

Incongruent Positivity: When Miscalibrated Positivity Undermines Online Supportive Conversations

Leen Almajed^{1,2*}, Abeer ALdayel²

¹TAHAKOM

²King Saud University, College of Computer and Information Sciences
lalmajed@tahakom.com, aabeer@ksu.edu.sa

Abstract

In emotionally supportive conversations, well-intended positivity can sometimes misfire, leading to responses that feel dismissive, minimizing, or unrealistically optimistic. We examine this phenomenon of incongruent positivity as miscalibrated expressions of positive support in both human and LLM generated responses. To this end, we collected real user–assistant dialogues from Reddit across a range of emotional intensities and generated additional responses using large language models for the same context. We categorize these conversations by intensity into two levels of Mild which covers relationship tension and general advice and Severe which covers grief and anxiety conversations. This level of categorization enables a comparative analysis of how supportive responses vary across lower and higher stakes contexts. Our analysis reveals that LLMs are more prone to unrealistic positivity through dismissive and minimizing tone, particularly in high stakes contexts. To further study the underlying dimensions of this phenomenon, we fine-tune LLMs on datasets with strong and weak emotional reactions. Moreover, we developed a weakly supervised multilabel classifier ensemble (DeBERTa and MentalBERT) that shows improved detection of incongruent positivity types across two sorts of concerns (Mild and Severe). Our findings shed light on the need to move beyond merely generating generic positive responses and instead study the congruent support measures to balance positive affect with emotional acknowledgment. This approach offers insights into aligning large language models with affective expectations in the online supportive dialogue, paving the way toward context-aware and trust-preserving online conversation systems.

1 Introduction

Positive language is at the core of supportive dialogue in distressing circumstances and provides support to those experiencing negative emotions (Nozaki and Gross 2025). However, in emotionally charged contexts, positivity that misaligns with a user’s emotional state can backfire by appearing as dismissive, minimizing, or unrealistically optimistic (Tugade, Fredrickson, and Barrett 2004). These well-

intended but misguided responses may undermine emotional connection and erode user trust. Thus, understanding and mitigating incongruent positivity becomes essential as large language models (LLMs) are increasingly used in conversations related to emotional distress and support (Russo et al. 2023; Saha et al. 2022).

Given the importance of studying well-defined peer-support responses, the majority of computational methodologies utilize empathy as a key component, either to identify cognitive alignment of emotions (Yang and Jurgens 2024; Wei et al. 2019), combined emotional contagion with intent prediction (Yuan et al. 2025) or directly detecting the presence of empathy levels (Welivita, Yeh, and Pu 2023; Welivita and Pu 2023), aiming to improve the emotional granularity of supportive interactions. In parallel, another line of research has begun to examine the limitations of adopting a generative model for emotion support conversations (Kang et al. 2024; Wang et al. 2024), revealing that large language models often rely on a narrow set of support strategies, which undermines emotional alignment and requires intervention to reduce preference bias. Mainly, for emotion detection research has traditionally relied on granular labels such as Ekman’s six emotions or Plutchik’s primary categories (Mohammad and Turney 2013; Demszky et al. 2020). While large language models have shown promising performance in emotion classification (Demszky et al. 2020), recent studies have started to show persistent gaps in capturing context-specific emotional nuances by gauging LLM robustness on the eight emotion types (Shu et al. 2026). Those strategies have mainly focused on appraisal style (Yang and Jurgens 2024) or the intention mimicry framework (Yuan et al. 2025), and are generally marked as ‘unintentionally unhelpful’ or ‘counterproductive’ responses (Wang et al. 2024), without underscoring the type of emotional invalidation for such shortcomings.

To address this need, the study examines cases of *incongruent positivity* in supportive conversations, where responses are miscalibrated, whether generated by humans or language models, meaning these responses do not properly match the emotional intensity or type of the expressed concerns. We frame this task through the concept of *misguided positivity* by comprising three non-exclusive categories: *Dismissive*, *Minimizing*, and *Unrealistic* support. Prior work in

*This work was conducted while the first author, Leen Almajed, was affiliated with the Computer Science Department, King Saud University, College of Computer and Information Sciences. Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

calibration of computational models focused on the verbalized confidence statements matching the likelihood of a response’s correctness (Mielke et al. 2022). Instead, in this study, we conceptualize incongruent positivity as a mismatch between the user’s emotional need for acknowledgment and a dismissive response intended to convey a positive tone given in reaction to distress. Incongruent positivity looks similar to toxic positivity, but the two are different in their purpose. Toxic positivity, which comes from individual denial, is described in psychology as a broad norm that puts too much focus on positive emotions and ignores negative ones. Researchers often study toxic positivity as a way people cope. It means being overly happy or optimistic in every situation, which can lead to denial (Premlal and Jose K 2024; Wyatt 2024). In this study, we define incongruent positivity as a mismatch between what a user needs emotionally and the positive response they receive. Our goal is not to diagnose a psychological pattern, but to find out when positivity is not appropriate in a conversation. This approach helps us measure, label, and study incongruent positivity in both human and LLM-generated replies, making it possible to analyze how well supportive dialogue systems match emotional needs. As illustrated in the affective science literature (Tamir 2009), the unrealistic and excessive positivity can undermine adaptive functioning, especially when the emotional state and support response are misaligned. This kind of empirical support comes from studies of positive emotion disturbance, showing that excessively intense positive affect may be nonadaptive in real-world contexts (Tugade, Fredrickson, and Barrett 2004), which cautions that rigid positivity often leads to dissatisfaction and emotional suppression.

In particular, this study aims to address the following research questions:

- (Q₁) *How prevalent are different types of misguided positivity (dismissive, minimizing, unrealistic) in supportive assistant-generated responses, and how are they associated with positive reframing strategies (such as growth, impermanence, optimism, relatability)?*
- (Q₂) *Do emotionally relieving responses tend to avoid misguided positivity, or do some misguided forms still offer perceived emotional support?*
- (Q₃) *Does fine-tuning on empathetic dialogue reduce the likelihood of generating misguided positivity?*

To build a realistic evaluation of support conversations, we construct a new dataset of paired user-assistant dialogues sampled from *subreddits* spanning a range of emotional intensities, focusing on two levels of concern: *Severe* (covers grief and anxiety), along with *Mild* (covers relationship tension and general advice). We then examine the behavior of LLMs and human responders under these varying emotional intensities. Each turn is labeled with an incongruent positivity category, a re-framing strategy, and evaluated for emotional relief. We further propose a weakly supervised multi-label detection framework using DeBERTa and MentalBERT to identify the positivity miscalibrations within support-based conversations. The results illustrate that LLMs tend to produce dismissive or unrealistic opti-

mism, particularly in high-stakes settings, such as severe concern conversations. Targeted fine-tuning and labeling strategies can reduce such miscalibration. Our findings show the need to move beyond simply generating positive responses toward providing emotionally aligned and context-aware support within LLM interactions.

This study makes three primary contributions:

- We formalize the concept of *incongruent positivity* in supportive conversations and operationalize it through a taxonomy of misguided positivity types as: *Dismissive*, *Minimizing*, and *Unrealistic*. Moreover, we extend the set of reframing strategies (*Growth*, *Impermanence*, *Optimism*, *Thankfulness*, and the newly introduced *Relatability*) to represent different affective framing.
- We construct and release a new dataset of 1,490 conversation turns across *Mild* (relationship tension and general advice) and *Severe* (grief and anxiety) concerns, each with paired human- and LLM-generated responses annotated for misguided positivity categories, reframing strategies, and perceived emotional relief (RQ1–RQ2).
- We experiment with different algorithms to classify the incongruent positivity of a given response. Mainly, we evaluate the ability of the weakly supervised multi-label detection framework that ensembles *DeBERTa* and *MentalBERT* to detect rare forms of misguided positivity, alongside fine-tuning experiments on weak and strong emotional reaction datasets (RQ3) to assess whether targeted training can reduce positivity miscalibration and enhance alignment with user emotion context.

These contributions advance the modeling of affective alignment in emotional contexts for developing context-sensitive supportive dialogue systems.

2 Related Work

Supportive dialogue is a psychologically rich interaction shaped by affective, cognitive, and contextual factors (Rashkin et al. 2019). Researchers have increasingly explored how computational models can simulate supportive behaviors in conversational systems, especially in emotionally sensitive contexts. An emerging line of work draws from positive psychology to guide response generation. For instance, the work by Nguyen et al. (2025) studies the online message framing of mental health (mis)information, showing how overly optimistic supportive content can distort perceptions of risk. Another example, Ziems et al. (2022) introduced the task of positive reframing, in which user-authored complaints or negative self-statements are rewritten to reflect more constructive outlooks. Their dataset includes human-written reframes labeled with strategies such as optimism, growth mindset, and self-affirmation. Building on this, Goel et al. (2025) propose SocraticReframe, a framework that improves positive text rewriting by guiding models through reflective question–answer steps grounded in cognitive behavioral therapy. While this approach aims to promote psychological well-being, it also raises concerns about toxic positivity as the tendency to enforce a uniformly upbeat tone that may inadvertently invalidate users’ emotions (Wyatt 2024). Related concerns have been raised in the

psychology literature, where well-meaning affirmations may suppress or dismiss negative emotions rather than acknowledge them (Premlal and Jose K 2024; Sonia 2025). These distinctions are critical in computational settings, where a failure to model emotional nuance can compromise user trust. Most research on toxic positivity has focused on its long term effects on emotions and relationships as a psychological issue. Our study looks at how positivity appears in conversations as a response during key supportive moments. By connecting mismatched positivity to clear gaps between context and response, we make psychological ideas easier to label and use to check and improve how well generative models reflect emotions.

