognition (multi-label score) and confidence calibration (position-weighted score), providing insights into complementary aspects of interpretive capability. Figure 3 illustrates these scoring calculations for an example prompt where ground truth is sadness.

Position-Weighted Circumplex Scoring Example			
<b>Scenario:</b> Ground Truth = "sadness" (valence = -0.70, arousal = -0.30)			
<b>Model Predictions:</b> ["disappointment", "grief", "anger", "fear"]			
<b>Step 1: Calculate Individual Circumplex Scores</b>			
Prediction	Coordinates	Calculation	
Disappointment	(0.60, -0.25)	Dist = $\sqrt{(-0.70 - 0.60)^2 + (-0.30 - (-0.25))^2}$ = $\sqrt{0.01 + 0.0025} = 0.11$ Score = $0.8 \times e^{-0.11} + 0.2 = 0.92$	=
Grief	(-0.80, -0.45)	Dist = 0.18, Same parent Score = $0.8 \times e^{-0.18} + 0.2 = 0.87$	
Anger	(-0.62, 0.38)	Dist = 0.68, Different parent Score = $0.6 \times e^{-0.68} = 0.31$	
Fear	(-0.60, 0.55)	Dist = 0.86, Different parent Score = $0.6 \times e^{-0.86} = 0.25$	
<b>Step 2: Multi-Label Score</b> = $(0.92 + 0.87 + 0.31 + 0.25) \div 4 = 0.59$ (59%)			
<b>Step 3: Apply Position Weights</b>			
Position	Weight	Score	Weighted
1st (disappointment)	1.0	0.92	0.92
2nd (grief)	0.75	0.87	0.65
3rd (anger)	0.50	0.31	0.16
4th (fear)	0.25	0.25	0.06
<b>Step 4: Position-Weighted Score</b> = $(0.92 + 0.65 + 0.16 + 0.06) \div 4 = 0.45$ (45%)			

**Figure 3: Circumplex-based scoring example for ground truth "sadness" (valence=-0.70, arousal=-0.30). Predictions: [disappointment, grief, anger, fear]. Disappointment (d=0.11, same parent) scores 0.917. Grief (d=0.18, same parent) scores 0.868. Anger (d=0.68, different parent) scores 0.304. Fear (d=0.86, different parent) scores 0.254. Multi-label: 0.586. Position-weighted: 0.446.**

The circumplex-based scoring system represents a methodological contribution that addresses limitations of binary classification but itself requires empirical validation. While grounded in established psychological theory (Russell's circumplex model [19], validated coordinates from Warriner et al. [24]), our specific implementation choices, particularly the exponential weighting parameters and position weights, were developed through iterative refinement rather than formal validation studies. We propose this methodology as a theoretically-motivated improvement over binary accuracy that better captures psychological plausibility of predictions. However, alternative implementations (different weighting schemes, distance metrics, or position decay functions) could be equally or more valid. Future work should validate these choices through: comparison with expert psychological judgments of interpretive quality, sensitivity analyses examining how findings change under alternative parameterizations, and criterion validation against independent measures (e.g., communication effectiveness, reader comprehension). The core findings of this work, that humans naturally embrace pluralistic interpretation (78% multi-label responses) and that this behavior converges with LLM approaches, remain evident regardless of specific scoring parameters.

### 3.3 Study Procedures

Twenty-five English-speaking participants were recruited from a university community through poster advertisements distributed across campus locations. The final sample included 14 females, 10 males, and one participant who identified as demigirl, with ages ranging from 18 to 64 (84% aged 18-24). Most participants reported daily social media usage (88%), with Instagram being the most popular platform (88%), followed by TikTok (56%) and Snapchat (44%). The majority identified primarily as content consumers (80%) rather than active posters.

Prior to scheduled sessions, participants completed the 33-item SSEIT [21] using Qualtrics, which took approximately 10-15 minutes. The SSEIT provides a validated measure of emotional intelligence across four domains: emotion perception, understanding, regulation, and expression. Participant scores ranged from 105 to 157 ( $M=132.6$ ,  $SD=14.8$ ), notably above the normative population mean of 124, suggesting above-average emotional intelligence in our sample.

Study sessions lasted 30-45 minutes and were conducted either remotely via Zoom (60%) or in-person in a university laboratory (40%), providing participants flexibility to choose their preferred modality. Each session followed a structured protocol designed to capture both quantitative emotion labelling data and qualitative insights into interpretive reasoning:

- (1) Participants reviewed the GoEmotions taxonomy, familiarising themselves with the 27 emotion labels organised under six parent categories
- (2) Participants completed the emotion labelling task, reviewing six randomly-assigned prompts from the full set of 12. The randomisation process operated at two levels: first, Qualtrics randomly selected two prompts from each parent emotion category for each participant, ensuring balanced coverage across all 12 prompts whilst maintaining manageable session length. Second, the order of prompt presentation was randomised to control for potential sequence effects. Emotion labels were presented in randomised order rather than alphabetically or by category, preventing ordering bias in selection patterns. Participants could select any number of emotions per text without restriction, enabling capture of pluralistic emotional interpretations.
- (3) Participants explained their reasoning aloud using a think-aloud protocol, articulating their interpretation process as they made selections. This provided real-time insights into the cognitive strategies employed during emotion detection.
- (4) Following the labelling task, participants engaged in a 10-15 minute semi-structured interview exploring their emotion interpretation approaches. The interview focused on understanding methods for identifying emotional cues in text, self-perceived accuracy in emotion detection across social media contexts, and experiences with emotional misinterpretation in digital communication.

Sessions were video and audio recorded with participant consent, and all participants received \$20 CAD Amazon gift card compensation for their time. The study received institutional ethics approval with key protections including comprehensive informed consent,

content screening for harmful material, flexible participation options, secure data anonymisation, and voluntary participation with withdrawal rights.

For Phase 2, the three LLMs received standardised prompts requesting their top four emotion predictions from the GoEmotions taxonomy. The prompting strategy differed between the two evaluation steps to reflect different research objectives. The initial 12-prompt replication employed zero-shot instructions with minimal guidance, mirroring the conditions humans faced and enabling direct comparison of interpretive approaches on identical text samples. The extended 570-prompt evaluation employed few-shot prompting with explicit role definition and five diverse examples, providing more structured guidance to assess performance at scale. This distinction allows examination of both inherent zero-shot emotional understanding (initial replication) and performance with explicit scaffolding (extended evaluation).

All LLM evaluations were conducted using Python scripts with API calls, and responses were recorded in a database for systematic analysis. Models received identical prompts within each evaluation step, ensuring consistent conditions for performance comparison.

### 3.4 Data Analysis

Quantitative analysis employed the circumplex-based and position-weighted scoring methodologies described in Section 3.2, calculating alignment scores for all human and LLM responses. We investigated the relationship between SSEIT scores and labelling patterns using Pearson correlation analysis, and compared performance across EI groups (Low: <110, Medium: 111-139, High: >140).

Qualitative analysis employed inductive thematic coding on interview transcripts to identify interpretive strategies and reasoning patterns. This methodology preserved the richness of participant insights whilst enabling systematic comparison across responses.

The analysis focused on three key dimensions: emotion detection strategies (how participants identified emotional cues), confidence patterns (how participants assessed their own accuracy), and multi-label selection reasoning (why participants chose multiple emotions rather than single classifications). This qualitative component provided essential context for understanding the quantitative performance patterns, revealing the sophisticated cognitive processes underlying human emotion interpretation.

