"Going to a trap house" conveys more fear than *"Going to a mall"*: Benchmarking Emotion Context Sensitivity for LLMs

Eojin Jeon[1*], **Mingyu Lee**[1*], **Sangyun Kim**[1],
Junho Kim[1], **Wanzee Cho**[1], **Tae-Eui Kam**[1], **SangKeun Lee**[1,2]
[1]Department of Artificial Intelligence [2]Department of Computer Science and Engineering
Korea University, Seoul, Republic of Korea
{skdlcm456, decon9201, silky, monocrat, meredith_cho}@korea.ac.kr
{kamte, yalphy}@korea.ac.kr

Abstract

Emotion context sensitivity—the ability to adjust emotional responses based on contexts—is a core component of human emotional intelligence. For example, being told, "You can come with me if you want," may elicit joy if the destination is a mall, but provoke fear if the destination is a trap house. As large language models (LLMs) are increasingly deployed in socially interactive settings, understanding this human ability becomes crucial for generating context-appropriate, emotion-aware responses. In this work, we introduce TRACE, a novel benchmark to evaluate LLMs' understanding of *emotion context sensitivity* of humans. This benchmark consists of 1,626 social scenarios and comprises two complementary tests: a sensitivity test, which measures whether models can detect emotional shifts caused by context changes, and a robustness test, which evaluates whether models can maintain stable emotion predictions when context changes are emotionally irrelevant. Each scenario pair keeps the core event constant while systematically varying contextual details—time, place, or agent—based on insights from behavioral theory and emotion psychology. Experimental results show that even the best-performing LLMs lag behind human performance by 20% in the sensitivity test and 15% in the robustness test, indicating substantial room for improvement in context-aware emotional reasoning.[1]

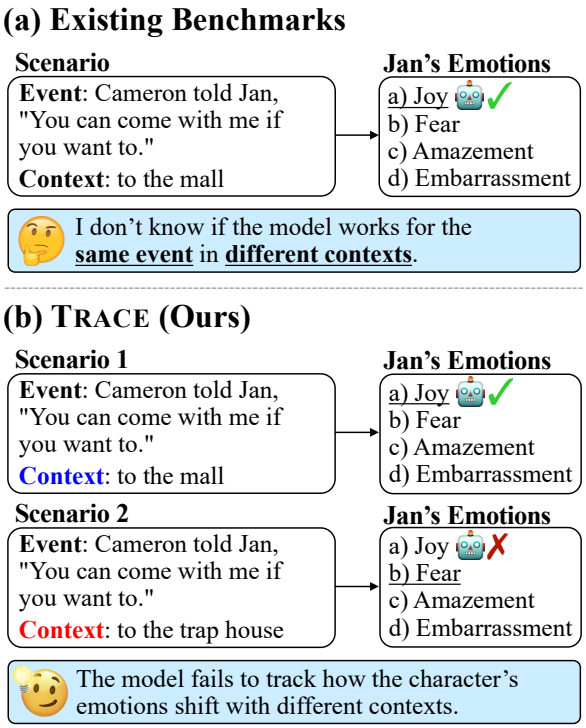


Figure 1: Comparison between existing emotional intelligence benchmarks and our TRACE. (a) Existing benchmarks pair each event with a single context. (b) TRACE pairs the same event with two different contexts.

1 Introduction

Emotion context sensitivity—the capacity for an individual to shift emotional responses according to contextual changes—is a foundational component of human emotional intelligence (Coifman and Bonanno, 2009, 2010). Emotional responses here encompass both the internal experience and external expression of emotion. This capacity enables adaptive behavior by eliciting emotions that are appropriate to contextual demands. For example, being told, "You can come with me if you want," may elicit joy if the destination is a mall, but provoke fear if the destination is a trap house. Such a fear emotion helps humans avoid potential danger or prepare for self-protection (Steimer, 2002). In other words, even when the core event (e.g., being offered to accompany someone) remains the same, emotional responses can differ depending on the context.