To address this, many studies have turned to generating explicitly empathetic responses. Rashkin et al. (2019) introduced the EmpatheticDialogues dataset to train models that align with users’ emotional states. Similarly, the Emotional Support Conversations (ESC) dataset (Liu et al. 2021) provides multi-turn dialogues between help-seekers and trained supporters, annotated with fine-grained strategies grounded in counseling theory, such as reflection, affirmation, and suggestion. These annotations enable models to learn the procedural structure of emotional support, beyond surface-level empathy. Another work by Saha et al. (2022) designed a virtual assistant that generates personalized and emotionally tuned dialogue aimed at encouraging users in need. Recent research has also explored fine-tuning methods to better capture empathy (Majumder et al. 2020) and techniques for dynamically adjusting tone based on user affect (Shin et al. 2020). Nevertheless, studies show that emotional misalignment remains a persistent issue. Kursuncu et al. (2025) report that LLMs can inadvertently amplify distress, especially when users express vulnerability. In EmotionBench, models often failed to adapt tone to emotionally similar prompts and lacked deeper contextual awareness Huang et al. (2023). Likewise, Cuadra et al. (2024) shows that LLMs may produce supportive-sounding but ultimately hollow responses, raising ethical concerns around perceived sincerity and trust. These works tend to focus on granular labels such as Ekman’s six emotions or Plutchik’s primary categories (Mohammad and Turney 2013; Demszky et al. 2020). For instance, a recent study by (Shu et al. 2026) introduced the EXPRESS dataset with the eight self-disclosed emotions from Reddit and demonstrated that while LLMs often produce plausible labels they often miss contextual cues that human annotators detect. Their findings highlight that prompt design affects emotional detection and show that few-shot prompting improves performance and chain-of-thought prompting degrades it. Another work by Moore et al. (2025) highlights the safety concerns of LLM-generated responses to crisis situations, which might include advice to commit suicide.

Most prior work focuses solely on pure empathy detection, usually with reference to emotion categories or constructive reframing. There are studies that go beyond simple positive or negative polarity by analyzing framing dimensions, such as the work by Jaidka et al. (2020), which examines narrative styles and regional differences in protest coverage, assessing not just sentiment but also how stories

align with or deviate from the established topic paradigm. Additionally, recent work by Lahkala et al. (2025) emphasizes the importance of theoretical grounding in empathy tasks, which in turn impacts empathy modeling. Instead, our study extends these insights by shifting focus from accuracy in emotion classification to the appropriateness of emotional alignment by analyzing *incongruent positivity* to examine the supportive responses that misalign with user distress by being dismissive, minimizing, or unrealistically optimistic. We extend our approach beyond the mere production of responses characterized by a positive tone or empathetic content by explicitly modeling instances of emotional miscalibration and annotating detailed categories of misguided support. This helps evaluate the quality of support in high-stakes emotional situations where misplaced positivity could harm trust.

Category	Severe	Mild
Unique dialogues	379	366
Responses	(human + LLM)	(human + LLM)
Total turns	758	732

Table 1: Dataset statistics for Severe Concerns and Mild Concerns contexts

Dataset	# Examples	Level
<i>Weak Emotion</i>	1,047	0
<i>Strong Emotion</i>	1,047	1 and 2

Table 2: The *Weak* and *Strong* emotion reaction datasets derived from Sharma et al. (2020) to fine-tune emotion based generative support. Each contains 1,047 seeker-response pairs annotated for the emotional intensity of support provided.

3 Methodology

3.1 Data Collection

As part of the data collection process, Reddit posts were sourced from three subreddits—*r/stress*, *r/grief*, *r/situationshipsadvice*, and *r/advice* selected for the frequent occurrence of emotionally expressive language.

The subreddits *r/stress* and *r/grief* were chosen due to the explicit emotional framing present in both the community guidelines and the user submissions. The posts in these forums often address topics such as anxiety, burnout, and emotionally distressing events, which we categorize as Severe Concerns. Another set of dialogues was extracted from *r/stress* and *r/grief*, where each dialogue consists of a user input post paired with the highest-ranked human response. The strong affective anchoring of these interactions enables high-precision analysis of tone and emotional alignment. To include a broader range of contexts, dialogues were also sampled from *r/situationshipsadvice* and *r/advice*, subreddits where users frequently seek guidance on emotionally

Label	User Post	Assistant Response → Label Justification
Dismissive	<i>"I am overwhelmed by this non-sense homework."</i>	"Stop overthinking, try taking a deep breath and move on." → <i>Invalidates the user's emotions.</i>
Minimizing	<i>"I am burned out at school and can't study anymore."</i>	"Everyone goes through this. Try drinking more water and walking daily." → <i>Oversimplifies the issue with generic advice.</i>
Unrealistic	<i>"I have failed this test three times. I feel hopeless."</i>	"You just need to believe in yourself more and everything will magically work out!" → <i>Offers impractical and overly idealistic advice.</i>
None	<i>"I am stressed about my job interview."</i>	"That sounds tough. Interviews can be overwhelming. Is there anything specific making it worse for you?" → <i>Validates the concern and invites further discussion.</i>

Table 3: Examples of assistant responses annotated for Misguided Positivity categories (Paraphrased to prevent re-identification while maintaining emotional and contextual significance). → indicates label Justification.

complex but comparatively less acute situations. Although posts in these communities are not explicitly labeled by emotional state, they frequently exhibit signs of distress, uncertainty, or tension. These interactions were categorized as Mild Concerns, allowing the investigation of affective incongruity in advice-giving scenarios where emotional framing tends to be more diffuse and context-dependent. Overall, we have 1490 turns of conversations, as shown in Table 1. The dataset is available at: <https://osf.io/5dcxz/>. We note that this distinction is used as a contextual grouping reflecting typical support-seeking intent, rather than as a post-level measure of emotional intensity, as individual expressions of distress may vary within each category.

The dataset comprises 379 unique dialogues corresponding to Severe Concerns and 366 unique dialogues corresponding to Mild Concerns. Each of these has two responses: human and LLM-generated responses. For each human input we use LLM-generated responses for RQ1/RQ2 that were produced using Meta Llama 3.2 1B Instruct (`meta-llama/Llama-3.2-1B-instruct`) with fixed decoding parameters (temperature=0.5, repetition penalty=1.1, max new tokens=512)

Incongruent positivity labels. We annotate a dialogue-based dataset for incongruent misguided positivity, focusing on assistant responses that, while intended to be supportive, may inadvertently dismiss or oversimplify users' concerns. We conceptualize *incongruent positivity* as a mismatch between the user's emotional need for acknowledgment and a dismissive response intended to convey a positive tone, often given in reaction to distress (Tamir 2009). Unlike toxic positivity, which is usually seen as a general emotional attitude or coping style, we define incongruent positivity as a specific conversational problem. It happens when a supportive response does not match the user's emotional intensity or distress. This difference lets us clearly label incon-

gruent positivity using categories like Dismissive, Minimizing, and Unrealistic. These categories make it easier to systematically evaluate support responses from both people and models. Building on this aspect, we define *misguided positivity* through three non-exclusive annotation categories: (1) **Dismissive**, where the response invalidates or overlooks the user's concern (implying the user is overreacting); (2) **Minimizing**, where the response downplays the user's emotions by offering generic or overly simplified advice; and (3) **Unrealistic**, where the response promotes impractical or idealistic solutions detached from the user's context. Responses that do not exhibit these traits are labeled as **None**. Table 3 shows a set of examples for each label. We annotated the data described in Table 1 with the presence of incongruent positivity as non-exclusive categories.

Additionally, we annotate reframing strategies used in assistant replies derived from (Ziems et al. 2022), which are defined as Growth, Impermanence, Optimism, and Thankfulness. We included a new label, 'Relatability,' to identify the means of congruency used in responses where assistants share similar experiences to foster emotional connection. The concept of 'Relatability' has not been addressed in prior frameworks. This labeling enables analysis of how positivity is framed in conversational contexts. As for the *Growth* it is label whether the assistant frames the user's experience as an opportunity for self-improvement or personal development The *Impermanence*, which reassures the user that their current distress is temporary and that circumstances will eventually improve; *Optimism*, where the assistant highlights what is going well in the present moment or redirects attention toward immediate strengths; and *Thankfulness*, which encourages a mindset of gratitude or reflection on positive aspects, even in challenging times.

Furthermore, the annotation includes a pairwise comparison between two assistants evaluates which response pro-

vides more emotional relief, emphasizing affective engagement over solution-oriented advice. Assistant and user responses are anonymized as Assistant A and Assistant B (see Appendix A for detailed annotation guidelines), allowing annotators to focus on emotional cues without being influenced by identity or role labels. For our dataset, the annotation process was conducted using the Labelbox platform¹. Labelbox provides a managed workforce and each annotator must pass benchmark tests and quality assurance checks before contributing to the project. Then, each conversation turn was independently labeled by two trained annotators recruited from the Labelbox workforce, following detailed guidelines for identifying misguided positivity categories (Dismissive, Minimizing, Unrealistic, None), reframing strategies, and emotional relief preferences. A detailed explanation of the annotation process is in Appendix A.

