**Online Supplementary Materials for the Manuscript Titled**

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

**Table of Contents**

Scenarios Presented to Participants in Studies 1 and 2……………………………………….3

Study 1: Different Prompts Provided to ChatGPT …………………………………………...5

Study 1: Reliability of Emotion Intensity Indices for Each Scenario………………………..21

Study 1: Comparisons of Strategy Frequency across Different Emotion Scenarios…………22

Study 1: Performance Indices for Each Emotion and Regulation Strategy………………….24

Study 1: Word Count and Its Link with LLM Performance…………………………………25

Study 1: Categorical Correspondence between Human and LLM-coded Strategies….......... 30

Study 2: Reliability of Emotion Intensity Indices for Each Scenario………………………..33

Study 2: Comparisons of Strategy Frequency across Different Emotion Scenarios…………34

Study 2: Performance Indices for Each Scenario in each LLM ……………………………..36

Study 2: Word Count and Its Link with LLM Performance…………………………………38

Study 2: Categorical Correspondence between Human and LLM-coded Strategies……….. 43

Free Customised Coders in ChatGPT 4.0……………………………………………………48

**Scenarios Presented to Participants in Studies 1 and 2**

**Anger 1 Scenario.** Your friend is working on an important project that needs to be completed soon. Despite their requests not to be interrupted, their partner continually distracts them, which negatively affects their work.

 What would you do to change how your friend is feeling?

**Anger 2 Scenario.** Your partner comes back from a long day at work and notices that the house is a mess. Their roommate who is also their best friend was supposed to clean up, but they did not.

What would you do to change how your friend is feeling?

**Anger 3 Scenario.** Your mum is waiting in a long line at the grocery store, and she is in a hurry to join her friend who is waiting for her. Suddenly, someone cuts the line and get ahead.

What would you do to change how your friend is feeling?

**Fear 1 Scenario\*\*** You are sitting with your father when you notice a spider crawling on a chair near him.

What would you do to change how your friend is feeling? 

**\*\***Discarded due to participants not producing any meaningful responses and suggesting there was no need for regulation in addition to low scores for perceived emotional intensity for fear.

**Fear 2 Scenario.** Your partner is about to present their ideas to a professional audience. Right before the presentation, a colleague tells them that the audience is twice as large as they expected.

What would you do to change how your friend is feeling?

**Fear 3 Scenario.** Your best friend has to complete a teamwork activity with their work colleagues in the woods. As part of this activity, they need to hike in a road that is known to be dangerous due to steep slopes and wild animals.

What would you do to change how your friend is feeling?

**Sadness 1 Scenario.** Your partner was hoping to get a promotion at work. She/He knows that the process is fair and there are other qualified candidates being considered as well. She/He learns from a close friend and colleague that another candidate won the promotion, and s/he did not get it.

What would you do to change how your friend is feeling?

**Sadness 2 Scenario.** Your mum tells you a very sad story that she recently read about the terrible tragedies that children in war zones experience.

What would you do to change how your friend is feeling?

**Sadness 3 Scenario.** You hear that a group of your friends and colleagues are going on a trip, but one of your best friends is not invited. Your best friend learns that they have not been invited to this trip.

What would you do to change how your friend is feeling?

**Study 1: Different Prompts Provided to ChatGPT**

Changes of prompts in each step are formatted in boldfaced type.

**Step 1: Initial Coding Prompts for ChatGPT based on the Interpersonal Affect Classification (Niven et al., 2009)**

I would like your help in coding the participant's qualitative responses (i.e., text) for the interpersonal emotion regulation strategy used, with a "1" indicating that a strategy was present and a "0" indicating that a strategy was not present. You must justify each coding to me. Below are definitions and examples for the different strategies.

Strategy 1: Affective Engagement (AE).

Definition: Talking to the target or allowing the target to talk/vent about the situation.

Providing reassurance to the target with the aim to validate the target’s thoughts or feelings.

The ultimate goal of this strategy is to engage with the feelings of the target.

Examples of coding: “Let them talk about how they feel so they can process their emotions.”  “Encourage her to discuss how she feels and try to reassure her that she is very capable and deserving of a promotion” “Talk to the target, make them realize their strengths.”

Strategy 2: Cognitive Engagement (CE).

Definition: Highlighting the positives of the situation and/or making the target think differently by, for example, rationalizing about what happened or providing advice. The ultimate goal of this strategy is to change the target’s mind/views.

Examples of coding: “I would say that there is not only one job and just because they didn't get it doesn't mean they were not qualified for it” “Talk about future opportunities” “Encourage them to tell their partner what the situation” “Give them advice”.

Strategy 3: Attention (A).

Definition: Spending time with the target and/or providing gifts/chocolates, display of physical affection. Employing behavioral strategies to help the target. The ultimate goal of this strategy is to be with the target.

Example of coding: “Make a celebration dinner or go to have a meal out.”

Strategy 4: Attention Deployment (AD).

Definition: Diverting the target’s attention away from the situation. It could include actions such as being with the target, but the ultimate goal is to take their mind off the situation.

Example of coding: “Go out for a long walk to take their mind off it.”

Strategy 5: Humor (H).

Definition: Using jokes/humor to make the target laugh about the situation.

Example of coding: “Try to make him laugh.”

**Step 2: Prompts Highlighting Differences between Strategies, Further Examples, and Key Indicators**

I would like your help in coding the participant's qualitative responses (i.e., text) for the interpersonal emotion regulation strategy used, with a "1" indicating that a strategy was present and a "0" indicating that a strategy was not present. You must justify each coding to me. Below are definitions and examples for the different strategies.

Strategy 1: Affective Engagement (AE).

Definition: **Engaging with the emotions of the target, either by talking or letting the target vent about the situation. It may involve providing verbal reassurance, comfort, and validating the target's feelings.**

**Key indicators:** **Talking to the target. Letting the target vent about their feelings or situation.  Expressions of empathy, validation, encouragement, comfort. Phrases that indicate a desire to improve or address the target's emotional state.**

**Distinctions from other strategies:**

1. **Affective Engagement VS Cognitive Engagement: Phrases like "everything is going to be fine" or "you'll get the next promotion" fall under AE, not CE. These phrases are more about offering emotional comfort and reassurance, which could be interpreted as forms of wishful thinking or hope, rather than offering concrete advice or alternative perspectives.**

**Goal: engage with the target’s emotion. It can be validating the emotion or improving the emotion.**

Strategy 2: Cognitive Engagement (CE).

Definition: Highlighting the positives of the situation and/or making the target think differently by, for example, rationalizing about what happened or providing advice. The ultimate goal of this strategy is to change the target’s mind/views.

**Key indicators: Highlighting positive aspects or silver linings of a situation. Offering advice, such as suggestions to do calming practice. Offering alternative perspectives to help the target see things differently. Providing rationalisation or broader context.**

**Distinction from other strategies:**

1. **Attention VS Cognitive Engagement: 'Attention' involves action on behalf of someone, such as removing the annoying partner, making alternative travel plans, etc. 'Cognitive Engagement' should not involve any direct action but merely an adjustment of perspective or thinking or giving advice and suggestions. To conclude, Attention includes actual behaviour, while Cognitive Engagement is more about offering advice in terms of words and changing the mindset.**
2. **Cognitive Engagement VS Affective Engagement: Affective Engagement aims at improving/validating/comforting the target’s emotion; thus, the focus is engaging in the emotion. AE keywords include “encouragement”, “validate”. However, Cognitive Engagement focuses on changing perspectives or providing advice, such as providing alternative approaches. It's more oriented toward rational and logical solutions or alternative viewpoints. AE is about "feeling with and trying to improve the emotions" of the person, and CE is about "thinking with and trying to change the mindset" of the person. When it comes to phrases like “encourage the target to think from the positive”, then it should be CE as it focuses on changing mindset instead of simple encouragement aiming at improving affect.  Phrases like "consider that there are other opportunities" or "you could aim for another promotion" are CE, not AE. These types of statements aim to offer rational advice or alternative perspectives, intending to change the target's mindset.**

**Goal: change the target's mindset and provide advice.**

Strategy 3: Attention Deployment (AD).

