AMERICAN
PSYCHOLOGICAL
ASSOCIATION

Emotion

© 2025 American Psychological Association
ISSN: 1528-3542

2025, Vol. 25, No. 7, 1653–1667
https://doi.org/10.1037/emo0001528

Exploring the Potential of Large Language Models to Understand Interpersonal Emotion Regulation Strategies From Narratives

Belén López-Pérez[1], Yuhui Chen[1], Xiuhui Li[1], Shixing Cheng[1], and Pooya Razavi[2, 3]
[1] School of Health Sciences, University of Manchester
[2] Department of Psychology, University of Oregon
[3] Edmentum, Bloomington, Minnesota, United States

Interpersonal emotion regulation involves using diverse strategies to influence others' emotions, commonly assessed with questionnaires. However, this method may be less effective for individuals with limited literacy or introspection skills. To address this, recent studies have adopted narrative-based approaches, though these require time-intensive qualitative analysis. Given the potential of artificial intelligence (AI) and large language models (LLM) for information classification, we evaluated the feasibility of using AI to categorize interpersonal emotion regulation strategies. We conducted two studies in which we compared AI performance against human coding in identifying regulation strategies from narrative data. In Study 1, with 2,824 responses, ChatGPT initially achieved Kappa values over .47. Refinements in prompts (i.e., coding instructions) led to improved consistency between ChatGPT and human coders ($\kappa > .79$). In Study 2, the refined prompts demonstrated comparable accuracy ($\kappa > .76$) when analyzing a new set of responses ($N =$ 2090), using both ChatGPT and Claude. Additional evaluations of LLMs' performance using different accuracy metrics pointed to notable variability in LLM's capability when interpreting narratives across different emotions and regulatory strategies. These results point to the strengths and limitations of LLMs in classifying regulation strategies, and the importance of prompt engineering and validation.

Keywords: interpersonal emotion regulation, artificial intelligence, ChatGPT, Claude, regulation strategies

Supplemental materials: https://doi.org/10.1037/emo0001528.supp

Emotion regulation entails processes aimed at changing emotional responses and how they are expressed (Gross, 2015). While some of these processes focus on managing one's own emotions (intrapersonal emotion regulation), others aim to influence other people's emotions, a process known as interpersonal emotion regulation (Zaki & Williams, 2013). The present research focuses on the latter—the strategies individuals use to regulate the emotions of others (Niven et al., 2009).

In everyday interactions, people shape others' emotions (Tran et al., 2023). For example, parents might comfort an upset child to ease feelings of sadness, or someone may try to calm a colleague who is stressed about an upcoming deadline. Interpersonal emotion regulation (Zaki & Williams, 2013) often encompasses different regulation strategies with distinct degrees of regulatory success (Sahi et al., 2023) and with varying effects on the regulator (Sahi et al., 2025). The classification of regulation strategies differs slightly depending on the theoretical model (Little et al., 2012; Niven et al., 2009). The process model of emotion regulation (Gross, 2015) outlines how individuals manage their emotions through a sequence of strategies that target different stages of the emotional process. It consists of five key stages: *Situation selection* (choosing environments that influence emotions), *situation modification* (changing the environment to alter its emotional impact), *attention deployment* (shifting focus to influence emotional responses), *cognitive change* (reframing a situation to change its emotional meaning), and *response modulation* (adjusting emotional expressions or physiological reactions). Although initially proposed to account for intrapersonal emotion regulation, the model has also been supported in the interpersonal domain (Little et al., 2012; MacCann et al., 2025). A slightly different categorization of regulation strategies is proposed by the interpersonal affect classification (IAC, Niven et al., 2009). This model makes an initial distinction between those strategies aimed at improving and worsening others' emotions. Within the affect improvement strategies (the focus of the current

This article was published Online First April 17, 2025.

Erik C. Nook served as action editor.

Data from Studies 1 and 2 as well as the instructions and scenarios are available at https://osf.io/ptyrk/. The authors thank Amelie Malandre, Aparna Menon, and Maria Muchon for their assistance with the thematic analysis of participants' responses as well as their help with ChatGPT and Claude coding.

Belén López-Pérez played a lead role in writing–original draft and writing–review and editing and an equal role in conceptualization, data curation, and formal analysis. Yuhui Chen played a lead role in formal analysis, a supporting role in writing–original draft and writing–review and editing, and an equal role in conceptualization and data curation. Xiuhui Li played a supporting role in data curation, formal analysis, and writing–review and editing. Shixing Cheng played a supporting role in data curation and formal analysis. Pooya Razavi played a lead role in conceptualization, a supporting role in formal analysis, and an equal role in writing–original draft and writing–review and editing.

Correspondence concerning this article should be addressed to Belén López-Pérez, School of Health Sciences, University of Manchester, Oxford Road, Manchester M139WJ, United Kingdom. Email: belen.lopez-perez@manchester.ac.uk

research) the classification considers the use of *affective engagement* (i.e., talking and providing a listening ear), *cognitive engagement* (i.e., reframing the situation so that the person can see it in a more positive light), *attention* (i.e., spending time with the target or diverting their attention), or *humor* (i.e., making the other person laugh). The present research draws on this theoretical framework as it was one of the first models explicitly developed to explain interpersonal emotion regulation (Niven et al., 2009) and has been previously applied in qualitative research (López-Pérez et al., 2016).

The assessment of regulation strategies has relied mainly on questionnaires, where participants indicated their tendency to use different regulation strategies to change the emotions of others (e.g., Little et al., 2012; López-Pérez et al., 2017; MacCann et al., 2025). However, there are times when the use of self-reports, while being cost-effective, is not the most appropriate or valid approach. The results from self-report questionnaires are dependent on people's ability to introspect (Bowling & Ebrahim, 2005). In addition, the close-ended structure of most questionnaires can be a threat to their external validity as participants are limited to choosing predetermined options that may not necessarily match how they would feel or act (e.g., Budd et al., 1981). The use of open-ended questions that elicit participants' narratives is a viable option to avoid these limitations. Previous research on interpersonal emotion regulation has shown the effectiveness of this approach, especially when working with children (Gummerum & López-Pérez, 2020; Kwon & López-Pérez, 2022; López-Pérez et al., 2016) or as a first step in questionnaire design (Hofmann et al., 2016; Swerdlow & Johnson, 2022) and experimental task development (Mittmann et al., 2021).

An important drawback of working with open-ended narratives is that coding participants' responses can be tedious and time-consuming (Powell & Guadagno, 2008). Further, although narrative coding by humans is considered the "gold standard" in qualitative research for its ability to capture the broader scope of language (Song et al., 2020), the subjective nature of human coders' interpretations can potentially lead to inconsistencies (Nowell et al., 2017). In this context, the emergence of accessible artificial intelligence (AI) might offer a new path for emotion researchers: Can narrative categorization using large language models (LLMs) match, or even surpass, human performance by reducing potential personal biases (Kocoń et al., 2023; Rathje et al., 2024)?

The use of publicly available LLMs, such as ChatGPT, has become increasingly popular due to their capacity to generate creative human-like text based on the input received (Kalla & Smith, 2023). To date, the limited research investigating LLMs' ability to make accurate inferences about emotions has generated mixed results (e.g., Elyoseph & Levkovich, 2023; Kocoń et al., 2023) with the greatest performance in the recognition between positive, negative, and neutral sentiment in written text (Rathje et al., 2024). When it comes to emotion regulation, ChatGPT has demonstrated some promise by accurately describing techniques to manage emotions. Although effective in many cases, it struggles with complex situations involving intricate motivational factors and detailed analyses of emotions, including secondary or mixed emotional responses (Vzorin et al., 2023).

The Present Research

Given the potential of LLMs in interpreting people's spontaneous responses, the current research's goal is to explore LLMs' performance when used to interpret and categorize people's interpersonal emotion regulation strategies. We argue that investigating the potential use of LLMs in the study of regulation strategies is important for different reasons. First, given the clinical implications of interpersonal emotion regulation (e.g., Messina et al., 2021), if LLMs show reasonable accuracy in categorizing emotion regulation strategies, they can potentially have applications in domains such as digital mental health by helping people and practitioners gain insight into interpersonal regulation strategies. Automated classification using LLMs could enhance mental health interventions and real-time support systems where LLMs may offer tailored feedback and help users reflect on their interpersonal emotion regulation strategies. Additionally, valid interpretations by LLMs can provide researchers with an alternative measurement approach where peoples' narratives (e.g., on social media or in conversations in laboratory or mental health settings) can be annotated and measured in terms of emotion regulation content and investigated in relation to other constructs. Overall, with the rapid rate with which LLMs are being incorporated into psychological research and practice, understanding the promises and limitations of these models can inform and calibrate the conclusions we make from LLM-based analyses.

Study 1

In the first study, we used the most commonly used publicly accessible generative LLM chatbot, the Chat Generative Pre-trained Transformer (ChatGPT), to categorize participants' open-ended narratives about regulating other people's emotions in specific scenarios. The aim of Study 1 was threefold. First, to evaluate whether ChatGPT could conduct similar categorizations of interpersonal emotion regulation strategies when compared with human coders. Second, to assess whether ChatGPT's performance could improve when provided with additional instructions and examples. Finally, to study the challenges encountered by ChatGPT in the categorization to learn about the boundaries between different regulation strategies.

To study the categorization, we relied on the IAC (Niven et al., 2009), as it is one of the first accounts that explicitly proposed regulation strategies for the interpersonal domain. This framework has been one of the most acknowledged in interpersonal emotion regulation research across different fields (i.e., developmental psychology, López-Pérez & Pacella, 2021; romantic relationships, Zhang & Chen, 2024; education, Jiang et al., 2023; as well as in the organizational domain, Troth et al., 2018). Specifically, we initially focused on the strategies of *affective engagement* (i.e., listening and talking to the target), *cognitive engagement* (i.e., helping the target to reframe their perspective), *attention* (i.e., valuing the target and taking their mind off the situation), and *humor* (i.e., acting in a humorous way to make the target laugh).

Method

Participants

Participants ($N = 353$, 64% women, $M_{\text{age}} = 19.46$, $SD = 2.14$) from the University of University of Oregon's Psychology and Linguistics subject pool completed the study as an option to fulfill course requirements (IRB# 09072010.006). Participants self-identified as: 66.5% White, 10.5% Asian, 11.7% Hispanic, Latinx, or Spanish Origin, 3.8% Black or African American, 2.3% Native Hawaiian or

Other Pacific Islander, 1.2% Middle Eastern or North African, 0.3% American Indian or Alaska Native, and 1.7% reported "Some other ethnicity or origin". Additionally, 2.0% preferred not to answer.