## 4 Human Emotion Interpretation Findings

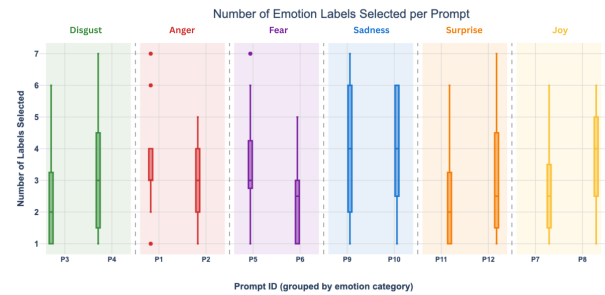
The human interpretation study revealed fundamental patterns in how individuals naturally approach emotion detection in text-based communication. Analysis of 25 participants' responses, combined with insights from semi-structured interviews, demonstrates that pluralistic interpretation represents the norm rather than the exception in authentic human emotional understanding.

### 4.1 Pluralistic Interpretation is the Norm

The analysis of participant responses revealed that humans naturally embrace emotional complexity rather than seeking singular classifications. Participants selected an average of 3.17 emotions per text (SD=1.8), with 78% of all responses involving multiple emotions. This multi-label behaviour emerged organically without any

instruction to select multiple emotions, yet only 22% of responses involved single-emotion classifications. This pattern provides strong empirical evidence that recognising emotional complexity is intrinsic to human emotional intelligence rather than a learned analytical strategy.

The distribution of label selections showed systematic patterns across emotion categories, as illustrated in Figure 4. Joy and surprise prompts generated the highest interpretive diversity, with participants frequently selecting 4–7 emotions per text. Conversely, disgust and fear prompts elicited more focused responses, typically generating 1–3 selections. This variation suggests that certain emotional expressions inherently contain overlapping dimensions or admit multiple valid interpretations, whilst others present clearer emotional signals that facilitate consensus among interpreters.



**Figure 4: Distribution of emotion labels selected per prompt, grouped by parent category. Box plots show median (line), quartiles (box), and range (whiskers). Joy and surprise prompts generated highest interpretive diversity (4–7 emotions), while disgust and fear prompts generated more focused responses (1–3 emotions).**

Overall alignment with GoEmotions ground truth labels reached 59% (SD=22.1%) using our circumplex-based scoring, with individual participants ranging from 44.4% to 81.4%. This moderate alignment level, combined with high multi-label frequency, indicates that human emotion interpretation often diverges from standardised dataset annotations not due to error, but due to sophisticated recognition of emotional complexity that single-label frameworks cannot capture. The substantial individual variation (coefficient of variation = 37.5%) demonstrates meaningful differences in interpretive approaches whilst maintaining structured patterns that our evaluation framework successfully captures.

### 4.2 Preliminary Evidence for an Emotional Intelligence Paradox

A suggestive but preliminary finding emerged in the relationship between EI and ground truth alignment, though there is a limited sample size. Participants completed the SSEIT prior to the labelling tasks, with scores ranging from 105 to 157 (M=132.6, SD=14.8), notably above the normative population mean of 124. When we examined alignment performance across EI groups (Low: <110, n=3; Medium: 111–139, n=12; High: >140, n=10), we found a weak

negative correlation between EI and ground truth alignment ( $r=-0.176$ , 95% CI  $[-0.52, 0.21]$ ,  $p=0.40$ ,  $n=25$ ). While this correlation did not reach statistical significance, potentially due to the small sample size and particularly the limited representation in the low EI group ( $n=3$ ), the direction of the relationship warrants further investigation with larger samples.

**Table 2: Performance by emotional intelligence group with confidence intervals**

EI Group	n	Mean Alignment	SD	95% CI
Low (<110)	3	0.649	0.143	[0.32, 0.98]
Medium (111-139)	12	0.573	0.076	[0.53, 0.62]
High (>140)	10	0.593	0.092	[0.53, 0.66]

*Note:* The large confidence interval for the Low EI group reflects the small sample size ( $n=3$ ), limiting the strength of conclusions drawn from this group.

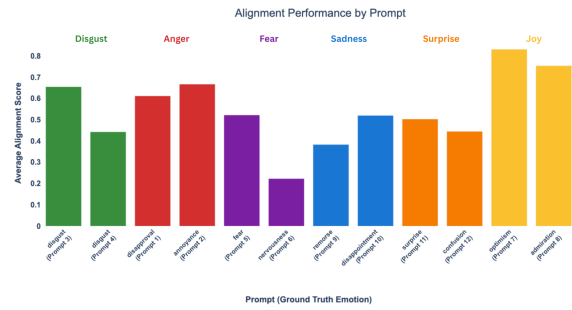
This pattern contradicts traditional assumptions that emotional expertise should improve accuracy on emotion detection tasks. However, qualitative analysis revealed the underlying mechanism. Participants with higher EI employed more sophisticated interpretive strategies, considering multiple contextual factors and emotional nuances. Rather than converging on single “correct” interpretations, emotionally intelligent participants embraced complexity.

Post-interview analysis reinforced this interpretation. High-EI participants consistently articulated nuanced reasoning: “I saw both disappointment and sadness—the disappointment is about the specific situation, but there’s an underlying sadness about the broader context” (P4, EI=141). In contrast, lower-EI participants more frequently sought single definitive answers: “I just picked the one that seemed most obvious” (P21, EI=107).

This “Emotional Intelligence Paradox” reveals a fundamental flaw in current evaluation frameworks. By rewarding consensus with single-label annotations, we systematically mischaracterise emotional sophistication as error. The participants who best understood emotional complexity performed worst on traditional metrics designed to reward alignment with oversimplified ground-truth labels. This finding suggests that the apparent “underperformance” of high-EI participants may actually reflect superior emotional understanding that recognises the inherent ambiguity in human emotional expression.

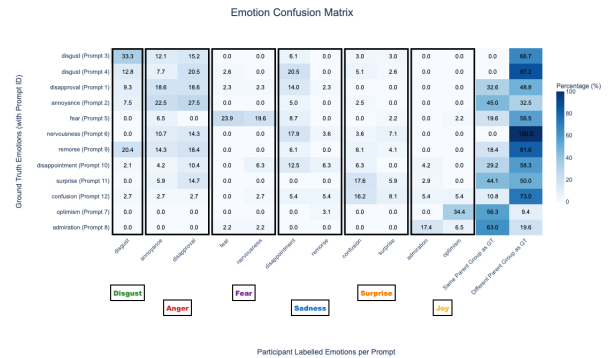
### 4.3 Emotion-Specific Challenges

Analysis by parent emotion category revealed systematic differences in interpretation difficulty, as shown in Figure 5. Individual prompt performance showed substantial variation, with standard deviations ranging from 0.18 (optimism) to 0.34 (nervousness), reflecting both the inherent difficulty of different emotions and individual differences in interpretive approaches. Joy emotions achieved the highest recognition rates (optimism: 87.4%, admiration: 76.6%), followed by anger emotions (annoyance: 74.3%, disapproval: 67.5%). These categories benefit from explicit linguistic markers and relatively consistent cultural expression patterns, making them more readily interpretable across participants.