Definition: **Efforts to divert the target’s attention away from the stressor. IMPORTANT: must include keywords like "Distract," "take their mind off," "divert attention," "draw away from," "sidetrack," etc. to be classified as attention deployment.**

**Distinction from other strategies:**

1. **Attention Deployment VS Cognitive Engagement: “suggesting” the target to move away from the distractor is not AD. Instead, “suggestion” should fall in cognitive engagement (CE).**
2. **Attention Deployment VS Attention: the answer “take them out for dinner” is not necessarily attention deployment. We need to see 1) contextual cues and 2) whether it is stated clearly that the goal is to “distract” attention.**

**Goal: The main aim is not just any intervention, but specifically an action that takes the target’s mind off their current concerns or stressors without necessarily directly addressing the root cause.**

**Strategy 4: Attention (A).**

Definition: Spending time with the target and/or providing gifts/chocolates, display of physical affection. **Employing behavior to help the target.** **Demonstrating support through actions, gestures, or physical affection.**

**Key indicators: Spending quality time with the target. It can include both direct interactions (like hanging out) and indirect gestures (like making plans or inviting others to spend time with the target). The intention is to make the target feel supported, valued, and not alone. Offering gifts, chocolates, or other tokens of appreciation. Displaying physical affection like hugs or holding hands. Help the target with actual behaviours, e.g., cleaning the room, removing distractions, directly speaking to a third party causing the target distress on the target’s behalf.**

**Distinction from other strategies:**

1. **Attention VS Affective Engagement: If someone says, "talk to them to comfort them," this should fall under AE and not necessarily represent an investment of time beyond the emotional connection. Simple conversations that serve to provide emotional support or comfort would be categorized here.  On the other hand, “Try to offer support”, or “do something to make the target” should be Attention. The key difference is, AE is something to do with affection, but A is more like actual actions taken.**

**Goal: spend time with the target, or make the target feel the presence of the subject is supportive. Use behavior to support the target.**

**Strategy 5: Humour (H).**

Definition: Using jokes/humor to make the target laugh about the situation.

**Key indicators: Making jokes or funny remarks. Efforts to evoke laughter in the target. Must be stated “jokes”, “humor” or similar words that indicate humor.  Unless there's a direct mention of jokes, humor, laughter, or similar indications, the statement shouldn't be coded under Humour.**

**Distinction from other strategies:**

1. **Humour VS Affective engagement: Humour should specifically mention keywords like "jokes", "humour", or similar words that clearly indicate an intent to make someone laugh using humour. If a response aims to lighten or improve the mood but doesn't use explicit words indicating humour (as defined), then it should be categorised under "Affective Engagement".**

**Goal: use humour to make the target laugh.**

**Additional Considerations**

**“Nothing”/inaction: Note that some participants may leave blank answers, in this case please just code them as all 0s. When a response primarily indicates not intervening, taking action, or changing the outlook, all strategies should be coded as 0s.  Also, when participants stated that “I will do nothing”, “nothing”, “their emotions are valid; I would not change anything” or something that stated that the agent would not do anything to change the target’s feeling, you need to code all strategies as 0s as no strategies were employed.**

**Vague responses: sometimes you can encounter vague responses, for example, “support them”, “Be as supportive as possible”, “talk” etc, please code them as all 0s as we cannot identify what specific strategies the agents were using.**

**Step 3: Prompts Adding Further Description of Scenarios to Distinguish the Target from Others**

I would like your help in coding the participant's qualitative responses (i.e., text) for the interpersonal emotion regulation strategy used, with a "1" indicating that a strategy was present and a "0" indicating that a strategy was not present. You must justify each coding to me. Below are definitions and examples for the different strategies.

Strategy 1: Affective Engagement (AE).

Definition: Engaging with the emotions of the target, either by talking or letting the target vent about the situation. It may involve providing verbal reassurance, comfort, and validating the target's feelings.

Key indicators: Talking to the target. Letting the target vent about their feelings or situation.  Expressions of empathy, validation, encouragement, comfort. Phrases that indicate a desire to improve or address the target's emotional state.

Distinctions from other strategies:

1. Affective Engagement VS Cognitive Engagement: Phrases like "everything is going to be fine" or "you'll get the next promotion" fall under AE, not CE. These phrases are more about offering emotional comfort and reassurance, which could be interpreted as forms of wishful thinking or hope, rather than offering concrete advice or alternative perspectives.

Goal: engage with the target’s emotion. It can be validating the emotion or improving the emotion.

Strategy 2: Cognitive Engagement (CE).

Definition: Highlighting the positives of the situation and/or making the target think differently by, for example, rationalizing about what happened or providing advice. The ultimate goal of this strategy is to change the target’s mind/views.

Key indicators: Highlighting positive aspects or silver linings of a situation. Offering advice, such as suggestions to do calming practice. Offering alternative perspectives to help the target see things differently. Providing rationalisation or broader context.

Distinction from other strategies:

1. Attention VS Cognitive Engagement: 'Attention' involves action on behalf of someone, such as removing the annoying partner, making alternative travel plans, etc. 'Cognitive Engagement' should not involve any direct action but merely an adjustment of perspective or thinking or giving advice and suggestions. To conclude, Attention includes actual behaviour, while Cognitive Engagement is more about offering advice in terms of words and changing the mindset.
2. Cognitive Engagement VS Affective Engagement: Affective Engagement aims at improving/validating/comforting the target’s emotion; thus, the focus is engaging in the emotion. AE keywords include “encouragement”, “validate”. However, Cognitive Engagement focuses on changing perspectives or providing advice, such as providing alternative approaches. It's more oriented toward rational and logical solutions or alternative viewpoints. AE is about "feeling with and trying to improve the emotions" of the person, and CE is about "thinking with and trying to change the mindset" of the person. When it comes to phrases like “encourage the target to think from the positive”, then it should be CE as it focuses on changing mindset instead of simple encouragement aiming at improving affect.  Phrases like "consider that there are other opportunities" or "you could aim for another promotion" are CE, not AE. These types of statements aim to offer rational advice or alternative perspectives, intending to change the target's mindset.

Goal: change the target's mindset and provide advice.

Strategy 3: Attention Deployment (AD).

Definition: Efforts to divert the target’s attention away from the stressor. IMPORTANT: must include keywords like "Distract," "take their mind off," "divert attention," "draw away from," "sidetrack," etc. to be classified as attention deployment.

Distinction from other strategies:

1. Attention Deployment VS Cognitive Engagement: “suggesting” the target to move away from the distractor is not AD. Instead, “suggestion” should fall in cognitive engagement (CE).
2. Attention Deployment VS Attention: the answer “take them out for dinner” is not necessarily attention deployment. We need to see 1) contextual cues and 2) whether it is stated clearly that the goal is to “distract” attention.

Goal: The main aim is not just any intervention, but specifically an action that takes the target’s mind off their current concerns or stressors without necessarily directly addressing the root cause.

Strategy 4: Attention (A).

Definition: Spending time with the target and/or providing gifts/chocolates, display of physical affection. Employing behavior to help the target. Demonstrating support through actions, gestures, or physical affection.

Key indicators: Spending quality time with the target. It can include both direct interactions (like hanging out) and indirect gestures (like making plans or inviting others to spend time with the target). The intention is to make the target feel supported, valued, and not alone. Offering gifts, chocolates, or other tokens of appreciation. Displaying physical affection like hugs or holding hands. Help the target with actual behaviours, e.g., cleaning the room, removing distractions, directly speaking to a third party causing the target distress on the target’s behalf.

Distinction from other strategies:

1. Attention VS Affective Engagement: If someone says, "talk to them to comfort them," this should fall under AE and not necessarily represent an investment of time beyond the emotional connection. Simple conversations that serve to provide emotional support or comfort would be categorized here.  On the other hand, “Try to offer support”, or “do something to make the target” should be Attention. The key difference is, AE is something to do with affection, but A is more like actual actions taken.

Goal: spend time with the target, or make the target feel the presence of the subject is supportive. Use behavior to support the target.

Strategy 5: Humour (H).

Definition: Using jokes/humor to make the target laugh about the situation.

Key indicators: Making jokes or funny remarks. Efforts to evoke laughter in the target. Must be stated “jokes”, “humor” or similar words that indicate humor.  Unless there's a direct mention of jokes, humor, laughter, or similar indications, the statement shouldn't be coded under Humour.

Distinction from other strategies:

1. Humour VS Affective engagement: Humour should specifically mention keywords like "jokes", "humour", or similar words that clearly indicate an intent to make someone laugh using humour. If a response aims to lighten or improve the mood but doesn't use explicit words indicating humour (as defined), then it should be categorised under "Affective Engagement".

Goal: use humour to make the target laugh.

Additional Considerations

“Nothing”/inaction: Note that some participants may leave blank answers, in this case please just code them as all 0s. When a response primarily indicates not intervening, taking action, or changing the outlook, all strategies should be coded as 0s.  Also, when participants stated that “I will do nothing”, “nothing”, “their emotions are valid; I would not change anything” or something that stated that the agent would not do anything to change the target’s feeling, you need to code all strategies as 0s as no strategies were employed.