Materials and Procedure

Participants were presented with nine scenarios in sequential order. Each scenario involved a situation where the protagonist would experience one of three emotions (sadness, anger, or fear; see Supplemental Materials for the full set of scenarios). For each scenario, they were instructed to imagine a protagonist (i.e., either their friend, a family member, or their romantic partner) in the described situation.

Following each scenario, participants reported how they would think the protagonist (e.g., their friend) would feel in this situation using a list of 10 emotions. The emotion list consisted of two emotion terms for the three target emotions, that is, sad and upset to evaluate sadness (average α across scenarios = .75); angry and irritated to assess anger (average α = .78); and fearful and scared to measure fear (average α = .82) and two positive emotions as reference points, that is, calm and peaceful to assess calm (average α = .76); and happy and joyful to evaluate happiness (α = .85). The selection of emotion terms was taken from prior research (Saija et al., 2024; Tamir & Ford, 2012). Participants reported their perception of the protagonist's emotional intensity on a 5-point Likert scale (1 = *Not at all* to 5 = *Extremely*).

These items were used as a test of validity for the scenarios. Specifically, we wanted to ensure the participants' perception of the protagonists' emotions matches the target emotion. For example, in a scenario intended to involve sadness, sadness should have the highest intensity. As shown in Table 1, this was the case for all but one scenario. For one of the fear scenarios (Scenario 1), *calm* had the highest endorsement, suggesting that the participants, on average, did not envision the target (i.e., their father) experiencing fear in the described situation (i.e., a close encounter with a spider). As such, this scenario was removed from further analyses. The remaining scenarios had reasonable validity (see Table 1 for the descriptive results and Supplemental Table S1 for the reliabilities).

Next, using an open-ended text box, participants described, without any time constraints, how they would make the protagonist feel better. The narratives were on average 17.17 words (*SD* = 10.29). Participants' narratives for anger (*M* = 19.49 words, *SD* = 12.26) were significantly longer than for sadness, *M* = 15.96 words, *SD* = 10.42; $t(299) = 7.919$, $p < .001$, $d = .46$, and fear, *M* = 16.17 words, *SD* = 10.64; $t(301) = 7.69$, $p < .001$, $d = .44$. Sadness and fear narratives did not differ in length, $t(298) = -.845$, $p = .199$, $d = -.049$. Data were collected between January to mid-March 2020.

Transparency and Openness Statement

Data from Studies 1 and 2 as well as the instructions and scenarios are available at https://osf.io/ptyrk/

Results

Analytic Approach

Human Coding Procedure. The open-ended responses were coded by two trained research assistants throughout the different steps. The two coders were knowledgeable of the IAC and followed an iterative coding process consistent with thematic analysis (e.g., Boyatzis, 1998; Joffe, 2011). The initial coding system contained the four regulation strategies outlined in the IAC (Niven et al., 2009). As a starting point, the two coders used the definitions and examples contained in the initial coding system. The first step involved coding 60 participants' responses for all the scenarios by two independent coders to assess whether the initial definitions and examples captured all strategies mentioned by participants. After this step, following a discussion among the research team, clarifying revisions were made to the definitions and examples, and a coding manual was established. As part of this revision, we noticed some challenges with the strategy of *attention*, as it could refer to providing attention to someone or distracting them. Hence, we decided to choose distinct labels to differentiate these two meanings: *attention* to refer to instances of "spending time with the target" and "doing things for the target," while *distraction*, to refer to instances of "distracting the other person" (see Supplemental Materials). Subsequently, the two coders coded 103 randomly selected narratives and convened to discuss and reach a consensus. Next, after 30% of the total number of responses (*n* = 847) were coded, we computed interrater agreement for each scenario. The interrater agreement for the different scenarios ranged from 94% to 99%, with Cohen's Kappa values consistently exceeding the .80 threshold (see Table 2). Given the high interrater reliability estimates, the remaining narratives were coded by one of the two coders. Note that coders could assign multiple regulation strategies to a narrative if the narrative reflected more than one strategy. The overall coding process took 13.5 hr for human coders.

Across all scenarios, the number of strategies identified ranged from 0 to 4. The number of strategies in the anger scenarios was significantly higher, followed by sadness, and lastly fear (statistical comparisons across scenarios can be found in Supplemental Materials).

The frequency of use of strategies varied across scenarios (Figure 1; Supplemental Table 2S). Collapsing across all emotions, *affective* and *cognitive engagement* were the most frequent strategies detected in the narratives, whereas *distraction* and *humor* had the lowest frequency.

Table 1
Mean and (Standard Deviation) of the Emotion Intensity Perception for Each Scenario in Study 1

Scenario	Sadness 1	Sadness 2	Sadness 3	Anger 1	Anger 2	Anger 3	Fear 1	Fear 2	Fear 3
Emotion									
Sadness	**3.43** (1.01)	**3.71** (0.99)	**3.57** (0.98)	2.59 (.83)	2.79 (.78)	2.47 (.76)	1.40 (0.74)	1.75 (0.80)	1.99 (0.97)
Anger	2.79 (0.99)	2.61 (1.06)	2.41 (1.03)	**3.53** (.84)	**3.59** (.84)	**3.59** (.99)	1.54 (0.84)	1.76 (0.79)	2.04 (1.03)
Fear	1.48 (0.81)	2.48 (1.14)	1.41 (0.77)	1.64 (.89)	1.31 (.68)	1.32 (.71)	1.96 (1.12)	**3.25** (1.02)	**3.30** (1.14)
Happiness	1.41 (0.75)	1.22 (0.62)	1.27 (0.66)	1.37 (.66)	1.30 (.67)	1.35 (.68)	1.66 (0.92)	1.91 (0.92)	1.93 (1.03)
Calm	1.83 (0.86)	1.55 (0.83)	1.51 (0.79)	1.66 (.72)	1.66 (.73)	1.69 (.82)	**2.51** (1.19)	1.63 (0.81)	1.83 (0.96)

Note. The highest emotion rating for each scenario is highlighted in bold.

Table 2

Cohen's Kappas and Percentage of Agreement for Each Scenario Between Different Coders in Studies 1 and 2

	Human coders (Study 1)	ChatGPT–Human (Study 1)	Human coders (Study 2)	ChatGPT–Human (Study 2)	Claude–Human (Study 2)	Claude–ChatGPT (Study 2)
Scenario	κ (%)	κ (%)	κ (%)	κ (%)	κ (%)	κ (%)
Anger 1	0.97 (98.70%)	0.87 (94.77%)	0.91 (96.15%)	0.84 (93.65%)	0.78 (91.19%)	0.64 (85.37%)
Anger 2	0.86 (93.91%)	0.80 (91.58%)	0.90 (95.69%)	0.90 (95.77%)	0.81 (92.18%)	0.81 (92.03%)
Anger 3	0.97 (98.91%)	0.79 (92.49%)	0.95 (98.30%)	0.81 (93.14%)	0.76 (91.287%)	0.70 (88.87%)
Fear 1	0.86 (95.00%)	0.81 (92.67%)	0.92 (96.40%)	0.78 (91.15%)	0.74 (90.54%)	0.65 (86.31%)
Fear 2	0.93 (97.39%)	0.79 (92.67%)	0.86 (95.00%)	0.79 (92.86%)	0.78 (93.33%)	0.64 (89.10%)
Sadness 1	0.83 (93.70%)	0.84 (93.65%)	0.83 (93.33%)	0.77 (90.86%)	0.77 (91.48%)	0.70 (88.28%)
Sadness 2	0.95 (98.48%)	0.81 (93.64%)	0.91 (97.08%)	0.82 (94.23%)	0.76 (92.77%)	0.70 (90.38%)
Sadness 3	0.85 (95.00%)	0.92 (96.78%)	0.84 (94.29%)	0.76 (90.59%)	0.78 (92.39%)	0.66 (86.90%)
Mean	0.90 (96.39%)	0.83 (93.53%)	0.89 (95.78%)	0.81 (92.78%)	0.77 (91.90%)	0.69 (88.40%)

Note. Human coders = Interrater reliability between two human coders; ChatGPT–human = Interrater reliability between ChatGPT and human coders; Claude–Human = Interrater reliability between Claude and human coders.

This pattern generalized to sadness and fear, but not to anger. For anger, *attention* was the most frequent regulatory strategy, followed by *affective* and *cognitive engagement* (for statistical comparisons see the Supplemental Materials).

ChatGPT Coding Procedure. We used ChatGPT Version 4.0 (OpenAI, 2023) to evaluate participants' narratives. The first round of coding was completed between June and July 2023. Prompt refinement and the second round of coding were conducted between July and September 2023.

The initial coding instructions (i.e., prompts) provided to ChatGPT consisted of instructions developed for human coders, including definitions, goals, and examples for each of the five emotional regulation strategy codes (i.e., *affective engagement, cognitive engagement, distraction, attention*, and *humor;* see Supplemental Materials). Consistent with the recent developments in prompt engineering (i.e., the practice of refining prompts or instructions given to LLMs with the goal of improving the model output; Meskó, 2023), we followed an iterative process that involved sequential modifications and testing to refine the coding instructions, thus increasing the model's coding coherence with human coders (see Figure 2). At each step, we measured the consistency between the codes generated by ChatGPT versus human coders, using Cohen's Kappa (κ) and the agreement percentage (%), to evaluate the effectiveness of the changes in the prompts. Prompts and participants' narratives were entered using the chat box. ChatGPT was required to return the response in a numerical table where each strategy received a value of 0 if it did not feature in the narrative or 1 if it was present in the narrative. As with human coding, ChatGPT could assign multiple regulation strategies to a narrative if the narrative reflected more than one strategy. In addition, ChatGPT was instructed to provide a rationale in text format for the coding of each strategy. This approach allowed us to verify its understanding and ensure that responses were not generated randomly, as recommended by prior research in prompt engineering (White et al., 2023).