**Figure 5: Alignment performance by individual prompts, grouped by parent emotion categories ( $n=25$  participants, with each participant rating 2 prompts per category via randomization). Joy and anger achieved highest recognition; fear (especially nervousness) proved most challenging. Bar heights represent mean alignment scores; variation across participants is reported in the text.**

Fear-related emotions presented the greatest challenges, with nervousness achieving only 31.3% alignment, the lowest of any emotion tested. Analysis of the confusion matrix (Figure 6) revealed that nervousness received zero exact matches, with participants instead frequently selecting disappointment (17.9%) and confusion (7.1%). This pattern indicates that low-arousal negative emotions with subtle manifestations resist clear interpretation in text-based contexts without additional contextual information.



**Figure 6: Emotion confusion matrix showing participant label selections for each ground truth emotion. Darker colors indicate higher selection frequency. Nervousness (row 6) shows 0% exact recognition and 100% cross-category confusion, primarily with disappointment and confusion.**

Surprise emotions demonstrated moderate but variable performance (confusion: 44.7%, surprise: 51.2%), with participants frequently confusing these labels with each other and with curiosity. The inherently context-dependent and ambiguous nature of surprise makes it challenging to interpret based on textual content

alone. Statistical analysis confirmed significant performance differences across emotion families (Kruskal-Wallis  $H=42.3$ ,  $p<0.001$ ), with post-hoc tests revealing the hierarchy: Joy > Anger > Disgust  $\approx$  Sadness  $\approx$  Surprise > Fear (all pairwise comparisons  $p<0.05$ ).

Within-category consistency varied substantially, revealing that emotion family membership alone does not guarantee similar interpretability. Whilst both disgust prompts achieved reasonable recognition (65.3%, 45.6%), fear prompts diverged dramatically (basic fear: 55.4% vs. nervousness: 31.3%). This highlights that subtle variations within emotion families can dramatically affect interpretability, suggesting that certain emotional subtypes require more contextual information or cultural knowledge for accurate interpretation.

#### 4.4 Interpretive Strategies

Semi-structured interviews revealed four primary strategies that participants employed when interpreting emotional content, demonstrating sophisticated reasoning processes that extend far beyond simple pattern matching. These strategies emerged from thematic analysis of participant responses to the question: “Were there any particular aspects of the text that you were looking for to indicate its intended emotion?”

The first strategy involved tonal interpretation through mental vocalisation, where participants mentally “heard” text in various emotional tones to gauge its emotional content. P1 explained the process as “taking them with different tones, depending on how it’s worded,” whilst P3 elaborated that she read text “in her own tone, the way the words run” or imagined “reading it from their perspective if I know them.” This reflects a complex cognitive process where participants simulate auditory delivery to compensate for the absence of vocal cues. P14 noted: “I don’t think about it consciously... I just assume it and I think I read it from their perspective and don’t read it from my own.” The effectiveness of this approach often depended on participants’ familiarity with the author, as mental vocalisation relies on assumptions about speaking style and emotional expression patterns.

A second strategy focused on punctuation analysis as prosodic cues, where participants consistently identified typographic elements as crucial emotional indicators. Punctuation provided primary cues, with P2 using “punctuation to get the pacing of the tone” and P4 noting that “if something had an exclamation mark, a more impacting tone” was indicated. P8 highlighted that capitalisation patterns indicated intensity, whilst P6 combined these elements with “strong” words like hate. Generational awareness shaped interpretation strategies, as P14 observed: “It changes... if someone wrote something in all capitals it’s like they’re screaming at you.” Emojis served as explicit emotional markers, with multiple participants indicating they make the text “more obvious.” This reveals how participants treat typographic elements as substitutes for intonation or tone, demonstrating sophisticated digital communication literacy.

The third strategy involved lexical analysis and register shifts, where participants systematically analysed word choice, emotional intensity, and language formality as emotional indicators. P1 looked for words that stuck out, whilst P4 noted that when a conversation gets more clinical, it becomes more impersonal as well. P11

focused on “really strong words like ‘hate’ or ‘love’ that show intense feelings,” and P25 identified “words that reflects a feeling like for example the words ‘Bloody Hell’ this indicated disgust, disappointment or anger.” Context and word combinations proved crucial for interpretation, with P5 describing “jumping to specific words” whilst noting that context is essential. P25 provided sophisticated analysis: “Or even words combined together mean certain things compared to alone. For the last prompt, when they say hey its not bad, the reason I put approval is because the words ‘Its not bad’, to me that shows the person writing agrees with them or it meets their criteria.” This attention to register and formality indicates participants understood emotional expression as fundamentally connected to communication style, requiring inferences about typical language patterns and deviations signalling emotional states.

The fourth strategy employed contextual inference and community validation, where participants drew upon broader contextual knowledge and situational understanding to interpret emotional content. P3 considered the reaction of both the author and the reader, whilst P4 noted the importance of “context and patterns in their behaviour and their lived experience” for familiar communicators. P7 provided platform-specific analysis: “It depends for each platform, for Instagram if its pictures I don’t connect with them unless its my friends... if its just text, if its sad or heartfelt you can connect with it.” Many participants actively sought external validation through community responses, with P13 describing: “as I go through I try to figure out is it negative positive then I go in the comments to see if people agree or not, what do they like what don’t they like etc.” P22 emphasised personal perspective: “I’m taking it as context and using my perspectives to enlighten me. The person who is saying vs the person who said it. Using my personal experience into it because I cannot disassociate.”

These strategies were not mutually exclusive, most participants employed multiple approaches simultaneously, with strategy selection adapting to text characteristics. High-EI participants demonstrated greater flexibility, seamlessly switching between strategies based on available cues, while lower-EI participants more rigidly applied single approaches. Critically, these sophisticated interpretive processes, particularly tonal simulation and contextual inference, require human cognitive capabilities that current AI systems lack. This suggests that human-AI collaboration in emotion detection should leverage these complementary strengths rather than seeking to replicate human interpretation through purely computational means.

## 5 Human-AI Comparison: Complementary Strengths

The second phase of this research evaluated three state-of-the-art LLMs (ChatGPT-4, Gemini 2.5 Flash, Cohere Command R+) on emotion detection tasks using zero-shot instructions, enabling systematic comparison with human interpretive approaches. This comparison reveals both striking convergences and critical divergences that illuminate the complementary nature of human and machine emotional understanding.

**Table 3: Human-AI performance comparison on identical 12-prompt task. Models excel at comprehensive recognition (multi-label); humans excel at confidence calibration (position-weighted).**

Metric	Humans	ChatGPT	Gemini	Cohere
Multi-label	59.0%	59.8%	67.1%	62.5%
Position-weighted	54.8%	41.3%	51.3%	44.2%
Avg. emotions/text	3.17	4.0	3.2	3.8
Variability (CV)	37.5%	38.2%	31.5%	35.7%

### 5.1 Convergent Multi-Label Behavior

The three LLMs were evaluated on the same 12 prompts from the human study using zero-shot instructions that mirrored the conditions participants faced. Remarkably, all three models independently adopted multi-label approaches, mirroring the human pluralistic tendency observed in Section 4.1.