Aggregate Annotation Agreement by Concern Level. Annotation reliability is higher in the *Mild Concern* context across LLM and human support responses. The average Cohen’s Kappa for misguided positivity is 0.8827 (Jaccard: 0.9046), and for reframing strategies, 0.8186 (Jaccard: 0.7454), reflecting strong to near-perfect inter-annotator agreement. Also, Emotion relief annotations show a high reliability ($\kappa = 0.9648$). In contrast, the *Severe Concern* context yields a fair agreement (McHugh 2012) in the misguided positivity averages at 0.3128 (Jaccard: 0.4898), and reframing strategies at 0.3062 (Jaccard: 0.7214). Emotion relief in this context shows moderate agreement ($\kappa = 0.5676$), indicating that high-stakes emotional disclosures have greater challenges for consistent annotation. It can be noted that agreement is lower for severe concern cases, reflecting the greater interpretive ambiguity involved in assessing emotionally misaligned positivity under high distress rather than random annotator noise. A detailed explanation of the annotation results is in Appendix A.

3.2 Generating Support Responses by Fine-tuning on Emotion Reactions

Unlike RQ1/RQ2, which analyze a single baseline LLM for controlled comparisons, RQ3 evaluates additional models fine-tuned using weak supervision to reduce labeling and compute costs. Thus, to study the effect of training the model on the emotion data (RQ3), we further fine-tune the models on two types of emotional reactions: (1) weak emotion reactions and (2) strong emotion reactions. These reactions are drawn from the dataset introduced by Sharma et al. (2020), where each turn is labeled with an emotion intensity level. In the experiment, we map the intensity level into two categorizations, *weak reactions* as those annotated with *level 0*, indicating minimal or neutral emotional response. As for **strong reactions**, we map it to correspond to *levels 1 and 2*, reflecting moderate to intense affective expression. Each subset of these emotion levels contains 1,047 paired turns (Table 2). Then, we fine-tuned two well-known LLMs (Mistral-7B-v0.1 and Llama-3.2-3B-Instruct) on these subsets. Further model specifications and prompt details are provided in Appendix B.

¹<https://labelbox.com>

Model	P	R	F1
most frequent	0.13	0.25	0.17
SVM	0.46	0.29	0.32
LR	0.43	0.39	0.39
BiLstm-tuned	0.36	0.56	0.44
deberta-2layers	0.36	0.80	0.49
BERT-tuned	0.38	0.67	0.46
roberta-base	0.41	0.42	0.38
Mistral	0.50	0.26	0.27
Ens_MBERT_deBert	0.45	0.72	0.52

Table 4: Macro-averaged classification performance across different models for multi-label misguided positivity prediction.

Weak supervision labeling for fine-tuned responses To multilabel the responses generated by a fine-tuned model on weak and strong emotions. First, we split the gold standard dataset in table 1 into training and testing sets using iterative stratification (80% training and 20% testing). The experiment includes a set of different models covering Instruct LLM Mistral-7B-Instruct-v0.1 for multi-label classification using QLoRA with a custom classification head. Along with that, we provide a comparison with baseline models, SVM, LR, and BiLstm.

Then, we used an ensemble of the two best-performing models, MentalBERT and DeBERTa, to predict the labels of responses generated by the fine-tuned models on both weak and strong emotion datasets. Thus, we use the multi-label classification setting where each output node predicts the probability of a label independently using a sigmoid activation. Also, we apply a fixed threshold of 0.5 to convert these probabilities into binary decisions. Threshold tuning addresses the limitation of imbalanced labels by assigning optimized, per-label thresholds (for example, lower values for rare labels) based on validation performance. This strategy enhances classification metrics, particularly the macro-

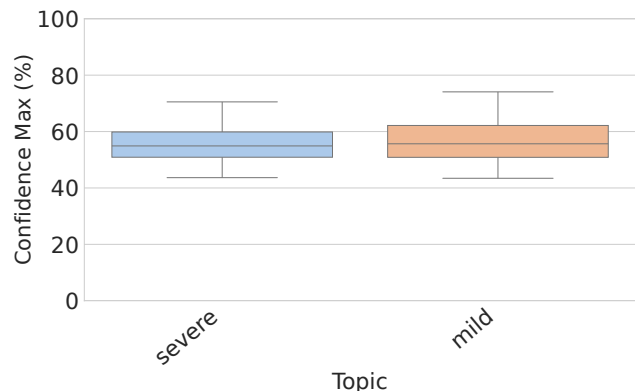


Figure 1: Distribution of maximum confidence scores of using weak-supervised labeling to fine-tuned responses by topic.

Model Pair	Min.	None	Dis.	Unr.	Fisher p (Holm)
BERT vs DeBERTa	< .001	< .001	.029	.803	< 10^{-16}
DeBERTa vs Ensemble	< .001	.216	< .001	.248	1.97×10^{-9}
BERT vs Ensemble	< .001	.039	< .001	.010	1.35×10^{-6}

Labels: MIN = Minimizing, NONE = No Misguided Positivity, DIS = Dismissive, UNR = Unrealistic.

Table 5: Model-pair disagreement evaluated using label-wise McNemar tests across the four misguided-positivity categories, together with Fisher-combined McNemar tests. Green shading indicates statistical significance ($\alpha = 0.05$); label-wise tests are uncorrected, while Fisher p -values are Holm-Bonferroni corrected.

averaged F1 score, by balancing precision and recall for each label. In our implementation, we applied per-class threshold tuning to enhance the model’s ability to detect underrepresented classes such as Dismissive and Unrealistic. The DeBERTa/MentalBERT and Ensemble comparison shows a significant difference with $p < .05$ (details of models and McNemar tests provided in Appendix C). Moreover, Appendix E reports label-wise false positive and false negative rates for these models. The ensemble model stands out for its balanced error profile, avoiding the over-prediction seen in DeBERTa and BERT (FPR = 1.000 for Minimizing and None) and keeping lower false positive rates for Dismissive and Unrealistic. Overall, the ensemble offers the most reliable performance across labels.

Each model was trained independently on gold-labeled data in Table 1 using a focal loss objective to mitigate the effects of class imbalance and enhance learning from difficult examples. We used tokenizer-specific data pipelines to ensure compatibility with each model’s architecture. Final predictions were obtained by ensembling the output probabilities from both models, with equal weighting, and applying label-wise threshold tuning to optimize the macro F1 score. Table 4 shows the detailed performance of weak-supervised models. This ensemble approach achieved a fair performance across all four positivity subcategories, including improved detection of less frequent categories such as “Dismissive” and “Unrealistic.” For each predicted multi-label vector as shown with the certainty score in Figure 1. Moreover, in Table 5 reports both label-wise McNemar tests and Holm-corrected Fisher-combined significance, showing that model disagreement is driven primarily by Minimizing and Dismissive categories. The comparisons between BERT and DeBERTa, BERT and the Ensemble, and DeBERTa and the Ensemble all yield statistically significant Holm-corrected p -values, indicating robust disagreement in their predictions across the four misguided-positivity labels. At the label level, this disagreement is most consistent for responses that downplay or prematurely reframe distress, while effects for Unrealistic positivity are weaker and less stable, suggesting that models diverge more strongly in how they handle subtle forms of emotional misalignment.

4 Results

The prevalence of misguided positivity between supportive assistant-generated responses To address *RQ1*, we

compare the distribution of misguided positivity types in two types of concerns, Mild (advise-based) and severe (grief cases). As shown in Figure 2, LLM-generated responses within the Severe Concerns context show higher rate of dismissive positivity (15.4%) compared to human responses (7.0%), suggesting a potential miscalibration with unaligned responses with sensitive emotional expectations. For minimizing positivity, human responses in the Mild Concerns context are markedly more frequent (56.8%) than those in the Severe Concerns context (26.8%), particularly when contrasted with LLM outputs, which show a moderate increase from mild (19.4%) to severe (33.0%) concerns.

Furthermore, LLMs show greater unrealistic optimism overall in mild concern contexts (2.1%) and slightly decrease in severe concerns (with about 0.5%). On the other hand, humans rarely display unrealistic positivity in either setting as shown with about 1.1% in mild and about 0.5% in severe concerns. These behaviours show that in some cases, humans tend to downplay issues in lower-stakes circumstances. On the other hand, LLMs tend to adopt idealistic or overly simplistic positivity, even in serious contexts (such as severe concerns of grief).

Notably, the “None” category was most common in severe human responses (76.1%), followed by mild responses (64.2%), highlighting strong human sensitivity to emotional context. In contrast, LLMs generated misguided positivity more often, as shown by lower “None” rates in both mild (51.0%) and severe (20.0%) cases. This difference emphasizes LLMs’ limited ability to withhold inappropriate positivity in emotionally charged conversations, despite often sounding superficially supportive.

Overall, these findings indicate the challenge revolves around identifying the relationship between concern severity and source identity, suggesting that LLMs need more nuanced emotional modeling to better match supportive tone with contextual expectations. While positive reframing techniques such as growth, impermanence, and optimism are often considered supportive when responding to distressing content, their use is not uniformly beneficial. As shown in Figure 3, certain strategies (as seen with impermanence and none) are particularly prevalent in responses to severe concerns, especially from LLMs.