ChatGPT's Performance

In Step 1, narratives were coded by ChatGPT using the same instructions (from the original manual) provided to human coders. A comparison of codes generated by ChatGPT and human coders

Figure 1

Frequency of Strategy Use Across Scenarios in Study 1

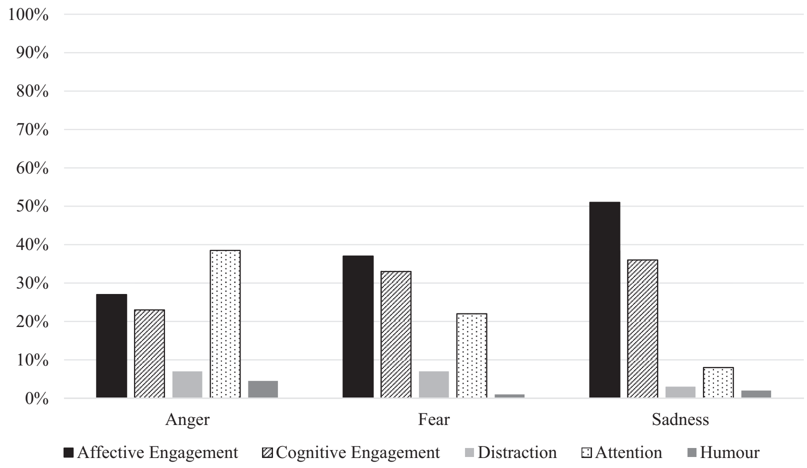
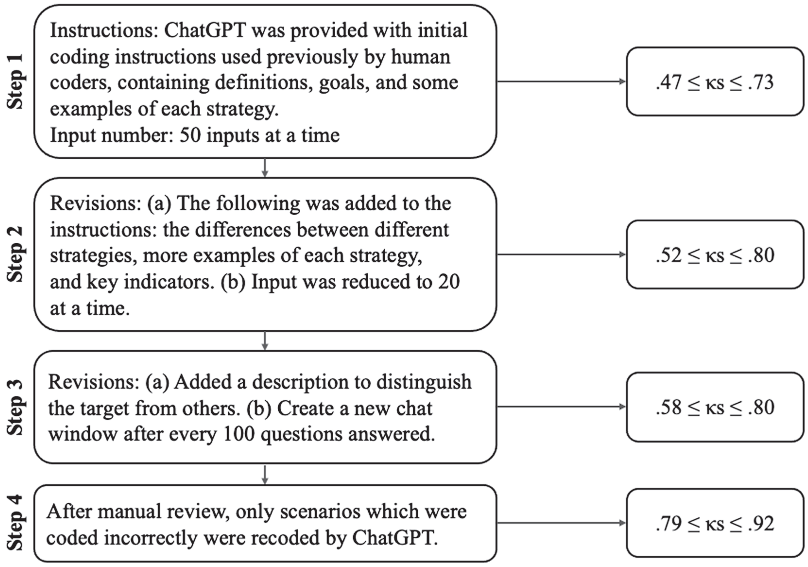


Figure 2

Different Iterations in the Prompts Provided to ChatGPT and Changes in the Coding Reliabilities


Step 1
Instructions: ChatGPT was provided with initial coding instructions used previously by human coders, containing definitions, goals, and some examples of each strategy.
Input number: 50 inputs at a time
$.47 \leq \kappa s \leq .73$
Step 2
Revisions: (a) The following was added to the instructions: the differences between different strategies, more examples of each strategy, and key indicators. (b) Input was reduced to 20 at a time.
$.52 \leq \kappa s \leq .80$
Step 3
Revisions: (a) Added a description to distinguish the target from others. (b) Create a new chat window after every 100 questions answered.
$.58 \leq \kappa s \leq .80$
Step 4
After manual review, only scenarios which were coded incorrectly were recoded by ChatGPT.
$.79 \leq \kappa s \leq .92$

revealed poor consistency between the two ($.47 \leq \kappa s \leq .73$; see Table 3 and Figure 2). Based on a thorough review of the cases where there was a discrepancy between ChatGPT and human coders, in Step 2, we added details such as comparisons between strategies (e.g., the distinction between "affective engagement" vs. "cognitive engagement"), more examples, and key indicators per strategy to the prompts provided to ChatGPT (see Supplemental Materials). We also reduced the number of inputs to 20 per coding session to avoid ChatGPT being overloaded with too many inputs in the same session. These improvements noticeably increased the consistency between ChatGPT and human coders ($.52 \leq \kappa s \leq .80$). During this phase, we noticed that ChatGPT sometimes confused the emotion regulation strategies employed on the "target" individual with those of other people in the scenarios. To address this, in Step 3, we clarified the instructions to better distinguish the target from other individuals in the scenarios (Supplemental Materials). Additionally, to prevent ChatGPT from making errors due to processing too many inputs, we started a new chat window and reentered the instructions after every 100 qualitative responses and kept adding only 20 narratives at a time. These adjustments led to an improvement in categorization accuracy ($.58 \leq \kappa s \leq .80$). This process took 4.71 hr to complete.

The results from Steps 1 to 3 showed the extent to which ChatGPT's performance can be enhanced through prompt refinement. At this stage, approximately 8% of the narratives (i.e., 230 out of 2,824) were misclassified by ChatGPT in nonsystematic ways. Multiple reviews of the misclassified narratives indicated that there was no discernible pattern in the classification errors for these narratives. As such, no additional effective prompt refinement was possible. Figure 3 represents heatmaps that display the proportional correspondence between human and LLM codes for each strategy.

The probabilistic nature of LLMs raises the possibility that the models are not consistent in their misclassifications over multiple attempts. If so, it should be possible to improve the classification accuracy for the incorrectly classified narratives by asking ChatGPT to recode them. To explore this possibility, in Step 4, we used the same prompts from Step 3 to code the misclassified narratives a second time (Supplemental Materials). As anticipated, this recoding process further enhanced the accuracy ($.79 \leq \kappa s \leq .92$) to the extent

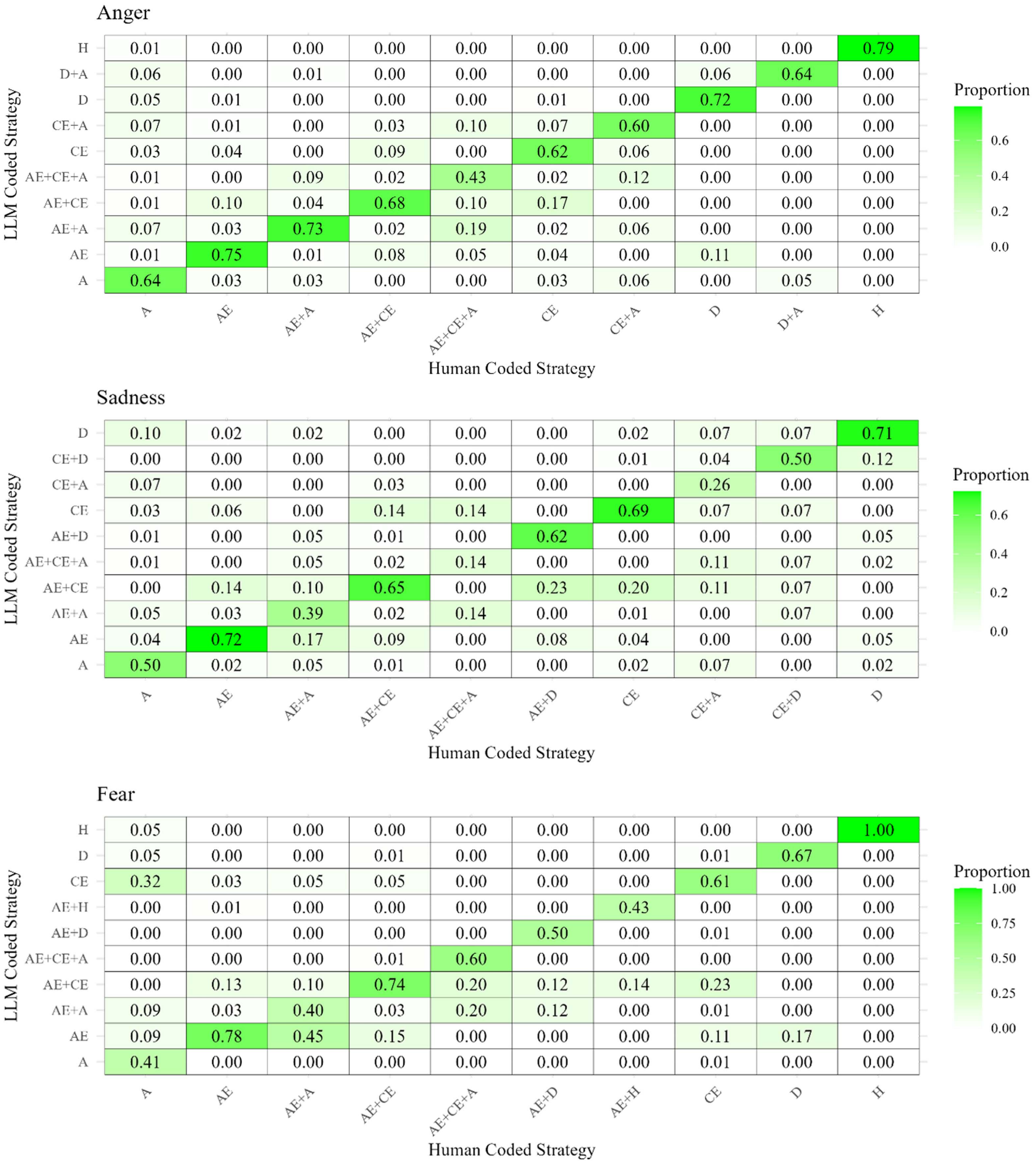
Table 3

Enhancements in Cohen's Kappa After Modified Prompts in Study 1 for ChatGPT

Scenario	Step 1 κ (% of agreement)	Step 2 κ (% of agreement)	Step 3 κ (% of agreement)	Step 4 κ (% of agreement)
Anger 1	0.47 (77.84%)	0.61 (84.40%)	0.71 (88.37%)	0.87 (94.77%)
Anger 2	0.49 (79.23%)	0.52 (79.35%)	0.77 (90.00%)	0.80 (91.58%)
Anger 3	0.49 (83.16%)	0.77 (91.69%)	0.80 (92.69%)	0.79 (92.49%)
Fear 1	0.73 (90.84%)	0.80 (92.73%)	0.74 (90.13%)	0.81 (92.67%)
Fear 2	0.66 (88.84%)	0.77 (92.08%)	0.73 (90.38%)	0.79 (92.67%)
Sadness 1	0.71 (88.65%)	0.74 (89.93%)	0.73(89.12%)	0.84 (93.65%)
Sadness 2	0.58 (85.74%)	0.64 (88.54%)	0.74 (91.02%)	0.81 (93.64%)
Sadness 3	0.57 (83.11%)	0.65 (87.08%)	0.58 (83.92%)	0.92 (96.78%)