Models selected an average of 3.2–4.0 emotions per prompt (ChatGPT: 4.0, Cohere: 3.8, Gemini: 3.2), closely matching human behaviour ( $M=3.17$ ,  $SD=1.8$ ). This convergence emerged despite fundamental differences in how humans and AI systems process language. Humans employ mental vocalisation, punctuation analysis, and contextual inference as described in Section 4.4, whilst LLMs rely on statistical patterns learned from vast text corpora. The spontaneous convergence on pluralistic interpretation suggests that recognising emotional complexity may be fundamental to sophisticated language understanding, regardless of whether that understanding is human or artificial.

Both humans and AI systems demonstrated similar variability in their responses. Human participants showed 37.5% coefficient of variation across prompts, whilst models clustered tightly around this benchmark (Gemini: 31.5%, Cohere: 35.7%, ChatGPT: 38.2%). However, this numerical similarity masks important qualitative differences. Whilst models maintained consistent response patterns, human variability reflected adaptive strategies tailored to different emotional contexts and individual interpretation styles. The models lacked the rich individual variation that characterises authentic human emotional understanding, following more predictable patterns than the nuanced variability observed in human responses.

The convergence on multi-label interpretation provides empirical support for pluralistic approaches to emotion detection, demonstrating that acknowledging multiple valid interpretations is not a concession to ambiguity but recognition of authentic emotional complexity in text-based communication.

### 5.2 Divergent Performance Patterns

Whilst humans and AI converged on multi-label behaviour, they diverged sharply in performance depending on the evaluation metric employed, as shown in Table 3. This divergence reveals complementary strengths that have important implications for emotion-aware interface design.

When examining multi-label scores that treat all selected emotions equally, AI models outperformed human participants. Gemini achieved 67.1%, Cohere 62.5%, and ChatGPT 59.8%, compared to humans' 59.0%. This suggests that AI systems excel at identifying

the full spectrum of plausible emotional interpretations within text, leveraging their exposure to vast corpora of emotional language during training.

However, Gemini's superior performance requires cautious interpretation. The evaluation used GoEmotions, developed by Google Research—the same organisation behind Gemini. This raises the possibility of data contamination or exposure to similar annotation patterns during training, potentially inflating Gemini's apparent advantage. Future work should employ datasets from diverse sources to ensure fair model comparisons and avoid such potential biases.

When applying position-weighted scoring that reflects the practical importance of prediction ranking, humans achieved superior performance (54.8%) compared to all AI models (Gemini: 51.3%, Cohere: 44.2%, ChatGPT: 41.3%). This hierarchy demonstrates that whilst AI models can identify multiple valid emotions, human interpretive sophistication and contextual understanding provide clear advantages in prioritising and ranking emotional interpretations according to their perceived salience.

The divergence between multi-label and position-weighted performance suggests distinct cognitive processes at work. AI models appear to employ statistical pattern matching that identifies multiple emotional associations without clear confidence hierarchies. Humans employ contextual reasoning, mental simulation, and metacognitive awareness to assess which interpretations are most salient, producing well-calibrated confidence rankings even when multiple emotions are plausible.

This complementary pattern (AI breadth versus human depth) suggests that optimal emotion-aware systems should leverage both capabilities rather than pursuing perfect human-AI alignment on simplified tasks. The question shifts from whether AI can match human performance on single metrics to how we can design systems that capitalise on the distinct strengths each brings to emotion interpretation.

### 5.3 Shared Interpretation Challenges

Analysis of emotion-specific performance revealed striking convergence between human and AI difficulties, suggesting that certain aspects of emotional interpretation present fundamental challenges that transcend the human-AI divide.

Both humans and AI achieved their highest performance on joy-related emotions (humans: approximately 82%, AI models: 51–66%). These emotions benefit from explicit positive linguistic markers and consistent cultural expression patterns that facilitate recognition across different processing mechanisms. Anger emotions also performed well (humans: approximately 71%, AI: 46–65%), likely due to clear negative language and explicit disapproval markers that provide strong textual cues.

Fear-related emotions, particularly nervousness, proved most difficult for all interpreters. Humans achieved only 31.3% alignment on nervousness, whilst AI models ranged from 22.8% (ChatGPT) to 34.6% (Cohere). This shared difficulty indicates that low-arousal negative emotions with subtle manifestations resist clear interpretation in text-based contexts regardless of whether the interpreter is human or artificial. Similarly, surprise emotions showed moderate performance across all interpreters (humans: approximately 48%,

**Table 4: Overall performance comparison across LLMs using position-weighted scoring on 570 prompts from three datasets. Cohere achieves highest position-weighted performance whilst Gemini excels at multi-label recognition and K-n metrics.**

Performance Metric	Cohere	Gemini	ChatGPT
Position Weighted	<b>34.3%</b>	33.1%	32.0%
Multi-Label Score	47.2%	<b>48.8%</b>	46.9%
K-1 Score	23.2%	23.5%	<b>23.9%</b>
K-2 Score	30.1%	<b>39.2%</b>	36.9%
K-3 Score	37.5%	<b>46.4%</b>	44.6%
K-4 Score	42.7%	<b>52.0%</b>	50.0%
Overall Ranking	1st	2nd	3rd

AI: 36–40%), reflecting surprise’s inherently context-dependent and ambiguous nature.

The convergence on challenging emotion categories indicates that these difficulties stem from inherent characteristics of emotional expression rather than interpreter-specific limitations. Low-arousal negative emotions like nervousness and disappointment with subtle manifestations resist clear interpretation in text-based contexts regardless of processing mechanism. Context-dependent emotions such as surprise and confusion require external information beyond immediate textual content, creating systematic challenges for both humans operating under time pressure and AI systems lacking world knowledge.

The one notable divergence emerged in sadness recognition, where Gemini (59.2%) approached human-level performance (53.7%), whilst ChatGPT struggled significantly (34.8%). This suggests that certain AI architectures may develop specialised capabilities for specific emotional patterns, though the mechanisms underlying these differences remain unclear and warrant further investigation.

The shared challenges point to domains where human-AI collaboration may be most beneficial. When interpreting low-arousal negative emotions or context-dependent expressions, neither humans nor AI demonstrate high confidence, suggesting value in presenting multiple interpretations for user consideration rather than forcing single classifications.

### 5.4 Extended Evaluation: Robustness and Scalability

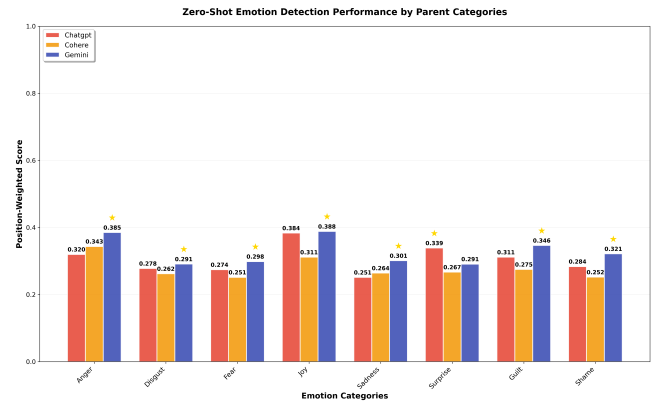
The second step of Phase 2 extended the evaluation to 570 prompts distributed across three datasets (GoEmotions, DailyDialog, and ISEAR) to assess whether the patterns observed in the initial 12-prompt comparison held at scale and across diverse communication contexts. This extended evaluation deliberately balanced single-labelled and multi-labelled samples (15 of each per parent emotion category), enabling assessment of how models handle both focused emotional expressions and complex multi-emotion scenarios.