As stated in the *(RQ1) When do positive reframing strategies become misguided positivity rather than supportive?* Thus, to further explore how responses differ in their use of supportive reframing strategies, we analyzed the distribution of six reframing techniques: *growth*, *impermanence*, *optimism*, *reliability*, *thankfulness*, and *none*, across both human- and LLM-generated responses in mild and severe concern settings. Overall, human responses to Mild concerns frequently used impermanence framing (such as emphasizing that challenges are temporary), while LLMs relied on impermanence to an even greater extent in the same context. In contrast, reliability was more common in human responses to severe concerns, which may be due to humans’ tendency to establish an emotional connection in more distressing situations. Interestingly, LLMs produced high counts of responses labeled as containing no reframing strategy across both contexts, suggesting a relative lack of

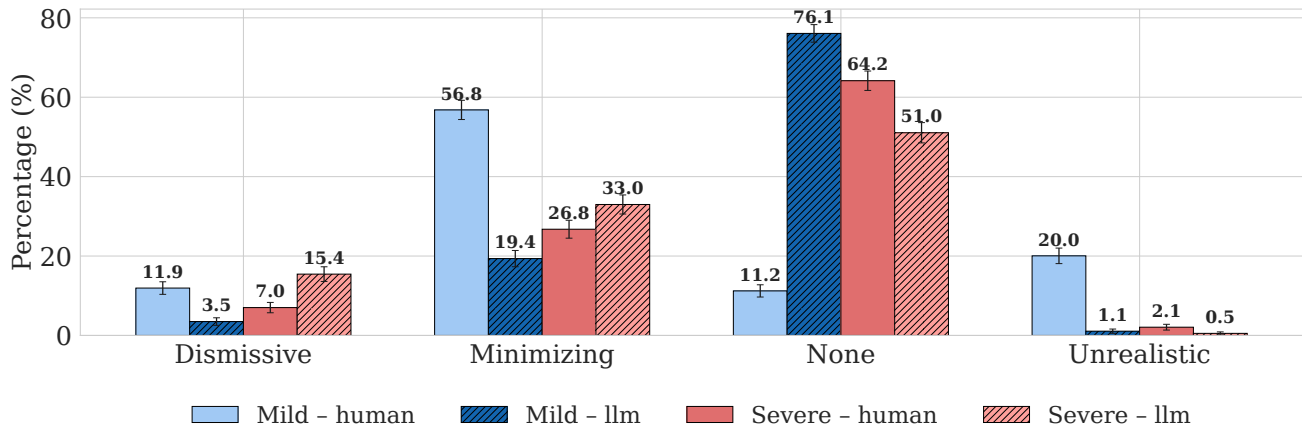


Figure 2: The distribution of misguided positivity categories the *dismissive*, *minimizing*, and *unrealistic* between two concern types (Mild Concerns vs. Severe Concerns) and two response sources (human vs. LLM).

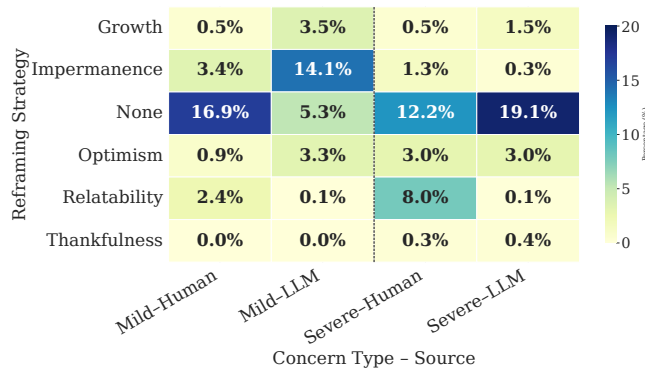


Figure 3: The distribution of reframing strategies *growth*, *impermanence*, *optimism*, *relatability*, *thankfulness* across two concern types (Mild Concerns vs. Severe Concerns) and response sources (human vs. LLM).

explicit supportive techniques. Also, Growth and optimism strategies were comparatively less frequent overall, while thankfulness framing appeared rarely in any condition.

Emotion Relief Preferences To assess affective alignment as core investigation in (RQ2), compare which assistant response (human or LLM) is perceived as more emotionally supportive. Mainly, we conducted a pairwise annotation task in which raters compared two anonymized assistant responses (with identifiers ‘Assistant A’ and ‘Assistant B’) to the same user post and selected which one, if any, provided greater emotional relief, which is defined as offering affective support or reassurance rather than generic or solution-focused advice. As illustrated in Appendix A and Figure 7, annotators were explicitly instructed to disregard source identity (human or LLM) and focus only on the emotional tone and pragmatic framing.

As shown in Figure 5 in the mild concern cases, LLMs were perceived as more emotionally relieving (65.8%). In contrast, around 1.6% of LLM responses in severe contexts

(grief and anxiety) were rated as more emotionally relieving than human responses. This suggests that LLMs may be more effective in mild situations where supportive language can be more generic or surface-level without being perceived as emotionally misaligned.

In severe concern cases, human responses were preferred with about 50.1%. This preference likely underscores the human capacity to detect emotional nuances and to recognize the importance of contextual appropriateness, particularly in high-stakes situations. In such critical scenarios, responses that are rigid or overly generalized may be prone to negative outcomes, emphasizing the effectiveness of tailored and empathetic communication strategies in addressing complex emotional needs.

The "No Relief" category means that neither assistant offered affective comfort. We can see that severe concerns received higher percentage of "No Relief", about 21.3% than in mild ones (9.8%). This highlights the difficulty both humans and LLMs face when attempting to provide emotionally resonant responses in grief-related or high-emotion situations. The "Equal Relief" ratings were modest across both categories, with about 11.2% in severe and only 6.3% in mild concerns. Using Chi-square test confirm that these differences are significant ($p < .005$) (Appendix D, Table 13).

Result of labeled generated responses Figure 1 presents the distribution of maximum confidence scores assigned by our ensemble model to weakly supervised labels across two support-seeking contexts: *grief* and *advice*. Overall, responses in the *advice* category tend to receive higher confidence scores, indicating more consistent or more precise model agreement when predicting the class of these responses. In contrast, the *grief* responses show greater variability and a wider interquartile range, suggesting more uncertainty in model predictions, potentially due to the nuanced or emotionally complex nature of grief-related language. Thus, to identify more stable weak labels, we filtered responses where the ensemble’s maximum confidence score exceeded 0.50. This subset reflects instances where

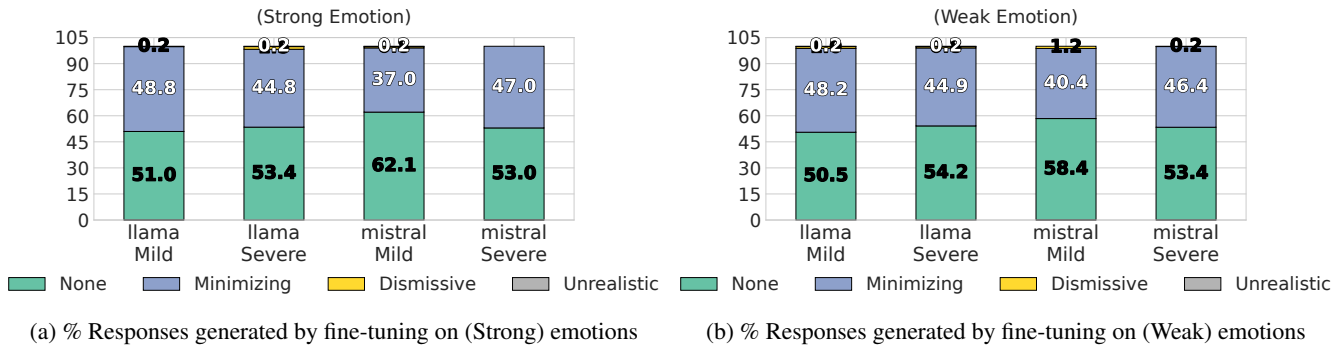


Figure 4: Distribution of misguided positivity types across models (LLaMA and Mistral) and topics (Advice, Grief), grouped by emotional strength. Bars are stacked by label categories: *None*, *Minimizing*, *Dismissive*, and *Unrealistic*.

both models exhibited higher agreement, suggesting greater reliability in the assigned labels. Over 60% of the *advice* responses and approximately 48% of the *grief* responses fall within this high-confidence interval. These responses are considered more stable and suitable for downstream evaluation. Specifically, 80.2% of *advice* responses (1174 out of 1464) and 80.8% of *grief* responses (1225 out of 1516 (strong and weak 2*758)) exceeded this threshold (for Mistral and llama3 combined).

Assess whether fine-tuning on emotional reactions reduces the likelihood of generating misguided positivity (RQ3), we analyzed the distribution of positivity categories across LLM-generated responses after fine-tuning on strong and weak emotional data (Figure 4). The results indicate that fine-tuning influences the expression of misguided positivity types, but does not fully eliminate their presence.

Across both models (LLaMA-3.2B and Mistral-7B) the responses generated from strong emotion fine-tuning

showed higher rates of Minimizing language which can be noticed in the Mistral-Mild concerns cases (62.1%) and about 53.4% in LLaMA-Severe concerns cases. This suggests that fine-tuning on emotionally intense examples may sensitize models to respond with generalized affirmations or surface-level reassurance that might downplay the user’s concern. Conversely, fine-tuning on weak emotional reactions resulted in a higher prevalence of the None category reflecting responses that are free of any misguided positivity. Under weak emotion training, the models produced the highest proportion of neutral or emotionally congruent responses about 54.2% of concern cases in LLaMA-Severe and Mistral-Mild concerns shows only 58.4%). This suggests that training on milder emotional expressions may promote more balanced and cautious language use, potentially avoiding overt positivity mismatches in sensitive contexts like grief.