Figure 3

Proportional Correspondence Between Human- and LLM-Coded Regulation Strategies



Anger

	A	AE	AE+A	AE+CE	AE+CE+A	CE	CE+A	D	D+A	H
H	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.79
D+A	0.06	0.00	0.01	0.00	0.00	0.00	0.00	0.06	0.64	0.00
D	0.05	0.01	0.00	0.00	0.00	0.01	0.00	0.72	0.00	0.00
CE+A	0.07	0.01	0.00	0.03	0.10	0.07	0.60	0.00	0.00	0.00
CE	0.03	0.04	0.00	0.09	0.00	0.62	0.06	0.00	0.00	0.00
AE+CE+A	0.01	0.00	0.09	0.02	0.43	0.02	0.12	0.00	0.00	0.00
AE+CE	0.01	0.10	0.04	0.68	0.10	0.17	0.00	0.00	0.00	0.00
AE+A	0.07	0.03	0.73	0.02	0.19	0.02	0.06	0.00	0.00	0.00
AE	0.01	0.75	0.01	0.08	0.05	0.04	0.00	0.11	0.00	0.00
A	0.64	0.03	0.03	0.00	0.00	0.03	0.06	0.00	0.05	0.00

Sadness

	A	AE	AE+A	AE+CE	AE+CE+A	AE+D	CE	CE+A	CE+D	D
D	0.10	0.02	0.02	0.00	0.00	0.00	0.02	0.07	0.07	0.71
CE+D	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.04	0.50	0.12
CE+A	0.07	0.00	0.00	0.03	0.00	0.00	0.00	0.26	0.00	0.00
CE	0.03	0.06	0.00	0.14	0.14	0.00	0.69	0.07	0.07	0.00
AE+D	0.01	0.00	0.05	0.01	0.00	0.62	0.00	0.00	0.00	0.05
AE+CE+A	0.01	0.00	0.05	0.02	0.14	0.00	0.00	0.11	0.07	0.02
AE+CE	0.00	0.14	0.10	0.65	0.00	0.23	0.20	0.11	0.07	0.00
AE+A	0.05	0.03	0.39	0.02	0.14	0.00	0.01	0.00	0.07	0.00
AE	0.04	0.72	0.17	0.09	0.00	0.08	0.04	0.00	0.00	0.05
A	0.50	0.02	0.05	0.01	0.00	0.00	0.02	0.07	0.00	0.02

Fear

	A	AE	AE+A	AE+CE	AE+CE+A	AE+D	AE+H	CE	D	H
H	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
D	0.05	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.67	0.00
CE	0.32	0.03	0.05	0.05	0.00	0.00	0.00	0.61	0.00	0.00
AE+H	0.00	0.01	0.00	0.00	0.00	0.00	0.43	0.00	0.00	0.00
AE+D	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.01	0.00	0.00
AE+CE+A	0.00	0.00	0.00	0.01	0.60	0.00	0.00	0.00	0.00	0.00
AE+CE	0.00	0.13	0.10	0.74	0.20	0.12	0.14	0.23	0.00	0.00
AE+A	0.09	0.03	0.40	0.03	0.20	0.12	0.00	0.01	0.00	0.00
AE	0.09	0.78	0.45	0.15	0.00	0.00	0.00	0.11	0.17	0.00
A	0.41	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00

Note. Each heatmap displays the 10 most frequent strategy combinations for the corresponding emotion. Each cell represents the proportion of strategies identified by human coders that the large language models (LLM) either matched (proportions displayed on the top-right to bottom-left diagonal) or mismatched (proportions displayed outside the diagonal). Acronyms for different strategies: AE = affective engagement; CE = cognitive engagement; D = distraction; A = attention; H = humor. See the online article for the color version of this figure.

that there was remarkable agreement (ranging from 91.58% to 96.78%; see Table 3) between emotion regulation strategies coded by ChatGPT and human coders. While these benchmarks are impressive, it is important to note that these results are dependent on manually checking ChatGPT's answers and do not represent performance levels for a fully automated process. Consequently, the analyses in the following section are based on the results obtained after Step 3.

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{True Positives} + \text{True Negatives} + \text{False Positives} + \text{False Negatives}}. \tag{1}$$

Category-Based Performance

Does ChatGPT's classification performance differ for different emotional scenarios and/or regulation strategies? To answer this question, we first calculated the frequency of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) classifications for each emotion regulation strategy combination (see Supplemental Materials). Using these values, we calculated three indices of classification accuracy that are commonly used to assess the performance of clinical tests (e.g., Shreffler & Huecker, 2023) and machine learning algorithms (Equation 1; e.g., Erickson & Kitamura, 2021). Specifically, we calculated a general index of classification accuracy using the equation below:

(see Equation 1 above)

This metric (ranging from 0 to 1) provides a general indication of how often the model correctly classifies the strategies expressed in the narratives. Note that "correct classification" in this case refers to both recognizing a strategy when it is present (TP) and not recognizing a strategy when it is absent (true negative). As shown in Figure 4a, ChatGPT had a generally high classification accuracy ($0.85 \leq$ accuracy ≤ 0.99), with some notable variations across the strategies. To further explore these variations, we calculated two additional metrics.

First, sensitivity was calculated using the following equation:

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}. \tag{2}$$

Sensitivity (ranging from 0 to 1) indicates ChatGPT's ability to detect a specific strategy when the strategy is actually present in the narratives (as determined by the human coders). As shown in Figure 4b, ChatGPT's sensitivity values for all strategies in response to anger ($0.82 \leq$ sensitivity ≤ 0.95) and sadness ($0.84 \leq$ sensitivity ≤ 0.96) were relatively high. Sensitivity estimates for fear scenarios were slightly lower with higher variability ($0.78 \leq$ sensitivity ≤ 0.94) scenarios.

In addition to detecting a strategy when it is present, it is equally important for a model to not detect a strategy when it is absent. To evaluate this aspect, we calculated ChatGPT's specificity using the following equation:

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}. \tag{3}$$

Specificity (ranging from 0 to 1) represents the model's performance in terms of *not* detecting a strategy when it is not present in a narrative. As shown in Figure 4c, specificity values were generally high ($0.81 \leq$ specificity ≤ 0.99) and had a similar pattern for all three emotions. Notably, the results suggest that, among all strategies, the model is more likely to detect *affective engagement* in narratives (where human coders do not detect it). It is also least likely to detect *humor* in its absence (Figure 4c).

Word Count and Its Link With LLM Performance

Given that the length of text can impact LLM's performance (e.g., Liu et al., 2024), we ran some exploratory analyses to determine whether the length of participant narratives, measured by word count, was associated with the LLM's performance in detecting interpersonal emotion regulation strategies compared to human coders. Partial correlations were calculated between word count and four key performance metrics (i.e., TP, TN, FP, and FN) while controlling for the total number of strategies identified by human coders. The results revealed a nuanced relation. Although some significant correlations were identified, suggesting that narrative length may influence LLM performance in specific contexts, many correlations were nonsignificant, indicating that narrative length does not consistently impact LLM accuracy. When significant, the effect sizes were small to moderate (R^2 = 0.03–0.13), suggesting these relations accounted for only a modest proportion of the variance in LLM performance.

Significant correlations were observed primarily for FP and TN for some scenarios, with longer narratives slightly increasing FP and decreasing TN, implying a minor tendency for LLMs to overidentify strategies in lengthier texts. However, the practical significance of these findings is limited due to the modest effect sizes. Conversely, many scenarios, particularly Anger 1 and Anger 3, showed no significant correlations across any performance metrics, and for sadness-related scenarios, significant correlations were mostly limited to TN, with minimal effects. Overall, these results suggest that while narrative length may influence LLM performance in specific situations, its impact is generally small and inconsistent. For a more detailed discussion, see the Supplemental Materials.

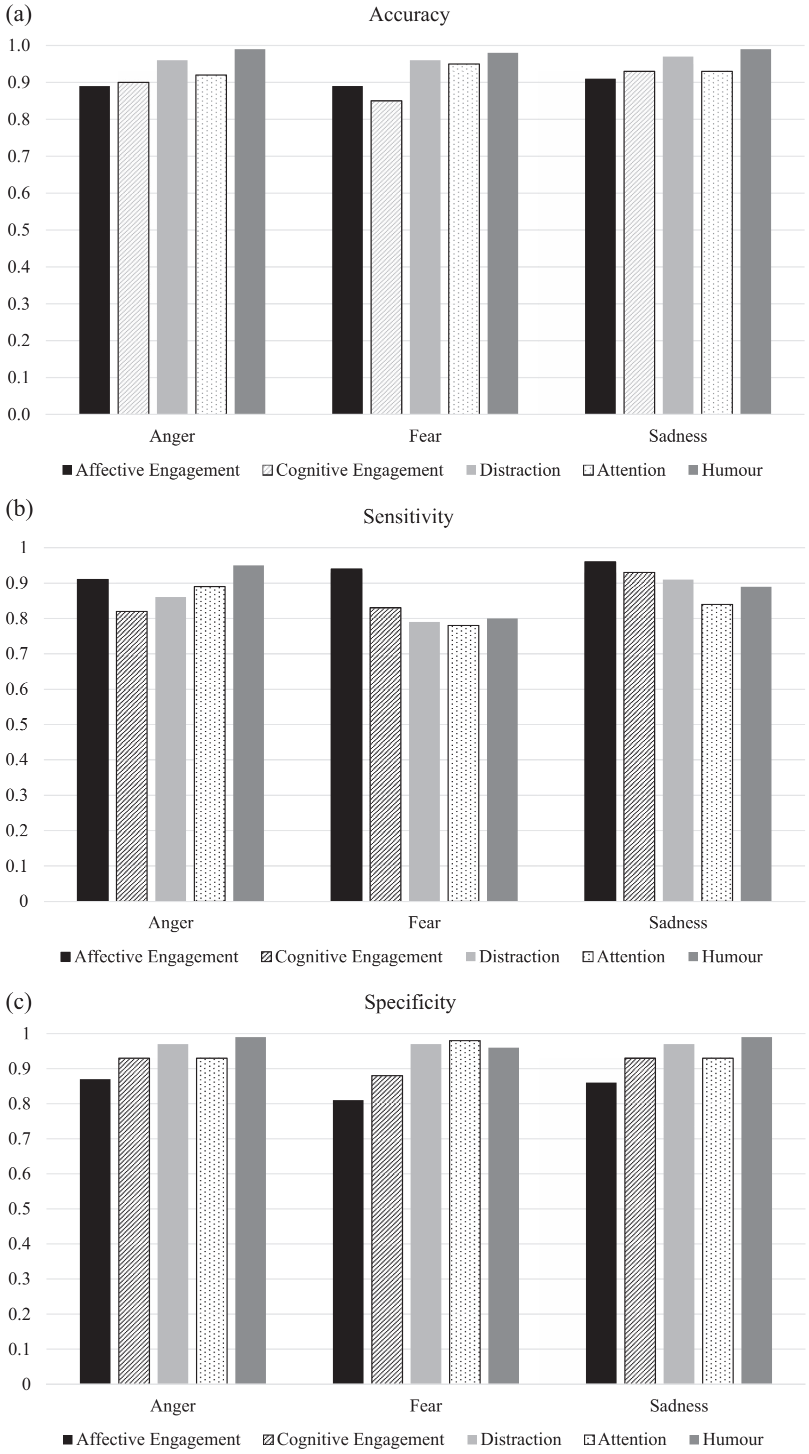
Discussion

In the present study, researchers coded a set of narratives consisting of interpersonal emotion regulation strategies and then used ChatGPT to code the same narratives. To evaluate ChatGPT's performance, we examined the consistency between ChatGPT and human coders' results. The results demonstrated that with a combination of prompt engineering and stepwise testing conducted by trained human coders, ChatGPT's performance in coding emotion regulation narratives can be substantially improved. Despite such systematic improvements, ChatGPT's performance did not match that of human coders. These results highlight the effectiveness of iterative prompt engineering, guided by domain-specific knowledge, in enhancing LLM performance. At the same time, the findings caution against expecting LLMs to perform sensitive research tasks with the same level of accuracy as human coders.