With the rapid progress of large language mod-

[*] These authors contributed equally to this work.
[1]Our code is available at https://github.com/jej127/Trace

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 14848–14869
November 4-9, 2025

els (LLMs), AI systems have been increasingly applied to tasks that involve social interaction with humans (Rashkin et al., 2019; Goubet and Chrysikou, 2019; Liu et al., 2020; Zhou et al., 2021; Kim et al., 2023). Consequently, it has become significantly important for these systems to understand human *emotion context sensitivity* in order to generate context-appropriate, emotion-aware responses. Despite the growing importance of this ability, no existing benchmark has been specifically designed to evaluate how well LLMs understand context-dependent emotional responses. Prior benchmarks (Demszky et al., 2020; Paech, 2023; Sabour et al., 2024) have primarily focused on general emotion recognition without explicitly disentangling the influence of contextual variations on emotional responses (See Figure 1).

As a first step toward understanding human *emotion context sensitivity*, we introduce TRACE—a novel benchmark designed to assess whether LLMs can accurately trace how emotions shift of humans in response to contextual changes, a core component of emotional intelligence. To comprehensively evaluate this capacity, TRACE comprises two complementary components: a sensitivity test and a robustness test. The sensitivity test assesses whether LLMs can correctly capture shifts in a character's emotion in response to changes in the surrounding context. In contrast, the robustness test evaluates whether models can maintain emotion predictions consistently when context changes are irrelevant to the character's emotional experience. Together, the two components provide a more comprehensive assessment of models' understanding of context-sensitive emotional responses, particularly those pertaining to emotional experience—our primary focus in this study.

Specifically, each example in both tests consists of a pair of scenarios in which the core event remains constant, while specific contextual details are varied. Building on prior work in behavioral theory and emotion research (Burke et al., 2009; Greenaway et al., 2018), we manipulate three key context dimensions—time, place, and agent—that have been identified as central in shaping human behavior, cognition, and emotional responses. Consequently, the sensitivity test includes scenario pairs where the contextual variation is intended to elicit different target emotions, whereas the robustness test includes pairs where the variation is designed to leave the target emotion unchanged.

We evaluate a range of open- and closed-source LLMs on TRACE. Results show that even the latest models fall significantly short of human performance in understanding *emotion context sensitivity*. Specifically, the best-performing model lags behind human accuracy by approximately 20% in the sensitivity test and 15% in the robustness test. To summarize, our contributions include:

- We introduce *emotion context sensitivity* as a new perspective in computational emotional understanding, inspired by psychological theory.
- We propose TRACE, a novel benchmark for systematically evaluating whether LLMs can track emotional responses across varying contexts.
- Our experimental results show that even state-of-the-art LLMs fall significantly short of human-level performance, highlighting the need for further advancement in emotional intelligence modeling.

2 Related Works

Our work is closely related to the prior works that have assessed the emotional understanding capability of models by constructing dedicated benchmarks. For example, DailyDialog (Li et al., 2017), EmotionLines (Hsu et al., 2018), and MELD (Poria et al., 2019) are dialogue datasets proposed to test the ability of models to recognize emotions in conversations. GoEmotions (Demszky et al., 2020) is a dataset covering a broader range of emotions, constructed to evaluate how well the models understand subtler differences between emotions. EmoBench (Sabour et al., 2024) contains QAs that require LLMs to predict emotions based on thorough reasoning rather than relying on frequent or explicit patterns. In addition, SECEU (Wang et al., 2023b) and EQ-Bench (Paech, 2023) focus on assessing emotional intelligence through compound emotion reasoning and scalar estimation, respectively. In contrast, benchmarks such as SocialIQA (Sap et al., 2019) and CICERO (Ghosal et al., 2022) evaluate models' social commonsense reasoning, where emotional inference arises as a secondary component within broader social contexts.