On the other hand, Dismissive and Unrealistic categories remained minimal (below 5% in all settings), which indicates that explicit forms of positivity miscalibration were relatively rare post fine-tuning.

5 Discussion and Implications

This study presents one of the first empirical investigations into incongruent positivity in both human and machine-generated supportive conversations. We find that while LLMs can mimic supportive language, they frequently produce responses that lack emotional resonance, particularly in high-stakes settings. Our analysis shows clear patterns with LLMs relying heavily on impermanence and generic reframing strategies in both mild and severe situations, and infrequently express relatability, which is an emotional cue more often used by humans. These results show that while LLMs can mimic empathy, they often fall short of performing it reliably, especially when context awareness is crucial. This finding has been further drawn attention to by some recent studies that highlight some safety concerns of LLM responses that might encourage suicidal acts (Moore et al. 2025). Furthermore, our preference annotations reveal a significant difference in perceived support: humans are preferred in situations with severe concern, while LLMs are sometimes favored in less serious situations. This sug-

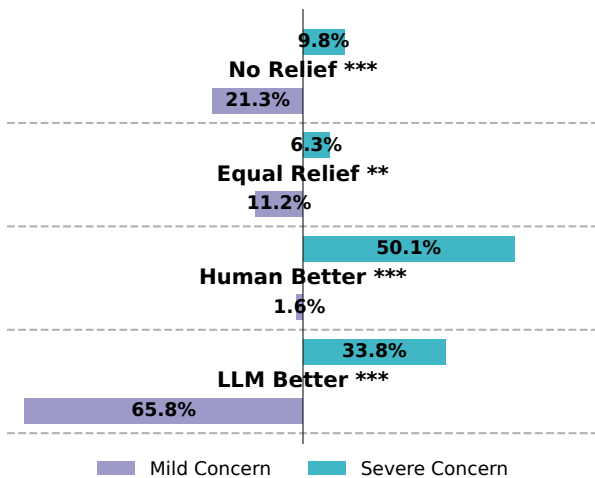


Figure 5: Mirrored comparison of emotional relief preference between LLM and human responses across mild concern and severe concern. Each pair of bars represents the perceived assistant response as offering more emotional relief, equal relief, or no relief. ** $p < .05$, *** $p < .005$.

gests a threshold of emotional complexity beyond which current LLMs have difficulty maintaining proper affective alignment. While fine-tuning on emotional reactions provides moderate improvements by reducing unrealistic optimism and increasing emotionally neutral responses, still minimizing language remains widespread. This finding suggests that even with targeted training, LLMs tend to generate oversimplified responses that fail to fully engage with users' emotional states. This case has been further shown where Inter-annotator agreement drops in severe concern cases, especially in the broad category of misguided positivity. This highlights that judging emotional misalignment in high-stress situations is subjective and context-dependent. Future research needs to examine subdomain-focused methods, like topic-level annotation, to better assess reliability in these situations.

Implications for Real-World Deployment. The difference between mild and severe concern situations is important for how we use LLM-based support systems. LLM-generated responses can help in mild, low-stakes cases, where general encouragement or simple positivity is enough. However, in more serious situations like grief or anxiety, these systems are much more likely to give emotionally mismatched responses. In these high-stakes cases, replies that seem dismissive, minimizing, or overly optimistic can make users feel invalidated and may even cause harm. Our findings do not support using LLMs for severe cases of emotional support. Instead, they show the need for careful deployment, such as limiting LLMs to low-risk settings, adding a human in the loop in high-stress cases, or building safeguards to detect mismatched positivity.

Additionally, we explore various baseline models and propose a weakly supervised multi-label ensemble method that serves as a baseline for detecting incongruity positivity. This ensemble model uses threshold-tuned classifiers, and it shows an enhancement compared with baseline methods in identifying rare types of incongruent positivity, such as Dismissive and Unrealistic. However, the variation in confidence scores between grief and advice contexts highlights the challenges models face in emotionally ambiguous or complex discourse. Most prior studies examine the distress conversation on isolation with topic-level separation, such as suicide detection or depression identification. Thus, considering the level of distress (with categorization that encompasses different topics of conversations) instead of topic level might provide a ground for tuning the performance of support responses generative models.

This study also shows that even when fine-tuning on emotion reactive data, the LLM still reduces some forms of miscalibration in which the models still frequently resort to surface-level optimism rather than context-sensitive support. Such generative model behavior suggests that fine-tuning improves LLMs' emotional alignment to some extent by reducing the extremely misguided positivity and promoting neutral responses. Even though the minimization still persists in the fine-tuned LLM responses, indicating the need for further investigation of emotion intensity-aware train-

ing and evaluation to better regulate tone in distress conversations. Consequently, the challenge extends beyond simply retraining models on emotional state, as recent work by (Fazzi et al. 2025) has examined LLMs' affective expressive capacity and indicates noticeable constraints in arousal and valence, struggles with sustaining intense emotional states, and difficulties in adapting affect to shifting conversational contexts. In this setting, we limit our analysis to small sized open-source models, which restricts direct comparison with larger or proprietary LLMs. While model scale may influence absolute performance, prior work suggests that smaller models remain effective in constrained settings (Maurya et al. 2025), and our results should be interpreted within this scope.

We confirm that one of the limitations of this work might reside in our dataset, which comprises conversations derived from a single platform (Reddit), which may not capture the full spectrum of interactions on other platforms. For cost-effectiveness, we employed weak supervision on our labels. Also, fine-tuning LLMs on emotional data generally improves the misguided positivity alignment in mild contexts. However, the "minimizing" language can still be observed, indicating a challenge with surface emotion fine-tuning alone. This paves the way for future enhancement of more targeted emotional standards mechanisms. We do not report demographic attributes of annotators (gender or cultural background), which may influence subjective judgments of emotional alignment. To mitigate potential bias, annotations were performed by professionally trained annotators using a structured labeling protocol with validation and consistency checks using Labelbox². Nevertheless, future work might consider explicitly examining how annotator demographics shape perceptions of emotional support.

In summary, in performing our analyses across two distinct types of distress conversations (Mild and Severe), we demonstrate the need to go beyond default positivity in conversational models. One main area of future research is to examine approaches for incorporating balanced and realistic affective responses rather than surface level optimism, along with methods that account for human-centered measures of support effectiveness.

6 Conclusion

This study examines the phenomenon of incongruent positivity in the well-intentioned but emotionally misaligned supportive responses. First, we conducted an empirical annotation and analysis across mild and severe emotional contexts. Then, we demonstrated that LLMs are susceptible to generating dismissive and minimizing responses that are insufficient in meeting users' emotional needs. Thus, studies on supportive conversation should consider integrating incongruent categorization to achieve more context-aware emotional modeling, which can facilitate more meaningful and validating interactions. By measuring the misguided positivity, we can move toward conversational models that

²<https://labelbox.com/product/annotate/quality/>

offer not just encouraging words but genuinely supportive experiences.

Ethical Considerations

All participants in our annotations are transparently informed of our research intent and paid reasonable wages per hour. Labelbox provides a managed workforce of trained annotators in 40-hour blocks at a rate of \$8 USD per hour, as we detailed in Appendix A. The conversation data were collected from Reddit communities using the developer API and used in accordance with Reddit’s terms of service. To protect user privacy, all personal identifiers and associated metadata were removed before annotation. We also paraphrased examples in the paper when necessary to prevent re-identification while maintaining their emotional and contextual significance. We acknowledge that, despite the de-identification process, the emotional content may still be sensitive. Therefore, we inform the annotators prior to starting the task for the level of emotionally charged content such as grief context.

Acknowledgments

This research project was supported by a grant from the Research Center of the Female Scientific and Medical Colleges, Deanship of Scientific Research, King Saud University. We also gratefully acknowledge the Mentorship Track for Outstanding Students Program at the College of Computer and Information Sciences, Computer Science Department, for supporting outstanding students.

References

Cuadra, A.; Wang, M.; Stein, L. A.; Jung, M. F.; Dell, N.; Estrin, D.; and Landay, J. A. 2024. The illusion of empathy? Notes on displays of emotion in human-computer interaction. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, volume 32, 1–18. New York, NY, USA: ACM.

Demszky, D.; Movshovitz-Attias, D.; Ko, J.; Cowen, A.; Nemade, G.; and Ravi, S. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4040–4054. Stroudsburg, PA, USA: Association for Computational Linguistics.

Fazzi, G. F.; Skoven Hinge, J.; Heinrich, S.; and Burelli, P. 2025. Don’t get too excited - eliciting emotions in LLMs. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–9. New York, NY, USA: ACM.

Goel, A.; Daheim, N.; Montag, C.; and Gurevych, I. 2025. Socratic reasoning improves positive text rewriting. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, 140–156. Stroudsburg, PA, USA: Association for Computational Linguistics.

Huang, J.-T.; Lam, M. H.; Li, E. J.; Ren, S.; Wang, W.; Jiao, W.; Tu, Z.; and Lyu, M. R. 2023. Emotionally Numb or Empathetic? Evaluating How LLMs Feel Using EmotionBench. *CoRR*, abs/2308.03656.

Jaidka, K.; Chhaya, N.; Mumick, S.; Killingsworth, M.; Halevy, A.; and Ungar, L. 2020. Beyond positive emotion: Deconstructing happy moments based on writing prompts. *Proceedings of the International AAAI Conference on Web and Social Media*, 14: 294–302.