Looking at ChatGPT's performance accuracy in terms of sensitivity and specificity, we observed some patterns worth further consideration. First, even though the overall performance accuracy metrics were high (above .80), the sensitivity and specificity analyses revealed notable variations across emotions and regulation strategies. For example, the model's ability to detect a specific strategy when was present in a narrative (i.e., sensitivity) was generally higher for

Figure 4

Differences in ChatGPT's (a) Accuracy, (b) Sensitivity, and (c) Specificity Across Emotions and Regulation Strategies in Study 1



the strategies regulating anger and sadness compared to fear. For fear, we observed lower sensitivity for strategies that were less frequent (i.e., *distraction*, *attention*, and *humor*). One possible explanation is that the lower frequency of these strategies provided fewer examples of LLM coding errors to inform prompt refinement, compared to more frequently occurring strategies (Azaria et al., 2024).

Finally, in terms of ChatGPT's ability to detect strategies not mentioned in the narratives, the results indicated that specificity was relatively lower for *affective engagement*. This strategy involves helping someone express and process their emotions through conversation and active listening (Niven et al., 2009). One possible reason for ChatGPT falsely detecting affective engagement could be the overlap between this strategy and others that also involve verbal interaction, such as *cognitive engagement* (e.g., advising someone to think more positively). This overlap may lead ChatGPT to misinterpret general conversational cues as indicators of affective engagement, especially when the context lacks clear emotional processing elements.

While these findings are encouraging, the iterative nature of the steps we took to improve the model performance on the same set of narratives raises an important question: Will the final prompts that generated relatively higher accuracy perform similarly on a new set of narratives? The next study aimed to address this question.

Study 2

The findings from Study 1 suggest the possibility that researchers and practitioners could use an iterative process on a subset of their data to refine their prompts and then apply these prompts to a much larger data set. However, before we can make such a claim, it is necessary to evaluate whether (or to what extent) prompts fine-tuned on a given data set can perform well when applied to "unseen" data. In Study 2, we examined this question by applying the final prompts developed in the first study to a new set of narratives.

In addition, we sought to examine the generalizability of this procedure beyond ChatGPT. Some previous studies have highlighted notable differences between ChatGPT 4.0 and other LLMs (Borji & Mohammadian, 2023), whereas others have reported similar performance levels across LLMs when decoding emotional information (Elsey, 2023). To further investigate the generalizability of our findings, we evaluated the performance of another high-performing LLM, that is, Claude (Anthropic, 2023), in detecting interpersonal emotion regulation strategies. Finally, we aimed to assess whether the categorization of specific regulation strategies presented similar or distinct challenges to the ones identified in Study 1.

Method

Participants

Two hundred eighty-eight participants (67.8% women, $M_{\text{age}} = 19.73$, $SD = 2.45$) from the University of Oregon's Psychology and Linguistics subject pool completed the study as an option to fulfill course requirements (IRB# 09072010.006). Participants identified as: 69.1% White, 10.7% Asian, 8.1% Hispanic, Latinx, or Spanish Origin, 4.0% Black or African American, 1.5% Middle Eastern or North African, 0.4% Native Hawaiian or Other Pacific Islander, and 0.7% American Indian or Alaska Native. An additional 4.0% identified with ethnicities not specified in the options provided, and 1.5% preferred not to answer.

Materials and Procedure

The data collection procedure and materials were identical to those in Study 1. As in the previous study, we evaluated the validity of the scenarios by analyzing participants' reports for perceived emotional intensity. Descriptive statistics (Table 4) showed the same pattern as in Study 1. As such, we excluded the first fear scenario from further analyses. For the remaining scenarios, participants' perception of the protagonists' most prominent emotion was consistent with the expected key emotion (see Table 4 and Supplemental Table 6S for reliabilities).

Participants' narratives were on average 18.84 words ($SD = 14.09$). The sadness narratives ($M = 16.27$ words, $SD = 12.44$) were significantly shorter than fear, $M = 18.44$ words, $SD = 12.53$; $t(246) = -4.43$, $p < .001$, $d = .28$, and anger, $M = 22.46$ words, $SD = 24.83$; $t(246) = -4.70$, $p < .001$, $d = .30$. The anger narratives were significantly longer than fear narratives; $t(248) = 3.11$, $p < .001$, $d = .20$.

Results

Analytic Approach

Human Coding Procedure. Participants' responses for each scenario ($N = 2{,}090$ total responses) were coded by the same two independent coders as in Study 1. The coders used the refined coding manual obtained in the previous study. As in Study 1, coders could assign multiple regulation strategies to a narrative if the narrative reflected more than one strategy. The interrater agreement for the different scenarios ranged from 93.33% to 98.30%, with Cohen's Kappas consistently exceeding .83 (see Table 2).

The number of strategies identified across scenarios ranged from 0 to 4. The number of strategies was highest for the anger scenarios, followed by sadness and fear (see Supplemental Table 7S).

The frequency of use of strategies varied across scenarios (Figure 5; Supplemental Table 7S). The obtained patterns were almost identical to Study 1. *Affective* and *cognitive engagement* were the most frequent strategies in response to sadness and fear, and *attention* was the most common regulatory strategy in response to anger (for statistical comparisons see the Supplemental Materials).

ChatGPT and Claude Coding Procedure. As in Study 1, we used ChatGPT Version 4.0 (OpenAI, 2023) to code participants' narratives. In addition, we used Claude Version 2 (Anthropic, 2023),

Table 4

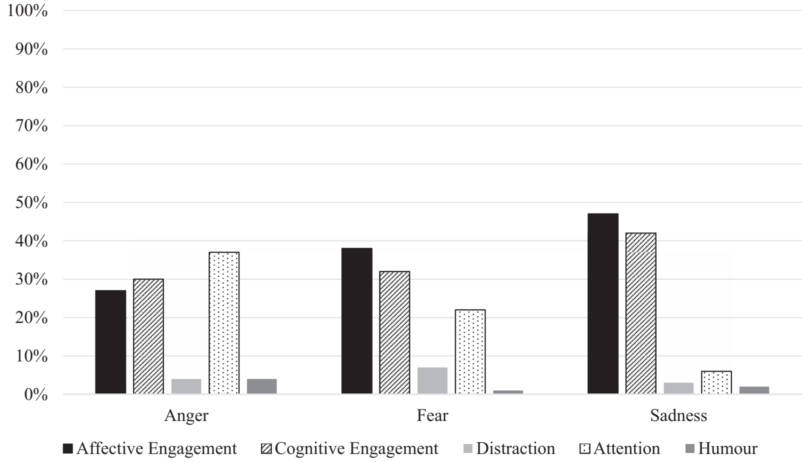
Mean and (Standard Deviation) of the Emotion Intensity Perception for Each Scenario in Study 2

Emotion	Sadness 1 scenario	Sadness 2 scenario	Sadness 3 scenario	Anger 1 scenario	Anger 2 scenario	Anger 3 scenario	Fear 1 scenario	Fear 2 scenario	Fear 3 scenario
Sadness	**3.27** (0.99)	**3.77** (1.03)	**3.57** (1.01)	2.63 (.83)	2.70 (.83)	2.50 (0.79)	1.45 (0.80)	1.71 (0.87)	1.88 (0.91)
Anger	2.78 (1.01)	2.65 (1.09)	3.32 (1.08)	**3.60** (.82)	**3.51** (.86)	**3.52** (1.02)	1.61 (0.93)	1.63 (0.86)	1.97 (0.94)
Fear	1.47 (0.85)	2.32 (1.10)	1.45 (0.84)	1.72 (.89)	1.28 (.72)	1.30 (0.70)	1.89 (1.09)	**3.30** (1.02)	**3.21** (1.08)
Happiness	1.41 (0.78)	1.29 (0.75)	1.35 (0.78)	1.29 (.66)	1.28 (.68)	1.28 (0.66)	1.55 (0.86)	1.96 (0.98)	1.88 (1.06)
Calm	1.84 (0.84)	1.52 (0.81)	1.51 (0.82)	1.58 (.72)	1.64 (.77)	1.68 (0.87)	**2.47** (1.17)	1.62 (0.77)	1.84 (0.94)

Note. The highest emotion rating for each scenario is highlighted in bold.

Figure 5

Frequency of Strategy Use Across Scenarios in Study 2



another AI language-based platform that is implemented in chat-based conversations. Claude was selected given its prominence and high performance among current LLMs (see Borji & Mohammadian, 2023). The narrative classification task with ChatGPT and Claude was conducted between September and October 2023, and between October and December 2023, respectively.

We employed the final version of the prompts from Study 1. This included detailed definitions, examples, and key indicators for each interpersonal emotion regulation strategy, and instructions to differentiate the target individual from others in each scenario. Concerning the coding approach, a limitation of 20 inputs per coding session and a requirement to use new chat windows after every 100 qualitative responses coded was implemented. This process was identical to Step 3 described in Study 1 and took 3.8 hr.