The extended evaluation revealed more modest performance levels that reflected the substantial challenges current zero-shot approaches face when confronted with comprehensive, real-world emotional expressions, as shown in Table 4. Cohere achieved the

highest position-weighted performance (34.3%), followed closely by Gemini (33.1%) and ChatGPT (32.0%), indicating remarkable consistency across models despite their different architectural approaches. The dramatic difference between the initial 12-prompt evaluation (position-weighted scores of 41.3–51.3%) and the extended 570-prompt evaluation (32.0–34.3%) highlights the importance of dataset diversity and scale in evaluating emotion detection capabilities. The 570-prompt evaluation encompasses a broader range of emotional expressions, more ambiguous contexts, and diverse communication styles that test the limits of zero-shot emotion detection.

When examining multi-label scores, which treat all predictions equally, models demonstrated stronger performance: Gemini achieved 48.8%, Cohere 47.2%, and ChatGPT 46.9%. This pattern mirrors the findings from the initial comparison, where models consistently performed better on multi-label metrics than position-weighted metrics, reflecting their strength in comprehensive emotion recognition rather than confidence calibration.

The progression from K-1 scores (23–24%) to K-4 scores (43–52%) demonstrated that whilst models struggle to identify exact ground truth emotions as their primary predictions, they often capture psychologically reasonable alternatives within their broader prediction sets. This finding suggests significant potential for human-AI collaborative emotion detection systems where AI presents multiple plausible interpretations for human consideration rather than making definitive classifications. Notably, Gemini demonstrated the strongest performance on K-n metrics, with K-4 reaching 52.0%, indicating superior capability at identifying ground truth within its top predictions even when not ranking it first.



**Figure 7: Extended evaluation performance by parent emotion categories across 570 prompts. Stars indicate highest-performing model for each category. Joy consistently achieves highest performance across all models, whilst sadness and fear present greatest challenges. Gemini demonstrates superior performance across six of eight emotion categories.**

Analysis by parent emotion category revealed consistent patterns with the initial evaluation, as illustrated in Figure 7. Joy emotions achieved the highest performance across all models (ChatGPT: 38.4%, Cohere: 31.1%, Gemini: 38.8%), suggesting that explicit

positive emotional markers facilitate recognition regardless of architectural approach. Conversely, sadness proved most challenging (ChatGPT: 25.1%, Cohere: 26.4%, Gemini: 30.1%), indicating that low-arousal negative emotions with subtle manifestations present consistent difficulties across different model architectures. Gemini demonstrated superior performance across six of the eight emotion categories (anger, joy, fear, surprise, guilt, and shame), establishing consistent strength across diverse emotional domains. ChatGPT showed particular strength in joy detection, whilst Cohere’s performance remained more uniform across categories without clear specialisation.

Individual emotion analysis revealed that Gemini achieved the best performance on 21 of the 28 emotions analysed, demonstrating remarkable consistency across the emotional spectrum. High-valence emotions with clear semantic markers consistently achieved better detection rates across all models. Love, pride, gratitude, and desire emerged as the most successfully detected emotions, with scores ranging from 34.7% to 48.2%. These emotions typically involve explicit positive language and clear contextual cues that facilitate identification. ChatGPT demonstrated particular specialisation in emotionally intense interpersonal emotions, achieving the highest scores for love (48.0%) and pride (48.2%), whilst Cohere excelled primarily in excitement detection.

Conversely, several emotions proved consistently difficult for all models, with performance scores clustering in the 20–30% range. Grief emerged as the most challenging emotion across all models (ChatGPT: 19.8%, Cohere: 20.9%, Gemini: 26.4%), likely due to its complex psychological nature involving loss, mourning, and culturally variable expression patterns. Embarrassment’s poor performance (ChatGPT: 23.5%, Cohere: 23.4%, Gemini: 27.3%) reflects the difficulty of inferring social self-consciousness from text alone, as it requires understanding implicit social dynamics and cultural norms that transcend surface-level linguistic patterns. The consistent struggle with nervousness (ChatGPT: 25.6%, Cohere: 23.8%, Gemini: 29.9%) and disappointment (ChatGPT: 25.6%, Cohere: 25.8%, Gemini: 27.2%) indicates that low-arousal negative emotions, which often manifest through subtle linguistic cues rather than explicit emotional language, present particular challenges for zero-shot detection approaches across all evaluated architectures.

The confusion analysis revealed a concerning pattern that persisted from the initial evaluation. All models demonstrated predominantly cross-family confusions (78–81%) compared to within-family confusions (19–22%). ChatGPT exhibited 20% within-family versus 80% cross-family confusions, Cohere showed 19% versus 81%, and Gemini demonstrated 21% versus 79%. This indicates that LLMs frequently make psychologically implausible errors even at scale, confusing emotions from entirely different emotional categories rather than making the subtle distinctions within emotion families that would be expected from human-like emotional understanding. Whilst Gemini performed slightly better at making psychologically reasonable errors, the overall pattern suggests that current zero-shot approaches lack the nuanced understanding of emotional relationships that characterises human emotional intelligence.

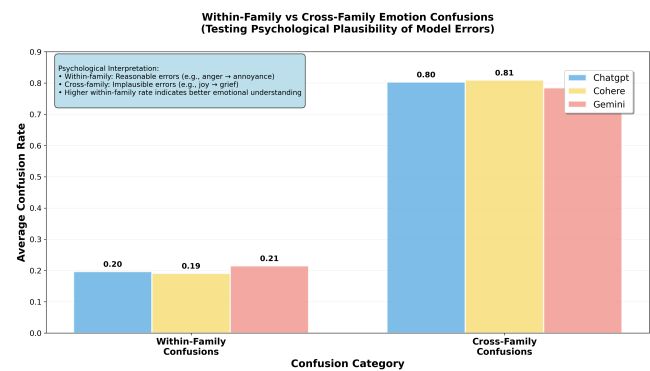
The extended evaluation confirms that the complementary human-AI pattern observed in the initial comparison (AI breadth versus human depth) holds robustly across diverse datasets and scales.

Whilst absolute performance scores decrease with dataset complexity and size, the fundamental strengths and limitations of each approach remain consistent. AI systems excel at identifying multiple plausible emotions but struggle with confidence calibration and psychological plausibility of errors. The finding that models perform better with multiple predictions than single predictions (K-1: 23–24% versus K-4: 43–52%) reinforces the value of pluralistic approaches to emotion detection, suggesting that optimal systems should present multiple interpretations with calibrated confidence levels rather than forcing single classifications. These patterns reinforce the case for collaborative systems that leverage both human contextual judgement and AI comprehensive recognition capabilities.

### 5.5 Error Analysis: Psychological Plausibility

Beyond overall performance levels, the types of errors made by humans versus AI reveal fundamental differences in emotional understanding. Analysis of model prediction errors in the extended 570-prompt evaluation revealed important insights into the psychological plausibility of LLM mistakes. Using the circumplex model’s hierarchical structure, we classified errors as within-family (psychologically plausible, such as confusing sadness with disappointment) versus cross-family (implausible, such as confusing sadness with joy).