In contrast to prior benchmarks, our work is the first to explicitly assess whether LLMs understand emotion context sensitivity—that is, how subtle contextual shifts can lead to meaningful emotional

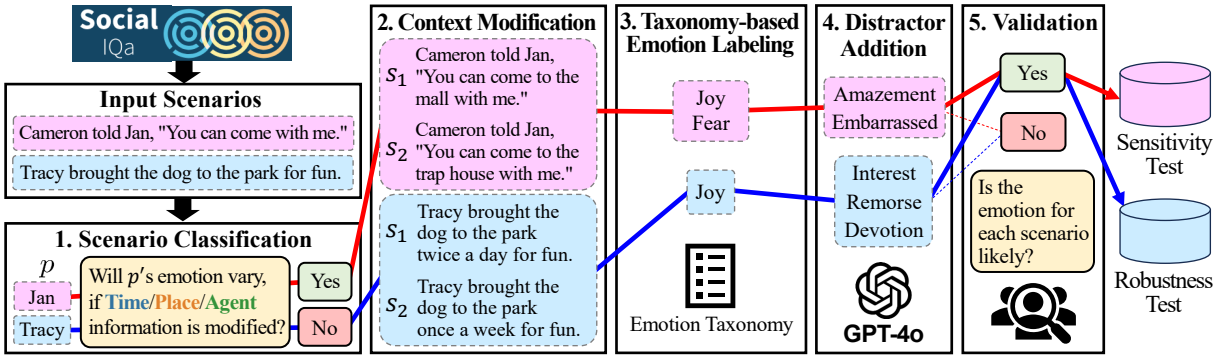


Figure 2: Data collection pipeline. Crowdworkers first identify the pivot character p in the scenario and determine if p's emotions could change when the contexts related to time, place, or agent in the scenario are altered (Step 1). In both cases, crowdworkers modify the contexts in the scenario to collect two distinct scenarios (s_1, s_2) (Step 2), along with the corresponding emotion(s) (Step 3). Next, we add adversarial distractors using GPT-4o (OpenAI, 2024a) via the OpenAI API (Step 4), and then the collected data is validated by other workers (Step 5). Data that passes this validation process belongs to either sensitivity or robustness test, depending on whether the emotion of p is deemed adjustable in Step 1.

Context	Sensitivity	Robustness	Total
Time	278	260	538
Place	242	206	448
Agent	398	242	640
Total	918	708	1,626

Table 1: Distribution of scenarios across context types. See Appendix F for specific subcategories of each context type.

changes. By isolating changes in time, place, or agent, TRACE assesses LLMs' ability to understand the emotional impact of specific types of contextual variation, identifying which dimensions pose the greatest challenge.

3 TRACE Benchmark

3.1 Overview of TRACE

TRACE is a benchmark designed to assess how well LLMs comprehend *emotion context sensitivity* of humans. To achieve this, our benchmark consists of two complementary tests: the sensitivity test and the robustness test. The sensitivity test examines whether a model correctly captures emotional shifts when meaningful context changes occur, while the robustness test ensures that the model maintains predictions when the context change is irrelevant to emotional shifts. Together, the two tests allow us to examine whether the model accurately tracks emotional shifts resulting from context changes, or merely reacts to the context change.

To support these evaluations, the benchmark comprises a total of 1,626 scenarios, systematically curated through structured annotation, adversarial distractor selection, and rigorous validation, as detailed in Section 3.2. Detailed dataset statistics are provided in Table 1.

3.2 Data Collection

Here, we detail the data collection process, which involved structured scenario annotation (Steps 1-3), adversarial distractor addition (Step 4), and rigorous validation (Step 5). To ensure high-quality annotations, we employed workers on Amazon Mechanical Turk (MTurk) with a qualification test. Further details on the quality control procedures can be found in Appendix A.

Step 1: Scenario Classification. We first extracted seed scenarios from the Social IQA dataset (Sap et al., 2019), which provides a diverse set of scenarios involving social interactions. Qualified crowdworkers then identified the pivot character p in each scenario and classified the scenario into one of two categories: those likely to result in an emotional change for p when a contextual dimension (time, place, or agent) is altered, and those unlikely to do so.

Step 2: Context Modification. To generate scenario pairs for both tests, workers either altered the contextual aspect of each scenario or added new contextual details to it, based on its classification in Step 1. The modification process was designed to align with the specific objectives of each test.

Step 2.1: Sensitivity Test. For the sensitivity test, workers modified contextual details such that

the p's emotional response would differ between the two versions. As a result, the two modified scenarios, s_1 and s_2, were annotated with **different** emotion labels. This design allows us to evaluate whether the model understands that changes in context can lead to changes in a p's emotional state.