ChatGPT's and Claude's Performance

The interrater reliability between ChatGPT and human coders (.77 ≤ κs ≤ .90) and the interrater reliability between Claude and human coders (.74 ≤ κs ≤ .81) was high. The consistency between ChatGPT and Claude across different emotions and regulatory strategies was moderate to high (.64 ≤ κs ≤ .81; see Table 2 for detailed information for each scenario).

As in Study 1, we looked at accuracy, sensitivity, and specificity to investigate each LLM's performance. The results for ChatGPT were consistent with the first study. ChatGPT had a generally high classification accuracy (0.85 ≤ accuracy ≤ 0.99; Figure 6a). Regarding sensitivity, results showed notable differences across emotion scenarios (Figure 6b), as observed in Study 1. Sensitivity was high for anger (0.88 ≤ sensitivity ≤ 0.95) but there was considerable variability in sensitivity across different strategies for sadness (0.67 ≤ sensitivity ≤ 1) and fear scenarios (0.70 ≤ sensitivity ≤ .98). Finally, ChatGPT's performance had high specificity (0.83 ≤ specificity ≤ 1) with a similar pattern for all three emotions (Figure 6c). Overall, these results indicate that the classification accuracy of prompts developed iteratively with one data set is likely generalizable to "unseen" data.

Claude's performance evaluation, using the same indices, generated similar results. Across different emotions and regulatory strategies, Claude showed high classification accuracy (0.84 ≤ accuracy ≤ 1; Figure 7a). Regarding sensitivity, as observed with ChatGPT, there were differences across emotion scenarios (Figure 7b). There was considerable variability in sensitivity across different strategies for anger (0.74 ≤ sensitivity ≤ 0.90), fear (0.71 ≤ sensitivity ≤ .99), and sadness (0.78 ≤ sensitivity ≤ 1) scenarios, with similar values to the ones obtained for ChatGPT. Finally, specificity values were generally high (0.81 ≤ specificity ≤ 1) and had a similar pattern for all three emotions (Figure 7c). Overall, these results suggest that, for the present task of detecting interpersonal emotion regulation strategies in narratives, prompts developed with ChatGPT lead to a similar performance when used with another high-performing LLM. For heatmaps representing the proportional correspondence between human and LLM codes, see Supplemental Figures 2S and 3S.

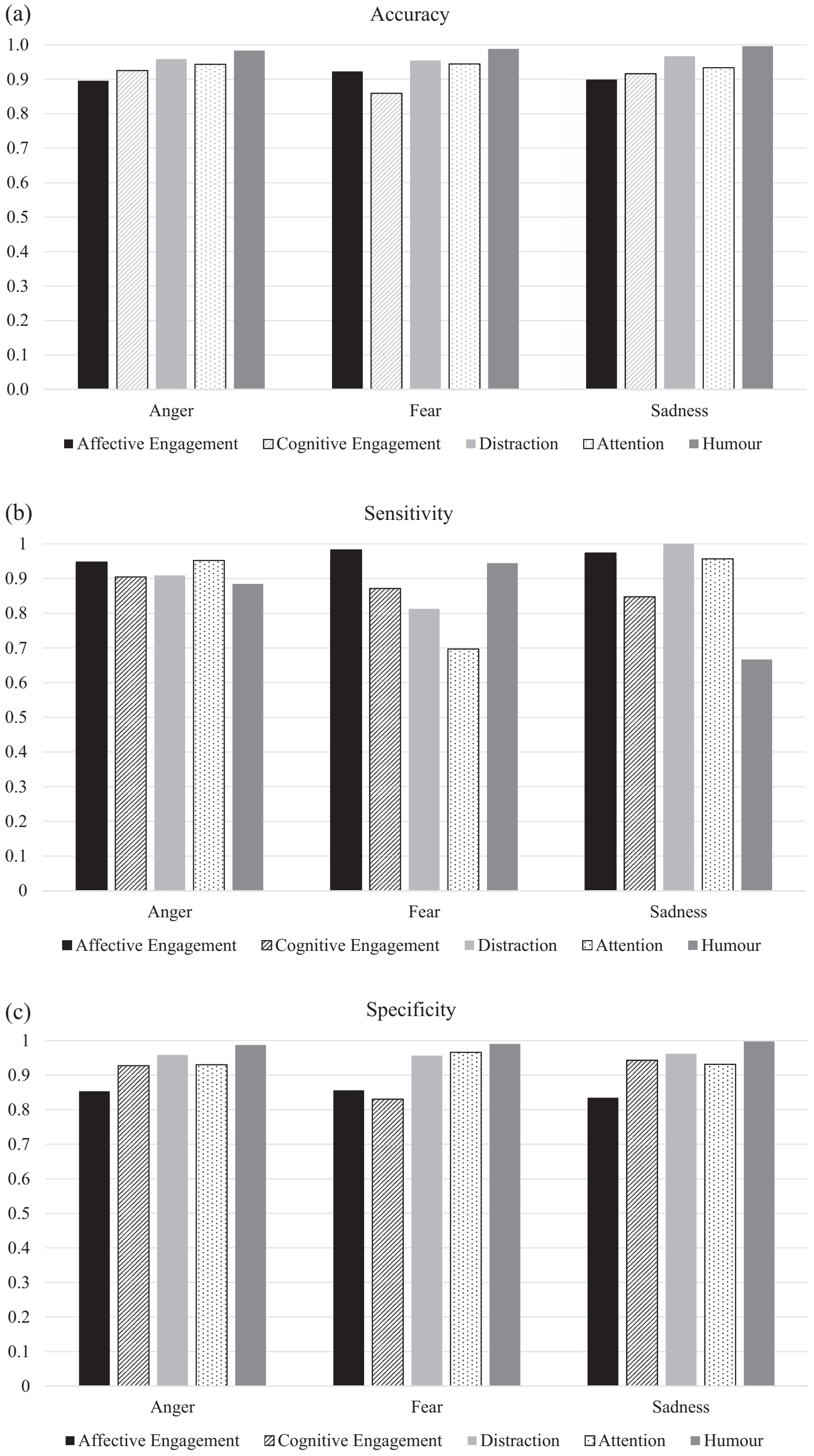
Word Count and Its Link With LLM Performance

As in Study 1, we conducted exploratory analyses to determine whether the length of participant narratives was associated with the LLMs' performance in interpreting interpersonal emotion regulation narratives compared to human coders. For each LLM, we calculated partial correlations between word count and four key performance metrics (i.e., TP, TN, FP, and FN) while controlling for the total number of strategies identified by human coders (for the full set of estimates, refer to the Supplemental Materials).

In terms of GPT's performance, we did not find significant correlations between word count and TP or FN for any of the scenarios, indicating that narrative length does not significantly relate to GPT's ability to correctly identify strategies recognized by human coders or miss strategies that are present. However, there was a pattern of significant negative correlations between word count and TNs across multiple scenarios (.04 ≤ R^2 ≤ .13), suggesting that as narrative length increases, there is a small-to-moderate decrease in the number of true negatives. Further, GPT results showed significant positive correlations between word count and FPs in some

Figure 6

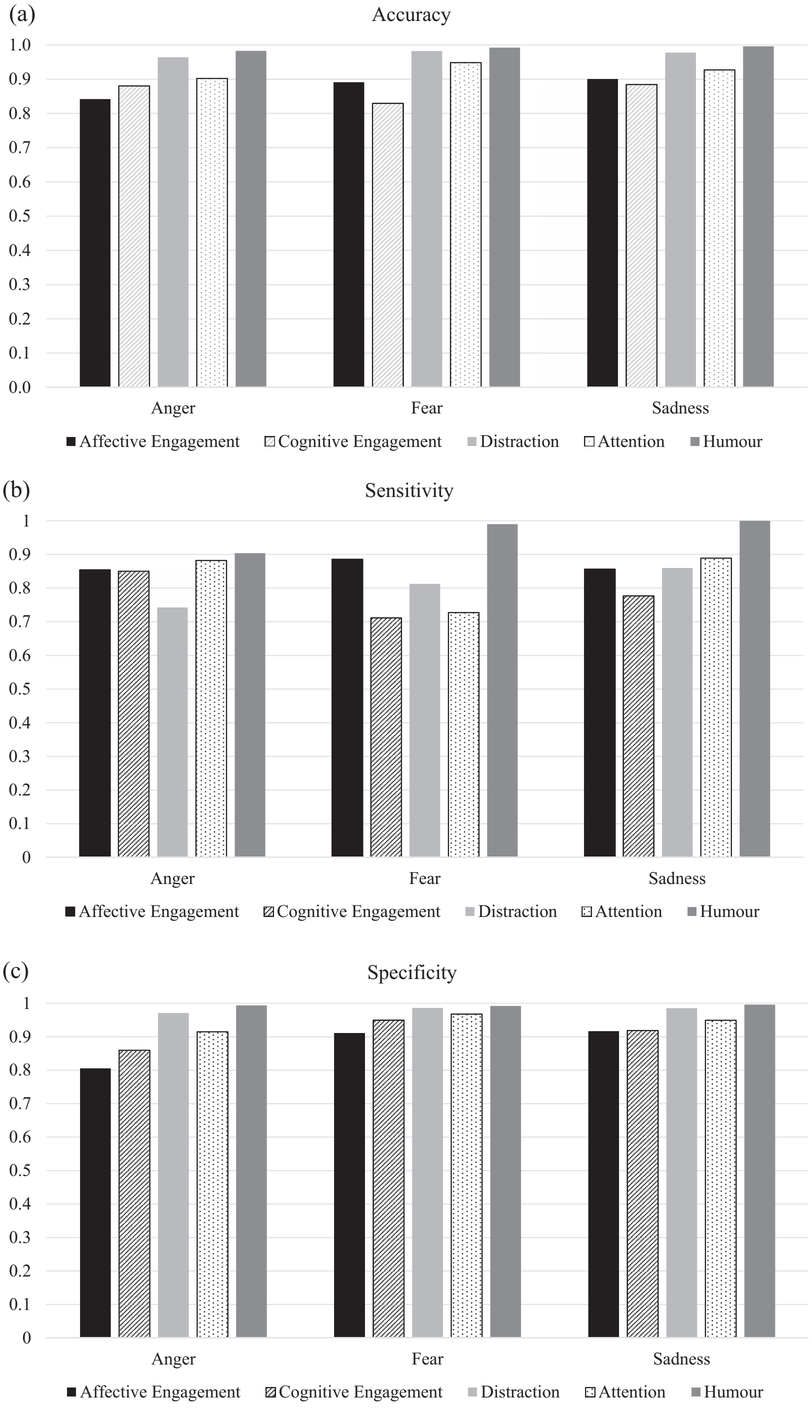
ChatGPT's (a) Accuracy, (b) Sensitivity, and (c) Specificity Across Emotions and Regulation Strategies in Study 2



scenarios (.04 ≤ R^2 ≤ .11), pointing to the possibility that longer narratives are associated with a small-to-moderate increase in FPs. In terms of Claude's performance, for the majority of cases, results did not show significant correlations between word count and any of the performance indices. In the few cases where there were significant correlations (four out of 32 correlations), results pointed to negative correlations between word count and TN (.04 ≤ R^2 ≤ .06) and positive correlations between word count and FP (.04 ≤ R^2 ≤ .05).