Vague responses: sometimes you can encounter vague responses, for example, “support them”, “Be as supportive as possible”, “talk” etc, please code them as all 0s as we cannot identify what specific strategies the agents were using.

**Anger 1**

**Please also note the target of the strategy. For example, in the Anger 1 scenario, “Your friend is working on an important project that needs to be completed soon. Despite their requests not to be interrupted, their partner continually distracts them, which negatively affects their work.” You should be careful whether the participants’ answer is aiming at “the friend” or “the friend’s partner”. If it is towards “the friend’s partner”, it should be different from the strategy aiming at “the friend”.**

**For example, “ask the person disturbing them to stop or invite my friend to leave the environment to be more productive” should be Attention but not Attention Deployment, because it is targeting at the “person disturbing”.**

**Anger 2**

**Please also note the target of the strategy. For example, in the Anger 2 scenario, “Your partner comes back from a long day at work and notices that the house is a mess. Their roommate who is also their best friend was supposed to clean up, but they did not.” You should be careful whether the participants’ answer is aiming at “the partner” or “the partner’s roommate”. If it is towards “the partner’s roommate”, it should be different from strategy aiming at “the partner”.**

**Anger 3**

**Please also note the target of the strategy. For example, in the Anger 3 scenario, “Your mum is waiting in a long line at the grocery store, and she is in a hurry to join her friend who is waiting for her. Suddenly, someone cuts the line and get ahead.” You should be careful whether the participants’ answer is aiming at “your mom” or “someone cuts the line”. If it is towards “someone cuts the line”, it should be different from strategy aiming at “your mom”. For example, if it is “to confront/talk to the one that cuts the line”, it should be attention, as the participant did something on behalf of the target.**

**Fear 1**

**Please also note the target of the strategy. For example, in the Fear 2 scenario, “Your partner is about to present their ideas to a professional audience. Right before the presentation, a colleague tells them that the audience is twice as large as they expected.” You should be careful whether the participants’ answer is aiming at “your partner” or anyone else.**

**Fear 2**

**Please also note the target of the strategy. For example, in the Fear 3 scenario, “Your best friend has to complete a teamwork activity with their work colleagues in the woods. As part of this activity, they need to hike in a road that is known to be dangerous due to steep slopes and wild animals.” You should be careful whether the participants’ answer is aiming at “your best friend” or anyone else, as it will be different strategies. Do you understand?**

**Sadness 1**

**Please also note the target of the strategy. For example, in the Sadness 1 scenario, “Your partner was hoping to get a promotion at work. She/He knows that the process is fair and there are other qualified candidates being considered as well. She/He learns from a close friend and colleague that another candidate won the promotion, and s/he did not get it.” You should be careful whether the participants’ answer is aiming at “your partner” or anyone else, as it will be different strategies. Do you understand?**

**Sadness 2**

**Please also note the target of the strategy. For example, in the Sadness 2 scenario, “Your mum tells you a very sad story that she recently read about the terrible tragedies that children in war zones experience.” You should be careful whether the participants’ answer is aiming at “your mom” or anyone else, such as children in war zones, as it will be different strategies. Do you understand?**

**Sadness 3**

**Please also note the target of the strategy. For example, in the Sadness 3 scenario, “You hear that a group of your friends and colleagues are going on a trip, but one of your best friends is not invited. Your best friend learns that they have not been invited to this trip.” You should be careful whether the participants’ answer is aiming at “one of your best friends not invited” or “a group of your friends”, as it will be different strategies. Do you understand?**

**Step 4: Manual Check of ChatGPT’s Response and Recode**

Upon reviewing the coding of the qualitative responses from our previous session, I've identified some discrepancies in your coding. Could you please revisit the specific coding prompts I previously supplied and focus on re-evaluating and adjusting the coding for responses numbered X, Y, and Z? Additionally, for each response, could you explain the rationale behind any changes made to the coding?

**Study 1: Reliability of Emotion Intensity Indices for Each Scenario**

**Table 1S**

*Reliabilities (Cronbach’s alpha) for Emotion Intensity Perception Indices in Study 1*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Scenarios | Anger | Fear | Sadness | Happiness | Calm |
| Anger 1 | .74 | .82 | .70 | .83 | .70 |
| Anger 2 | .78 | .86 | .78 | .90 | .67 |
| Anger 3 | .83 | .84 | .77 | .85 | .90 |
| Fear 1 | .75 | .92 | .73 | .89 | .74 |
| Fear 2 | .72 | .84 | .61 | .88 | .78 |
| Fear 3 | .80 | .93 | .71 | .71 | .87 |
| Sadness 1 | .81 | .70 | .79 | .89 | .70 |
| Sadness 2 | .78 | .75 | .78 | .90 | .73 |
| Sadness 3 | .78 | .76 | .84 | .83 | .75 |

**Study 1: Comparisons of Strategy Frequency across Different Emotion Scenarios**

The number of strategies in the anger scenarios was significantly higher (*M* = 3.27, *SD* = 1.78) than in the fear (*M* = 1.86, *SD* = 1.11; *t*(352) = 19.42, *p* < .001, *d* = 1.03) and sadness scenarios (*M* = 2.82, *SD* = 1.72; *t*(352) = 5.87, *p* < .001, *d* = 0.31). The number of strategies identified for sadness was significantly higher than in the fear scenarios (*t*(352) = 13.92, *p* <.001, *d* = 0.74).

**Table 2S**

*Frequency of Strategy Use across Emotion Scenarios in Study 1*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Anger | Sadness | Fear | Total |  |  |  |
| Affective engagement | 324 (27%) | 374 (37%) | 355 (51%) | 1053 |  |  |  |
| Cognitive engagement | 280 (23%) | 334 (33%) | 252 (36%) | 866 |  |  |  |
| Distraction | 79 (7%) | 83 (7%) | 19 (3%) | 181 |  |  |  |
| Attention | 462 (38.5%) | 223 (22%) | 52 (8%) | 737 |  |  |  |
| Humour | 55 (4.5%) | 9 (1%) | 15 (2%) | 79 |  |  |  |
| Total | 1200 | 1023 | 693 | 2916 |  |  |  |

For anger, attention was used significantly more than any other strategy (*ꭓ*2(1) > 36.01, *ps* < .001). For sadness (*ꭓ*2(1) > 3.50, *ps* < .061) and fear (*ꭓ*2(1) > 31.70, *ps* < .001) scenarios, affective engagement was used significantly more than any other strategy (sadness: *ꭓ*2(1) > 3.50, *ps* < .062 fear: *ꭓ*2(1) > 31.70, *ps* < .001). Across emotions, affective engagement was significantly more common in response to fear, compared to sadness (*ꭓ*2(1) = 33.09, *p* < .001) and anger (*ꭓ*2(1) = 110.06, *p* < .001). Cognitive engagement was significantly less common in response to anger compared to fear (*ꭓ*2(1) = 37.003, *p* < .001) and sadness (*ꭓ*2(1) = 27.62, *p* < .001) scenarios; but there was no significant difference between fear and sadness scenarios (*ꭓ*2(1) = 1.65, *p* = .201). Distraction was used significantly less in fear scenarios compared to anger (*ꭓ*2(1) = 13.43, *p* < .001) and sadness scenarios (*ꭓ*2(1) = 12.97, *p* < .001). There were no significant differences between the sadness and anger scenarios (*ꭓ*2(1) = 0.01, *p* = .992). Attention was less used in the fear scenarios as compared to sadness (*ꭓ*2(1) = 59.18, *p* < .001) and anger (*ꭓ*2(1) = 205.63, *p* < .001). The use of attention was significantly higher in response to anger compared to sadness (*ꭓ*2(1) = 70.37, *p* < .001). Finally, humour was used more often in the anger scenario compared to sadness (*ꭓ*2(1) = 7.94, *p* = .001) and fear (*ꭓ*2(1) = 10.66, *p* < .001). There were no significant differences between sadness and fear scenarios (*ꭓ*2(1) = 2.98, *p* = .083; Table 2S).