When humans made incorrect predictions in the initial study, they predominantly selected emotions from the same parent category or adjacent regions of the circumplex space. The confusion matrix shown in Figure 6 (Section 4.3) demonstrated that even when participants failed to identify exact ground truth labels, their alternatives respected emotional similarity. For instance, when nervousness was misidentified, participants selected disappointment—sharing similar negative valence and arousal levels—rather than joy or excitement, which occupy opposite regions of the emotional space. This pattern reveals that human errors, whilst technically incorrect, maintain psychological coherence by respecting the dimensional structure of emotions.



**Figure 8: Within-family versus cross-family confusion rates for AI models across 570-prompt extended evaluation. All three models show 78–81% cross-family confusions, indicating errors that violate psychological relationships between emotions. Gemini performs slightly better (21% within-family) than ChatGPT (20%) or Cohere (19%).**

In stark contrast, all three LLM models demonstrated predominantly cross-family confusions (78–81%) compared to within-family errors (19–22%) in the extended evaluation, as illustrated in Figure 8. This pattern indicates that when AI systems make mistakes, they frequently make psychologically implausible errors, confusing emotions from entirely different emotional categories rather than making subtle distinctions within related emotion families. Specifically, ChatGPT exhibited 20% within-family versus 80% cross-family confusions, Cohere showed 19% versus 81%, and Gemini demonstrated 21% versus 79%. Whilst Gemini performed marginally better at making psychologically reasonable errors, the overall pattern reveals that current zero-shot approaches lack the nuanced understanding of emotional relationships that characterises human emotional intelligence.

This finding has critical implications for emotion-aware systems. When AI models make errors, users cannot rely on those errors being “close” to correct—a prediction of joy when the true emotion is grief represents a dramatic misunderstanding, not a subtle misclassification. The high rate of cross-family confusions suggests that AI systems may lack the embodied emotional experience and contextual understanding that enables humans to make psychologically plausible interpretive errors.

These patterns suggest several design implications for emotion-aware interfaces. First, systems should present AI predictions with explicit confidence levels that reflect not only the probability of correctness but also the psychological distance between predicted and alternative emotions. Second, interfaces should highlight when predictions span distant emotional categories, warning users of potential implausibility and prompting them to apply their own contextual judgement. Third, systems should provide opportunities for human validation, particularly when AI confidence is distributed across psychologically distant emotions, as this pattern often indicates fundamental interpretive confusion rather than legitimate emotional ambiguity. Finally, collaborative systems should leverage human judgement to filter psychologically implausible AI suggestions, positioning AI as a comprehensive detector that identifies possibilities whilst humans provide the psychological grounding needed to assess plausibility.

The error analysis reinforces that optimal human-AI collaboration should position AI as a broad but sometimes confused detector, with humans providing the psychological coherence needed to identify truly plausible interpretations. Rather than treating all AI errors as equivalent, emotion-aware systems should distinguish between psychologically reasonable alternatives and implausible confusions, adjusting their presentation and confidence indicators accordingly.

## 5.6 Summary: Toward Collaborative Emotion Detection

The human-AI comparison revealed three key insights that challenge conventional approaches to emotion detection system design.

First, both humans and AI naturally embrace pluralistic interpretation, with 78% of human responses and 100% of AI responses involving multiple emotions. This convergence, emerging without explicit instruction, provides strong evidence that acknowledging interpretive diversity is fundamental to sophisticated emotional understanding rather than a technical limitation to overcome.

Second, humans and AI demonstrate complementary strengths that suggest collaborative rather than competitive approaches. AI excels at comprehensive emotion recognition, identifying the full spectrum of plausible interpretations (Gemini: 67.1% multi-label score, though this may reflect potential exposure to GoEmotions during training). Humans excel at confidence calibration and contextual prioritisation (54.8% position-weighted score), leveraging sophisticated reasoning processes that incorporate tonal simulation, contextual inference, and metacognitive awareness. This complementarity suggests that optimal emotion-aware systems should leverage both capabilities rather than pursuing perfect human-AI alignment on oversimplified tasks.

Third, the psychological plausibility of errors matters as much as overall accuracy. Whilst AI systems achieve competitive aggregate performance, their predominantly cross-family confusions (78–81%) reveal fundamental gaps in understanding emotional relationships. Humans make psychologically coherent errors even when incorrect, respecting the dimensional structure of emotions and selecting alternatives from related emotional categories. This difference stems from humans’ embodied emotional experience and contextual understanding, which enables psychologically grounded interpretive errors that AI systems currently lack.

These findings motivate a paradigm shift from consensus-seeking to plurality, honouring emotion detection. Rather than training systems to replicate human consensus on single labels, we should develop interfaces that present multiple plausible interpretations calibrated by both AI breadth and human contextual judgement. The path forward lies not in achieving perfect human-AI alignment on oversimplified benchmarks, but in creating collaborative systems that leverage the distinct strengths each brings to emotion interpretation whilst acknowledging their respective limitations.

## 6 Discussion

### 6.1 The Pluralistic Paradigm

This work challenges a foundational assumption in emotion detection research: that consensus on emotional interpretation can and should be achieved. Our findings demonstrate that pluralistic interpretation is not a measurement problem to be solved but an authentic feature of human emotional understanding.

Three converging lines of evidence support this paradigm shift. First, 78% of human responses naturally involved multiple emotions without instruction, with emotionally intelligent participants showing lower agreement with single-label ground truth ( $r=-0.176$ ). This “Emotional Intelligence Paradox” reveals that current evaluation frameworks systematically mischaracterise sophistication as error. Participants with higher SSEIT scores employed more sophisticated interpretive strategies, considering multiple contextual factors and emotional nuances rather than seeking single definitive answers. The qualitative analysis revealed that high-EI participants consistently articulated nuanced reasoning—recognising that emotional expressions often contain primary and secondary content, conflicting or mixed emotional states, and emotional progression that single-label classifications cannot capture.

Second, state-of-the-art LLMs independently adopted multi-label approaches in zero-shot scenarios, suggesting that recognising emotional complexity is fundamental to advanced language understanding regardless of processing mechanism. All three evaluated models (ChatGPT, Gemini, Cohere) selected an average of 3.2–4.0 emotions per prompt without explicit instruction to do so, closely matching human behaviour ( $M=3.17$ ). This convergence emerged despite fundamental differences in how humans and AI systems process language—humans employing mental vocalisation, punctuation analysis, and contextual inference whilst LLMs rely on statistical patterns learned from vast text corpora.

Third, both humans and AI achieved dramatically better performance when evaluation honoured interpretive diversity. Humans reached 59% multi-label scores versus 54.8% position-weighted, whilst AI models similarly showed stronger multi-label (59.8–67.1%) than position-weighted (41.3–51.3%) performance in the initial 12-prompt comparison. This pattern held robustly in the extended 570-prompt evaluation, where multi-label scores (46.9–48.8%) consistently exceeded position-weighted scores (32.0–34.3%). The progression from K-1 scores (23–24%) to K-4 scores (43–52%) in the extended evaluation demonstrated that models capture multiple plausible interpretations even when missing the exact ground truth as their primary prediction.

The implication is clear: emotion detection systems should stop pursuing consensus on single "correct" labels and instead embrace the legitimate diversity in how different individuals—and different AI systems—interpret emotional content. The question shifts from "which emotion is correct?" to "what interpretations are psychologically plausible, and how confident should we be in each?" This paradigm shift has profound implications for how we design, evaluate, and deploy emotion-aware systems in real-world contexts.