Figure 7

Claude's (a) Accuracy, (b) Sensitivity, and (c) Specificity Across Emotions and Regulation Strategies in Study 2



The overall pattern of correlations between word counts and TN and FP suggests that as narratives become longer, LLMs are slightly more prone to incorrectly identifying emotion regulation strategies not recognized by human coders (higher FP) and less likely to agree with human coders when no strategy is present (lower TN). The lack of significant correlations between word count and TP or FN for both models suggests that the LLMs' ability to correctly identify strategies (TP) or their tendency to miss strategies (FN) may be unrelated to the narrative length. When comparing the two models, GPT appears to be more sensitive to narrative length, as evidenced by a greater number of significant correlations. Claude's performance is less consistently associated with word count, suggesting potential differences in how the two models process longer texts.

Discussion

Results of Study 2, for both ChatGPT and Claude, were generally similar to the ones observed in Study 1. As a whole, these results suggest that when coding participants' narratives, researchers can refine their prompts through an iterative process on a subset of the data (as shown in Study 1), evaluate its performance accuracy, and expect a similar performance level when the same prompts are applied to new data sampled from a similar population. This is an encouraging step toward the automation of the time-consuming task of qualitative data coding.

Additionally, present findings reaffirm those from Study 1 which point to notable variability in LLM's performance depending on the emotion and the regulatory strategy. As researchers refine their prompts, it is critical to take steps to rigorously evaluate LLM's performance for different categories (here, emotions and regulatory strategies) and be particularly mindful of the categories for which LLM interpretations have higher inconsistency with human coders.

General Discussion

Through systematic tests of LLMs' capabilities when interpreting participants' narratives about emotion regulation strategies, the present studies provided important insights about the strengths and limitations of LLMs as tools that can augment theoretical and applied research on emotional processes. One of the primary findings of these studies was that researchers aiming to use LLMs for automating the time-intensive task of qualitative data coding can refine their prompts through an iterative process on a subset of participant narratives, assess the accuracy of these prompts, and achieve comparable performance when applying them to new data from a similar population.[1] The present studies also highlight notable variability in LLM's ability to interpret narratives across different regulation strategies and emotions, as well as the potential influence of narrative length on LLM's performance. Below, we elaborate on these points and explore how they can inform future research.

Variability Across Regulation Strategies

Results of Studies 1 and 2 showed persistent variability in LLMs' classification of different strategies. The challenge to accurately identify some strategies may stem from literal decoding of the information rather than thinking about what the broader context, the motivations behind, and the aim of the strategy described by the participants. For example, LLMs coded responses had lower specificity when it came to coding *affective engagement*. The difficulty in distinguishing between *affective* and *cognitive engagement* strategies aligns with previous research in interpersonal emotion regulation scale development, which supported only an overall engagement dimension (e.g., positive engagement) but not the differentiation of the specific strategies within this dimension (i.e., affective and cognitive engagement; López-Pérez et al., 2019).

Similar misidentifications were also observed for *attention*, which broadly refers to spending time with the target and making them feel valued (Niven et al., 2009). Several factors may account for this. First, in some cases (e.g., fear narratives), *attention* was one of the relatively less frequently mentioned strategies by the participants. This led to limited data availability during the iterative prompt refinement process and lower improvement in ChatGPT and Claude's performance (Azaria et al., 2024). Additionally, distinguishing a*ttention* from other strategies can sometimes be difficult when the motivation of the actor is not explicitly described. For instance, a statement like "taking someone out for lunch" could be interpreted as a*ttention* (spending time with the target) and/or as *distraction* (taking the target's mind off the emotion-eliciting event). This ambiguity poses challenges for LLMs, which often rely on literal interpretations of the presented information (Kocoń et al., 2023). In fact, ChatGPT has demonstrated lower performance in intrapersonal emotion regulation tasks when required to consider motivational aspects (Vzorin et al., 2023).

These findings highlight two important considerations for researchers using LLMs to code participants' narrative data. First, any ambiguity in how categories (e.g., emotion regulation strategies) are defined will likely be mirrored in the LLM's results, making it imperative to clarify such categories and address potential overlaps at the outset. Second, contextual and motivational details appear central to distinguishing some strategies from others. As such, future research designs should explicitly prompt participants to provide such information. Doing so can reduce misclassification of strategies, particularly those that hinge on understanding the actor's underlying motivation or the broader context, ultimately improving the efficacy of LLM-assisted analyses.

Variability Across Emotion Categories

Some of the results pointed to higher sensitivity and specificity for anger scenarios compared to the other emotions, especially fear. The classification challenges observed in the fear and sadness scenarios can be due to participants frequently reporting strategies that elaborate on the target's emotional experience (e.g., *affective engagement*), which are more difficult to identify, as noted in the previous section. Additionally, fewer instances of certain strategies (e.g., three of the five regulation strategies appeared in fewer than 10% of the narratives) constrained data availability during iterative prompt refinements. For example, in Study 1, the number of strategies described for fear scenarios (693) was nearly half that of anger scenarios (1,020), providing fewer opportunities for the researchers to find patterns in LLM's misclassification of fear narratives and refine the prompts.

This pattern underscores a key consideration for future research: Analogous to the importance of having enough data for prompt engineering, the sample used for refining the prompts should be large enough to be able to capture a considerable number of misidentifications. This will allow researchers to find patterns in the misidentified narratives and improve their prompts accordingly.

The Role of Word Count

In both studies, we explored the association between narrative lengths and LLMs' performance accuracy by examining the partial correlations between word count and four performance indices (i.e., TP, TN, FP, and FN) while controlling for the total number of strategies

[1] To facilitate future research on participants' narratives about interpersonal emotion regulation, we have developed two free customized versions within ChatGPT 4.0 that may support the narrative coding within the Interpersonal Affect Classification (please see Supplemental Materials for details and access).

identified by human coders. The lack of significant correlations between word count and TP or FN for both models suggests that the LLMs' ability to correctly identify strategies (TP) or their tendency to miss strategies (FN) may be unrelated to the narrative length. However, the overall pattern of correlations between word counts and TN and FP suggests that as narratives become longer, LLMs are slightly more prone to incorrectly identifying emotion regulation strategies not recognized by human coders (higher FP) and less likely to agree with human coders when no strategy is present (lower TN).

We speculate that LLMs may be more prone to overidentifying emotion regulation strategies in longer narratives due to the increased amount of text, which provides more opportunities for the models to detect patterns that they interpret as strategies, even if human coders do not. This could result in an increase in FPs. While LLMs may slightly overidentify strategies in longer narratives, the partial correlation results did not suggest that the model's ability to correctly identify strategies (TP) or avoid missing strategies (FN) is related to word count.

Limitations

Despite the novelty, the present research is not without limitations. First, we only considered strategies comprised in the IAC (Niven et al., 2009) but there are potentially other models that could be considered when looking at people's responses. For example, although not formulated explicitly to explain interpersonal regulation strategies, the process model of emotion regulation (Gross, 2002) could be well used to evaluate whether the language-based models also present challenges in the categorization of particular strategies. Second, because large language models and other deep neural network architectures are highly complex, it can be difficult to pinpoint precisely how they arrive at their outputs (Li et al., 2024). In this research, we only know what mistakes LLMs made but we can only guess why they happened. In fact, these errors motivated different changes and iterations described in Study 1's coding instructions. Future research should systematically explore the underlying factors behind these misclassifications through studies designed to enhance interpretability and error analysis. Knowing more about LLMs' logic would be important not only for research but also for studies aiming to implement interventions to anticipate where possible flaws might be present when categorizing or proposing regulation strategies to participants. Finally, this research has only focused on the interpersonal domain of emotion regulation. Future research could investigate if similar strengths and limitations are also apparent when LLMs code narratives about intrapersonal emotion regulation.

Implications

A key implication of the present findings is that researchers seeking to automate qualitative data coding with large language models can iteratively refine their prompts using a subset of participant narratives, evaluate the accuracy of these prompts, and then apply them to new data from a comparable population and expect similar performance. Through this process, researchers can rigorously evaluate the performance of LLMs and calibrate their confidence in the reliability and generalizability of conclusions drawn from LLM-generated results accordingly.

Although the coding completed by LLMs was faster than human coders and might be able to overcome some of the biases present in human coding (Rathje et al., 2024), the use of LLMs still presents significant challenges. It requires considerable prior prompt engineering and consistent human supervision to evaluate the accuracy of the responses. Moreover, researchers seeking higher precision must often resort to additional manual recoding, which can diminish the advantages of using LLMs in the first place. These issues raise a critical question of how AI can evolve to become substantially more beneficial for emotion regulation research and intervention.

References

Anthropic. (2023). *Claude* (Version 1) [Computer software]. https://www.anthropic.com

Azaria, A., Azoulay, R., & Reches, S. (2024). ChatGPT is a remarkable tool—for experts. *Data Intelligence*, *6*(1), 240–296. https://doi.org/10.1162/dint_a_00235

Borji, A., & Mohammadian, M. (2023). *Battle of the Wordsmiths: Comparing ChatGPT, GPT-4, Claude, and Bard*. SSRN. https://doi.org/10.2139/ssrn.4476855

Bowling, A., & Ebrahim, S. (2005). *Handbook of health research methods: Investigation, measurement and analysis*. Open University Press.

Boyatzis, R. E. (1998). *Transforming qualitative information: Thematic analysis and code development*. Sage Publications.