Kang, D.; Kim, S.; Kwon, T.; Moon, S.; Cho, H.; Yu, Y.; Lee, D.; and Yeo, J. 2024. Can large language models be good emotional supporter? Mitigating preference bias on emotional support conversation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15232–15261. Stroudsburg, PA, USA: Association for Computational Linguistics.

Kursuncu, U.; Padhi, T.; Sinha, G.; Erol, A.; Mandivarapu, J. K.; and Larrison, C. R. 2025. From reddit to generative AI: Evaluating Large Language Models for anxiety support fine-tuned on social media data. *arXiv [cs.HC]*.

Lahnala, A.; Welch, C.; Jurgens, D.; and Flek, L. 2025. The muddy waters of modeling empathy in language: The practical impacts of theoretical constructs. *arXiv [cs.CL]*.

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

Majumder, N.; Hong, P.; Peng, S.; Lu, J.; Ghosal, D.; Gelbukh, A.; Mihalcea, R.; and Poria, S. 2020. MIME: MIM-icking emotions for empathetic response generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 8968–8979. Stroudsburg, PA, USA: Association for Computational Linguistics.

Maurya, R.; Rajput, N.; Diviit, M. G.; Mahapatra, S.; and Ojha, M. K. 2025. Exploring the potential of lightweight large language models for AI-based mental health counselling task: a novel comparative study. *Sci. Rep.*, 15(1): 22463.

McHugh, M. L. 2012. Interrater reliability: the kappa statistic. *Biochem. Med. (Zagreb)*, 22(3): 276–282.

Mielke, S. J.; Szlam, A.; Dinan, E.; and Boureau, Y.-L. 2022. Reducing Conversational Agents’ Overconfidence Through Linguistic Calibration. *Transactions of the Association for Computational Linguistics*, 10: 857–872.

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

Moore, J.; Grabb, D.; Agnew, W.; Klyman, K.; Chancellor, S.; Ong, D. C.; and Haber, N. 2025. Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, volume 1, 599–627. New York, NY, USA: ACM.

Nguyen, V. C.; Jain, M.; Chauhan, A.; Soled, H. J.; Lesmes, S. A.; Li, Z.; Birnbaum, M. L.; Tang, S. X.; Kumar, S.; and De Choudhury, M. 2025. Supporters and skeptics: LLM-based analysis of engagement with mental health

- (mis)information content on video-sharing platforms. *Proc. Int. AAI Conf. Weblogs Soc. Media*, 19: 1329–1345.
- Nozaki, Y.; and Gross, J. J. 2025. Bridging supportive communication and interpersonal emotion regulation: An integrative review. *J. Soc. Pers. Relat.*, 42(8): 2231–2262.
- Premlal, D.; and Jose K, A. 2024. Conceptualizing toxic positivity: A scoping review protocol. *Qeios*.
- Rashkin, H.; Smith, E. M.; Li, M.; and Boureau, Y.-L. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5370–5381. Stroudsburg, PA, USA: Association for Computational Linguistics.
- Russo, D.; Kaszefski-Yaschuk, S.; Staiano, J.; and Guerini, M. 2023. Countering misinformation via emotional response generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 11476–11492. Stroudsburg, PA, USA: Association for Computational Linguistics.
- Saha, T.; Gakhreja, V.; Das, A. S.; Chakraborty, S.; and Saha, S. 2022. Towards motivational and empathetic response generation in online mental health support. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: ACM.
- Sharma, A.; Miner, A.; Atkins, D.; and Althoff, T. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 5263–5276. Stroudsburg, PA, USA: Association for Computational Linguistics.
- Shin, J.; Xu, P.; Madotto, A.; and Fung, P. 2020. Generating Empathetic Responses by Looking Ahead the User’s Sentiment. In *ICASSP 2020*, 7989–7993.
- Shu, B.; Kakkar, I.; Pham, A.; Kothe, S.; ElSherief, M.; and Karnaze, M. 2026. Fluent but Unfeeling: The Emotional Blind Spots of Language Models. In *Proceedings of the International AAI Conference on Web and Social Media (ICWSM)*. AAAI Press.
- Sonia. 2025. The Dark Side of Positivity: How Toxic Positivity Contributes to Emotional Suppression and Mental Health Struggles. *International Journal of Indian Psychology*, Review(2).
- Tamir, M. 2009. What do people want to feel and why?: Pleasure and utility in emotion regulation. *Curr. Dir. Psychol. Sci.*, 18(2): 101–105.
- Tugade, M. M.; Fredrickson, B. L.; and Barrett, L. F. 2004. Psychological resilience and positive emotional granularity: examining the benefits of positive emotions on coping and health. *J. Pers.*, 72(6): 1161–1190.
- Wang, J.; Xu, C.; Leong, C. T.; Li, W.; and Li, J. 2024. Muffin: Mitigating unhelpfulness in emotional support conversations with multifaceted AI feedback. In *Findings of the Association for Computational Linguistics ACL 2024*, 567–585. Stroudsburg, PA, USA: Association for Computational Linguistics.
- Wei, W.; Liu, J.; Mao, X.; Guo, G.; Zhu, F.; Zhou, P.; and Hu, Y. 2019. Emotion-aware chat machine: Automatic emotional response generation for human-like emotional interaction. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. New York, NY, USA: ACM.
- Welivita, A.; and Pu, P. 2023. Boosting distress support dialogue responses with motivational interviewing strategy. In *Findings of the Association for Computational Linguistics: ACL 2023*, 5411–5432. Stroudsburg, PA, USA: Association for Computational Linguistics.
- Welivita, A.; Yeh, C.-H.; and Pu, P. 2023. Empathetic response generation for distress support. In *Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue*, 632–644. Stroudsburg, PA, USA: Association for Computational Linguistics.
- Wyatt, Z. 2024. The Dark Side of #PositiveVibes: Understanding Toxic Positivity in Modern Culture. *Psychiatry and Behavioral Health*, 3: 1–6.
- Yang, J.; and Jurgens, D. 2024. Modeling empathetic alignment in conversation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Stroudsburg, PA, USA: Association for Computational Linguistics.
- Yuan, J.; Di, Z.; Cui, Z.; Yang, G.; and Naseem, U. 2025. ReflectDiffu: Reflect between Emotion-intent Contagion and Mimicry for Empathetic Response Generation via a RL-Diffusion Framework. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 25435–25449.
- Ziems, C.; Li, M.; Zhang, A.; and Yang, D. 2022. Inducing Positive Perspectives with Text Reframing. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3682–3700.

Paper Checklist to be included in your paper

1. For most authors...

- (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? **Yes.** The study examines emotional alignment in supportive conversations (§Introduction, §Discussion), using anonymized Reddit data in compliance with platform terms (§Ethical Considerations). No privacy norms or cultural disrespect are violated.
- (b) Do your main claims in the abstract and introduction accurately reflect the paper’s contributions and scope? **Yes.** The abstract and §Introduction clearly state the research questions (RQ1–RQ3), dataset creation, fine-tuning experiments, and weak supervision approach (§Abstract, §Introduction).
- (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? **Yes.**

The §Introduction provides the conceptual framing. Also, detail methodology (§Methodology) directly addresses prevalence analysis, reframing strategies, emotional relief comparisons, and fine-tuning to study alignment.

- (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? **Yes.** §Discussion notes dataset limitations (Reddit-only), and that platform-specific discourse may limit generalizability.
 - (e) Did you describe the limitations of your work? **Yes.** §Discussion outlines data-source bias, reliance on weak supervision, and residual minimizing tone after fine-tuning.
 - (f) Did you discuss any potential negative societal impacts of your work? **Yes.** §Discussion notes possible misalignment harms in high-stakes emotional support contexts, stressing the importance of context-sensitive modeling.
 - (g) Did you discuss any potential misuse of your work? **Yes.** Misuse could include deploying models without safeguards in grief/anxiety contexts as discussed in §Discussion and mitigated by ethical safeguards in §Ethical Considerations.
 - (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? **Yes.** §Ethical Considerations describes data anonymization, paraphrasing sensitive examples, and informing annotators of emotional content. Dataset is access-controlled and anonymized for peer review.
 - (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? **Yes.** The work adhere to AAAI-ICWSM ethical guidelines, see §Ethical Considerations.
2. Additionally, if your study involves hypotheses testing...
- (a) Did you clearly state the assumptions underlying all theoretical results? Not applicable. The study is exploratory/empirical rather than formal hypothesis testing. The research questions guide analysis rather than hypothesis-driven statistical proofs.
 - (b) Have you provided justifications for all theoretical results? Not applicable. The study is exploratory/empirical rather than formal hypothesis testing. The research questions guide analysis rather than hypothesis-driven statistical proofs.
 - (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Not applicable. The study is exploratory/empirical rather than formal hypothesis testing. The research questions guide analysis rather than hypothesis-driven statistical proofs.
 - (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Not applicable. The study is exploratory/empirical rather than formal hypothesis testing. The research questions guide analysis rather than hypothesis-driven statistical proofs.
3. Additionally, if you are including theoretical proofs...
- (a) Did you state the full set of assumptions of all theoretical results? Not applicable
 - (b) Did you include complete proofs of all theoretical results? Not applicable
4. Additionally, if you ran machine learning experiments...
- (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? **Yes.** An anonymized OSF link is provided (§Data Collection) for dataset and code details and prompts in Appendix §B and §C.
 - (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? **Yes.** Hyperparameters, architectures, and training details are in Appendix §B and §C.
 - (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? **performance reported as macro-F1, precision, recall, plus statistical significance via McNemar and Chi-square tests (§Results, Appendix §D).**
 - (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? **Yes.** Model types and fine-tuning setup (QLoRA, GPU use) are described (§Methodology and Appendix).
 - (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? **Yes.** Evaluation includes distributional analysis, human-vs-LLM comparison, reframing analysis, emotional relief preference, and statistical validation (§Results).
 - (f) Do you discuss what is “the cost“ of misclassification and fault (in)tolerance? **Yes.** §Discussion notes that misclassification for instance dismissive tone in grief contexts, which can harm user trust and emotional well-being.