**Study 1: Performance Indices for Each Emotion and Regulation Strategy**

**Table 3S**

*Performance Indices for Each Emotion in Study 1*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | Affective Engagement | Cognitive Engagement | Distraction | Attention | Humour |
| Accuracy | Anger | .89 | .90 | .96 | .92 | .99 |
| Fear | .89 | .85 | .96 | .95 | .98 |
| Sadness | .91 | .93 | .97 | .93 | .99 |
| Sensitivity | Anger | .91 | .82 | .86 | .89 | .95 |
|  | Fear | .94 | .83 | .79 | .78 | .80 |
|  | Sadness | .96 | .93 | .91 | .84 | .89 |
| Specificity | Anger | .87 | .93 | .97 | .93 | .99 |
|  | Fear | .81 | .88 | .97 | .98 | .96 |
|  | Sadness | .86 | .93 | .97 | .93 | .99 |

**Study 1: Word Count and Its Link with LLM Performance**

The purpose of these analyses was to determine whether the length of participant narratives, as measured by word count, is associated with the performance of Large Language Models (LLMs) in detecting interpersonal emotion regulation strategies, relative to human coders. Specifically, we examined partial correlations between word count and four key performance indices—True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN)—while controlling for the total number of strategies identified by human coders. The results across eight emotional scenarios provide insights into how narrative length relates to LLM accuracy.

The results (Table 4S) present a nuanced picture. While some significant correlations were observed, indicating that word count can impact LLM performance in specific contexts, many correlations were non-significant, suggesting that narrative length does not universally affect LLM accuracy. Furthermore, the effect sizes for the significant associations, as indicated by *R*2 values, were generally small-to-moderate, ranging from 0.03 to 0.13, implying that when significant relationships were found, they accounted for a relatively small proportion of variance in the performance indices. Below, we will discuss the patterns in significant and non-significant findings, and our interpretation.

**Table 4S**

*Partial Correlations between Word Count and Different Accuracy Indices (Study 1)*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Scenario | Index1 | Estimate2 | *p* (adjusted3) | *R*2 |
| Anger 1 | TP | 0.11 | 0.671 | 0.01 |
| Anger 1 | TN | -0.14 | 0.224 | 0.02 |
| Anger 1 | FP | 0.13 | 0.322 | 0.02 |
| Anger 1 | FN | -0.13 | 0.300 | 0.02 |
| Anger 2 | TP | 0.02 | > 0.999 | 0.00 |
| Anger 2 | TN | -0.35 | <0.001 | **0.12** |
| Anger 2 | FP | 0.36 | <0.001 | **0.13** |
| Anger 2 | FN | 0.00 | > 0.999 | 0.00 |
| Anger 3 | TP | 0.15 | 0.171 | 0.02 |
| Anger 3 | TN | -0.13 | 0.322 | 0.02 |
| Anger 3 | FP | 0.13 | 0.322 | 0.02 |
| Anger 3 | FN | -0.14 | 0.221 | 0.02 |
| Fear 2 | TP | 0.20 | 0.025 | 0.04 |
| Fear 2 | TN | -0.27 | <0.001 | **0.07** |
| Fear 2 | FP | 0.26 | <0.001 | **0.07** |
| Fear 2 | FN | -0.19 | 0.025 | **0.04** |
| Fear 3 | TP | -0.01 | > 0.999 | 0.00 |
| Fear 3 | TN | -0.22 | <0.001 | **0.05** |
| Fear 3 | FP | 0.18 | 0.044 | **0.03** |
| Fear 3 | FN | 0.01 | > 0.999 | 0.00 |
| Sadness 1 | TP | 0.10 | 0.780 | 0.01 |
| Sadness 1 | TN | -0.28 | <0.001 | **0.08** |
| Sadness 1 | FP | 0.27 | <0.001 | **0.07** |
| Sadness 1 | FN | -0.09 | > 0.999 | 0.01 |
| Sadness 2 | TP | 0.06 | > 0.999 | 0.00 |
| Sadness 2 | TN | -0.18 | 0.044 | **0.03** |
| Sadness 2 | FP | 0.18 | 0.060 | 0.03 |
| Sadness 2 | FN | -0.08 | > 0.999 | 0.01 |
| Sadness 3 | TP | 0.00 | > 0.999 | 0.00 |
| Sadness 3 | TN | -0.19 | 0.025 | **0.04** |
| Sadness 3 | FP | 0.15 | 0.216 | 0.02 |
| Sadness 3 | FN | 0.02 | > 0.999 | 0.00 |

*Note.* *R2* values for the significant partial correlations are in bold.

1TP: True Positive, TN: True Negative, FP: False Positive, and FN: False Negative.

2The estimates represent Spearman *ρ*.

3To control the family-wise error rate, p-values are Bonferroni-Holm adjusted.

**Significant Findings**

The majority of partial correlations were not significant. In cases where there was a significant correlation (11 out of 32 correlations), we observed the following patterns:

**False Positives (FP) and True Negatives (TN):** In several scenarios, specifically Anger 2, Fear 2, Fear 3, Sadness 1, Sadness 2, and Sadness 3, significant correlations were found between word count and FP (positive correlation) and TN (negative correlation). The *R2* values for these correlations ranged from 0.03 to 0.13, indicating small to moderate effect sizes. These findings suggest that in longer narratives, LLMs may slightly increase the identification of emotion regulation strategies not recognized by human coders (higher FP) and slightly decrease the agreement with human coders on the absence of strategies (lower TN). However, given the small effect sizes, the practical impact of narrative length on these indices is limited.

**True Positives (TP) and False Negatives (FN):** Significant correlations were observed in the Fear 2 scenario, where TP had a positive correlation (*ρ* = 0.20, *p* = 0.025) and FN had a negative correlation (*ρ* = -0.19, *p* = 0.025) with word count. In this specific context, longer fear narratives slightly enhanced the LLM's ability to detect strategies that human coders also identified (higher TP) and reduced missed detections (lower FN). Again, the small effect sizes indicate that while the relationship is statistically significant, the magnitude of the effect is modest.

**Non-Significant Findings**

**Anger Scenarios:** ForAnger 1 and Anger 3, all correlations between word count and TP, TN, FP, and FN were non-significant, with *R2* values ranging from 0.01 to 0.02.These results point to an absence of evidence in support of the relationship between narrative length and the LLM's performance in detecting emotion regulation strategies.

**Sadness Scenarios:** ForSadness 2 and Sadness 3, TP and FN correlations were non-significant, and while FP and TN showed significant correlations, the effect sizes were small (*R2* = 0.03 to 0.04). The non-significant TP and FN correlations, for these sadness scenarios, point to an absence of evidence in support of the relation between narrative length and the LLM's ability to correctly identify strategies (TP) or miss strategies (FN).

**Overall Interpretations**

Considering both the significant and non-significant results, it appears that while word count can have some influence on LLM performance, the overall impact is limited. The significant correlations, where present, often show small effect sizes, suggesting that narrative length accounts for a small proportion of the variance in LLM performance indices.

The tendency for longer narratives to be associated with a slight increase in false positives may reflect the LLM's propensity to identify more potential strategies simply because there is more text to analyze. However, the small effect sizes indicate that this effect is not strong. The decrease in true negatives associated with longer narratives suggests a slight reduction in instances where both LLMs and human coders agree that no strategy is present. Again, the effect is modest. The general lack of significant relationships between word count and TP or FN in most scenarios implies that narrative length does not significantly enhance or impair the LLM's ability to correctly identify strategies or miss them.

In interpreting these findings, we believe two possible explanations are worth considering.First, the small increases in false positives with longer narratives may be due to LLMs overgeneralizing when presented with more text. Longer narratives provide more textual data, which may lead LLMs to identify patterns or language cues that they associate with emotion regulation strategies, even when such strategies are not recognized by human coders. This could be due to the LLMs' training data, which might include varied expressions of emotion regulation that are not universally accepted or are context dependent. However, given the small effect sizes, this tendency is not strong and may not substantially impact the overall accuracy of the LLMs.

Second, longer texts may contain more complex language structures and nuanced expressions, increasing the likelihood of misinterpretation by LLMs. The models might struggle to accurately parse and interpret these nuances compared to human coders, leading to discrepancies. In the case of narratives studied here, the minimal effect sizes suggest that LLMs are not substantially more prone to errors in longer texts compared to shorter ones. However, this is a possibility that researchers should take into consideration, especially when investigating longer texts.

**Study 1: Categorical Correspondence between Human and LLM-coded Strategies**

We generated a series of heatmaps to visually represent the correspondence between strategies coded by human coders and those identified by the LLM. Since each narrative could be assigned to one or more strategies, we first identified the ten most frequently observed strategies or strategy combinations for each emotion (see Table 5S). We then calculated the correspondence between human-coded and LLM-coded strategies. In the heatmaps below, each cell represents the proportion of strategies identified by human coders that the LLM either matched (proportions shown along the top-right to bottom-left diagonal) or mismatched (proportions shown outside the diagonal). For example, in Figure 1S top panel, 64% of anger narratives coded by human coders as using only the Attention strategy (A) were also categorized as Attention by the LLM (i.e., a full match between human and LLM coding). However, 1% of these narratives were coded by the LLM as Affective Engagement (AE), 7% as Affective Engagement and Attention (AE+A), and so on.