## 6.2 Implications for Human-AI Collaboration

Our findings reveal complementary human-AI strengths that suggest new interaction paradigms for emotion-aware systems. AI excels at comprehensive emotion recognition, identifying the full spectrum of plausible interpretations (Gemini: 67.1% multi-label score in the initial comparison, though this may reflect potential exposure to GoEmotions during training). Humans excel at confidence calibration and contextual prioritisation (54.8% position-weighted score), leveraging sophisticated reasoning processes that incorporate tonal simulation, contextual inference, and metacognitive awareness.

However, the psychological plausibility of errors reveals a critical limitation in current AI approaches. AI models make predominantly cross-family confusions (78–81%), producing psychologically implausible errors that confuse emotions from entirely different emotional categories. When humans made errors, they maintained psychological coherence by selecting alternatives from the same parent category or adjacent regions of the circumplex space. This fundamental difference suggests that humans possess embodied emotional experience and contextual understanding that enables psychologically grounded interpretations—capabilities that current AI systems lack.

These complementary strengths and limitations suggest that optimal emotion-aware systems should position AI as a comprehensive

detector that identifies multiple possibilities whilst humans provide the psychological grounding and contextual judgement needed to assess plausibility and prioritise interpretations. Rather than pursuing perfect human-AI alignment on oversimplified tasks, we should develop collaborative approaches that leverage AI's breadth of pattern recognition with human depth of emotional understanding.

The convergence on challenging emotion categories—where both humans and AI struggled with low-arousal negative emotions like nervousness and context-dependent emotions like surprise—points to domains where collaboration may be most beneficial. When neither humans nor AI demonstrate high confidence, presenting multiple interpretations for consideration rather than forcing single classifications may better serve users' needs whilst acknowledging the legitimate ambiguity in emotional expression.

## 6.3 Limitations and Future Directions

Several limitations affect the generalizability of our findings.

**6.3.1 Sample Size and Statistical Power.** The human interpretation study involved 25 participants, with EI group imbalances (Low:  $n=3$ , Medium:  $n=12$ , High:  $n=10$ ) that limit statistical power for subgroup analyses. The observed negative correlation between EI and ground truth alignment did not reach statistical significance. Post-hoc power analysis indicates that detecting medium effects ( $d=0.5$ ) with 80% power would require approximately  $n=64$  per group. While this underpowered design means EI-related findings should be interpreted as preliminary evidence requiring replication, the core finding of pluralistic interpretation (78% multi-label responses,  $M=3.17$  emotions per text) remains robust across all participants and converged with LLM behavior. The demographic profile, predominantly young (84% aged 18–24), university-affiliated, with above-average EI ( $M=132.6$  vs. population norm=124), suggests findings may better characterize emotionally sophisticated, digitally, literate populations. Cross-cultural validation with diverse age groups and cultural backgrounds is needed to assess broader generalizability.

**6.3.2 Dataset and Model Limitations.** Our dataset focus on informal social media text (GoEmotions) for human-AI comparison may not generalise to formal documents, professional communications, or literary texts where emotional expression follows different conventions.

**6.3.3 Writer-Reader Perspective Gap.** A critical limitation acknowledged but not addressed in this research concerns the fundamental distinction between emotional expression (writer's intent) and emotional interpretation (reader's perception). This study focused exclusively on reader interpretation without examining how writers' intended emotional messages align with reader perceptions. Research demonstrates that emotions expressed by writers are significantly more complex than those perceived by readers [1], yet most emotion detection systems optimize for reader interpretations rather than writer intentions.

This writer-reader disconnect has profound implications for emotion detection applications. Systems designed for mental health monitoring might need to capture writer intentions to assess emotional states accurately, while content moderation systems might prioritize reader interpretations to predict community responses.

Future research must develop methodologies that can capture both perspectives and understand how the gap between emotional expression and interpretation varies across individuals, contexts, and cultural backgrounds.

## 7 Conclusion

This work challenges the consensus-seeking paradigm that has dominated emotion detection research, demonstrating that pluralistic interpretation is fundamental to sophisticated emotional understanding. Through systematic comparison of human participants ( $n=25$ ) and state-of-the-art LLMs across identical emotion labelling tasks, we reveal three critical insights that transform how we should approach emotion detection in digital contexts.

First, both humans and AI naturally embrace interpretive plurality. Without instruction, 78% of human responses involved multiple emotions ( $M=3.17$  per text), mirrored by AI systems selecting 3.2–4.0 emotions per text. This convergence across fundamentally different processing mechanisms—humans employing mental vocalisation and contextual inference, AI relying on statistical patterns—suggests that acknowledging emotional complexity is essential rather than optional for sophisticated language understanding.

Second, the “Emotional Intelligence Paradox” exposes fundamental flaws in current evaluation approaches. Participants with higher EI demonstrated lower agreement with single-label ground truth ( $r=-0.176$ ), not because they understood emotions less well, but because they recognised legitimate complexity that simplified annotations systematically eliminate. The participants who best understood emotional complexity performed worst on traditional metrics, revealing that current evaluation frameworks mischaracterise emotional sophistication as error. This counterintuitive finding challenges the assumption that accuracy on consensus-based benchmarks reflects genuine emotional understanding.

Third, humans and AI demonstrate complementary strengths that suggest collaborative rather than competitive approaches. AI excels at comprehensive emotion recognition (Gemini: 67.1% multi-label score in initial comparison, though this may reflect potential exposure to GoEmotions during training), identifying the full spectrum of plausible interpretations. Humans excel at confidence calibration and contextual prioritisation (54.8% position-weighted score), leveraging sophisticated reasoning processes and metacognitive awareness to assess which interpretations are most salient. However, the psychological plausibility of errors reveals a critical limitation: AI models make predominantly cross-family confusions (78–81%), whilst humans maintain psychological coherence even when incorrect.

We introduce methodological innovations—circumplex-based scoring and position-weighted evaluation—that honour rather than penalise interpretive sophistication. These frameworks recognise that emotional “errors” are not equivalent, distinguishing between psychologically plausible alternatives and implausible confusions. The progression from K-1 scores (23–24%) to K-4 scores (43–52%) in our extended 570-prompt evaluation demonstrates that both humans and AI capture multiple valid interpretations, reinforcing the value of pluralistic approaches.

The path forward lies not in achieving perfect human-AI alignment on oversimplified tasks, but in developing emotion-aware

systems that embrace the complexity of human emotional understanding whilst leveraging AI’s pattern recognition capabilities. By honouring interpretive plurality, we can create systems that genuinely support rather than constrain human emotional communication in digital contexts. Rather than training models to replicate consensus on single labels, we should develop collaborative approaches where AI presents multiple plausible interpretations and humans provide the contextual judgement needed to assess their relevance. This paradigm shift opens new possibilities for emotion-aware interfaces that acknowledge the beautiful complexity of human emotional experience whilst harnessing the complementary strengths of human and artificial intelligence.

## 8 GenAI Usage Disclosure

This research involved the use of generative AI tools as follows:

### 8.1 Research Design and Data Collection

No generative AI tools were used in the design of the research methodology, selection of datasets, recruitment of participants, or collection of human study data. All research design decisions were made by the human researchers.