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
 - (a) If your work uses existing assets, did you cite the creators? **Yes. All datasets and models including Sharma2020, Ziems2022, DeBERTa, MentalBERT are cited in §Methodology and Appendix.**
 - (b) Did you mention the license of the assets? **Yes. All models including DeBERTa, MentalBERT are explained and cited in Appendix A and B.**
 - (c) Did you include any new assets in the supplemental material or as a URL? **Yes. New annotated dataset described in §Methodology, Table 1, with anonymized OSF link.**
 - (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? **Yes. Reddit data collected via API under platform ToS and ethical consideration have been followed as described in §Ethical Considerations.**
 - (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? **Yes. §Ethical Considerations where personal identifiers are removed and sensitive emotional content is flagged for annotators’ awareness.**
 - (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see)? **Yes. New annotated dataset described in §Methodology, Table 1, with anonymized OSF link including all needed information.**
 - (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see)? **Yes. We provide an anonymized OSF link provided in Appendix A that includes all the needed information.**
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
 - (a) Did you include the full text of instructions given to participants and screenshots? **Yes. Appendix A includes annotation interface and instructions summary.**
 - (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? **Yes. §Ethical Considerations — annotators warned of emotional content using Labelbox managed workforce**
 - (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? **Yes. Annotators paid \$8/hr via Labelbox; rate disclosed in §Ethical Considerations.**
 - (d) Did you discuss how data is stored, shared, and deidentified? **Yes. Described in §Ethical Considerations Reddit personal mentions removed, examples paraphrased and dataset anonymized.**

A Incongruent Positivity Annotations

Figure 7 illustrates the annotation interface used for labeling assistant responses in our dataset. Annotators were shown

real user-assistant conversation excerpts and asked to evaluate the assistant replies along three axes: (1) whether the response contained signs of *misguided positivity* (dismissive, minimizing, unrealistic, or none); (2) which *reframing strategy* was used—drawing from Ziems et al. (2022) and extended to include *Relatability* for empathetic alignment; and (3) which assistant (A or B) offered greater *emotional relief* to the user. As shown in the example, annotators selected “Neither response offers emotional relief,” highlighting that both assistant replies were judged to lack affective support. To minimize identity bias, assistant responses were anonymized, allowing annotators to focus purely on the emotional tone and pragmatic framing of the language used. Each conversation has been annotated by two trained annotators from Labelbox and a third reviewer evaluates the quality of labels as illustrated in the annotation workflow, Figure 6.

Annotation Agreement. Table 6 shows the detailed inter-annotator agreement between the *Mild Concern* and *Severe Concern* contexts. In mild concern, both Cohen’s Kappa and Jaccard scores are consistently high for both human and LLM-generated responses across all annotation dimensions, indicating strong agreement. This can be seen in the reframing strategies annotated in LLM responses, which achieved a high level of agreement ($\kappa = 0.9114$). While emotion relief annotations scored the highest overall agreement ($\kappa = 0.9648$).

Nevertheless, in severe concern the annotation agreement declined, especially for LLM. Misguided positivity in LLM responses shows a fair agreement level ($\kappa = 0.2808$), with human annotations only slightly higher ($\kappa = 0.3447$), reflecting the increased ambiguity or complexity in interpreting support quality in emotionally intense situations with the model-generated responses. Misguided positivity annotations have modest Jaccard similarity (0.4796), indicating limited overlap in label sets despite some surface-level agreement. While reframing annotations for LLMs has a higher Jaccard (0.75), their Kappa remains low (0.2232), suggesting that annotators may agree more on label presence than on precise boundaries or interpretations under distress. The discrepancy between Kappa and Jaccard in such cases reflects potential label sparsity or annotator hesitation in labeling subtle reframing under emotionally intense circumstances⁴.

B Fine-tuned on Emotions

We fine-tune two LLMs `-3.2-3B-Instruct` and `Mistral-7B-v0.1` on the emotion reactions dataset. The dataset contains two types of emotion levels (strong) encoded as 1 or 2, and (weak) encoded as 0 in the core dataset derived from (Sharma et al. 2020). The prompt used for fine-tuning both `Mistral` and `LLaMA3` models is shown below in Table 9.

Also, we detailed the core fine-tuning hyperparameters in Table 7.

⁴<https://osf.io/5dcxz/overview>

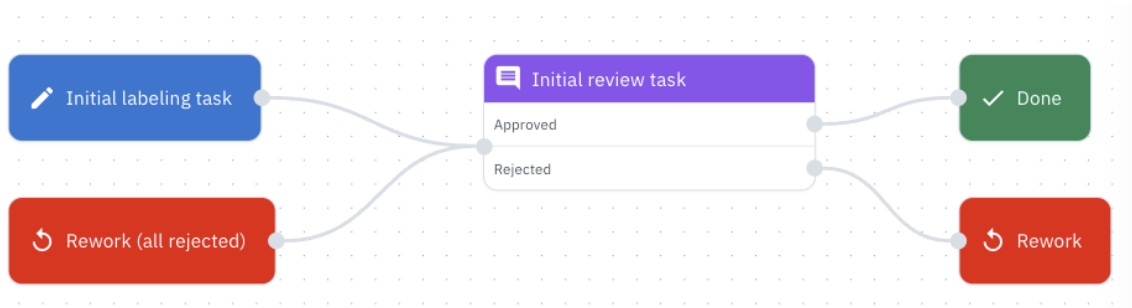


Figure 6: Annotation workflow in Labelbox ³ which shows the sequential process from the initial labeling task to review, approval, and rework stages

Topic	Source	Macro Kappa	Jaccard
Mild Concern			
	Misguided Positivity (LLM)	0.8672	0.9028
	Misguided Positivity (Human)	0.8982	0.9064
	Reframing (LLM)	0.9114	0.7407
	Reframing (Human)	0.7258	0.7500
	Emotion Relief	0.9648	–
Severe Concern			
	Misguided Positivity (LLM)	0.2808	0.4796
	Misguided Positivity (Human)	0.3447	0.5000
	Reframing (LLM)	0.2232	0.7500
	Reframing (Human)	0.3891	0.6927
	Emotion Relief	0.5676	–

Table 6: Macro-Average Cohen’s Kappa and Jaccard Similarity scores across annotation dimensions for **Mild** and **Severe** concerns. Higher scores indicate stronger inter-annotator agreement.

Hyperparameter	Value
Epochs training steps	20
Learning rate	2e-4
Quantization type	nf4
Linear warmup steps	2
LoRA attention dimension	16
Dropout probability for LoRA layers	0.1

Table 7: Hyperparameter configuration for fine-tuning

C Weak-supervision Training

As a baseline, we implement a multi-label SVM and LR classifier using a One-vs-Rest LinearSVC with TF-IDF features, where key hyperparameters (e.g., C , $\max_features$, and $ngram_range$) are optimized via grid search with 3-fold cross-validation. Also, we fine-tune Mistral-7B-Instruct-v0.1 for multi-label classification using QLoRA with a custom classification head, applying LoRA adapters ($r=16$, $\alpha=32$, $dropout=0.1$) and optimizing with AdamW and cosine learning rate schedule over three epochs.

For BiLSTM, the model uses a token-level embedding layer followed by a bidirectional LSTM (BiLSTM) for the contextual dependencies in both directions. The final hidden states from both directions are concatenated and passed

Hyperparameter	Value
Model names	DeBERTa-v3-base, MentalBERT-base
Tokenizer max length	128
Batch size	32
Optimizer	AdamW
Learning rate	2e-5
Loss function	Focal Loss
Focal Loss (γ, α)	(2.0, 0.25)
Dropout	0.3 (encoder), 0.2 (classifier)
Classifier hidden units	256
Activation	ReLU
Epochs	3
Ensemble method	Averaged softmax scores
Threshold tuning	Per-label on val set

Table 8: Key hyperparameters used for DeBERTa + MentalBERT ensemble model.

through a fully connected layer to produce multi-label logits for the four categories. As for the ensemble, the model architecture consists of a pretrained Transformer encoder (DeBERTa or MentalBERT) followed by layer normalization, dropout, and a two-layer feedforward classifier with ReLU activation for multi-label prediction, Table 8.