**Table 5S**

*Top 10 Most Frequently Observed Strategies According to Human Coders* (Study 1)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Anger Narratives | | Sadness Narratives | | Fear Narratives | |
| Human Coded Strategy | % Coded by human coders | Human Coded Strategy | % Coded by human coders | Human Coded Strategy | % Coded by human coders |
| A | 22.85 | AE | 17.66 | AE | 29.89 |
| CE | 11.43 | CE | 15.39 | CE | 19.83 |
| AE | 10.1 | A | 12.75 | AE+CE | 14.16 |
| AE+A | 9.16 | AE+CE | 10.86 | A | 3.12 |
| AE+CE | 6.14 | AE+A | 3.87 | AE+A | 2.83 |
| CE+A | 4.72 | D | 3.87 | AE+D | 1.13 |
| D+A | 2.08 | CE+A | 2.55 | AE+H | 0.99 |
| AE+CE+A | 1.98 | CE+D | 1.32 | D | 0.85 |
| H | 1.79 | AE+D | 1.23 | AE+CE+A | 0.71 |
| D | 1.70 | AE+CE+A | 0.66 | H | 0.57 |

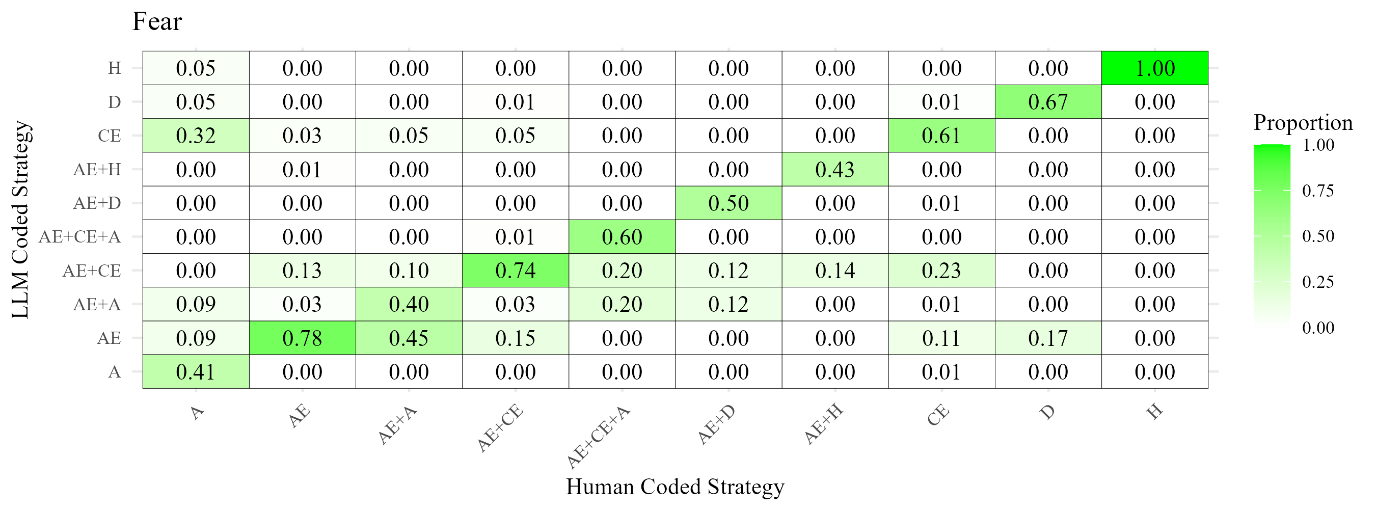
*Note.* Acronyms for different strategies: AE = Affective Engagement; CE = Cognitive Engagement; D = Distraction; A = Attention; H = Humour

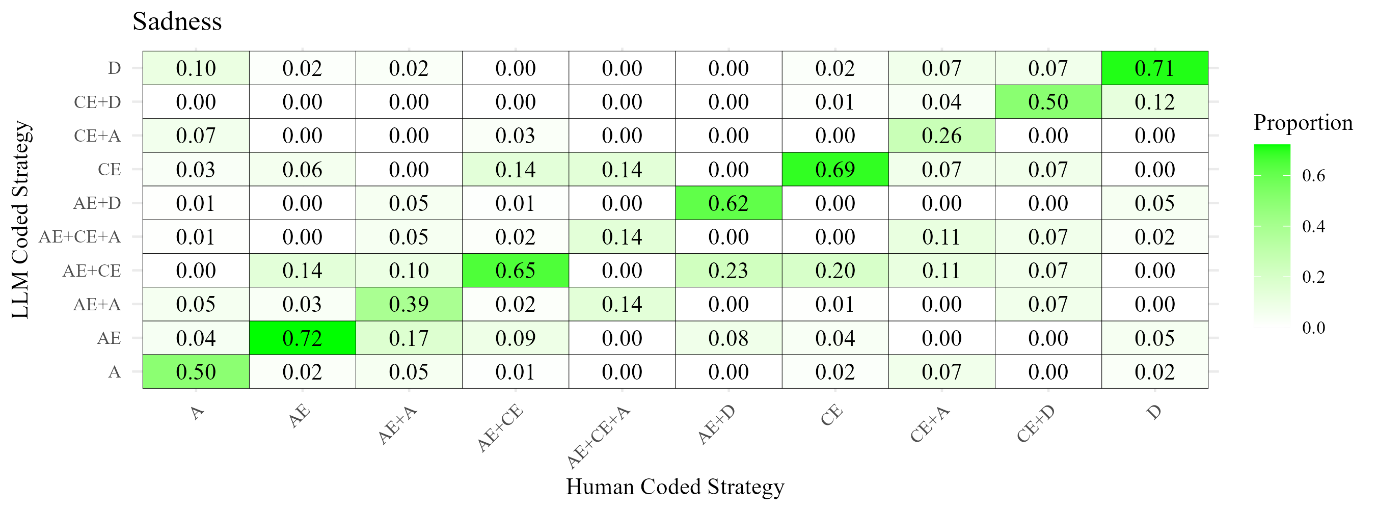
As shown in Figure 1S top panel, for all ten of the most frequent strategy combinations in the anger narratives, the LLM’s most frequent codes matched those of the human coders (see the proportions along the top-right to bottom-left diagonal). The same was true for the sadness narratives (Figure 1S middle panel), with one exception: For sadness narratives where human coders identified a combination of Affective Engagement, Cognitive Engagement, and Attention (i.e., AE+CE+A), 14% of the LLM codes matched this exact combination, while another 14% detected only CE, and another 14% detected AE+A. For fear narratives (Figure 1S bottom panel), the LLM’s most frequent codes matched the human codes for 9 out of 10 combinations. The exception was narratives coded by humans as AE+A, where 40% were coded as AE+A by the LLM (i.e., a match), while 45% were coded as AE only.

**Figure 1S**

*Proportional Correspondence between Human- and LLM-Coded Regulation Strategies*







*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 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 = Humour

**Study 2: Reliability of Emotion Intensity Indices for Each Scenario**

**Table 6S**

*Reliabilities (Cronbach’s alpha) for Emotion Intensity Perception Indices in Study 2*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Scenarios | Anger | Fear | Sadness | Happiness | Calm |
| Anger 1 | .70 | .87 | .63 | .86 | .67 |
| Anger 2 | .77 | .92 | .62 | .94 | .79 |
| Anger 3 | .78 | .88 | .66 | .89 | .78 |
| Fear 1 | .80 | .95 | .70 | .84 | .76 |
| Fear 2 | .82 | .86 | .71 | .88 | .77 |
| Fear 3 | .76 | .93 | .75 | .92 | .84 |
| Sadness 1 | .84 | .75 | .74 | .90 | .72 |
| Sadness 2 | .74 | .88 | .77 | .93 | .82 |
| Sadness 3 | .86 | .89 | .77 | .93 | .88 |

**Study 2: Comparisons of Strategy Frequency across Different Emotion Scenarios**

The number of strategies in the anger scenarios (*M* = 3.53, *SD* = 1.76) than in the fear (*M* = 2.17, *SD* = 1.21; *t*(287) = 16.65, *p* < .001, *d* = .98) and sadness scenarios (M = 3.05, SD = 1.70; *t*(287) = 5.51, *p* < .001, *d* = .32). The number of strategies identified for sadness was significantly higher than the fear scenarios (*t*(287) = 11.39, *p* < .001, *d* = .67).