### 8.2 Data Analysis

No generative AI tools were used in the quantitative analysis of results, statistical testing, or calculation of evaluation metrics. All analyses were conducted using standard statistical software and custom Python scripts written by the researchers.

The qualitative analysis of interview transcripts employed human-driven inductive thematic coding. No AI tools were used to generate themes or analyze participant responses. All coding and interpretation were performed by the researchers.

### 8.3 Large Language Model Evaluation

The research explicitly evaluated three large language models (ChatGPT-4, Gemini 2.5 Flash, Cohere Command R+) as research subjects in zero-shot emotion detection tasks. These models were provided with text prompts and asked to identify emotions, with their outputs forming the basis of the human-AI comparison analysis.

### 8.4 Writing and Editing

Generative AI tools (specifically Grammarly) were used to assist with manuscript preparation in the following ways:

- Improving prose clarity and flow whilst maintaining the authors’ voice and arguments
- Restructuring sentences for better readability
- Suggesting alternative phrasings when concepts were unclear

All substantive content, arguments, interpretations, and conclusions are the original work of the human authors. The AI assistant did not generate novel research ideas, create data, or make analytical decisions. All AI-suggested revisions were reviewed and approved by the authors, who take full responsibility for the accuracy and integrity of the work.

## 8.5 Code and Supplementary Materials

No generative AI tools were used in the development of analysis scripts, evaluation frameworks, or data processing pipelines. All code was written by the researchers.

The circumplex-based scoring methodology and position-weighted evaluation frameworks represent novel contributions developed by the researchers without AI assistance.

## References

- [1] Nurudin Alvarez-Gonzalez, Andreas Kaltenbrunner, and Vicenç Gómez. 2021. Uncovering the Limits of Text-based Emotion Detection. In *Findings of the Association for Computational Linguistics: EMNLP 2021* (Stroudsburg, PA, USA). Association for Computational Linguistics, 2560–2583. doi:10.18653/v1/2021.findings-emnlp.219
- [2] Lisa Feldman Barrett. 2006. Solving the Emotion Paradox: Categorization and the Experience of Emotion. *Personality and Social Psychology Review* 10 (2 2006), 20–46. Issue 1. doi:10.1207/s15327957pspr1001\_2
- [3] Laura Ana Bostan, ; Maria, and Roman Klinger. 2018. *An Analysis of Annotated Corpora for Emotion Classification in Text*. Technical Report. 2104–2119 pages. <https://creativecommons.org/licenses/by/4.0/legalcodehttp://www.ims.uni-stuttgart.de/data/unifyemotion.http://www.ims.uni-stuttgart.de/data/unifyemotion>.
- [4] Sven Buechel and Udo Hahn. 2016. Emotion analysis as a regression problem-dimensional models and their implications on Emotion representation and metrical evaluation. In *Frontiers in Artificial Intelligence and Applications*, Vol. 285. IOS Press BV, 1114–1122. doi:10.3233/978-1-61499-672-9-1114
- [5] Sven Buechel and Udo Hahn. 2017. EmoBank: Studying the Impact of Annotation Perspective and Representation Format on Dimensional Emotion Analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers* (Stroudsburg, PA, USA). Association for Computational Linguistics, 578–585. doi:10.18653/v1/E17-2092
- [6] Rafael A. Calvo and Sunghwan Mac Kim. 2013. Emotions in text: Dimensional and categorical models. In *Computational Intelligence*, Vol. 29. 527–543. Issue 3. doi:10.1111/j.1467-8640.2012.00456.x
- [7] Dorotyia Demsky, 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* (Stroudsburg, PA, USA). Association for Computational Linguistics, 4040–4054. doi:10.18653/v1/2020.acl-main.372
- [8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North* (Stroudsburg, PA, USA). Association for Computational Linguistics, 4171–4186. doi:10.18653/v1/N19-1423
- [9] Paul Ekman, I Thank, Richard Davidson, Phoebe Ellsworth, Wallace V Friesen, Robert Levenson, Harriet Oster, and Erika Rosenberg. 1992. *Are There Basic Emotions?* Technical Report. 550–553 pages. Issue 3.
- [10] Elaine Fox. 2008. *Emotion Science*. Macmillan Education UK. doi:10.1007/978-1-137-07946-6
- [11] Stephan Hamann and Turhan Canli. 2004. Individual differences in emotion processing. *Current Opinion in Neurobiology* 14 (4 2004), 233–238. Issue 2. doi:10.1016/j.conb.2004.03.010
- [12] Dacher Keltner, Disa Sauter, Jessica Tracy, and Alan Cowen. 2019. Emotional Expression: Advances in Basic Emotion Theory. *Journal of Nonverbal Behavior* 43 (6 2019), 133–160. Issue 2. doi:10.1007/s10919-019-00293-3
- [13] Xuan Liu, Tianyi Shi, Guohui Zhou, Mingzhe Liu, Zhengtong Yin, Lirong Yin, and Wenfeng Zheng. 2023. Emotion classification for short texts: an improved multi-label method. *Humanities and Social Sciences Communications* 10, 1 (12 2023). doi:10.1057/s41599-023-01816-6
- [14] Zhe Liu, Anbang Xu, Yufan Guo, Jalal Mahmud, Haibin Liu, and Rama Akkiraju. 2018. Seemo: A Computational Approach to See Emotions. 1–12. doi:10.1145/3173574.3173938
- [15] Christopher D. Manning. 2022. Human Language Understanding & Reasoning. *Daedalus* 151 (5 2022), 127–138. Issue 2. doi:10.1162/daed\_a\_01905
- [16] Abdullah Al Maruf, Fahima Khanam, Md Mahmudul Haque, Zakaria Masud Jiyad, M. F. Mridha, and Zeyar Aung. 2024. Challenges and Opportunities of Text-Based Emotion Detection: A Survey. *IEEE Access* 12 (2024), 18416–18450. doi:10.1109/ACCESS.2024.3356357
- [17] Saif Mohammad. 2018. Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Stroudsburg, PA, USA). Association for Computational Linguistics, 174–184. doi:10.18653/v1/P18-1017
- [18] JONATHAN POSNER, JAMES A. RUSSELL, and BRADLEY S. PETERSON. 2005. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology* 17 (9 2005), Issue 03. doi:10.1017/S0954579405050340
- [19] James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology* 39 (12 1980), 1161–1178. Issue 6. doi:10.1037/h0077714
- [20] Klaus R. Scherer. 2005. What are emotions? And how can they be measured? *Social Science Information* 44 (12 2005), 695–729. Issue 4. doi:10.1177/0539018405058216
- [21] Nicola Schutte. 1998. *Statistics Solutions Advancement Through Clarity Schutte Self-Report Emotional Intelligence Test (SSEIT)*. Technical Report. <http://www.statisticssolutions.com>
- [22] Julian Striegl, Jordan Wenzel Richter, Leoni Grossmann, Björn Bråstad, Marie Gotthardt, Christian Rück, John Wallert, and Claudia Loitsch. 2024. Deep learning-based dimensional emotion recognition for conversational agent-based cognitive behavioral therapy. *PeerJ Computer Science* 10 (2024). doi:10.7717/peerj-cs.2104
- [23] Ashish Vaswani, Google Brain, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. *Attention Is All You Need*. Technical Report.
- [24] Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods* 45 (12 2013), 1191–1207. Issue 4. doi:10.3758/s13428-012-0314-x