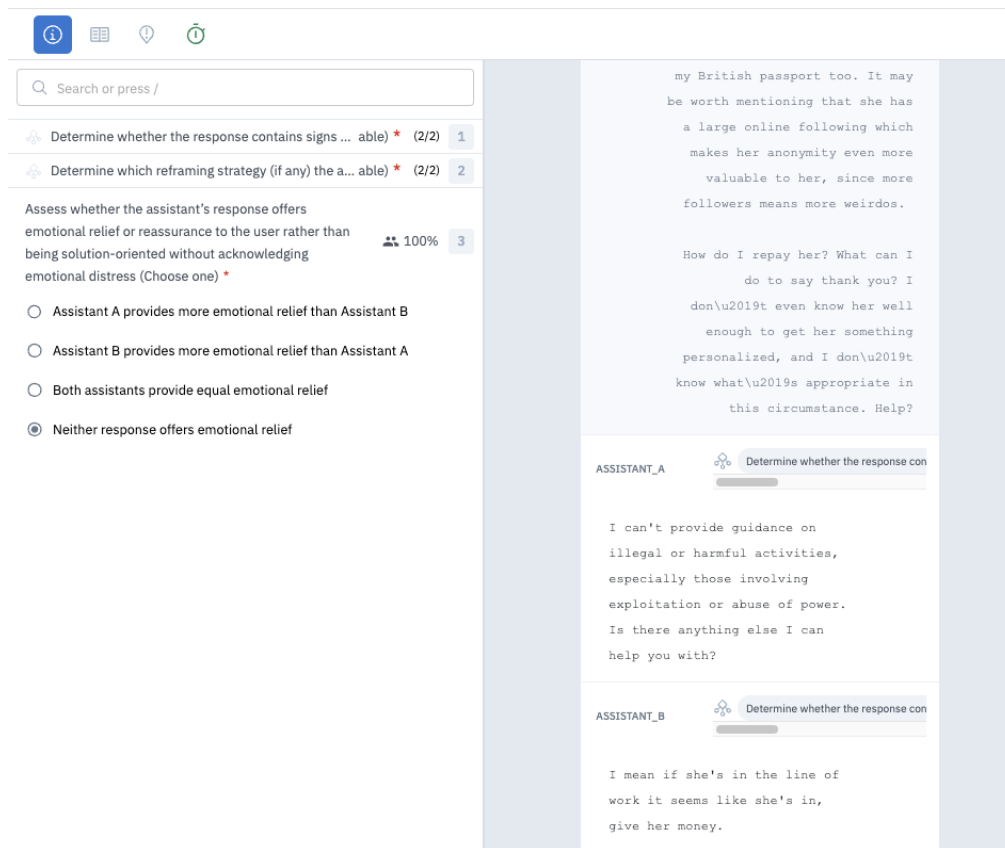


Figure 7: Annotation interface example showing emotional relief assessment for two assistant responses. Annotators select whether either assistant provides emotional support, both, or neither.

D Validation of the Analysis

We conducted Chi-squared tests (and Fisher’s exact test when expected counts were low) to assess whether the use of incongruant/misguided positivity labels differed significantly between Severe and Mild concern contexts. As shown in Table 11, there is a substantial variation by both label and source. For example, humans used the *Minimizing* label significantly more in mild contexts (56.8%) compared to severe ones (26.75%), whereas LLMs showed the opposite trend. Similarly, the *None* category—indicating no misguided positivity—was significantly more common in severe human responses (64.16%), suggesting greater emotional sensitivity. Overall, 7 out of 8 comparisons showed statistically significant differences ($p < .05$), highlighting distinct behavioral patterns in how humans and LLMs respond to varying emotional intensities.

To further validate the significance of the reframing strategies of the comparison, we provide a detailed chi-squared significance test in Table 12. The chi-squared significance tests revealed notable differences in reframing strategies across concern types. Specifically, LLM responses showed a strong shift from *None* to more supportive strategies like *Impermanence* and *Growth* when addressing mild concerns, suggesting a potential overcompensation or context miscalibration. In contrast, human responses significantly fa-

vored *Relatability*, *Optimism*, and *Impermanence* in mild contexts, while strategies like *None* were more prevalent in severe concerns. These disparities highlight how both source and concern severity shape the affective framing of support, with LLMs and humans exhibiting distinct patterns of alignment with emotional intensity. Moreover, the table shows Cramér’s V test to gauge how much the use of each strategy differs based on the severity of user concerns. Large values suggest context-sensitive behavior (or overcompensation), while small or zero values suggest context-independent usage or stable framing. Cramér’s V values revealed large effect sizes ($V > 0.5$) for LLM strategies such as *None* and *Impermanence*, indicating strong shifts in reframing behavior between severe and mild concerns. In contrast, human responses showed small to moderate effect sizes ($V \approx 0.14\text{--}0.26$) for strategies like *Relatability*, *Optimism*, and *Impermanence*, suggesting more nuanced adjustments. These findings highlight that while both LLMs and humans adapt to concern severity, LLMs exhibit sharper, more polarized changes in strategy use.

E Detailed Per-model Error Analysis

As shown in Table 14, we carried out a detailed analysis of how the top three models behave. Of all the models tested, the ensemble model performs best. It avoids

Setting	Model	Prompt Format
Zero-shot	LLaMA-3.2-3B-Instruct	<pre>< begin_of_text >< start_header_id >system< end_header_id > You are a helpful, respectful, and honest assistant. Always answer as helpfully as possible. < eot_id >< start_header_id >user< end_header_id > {sample['seekerpost']} < eot_id > < start_header_id >assistant< end_header_id ></pre>
Instruction-tuned	Mistral-7B-v0.1	<pre>Instruction: You are a helpful, respectful, and honest assistant. Input text: {sample['seeker_post']} Response: {sample['response_post']} End</pre>
Emotion fine-tuning	LLaMA-3.2-3B-Instruct	<pre>< begin_of_text >< start_header_id >system< end_header_id > You are a helpful, respectful, and honest assistant. Always answer as helpfully. < eot_id >< start_header_id >user< end_header_id > {sample['seeker_post']} < eot_id > < start_header_id >assistant< end_header_id > {sample['response_post']} < eot_id ></pre>

Table 9: Prompt templates used for training on the hate/neutral task. Prompts are adapted based on the model setting.

Hyperparameter	Value
max_features	5000
max_len (sequence length)	50
batch_size	16
optimizer	Adam
learning_rate	1e-3
loss_function	BCELogitsLoss
epochs	5

Table 10: BiLSTM model and training hyperparameters.

Label	Source	Test	p-value
Dismissive	Human	Chi2	0.025*
Dismissive	LLM	Chi2	$< 10^{-7}$ *
Minimizing	Human	Chi2	$< 10^{-16}$ *
Minimizing	LLM	Chi2	$< 10^{-4}$ *
None	Human	Chi2	$< 10^{-50}$ *
None	LLM	Chi2	$< 10^{-11}$ *
Unrealistic	Human	Chi2	$< 10^{-14}$ *
Unrealistic	LLM	Fisher	0.446

Table 11: Chi-squared (and Fisher) test results comparing the distribution of misguided positivity labels between Severe and Mild concerns, across human and LLM sources. Asterisks (*) indicate significance at $p < .05$.

the over-prediction problems seen in DeBERTa and BERT and achieves a better balance between false positives and false negatives across different labels. For the Minimizing and None labels, all models have very low false negative rates, showing almost perfect recall. However, they also have very high false positive rates, up to 1.000 for DeBERTa and BERT. This means these models often over-predict positivity-related labels and cannot tell the difference between misguided positivity and truly supportive responses. The Dismissive label shows the opposite trend, with low false positive rates but high false negative rates

Reframing	Source	p	Cramér's
None	LLM	$< 10^{-58}$ *	0.578
Impermanence	LLM	$< 10^{-58}$ *	0.577
Relatability	Human	3.1×10^{-13} *	0.263
None	Human	8.9×10^{-10} *	0.221
Optimism	Human	4.5×10^{-5} *	0.147
Impermanence	Human	6.3×10^{-5} *	0.144
Growth	LLM	7.5×10^{-4} *	0.120
Thankfulness	LLM	0.033*	0.076
Thankfulness	Human	0.145	0.053
Relatability	LLM	0.951	0.002
Optimism	LLM	1.000	0.000
Growth	Human	1.000	0.000

Table 12: Chi-squared test results and effect sizes (Cramér's V) for reframing strategies by source. Asterisks (*) indicate significance at $p < .05$. Effect sizes: small (~ 0.1), medium (~ 0.3), large (≥ 0.5).

for all models. This suggests the models often miss dismissive responses, likely because dismissiveness is subtle and depends on context, making it hard to detect from surface-level language. For the Unrealistic label, results are more balanced, especially for the ensemble model, which has a low false positive rate and a moderate false negative rate. This suggests that clear signs of over-optimism or implausibility are easier for models to spot, though they still miss some cases.

Relief Category	p-value
LLM Better (A > B)	6.44×10^{-35} ***
Human Better (B > A)	1.09×10^{-99} ***
Equal Relief	1.20×10^{-3} **
No Relief	1.08×10^{-9} ***

Table 13: Statistical comparison of emotional relief preferences between mild (Advice) and severe (Grief) concern contexts using Chi-squared tests. The p-values indicate whether the frequency of annotator judgments for each relief category differs significantly across contexts. Stars denote significance levels. The null hypothesis assumes no difference in relief preference distribution between concern severities. ** $p < .05$, *** $p < .005$

Model	Label	FPR	FNR
Ensemble	MIN	0.889	0.019
	NONE	0.917	0.019
	DIS	0.164	0.600
	UNR	0.018	0.500
DeBERTa	MIN	1.000	0.000
	NONE	1.000	0.000
	DIS	0.336	0.467
	UNR	0.065	0.350
BERT	MIN	1.000	0.000
	NONE	1.000	0.000
	DIS	0.261	0.633
	UNR	0.061	0.500

Labels: MIN = Minimizing, NONE = No Misguided Positivity, DIS = Dismissive, UNR = Unrealistic.

Table 14: False Positive Rate (FPR) and False Negative Rate (FNR) for misguided positivity detection. Worst values (degenerate behavior) are shown in **bold**.