The frequency of use varied across scenarios. For anger, ‘attention’ was significantly more used than any other strategy (*ꭓ*2(1) > 11.19, *ps* < .001). For sadness scenarios, ‘affective engagement’ was significantly more used than any other strategy (*ꭓ*2(1) > 7.24, *ps* < .001). For fear scenarios, both ‘affective engagement’ (*ꭓ*2(1) > 271.21, *ps* < .001) and ‘cognitive engagement’ (*ꭓ*2(1) > 223.28, *ps* < .001) were significantly more used than the rest of the strategies, but they did not differ between them as they were equally used (*ꭓ*2(1) = 3.28, *p* = .072; Table 7S).

Across emotions, ‘affective engagement’ was significantly more used in fear than sadness (*ꭓ*2(1) = 12.42, *p* < .001) and anger (*ꭓ*2(1) = 68.65, *p* < .001); in addition, it was more used in sadness than anger scenarios (*ꭓ*2(1) = 26.71, *p* < .001). ‘Cognitive engagement’ was more used in the fear scenarios than in the sadness (*ꭓ*2(1) = 16.16, *p* < .001) and anger scenarios (*ꭓ*2(1) = 27.73, *p* < .001), these two did not differ between them (*ꭓ*2(1) = .90, *p* = .342). ‘Attention deployment’ was hardly used across scenarios but was used significantly more used in the sadness scenarios than in the anger (*ꭓ*2(1) = 8.46, *p* < .001) and fear scenarios (*ꭓ*2(1) = 11.73, *p* < .001), these two did not differ between them (*ꭓ*2(1) = 1.11, *p* = .293). ‘Attention’ was less used in the fear scenarios as compared to sadness (*ꭓ*2(1) = 72.90, *p* < .001) and anger (*ꭓ*2(1) = 198.29, *p* < .001). The use of attention was significantly higher in anger than in sadness (*ꭓ*2(1) = 51.74, *p* < .001). Finally, ‘humour’ was also hardly used but was more used in the anger scenario compared to sadness (*ꭓ*2(1) = 17.26, *p* = .001) and fear (*ꭓ*2(1) = 4.96, *p* = .02). There were no differences in the use of ‘humour’ in sadness and fear scenarios (*ꭓ*2(1) = 2.69, *p* = .104; Table 7S).

**Table 7S**

*Frequency of Strategy Use in Study 2 across Emotion Scenarios*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Anger | Sadness | Fear | Total |
| Affective engagement | 271 (27%) | 349 (38%) | 295 (47%) | 915 |
| Cognitive engagement | 299 (30%) | 296 (32%) | 263 (42%) | 858 |
| Distraction | 38 (4%) | 61 (7%) | 17 (3%) | 116 |
| Attention | 375 (37%) | 203 (22%) | 39 (6%) | 617 |
| Humour | 35 (4%) | 7 (1%) | 12 (2%) | 54 |
| Total | 1018 | 916 | 629 | 2563 |

**Study 2: Performance Indices for Each Scenario in each LLM**

**Table 8S**

*Performance Indices for Each Emotion for ChatGPT*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | Affective Engagement | Cognitive Engagement | Distraction | Attention | Humour |
| Accuracy | Anger | .90 | .93 | .96 | .94 | .98 |
| Fear | .93 | .86 | .95 | .94 | .99 |
| Sadness | .90 | .92 | .97 | .93 | .99 |
| Sensitivity | Anger | .95 | .90 | .91 | .95 | .88 |
|  | Fear | .98 | .87 | .81 | .70 | .94 |
|  | Sadness | .97 | .85 | 1 | .96 | .67 |
| Specificity | Anger | .85 | .93 | .96 | .93 | .99 |
|  | Fear | .86 | .83 | .96 | .97 | .99 |
|  | Sadness | .83 | .94 | .96 | .93 | 1 |

**Table 9S**

*Performance Indices for Each Emotion for Claude*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | Affective Engagement | Cognitive Engagement | Distraction | Attention | Humour |
| Accuracy | Anger | .84 | .88 | .96 | .90 | .98 |
| Fear | .89 | .83 | .98 | .95 | .99 |
| Sadness | .90 | .88 | .98 | .93 | 1 |
| Sensitivity | Anger | .85 | .85 | .74 | .88 | .90 |
|  | Fear | .89 | .71 | .81 | .73 | .99 |
|  | Sadness | .86 | .78 | .86 | .89 | 1 |
| Specificity | Anger | .81 | .86 | .97 | .91 | .99 |
|  | Fear | .91 | .95 | .99 | .97 | .99 |
|  | Sadness | .92 | .92 | .99 | .95 | 1 |

**Study 2: Word Count and Its Link with LLM Performance**

The present analyses aimed to determine whether the length of participant narratives, measured by word count, is associated with the performance accuracy of two LLMs—GPT and Claude—in detecting interpersonal emotion regulation strategies compared to human coders. By examining partial correlations between word count and four performance indices while controlling for the total number of strategies identified by human coders, we sought to understand the relationship between narrative length and LLM accuracy across eight emotion scenarios (three anger, two fear, and three sadness).

The findings (Table 10S) reveal nuanced patterns of association between word count and LLM performance, with variations across different scenarios and between the two LLMs. Significant correlations were observed primarily with the TN and FP indices, while correlations with TP and FN indices were generally non-significant. The effect sizes, as indicated by the magnitude of the correlation estimates, ranged from small to moderate, suggesting that while statistically significant relationships exist, the practical impact may be limited. Below, we will discuss the patterns in significant and non-significant findings and compare the two LLMs.

**Table 10S**

*Partial Correlations between Word Count and Different Accuracy Indices (Study 2)*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | GPT | | | Claude | | |
| Scenario | Index1 | Estimate2 | *p* (adjusted3) | *R*2 | Estimate2 | *p* (adjusted3) | *R*2 |
| Anger 1 | TP | 0.06 | > 0.999 | 0.00 | 0.06 | > 0.999 | 0.00 |
| Anger 1 | TN | -0.21 | 0.022 | **0.04** | -0.21 | 0.029 | **0.04** |
| Anger 1 | FP | 0.17 | 0.095 | 0.03 | 0.17 | 0.13 | 0.03 |
| Anger 1 | FN | -0.02 | > 0.999 | 0.00 | -0.02 | > 0.999 | 0.00 |
| Anger 2 | TP | 0.13 | 0.63 | 0.02 | 0.00 | > 0.999 | 0.00 |
| Anger 2 | TN | -0.08 | > 0.999 | 0.01 | -0.23 | <0.001 | 0.05 |
| Anger 2 | FP | 0.08 | > 0.999 | 0.01 | 0.21 | 0.029 | **0.04** |
| Anger 2 | FN | -0.12 | 0.833 | 0.01 | 0.03 | > 0.999 | 0.00 |
| Anger 3 | TP | -0.04 | > 0.999 | 0.00 | 0.05 | > 0.999 | 0.00 |
| Anger 3 | TN | -0.35 | <0.001 | **0.12** | -0.16 | 0.225 | 0.03 |
| Anger 3 | FP | 0.33 | <0.001 | **0.11** | 0.15 | 0.391 | 0.02 |
| Anger 3 | FN | 0.04 | > 0.999 | 0.00 | -0.04 | > 0.999 | 0.00 |
| Fear 2 | TP | 0.04 | > 0.999 | 0.00 | -0.01 | > 0.999 | 0.00 |
| Fear 2 | TN | -0.29 | <0.001 | **0.08** | -0.25 | <0.001 | **0.06** |
| Fear 2 | FP | 0.25 | <0.001 | **0.06** | 0.23 | <0.001 | **0.05** |
| Fear 2 | FN | -0.02 | > 0.999 | 0.00 | 0.01 | > 0.999 | 0.00 |
| Fear 3 | TP | 0.06 | > 0.999 | 0.00 | 0.18 | 0.108 | 0.03 |
| Fear 3 | TN | -0.20 | 0.022 | **0.04** | -0.03 | > 0.999 | 0.00 |
| Fear 3 | FP | 0.20 | 0.022 | **0.04** | 0.01 | > 0.999 | 0.00 |
| Fear 3 | FN | -0.07 | > 0.999 | 0.00 | -0.16 | 0.264 | 0.03 |
| Sadness 1 | TP | 0.09 | > 0.999 | 0.01 | 0.12 | > 0.999 | 0.01 |
| Sadness 1 | TN | -0.36 | <0.001 | **0.13** | 0.02 | > 0.999 | 0.00 |
| Sadness 1 | FP | 0.31 | <0.001 | **0.10** | -0.03 | > 0.999 | 0.00 |
| Sadness 1 | FN | -0.10 | > 0.999 | 0.01 | -0.11 | > 0.999 | 0.01 |
| Sadness 2 | TP | 0.11 | > 0.999 | 0.01 | -0.03 | > 0.999 | 0.00 |
| Sadness 2 | TN | -0.27 | <0.001 | **0.07** | -0.14 | 0.616 | 0.02 |
| Sadness 2 | FP | 0.26 | <0.001 | **0.07** | 0.11 | > 0.999 | 0.01 |
| Sadness 2 | FN | -0.12 | 0.896 | 0.01 | 0.03 | > 0.999 | 0.00 |
| Sadness 3 | TP | 0.02 | > 0.999 | 0.00 | 0.02 | > 0.999 | 0.00 |
| Sadness 3 | TN | -0.35 | <0.001 | **0.12** | -0.07 | > 0.999 | 0.00 |
| Sadness 3 | FP | 0.30 | <0.001 | **0.09** | 0.05 | > 0.999 | 0.00 |
| Sadness 3 | FN | 0.00 | > 0.999 | 0.00 | 0.00 | > 0.999 | 0.00 |