Budd, E. C., Sigelman, C. K., & Sigelman, L. (1981). Exploring the outer limits of response bias. *Sociological Focus*, *14*(4), 297–307. https://doi.org/10.1080/00380237.1981.10570403

Elsey, J. (2023). *Validity of large language models for sentiment analysis: Evidence of performance comparable to human coders*. PsyArXiv. https://doi.org/10.31234/osf.io/kcuwy

Elyoseph, Z., & Levkovich, I. (2023). Beyond human expertise: The promise and limitations of ChatGPT in suicide risk assessment. *Frontiers in Psychiatry*, *14*, Article 1213141. https://doi.org/10.3389/fpsyt.2023.1213141

Erickson, B. J., & Kitamura, F. (2021). Magician's corner: 9. Performance metrics for machine learning models. *Radiology Artificial Intelligence*, *3*(3), e200126. https://doi.org/10.1148/ryai.2021200126

Gross, J. J. (2002). Emotion regulation: Affective, cognitive, and social consequences. *Psychophysiology*, *39*(3), 281–291. https://doi.org/10.1017/S0048577201393198

Gross, J. J. (2015). Emotion regulation: Current status and future prospects. *Psychological Inquiry*, *26*(1), 1–26. https://doi.org/10.1080/1047840X.2014.940781

Gummerum, M., & López-Pérez, B. (2020). "You shouldn't feel this way!" Children's and adolescents' interpersonal emotion regulation of victims' and violators' feelings after social exclusion. *Cognitive Development*, *54*, Article 100874. https://doi.org/10.1016/j.cogdev.2020.100874

Hofmann, S. G., Carpenter, J. K., & Curtiss, J. (2016). Interpersonal Emotion Regulation Questionnaire (IERQ): Scale development and psychometric characteristics. *Cognitive Therapy and Research*, *40*(3), 341–356. https://doi.org/10.1007/s10608-016-9756-2

Jiang, W., Wang, W., & Yin, H. (2023). 'Being happy means doing it together': Exploring the interpersonal emotion regulation of kindergarten principals. *Educational Management Administration & Leadership*. Advance online publication. https://doi.org/10.1177/17411432231206623

Joffe, H. (2011). Thematic analysis. In D. Harper & A. R. Thompson (Eds.), *Qualitative research methods in mental health and psychotherapy: A guide for students and practitioners* (pp. 209–223). Wiley.

Kalla, D., & Smith, N. (2023). Study and analysis of chat GPT and its impact on different fields of study. *International Journal of Innovative Science and Research Technology*, *8*(3), 827–833. https://ssrn.com/abstract=4402499

Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., Bielaniewicz, J., Gruza, M., Janz, A., Kanclerz, K., Kocoń, A., Koptyra, B., Mieleszczenko-Kowszewicz, W., Miłkowski, P., Oleksy, M., Piasecki, M., Radliński, Ł., Wojtasik, K., Woźniak, S., & Kazienko, P. (2023).

ChatGPT: Jack of all trades, master of none. *Information Fusion*, *99*, Article 101861. https://doi.org/10.1016/j.inffus.2023.101861

Kwon, K., & López-Pérez, B. (2022). Cheering my friends up: The unique role of Interpersonal emotion regulation strategies in social competence. *Journal of Social and Personal Relationships*, *39*(4), 1154–1174. https://doi.org/10.1177/02654075211054202

Li, Y., Chan, J., Peko, G., & Sundaram, D. (2024). An explanation framework and method for AI-based text emotion analysis and visualisation. *Decision Support Systems*, *178*, Article 114121. https://doi.org/10.1016/j.dss.2023.114121

Little, L. M., Kluemper, D., Nelson, D. L., & Gooty, J. (2012). Development and validation of the Interpersonal Emotion Management Scale. *Journal of Occupational and Organizational Psychology*, *85*(2), 407–420. https://doi.org/10.1111/j.2044-8325.2011.02042.x

Liu, M., Okuhara, T., Chang, X., Shirabe, R., Nishiie, Y., Okada, H., & Kiuchi, T. (2024). Performance of ChatGPT across different versions in medical licensing examinations worldwide: Systematic review and meta-analysis. *Journal of Medical Internet Research*, *26*, e60807. https://doi.org/10.2196/60807

López-Pérez, B., Howells, L., & Gummerum, M. (2017). Cruel to be kind: Factors underlying altruistic efforts to worsen another person’s mood. *Psychological Science*, *28*(7), 862–871. https://doi.org/10.1177/0956797617696312

López-Pérez, B., Morillo, D., & Wilson, E. (2019). Development and validation of the Interpersonal Affect Improvement Strategies Questionnaire. *European Journal of Psychological Assessment*, *35*(2), 280–294. https://doi.org/10.1027/1015-5759/a000394

López-Pérez, B., & Pacella, D. (2021). Interpersonal emotion regulation in children: Age, gender, and cross-cultural differences using a serious game. *Emotion*, *21*(1), 17–27. https://doi.org/10.1037/emo0000690

López-Pérez, B., Wilson, E. L., Dellaria, G., & Gummerum, M. (2016). Developmental differences in children’s interpersonal emotion regulation. *Motivation and Emotion*, *40*(5), 767–780. https://doi.org/10.1007/s11031-016-9569-3

MacCann, C., Double, K. S., Olderbak, S., Austin, E. J., Pinkus, R. T., Walker, S. A., Kunst, H., & Niven, K. (2025). What do we do to help others feel better? The eight strategies of the Regulating Others’ Emotions Scale (ROES). *Emotion*, *25*(2), 410–429. https://doi.org/10.1037/emo0001459

Meskó, B. (2023). Prompt engineering as an important emerging skill for medical professionals: Tutorial. *Journal of Medical Internet Research*, *25*, Article e50638. https://doi.org/10.2196/50638

Messina, I., Calvo, V., Masaro, C., Ghedin, S., & Marogna, C. (2021). Interpersonal emotion regulation: From research to group therapy. *Frontiers in Psychology*, *12*, Article 636919. https://doi.org/10.3389/fpsyg.2021.636919

Mittmann, G., Zehetmayer, S., & Schrank, B. (2021). Study protocol for a randomised controlled trial to evaluate the effectiveness of a serious game targeting interpersonal emotion regulation in early adolescents. *Trials*, *22*(1), Article 741. https://doi.org/10.1186/s13063-021-05706-7

Niven, K., Totterdell, P., & Holman, D. (2009). A classification of controlled interpersonal affect regulation strategies. *Emotion*, *9*(4), 498–509. https://doi.org/10.1037/a0015962

Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic analysis: Striving to meet the trustworthiness criteria. *International Journal of Qualitative Methods*, *16*(1), 1–13. https://doi.org/10.1177/1609406917733847

OpenAI. (2023). *ChatGPT* (Version 4.0) [Computer software]. https://www.openai.com/

Powell, M. B., & Guadagno, B. (2008). An examination of the limitations in investigative interviewers’ use of open-ended questions. *Psychiatry, Psychology, and Law: An Interdisciplinary Journal of the Australian and New Zealand Association of Psychiatry, Psychology and Law*, *15*(3), 382–395. https://doi.org/10.1080/13218710802101621

Rathje, S., Mirea, D. M., Sucholutsky, I., Marjieh, R., Robertson, C. E., & Van Bavel, J. J. (2024). GPT is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences*, *121*(34), e2308950121. https://doi.org/10.1073/pnas.2308950121

Sahi, R. S., Gaines, E. M., Nussbaum, S. G., Lee, D., Lieberman, M. D., Eisenberger, N. I., & Silvers, J. A. (2025). You changed my mind: Immediate and enduring impacts of social emotion regulation. *Emotion*, *25*(2), 330–339. https://doi.org/10.1037/emo0001284

Sahi, R. S., He, Z., Silvers, J. A., & Eisenberger, N. I. (2023). One size does not fit all: Decomposing the implementation and differential benefits of social emotion regulation strategies. *Emotion*, *23*(6), 1522–1535. https://doi.org/10.1037/emo0001194

Saija, E., Cervin, M., Baiocco, R., Barcaccia, B., Ioverno, S., & Pallini, S. (2024). Dispositional and state sadness, interpersonal features, and internalizing/externalizing symptoms: A network analysis. *Journal of Applied Developmental Psychology*, *94*, 101678. https://doi.org/10.1016/j.appdev.2024.101678

Shreffler, J., & Huecker, M. R. (2023). Diagnostic testing accuracy: Sensitivity, specificity, predictive values and likelihood ratios. *StatPearls*. StatPearls Publishing. https://europepmc.org/article/nbk/nbk557491

Song, H., Tolochko, P., Eberl, J. M., Eisele, O., Greussing, E., Heidenreich, T., Lind, F., Galyga, S., & Boomgaarden, H. G. (2020). In validations we trust? The impact of imperfect human annotations as a gold standard on the quality of validation of automated content analysis. *Political Communication*, *37*(4), 550–572. https://doi.org/10.1080/10584609.2020.1723752

Swerdlow, B. A., & Johnson, S. L. (2022). The Interpersonal Regulation Interaction Scale (IRIS): A multistudy investigation of receivers’ retrospective evaluations of interpersonal emotion regulation interactions. *Emotion*, *22*(6), 1119–1136. https://doi.org/10.1037/emo0000927

Tamir, M., & Ford, B. Q. (2012). When feeling bad is expected to be good: Emotion regulation and outcome expectancies in social conflicts. *Emotion*, *12*(4), 807–816. https://doi.org/10.1037/a0024443

Tran, A., Greenaway, K. H., Kostopoulos, J., O’Brien, S. T., & Kalokerinos, E. K. (2023). Mapping interpersonal emotion regulation in everyday life. *Affective Science*, *4*(4), 672–683. https://doi.org/10.1007/s42761-023-00223-z

Troth, A. C., Lawrence, S. A., Jordan, P. J., & Ashkanasy, N. M. (2018). Interpersonal emotion regulation in the workplace: A conceptual and operational review and future research agenda. *International Journal of Management Reviews*, *20*(2), 523–543. https://doi.org/10.1111/ijmr.12144

Vzorin, G., Bukinich, A., Sedykh, A., Vetrova, I., & Sergienko, E. (2023). *Emotional intelligence of GPT-4 large language model*. Preprints. https://doi.org/10.20944/preprints202310.1458.v1

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). *A prompt pattern catalog to enhance prompt engineering with ChatGPT*. arXiv preprint arXiv:2302.11382. https://file.mixpaper.cn/paper_store/2023/681177f8-cd15-4e0f-a23b-997c6b9f9dd2.pdf

Zaki, J., & Williams, W. C. (2013). Interpersonal emotion regulation. *Emotion*, *13*(5), 803–810. https://doi.org/10.1037/a0033839

Zhang, L. R., & Chen, W. W. (2024). Is it harmful to me or to us? A dyadic analysis of Chinese young adults’ dysfunctional individuation and romantic relationship satisfaction. *International Journal of Psychology*, *59*(2), 279–287. https://doi.org/10.1002/ijop.13092

Received July 4, 2024
Revision received March 3, 2025
Accepted March 4, 2025 ■