*Note.* *R2* values for the significant partial correlations are in bold.

1TP: True Positive, TN: True Negative, FP: False Positive, and FN: False Negative.

2The estimates represent Spearman *ρ*.

3To control the family-wise error rate, p-values are Bonferroni-Holm adjusted.

**Associations for GPT**

**Non-Significant Correlations with TP and FN:** Across all scenarios, we did not find significant correlations between word count and TP or FN, 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. In cases where there were significant correlations (13 out of 32 correlations), we observed the following two patterns:

**Significant Negative Correlations with TN:** There was a pattern of significant negative correlations between word count and TNs across multiple scenarios, including Anger 1, Anger 3, Fear 2, Fear 3, Sadness 1, Sadness 2, and Sadness 3. The *R2* values for these correlations ranged from 0.04 to 0.13, suggesting that as narrative length increases, there is a small-to-moderate decrease in the number of true negatives.

**Significant Positive Correlations with FP:** Corresponding to the decrease in TN, GPT results showed significant positive correlations between word count and FPs in scenarios such as Anger 3, Fear 2, Fear 3, Sadness 1, Sadness 2, and Sadness 3, with *R2* values ranging from 0.04 to 0.11. These findings suggest that longer narratives are associated with a small-to-moderate increase in false positives.

**Associations for Claude**

In the majority of scenarios, results did not show significant correlations between word count and any of the performance indices, including TP, TN, FP, and FN. In cases where there were significant correlations (four out of 32 correlations), we observed the following two patterns:

**Significant Negative Correlations with TN:** Results showed significant negative correlations between word count and TNs in the Anger 1, Anger 2, and Fear 2 scenarios. The *R2* values for these correlations ranged from 0.04 to 0.06, suggesting that as narrative length increases, there is a modest decrease in the number of true negatives.

**Significant Positive Correlations with FP:** Results also indicated significant positive correlations between word count and FPs in Anger 2 and Fear 2. The *R2* values for these correlations ranged from 0.04 to 0.05, pointing to a small-to-moderate increase in false positives with longer narratives in these scenarios.

**Overall Interpretations**

A comparison of the GPT and Claude results indicates that GPT showed more consistent significant associations between word count and TN and FP indices across multiple scenarios, whereas Claude's significant associations were fewer and limited to specific scenarios. The effect sizes were generally small to moderate for both LLMs. GPT had *R²* values ranging from 0.04 to 0.13 in significant correlations, while Claude's *R²* values ranged from 0.04 to 0.06. As a whole, 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.

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.

Similar to our conclusion for Study 1, we speculate that LLMs may be more prone to over-identifying 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 false positives. While LLMs may slightly over-identify strategies in longer narratives, their ability to correctly identify true strategies (TP) or avoid missing strategies (FN) remains stable.

GPT showed a more consistent association between word count and performance indices across scenarios, while Claude's associations were less frequent and scenario dependent. This indicates that GPT may be more sensitive to narrative length than Claude. The differences between the models may be due to variations in their training data, algorithms, or sensitivity to textual features like text structure.

**Study 2: Categorical Correspondence between Human and LLM-coded Strategies**

As in Study 1, we generated a series of heatmaps to visually represent the correspondence between strategies coded by human coders and those identified by the LLMs. We first identified the ten most frequently observed strategies or strategy combinations for each emotion (see Table 11S). We then calculated the correspondence between human-coded and LLM-coded strategies. In the heatmaps below, each cell represents the proportion of strategies identified by human coders that the LLM either matched (proportions shown along the top-right to bottom-left diagonal) or mismatched (proportions shown outside the diagonal).

**Table 11S**

*Top 10 Most Frequently Observed Strategies According to Human Coders* (Study 2)

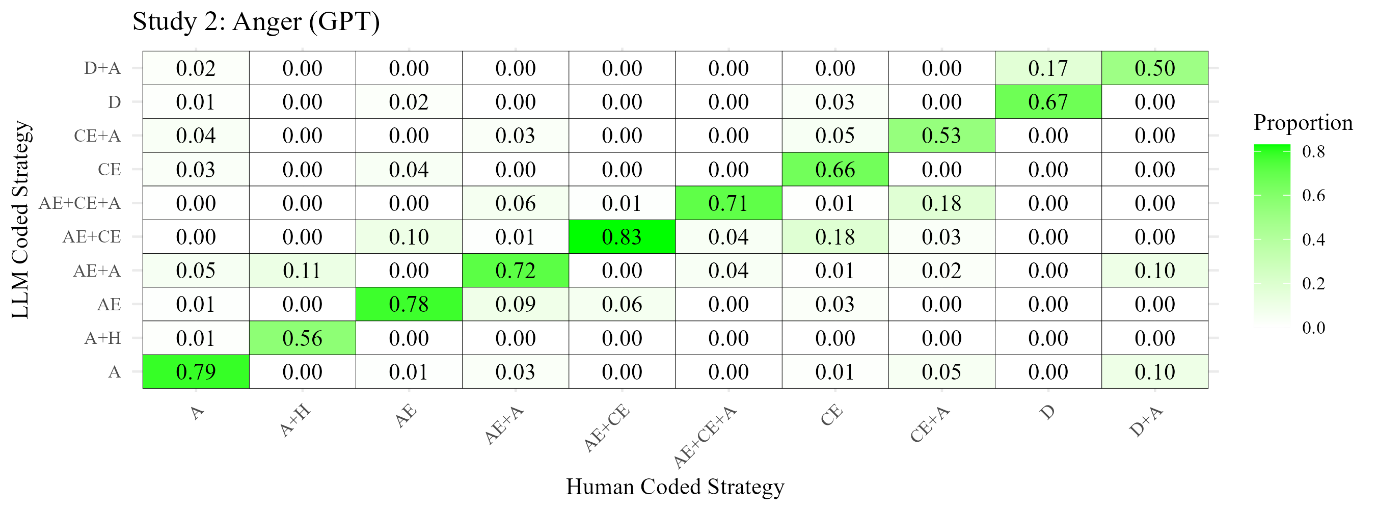
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Anger Narratives | | Sadness Narratives | | Fear Narratives | |
| Human Coded Strategy | % Coded by human coders | Human Coded Strategy | % Coded by human coders | Human Coded Strategy | % Coded by human coders |
| A | 24.77 | AE | 18.98 | AE | 27.78 |
| CE | 17.82 | CE | 15.05 | CE | 24.13 |
| AE | 10.88 | A | 13.66 | AE+CE | 17.88 |
| AE+A | 7.99 | AE+CE | 13.19 | A | 3.3 |
| AE+CE | 7.41 | AE+A | 3.94 | AE+A | 1.74 |
| CE+A | 4.28 | D | 3.12 | AE+D | 1.04 |
| AE+CE+A | 2.78 | CE+A | 2.55 | AE+CE+A | 0.69 |
| A+H | 1.27 | AE+CE+A | 1.74 | AE+CE+H | 0.69 |
| D+A | 1.04 | AE+D | 1.27 | D | 0.69 |
| D | 1.04 | CE+D | 0.93 | CE+D | 0.69 |

*Note.* Acronyms for different strategies: AE = Affective Engagement; CE = Cognitive Engagement; D = Distraction; A = Attention; H = Humour

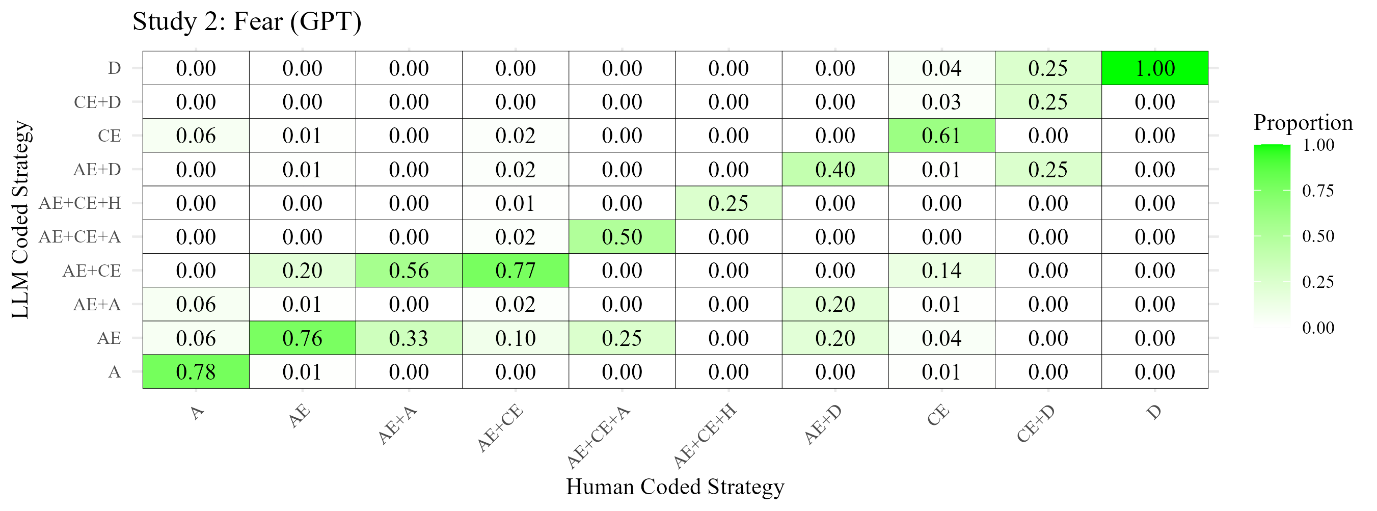
**GPT Results.** As demonstrated in Figure 2S top panel, for all ten of the most frequent strategy combinations in the anger narratives, GPT’s most frequent codes matched those of the human coders (see the proportions along the top-right to bottom-left diagonal). The same was true for the sadness narratives (Figure 2S middle panel), with one exception: For sadness narratives where human coders identified a combination of CE+A, 29% of the LLM codes matched this exact combination, while another 29% detected AE+ CE+A. For fear narratives (Figure 2S bottom panel), GPT’s most frequent codes matched the human codes for 8 out of 10 combinations. One significant exception was narratives coded by humans as AE+A, where 56% were coded by GPT as AE+CE and 33% were coded as AE only.

**Figure 2S**

*Proportional Correspondence between Human- and GPT-Coded Regulation Strategies*



****

****

*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 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 = Humour

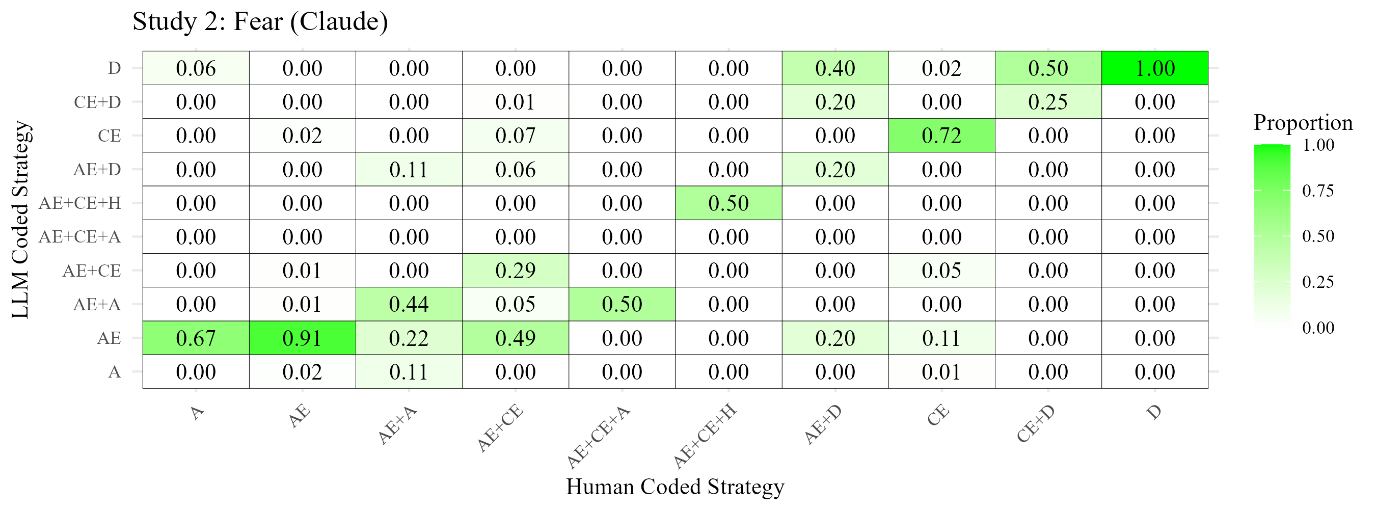
**Claude Results.** As demonstrated in Figure 3S top panel, for all ten of the most frequent strategy combinations in the anger narratives, Claude’s most frequent codes matched those of the human coders. For the sadness narratives (Figure 3S middle panel), Claude’s most frequent codes matched the human codes for 8 out of 10 combinations. For sadness narratives where human coders identified a combination of CE+D, only 14% of the LLM codes matched this exact combination, while 71% detected only D. For narratives where human coders identified D, 96% were identified as AE and 4% were identified as A. For fear narratives (Figure 3S, bottom panel), the results were notably different. Claude’s most frequent codes matched the human codes for only 5 out of 10 strategy combinations. Notably, the two most frequent strategies among fear narratives—Affective Engagement (AE) and Cognitive Engagement (CE)—were included in this group. For the remaining five combinations (all of which, except AE+CE, were low-frequency strategies; see Table 11S), Claude’s most frequent codes did not align with the strategies coded by human coders.

**Figure 3S**

*Proportional Correspondence between Human- and Claude-Coded Regulation Strategies*

**

A screenshot of a computer

AI-generated content may be incorrect.

*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 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 = Humour

**Free Customised Coders in ChatGPT 4.0**

OpenAI released a customizable iteration of ChatGPT-4 titled Custom GPT. Designed to cater to unique user needs, Custom GPT integrates instructions, additional knowledge, and diverse skill sets. As a result of this research, we developed 'EmoReg Coder'—a specialized version of Custom GPT aimed at interpersonal emotion regulation data coding including the scenarios used in this research and ‘Emotion Regulation Coder’ to accurately categorise regulation strategies only considering the coding instructions focusing on the Interpersonal Affect Classification without considering the scenarios. Both coders maintain the core capabilities and accuracy of ChatGPT-4 but offers enhanced efficiency and accuracy for specific tasks. Key improvements include:

Integrated Knowledge: Both coders are preloaded with relevant coding information, eliminating the repetitive task of providing instructions for each session, thereby streamlining the coding process.

Structured Output: The AI is programmed to present its findings in an organized table format, detailing the original qualitative responses, the corresponding coding output, and an explanatory rationale for its coding decisions. This structure is a significant advancement over the previous ChatGPT-4 outputs, which could be inconsistent and disorganised.

Human-Coded Training: By learning from an extensive array of human-coded examples, both coders are designed to emulate human-like coding patterns. Since generative AI technologies enhance their performance with data exposure, this training approach aims to refine the accuracy of the coders’ outputs.

'EmoReg Coder' (coder that is based on our scenarios and the Interpersonal Affect Classification refined strategies) is now accessible to all ChatGPT Plus members at <https://chat.openai.com/g/g-jThYcC0MG-emoreg-coder>. ‘Emotion Regulation Coder’ (general coder that considers only the Interpersonal Affect Classification framework with the refined instructions) can be accessed here <https://chat.openai.com/g/g-2QxGg3TAg-emotion-regulation-coder>. We invite users to employ the coders in their research and share feedback with us to further its development.