Step 2.2: Robustness Test. For the robustness test, workers modified contextual details while preserving the p's emotional state. Thus, s_1 and s_2 in these pairs shared the **same** correct emotion label. This design evaluates whether the model can maintain consistent emotion predictions when context changes are irrelevant to the p's emotional response.

Step 3: Taxonomy-based Emotion Labeling. To ensure consistent emotion labeling, we provided crowdworkers with a predefined emotion taxonomy rather than allowing free-form responses. Free-form labeling often leads to inconsistencies, as a single emotional state can be described in diverse ways (e.g., "sad," "down," "depressed"), leading to lower agreement and ambiguous evaluation. By constraining the label space, we improved annotation agreements among workers and enabled clear comparisons between model predictions and ground truth.

We initially adopted *Plutchik's Wheel of Emotions* (Plutchik, 1982), a structured taxonomy of eight basic emotions with varying intensities. While this framework also includes compound emotional states, only the basic emotions are differentiated by intensity. This lack of finer-grained distinctions of compound emotions could lead to ambiguous labeling, as annotators are forced to map nuanced emotions to rough categories. To address this limitation, we expanded the taxonomy to include 64 emotions, enabling more nuanced differentiation across a wider range of emotional states. Crowdworkers annotated p's emotional response for each scenario based on this extended taxonomy, resulting in more consistent and fine-grained emotion labels across the dataset. See Appendix B for details.

Step 4: Distractor Addition. To create a multiple-choice question using the scenario, we added distractor labels to the scenario. To make the task more challenging and discourage cue-based shortcuts, inspired by the adversarial framework of Zellers et al. (2019), we balanced two competing objectives in selecting distractors: (1) they should be emotionally distinct from the correct label, and (2) plausible enough to mislead the model.

As a first step, we organized the 64-category emotion taxonomy into a tabular format. We then applied a rule-based filtering method to remove distractor candidates that were too close to the correct emotion—specifically, any emotions within two adjacent cells were excluded.

Next, we incorporated a model-based selection step to identify distractors likely to mislead the model. We provided the filtered emotion taxonomy and the scenario to the LLM, and prompted it to select the top three plausible emotions. This process was repeated three times, and the most frequently selected emotions were used as distractor candidates. Following prior work (Sap et al., 2019), we additionally included one distractor drawn from the other scenario in the same pair, as the two are contextually similar and thus emotionally confusable. As a result, we used the top two model-selected emotions for the sensitivity test, and the top three for the robustness test. Detailed prompts are provided in Table 9.

Step 5: Validation. To ensure the benchmark's quality and reliability, we conducted a two-round rigorous validation process on the modified scenario pairs. In the first round, three carefully selected outstanding crowdworkers independently answered each multiple-choice question. If all three unanimously selected the correct label, the scenario was accepted. Otherwise, the emotion label and context were reviewed and revised through expert consensus.

In the second round, revised scenarios were re-evaluated by a new group of crowdworkers. If the correct label was unanimously selected, the scenario was retained in the final benchmark; otherwise, it was discarded. By doing so, while emotional responses may vary across individuals, we constructed the benchmark to minimize such subjectivity and enable consistent and reliable evaluation.

4 Experiment

In this section, we evaluate recent LLMs on the TRACE benchmark and thoroughly investigate their limitations.

4.1 Baseline Models

In the experiment, we employ 9 widely used LLMs to establish our baselines. For open-source LLMs,

Model	Prompt	Sensitivity		Robustness		Avg.	
		Acc_p	Acc_q	Acc_p	Acc_q	Acc_p	Acc_q
Random	-	6.3	25.0	6.3	25.0	6.3	25.0
Open-source LLMs							
Llama-3.2-1B (MetaAI, 2024)	Base	6.5	25.4	8.2	27.0	7.3	26.1
Llama-3.2-3B (MetaAI, 2024)		34.0	60.6	54.2	66.4	42.8	63.1
Llama-3.1-8B (Dubey et al., 2024)		44.9	68.7	66.4	76.7	54.2	72.2
Gemma-2-2B (Rivière et al., 2024)		11.8	34.3	12.7	32.3	12.2	33.5
Gemma-2-9B (Rivière et al., 2024)		55.6	74.9	66.7	77.4	60.4	76.0
s1-32B (Muennighoff et al., 2025)		52.3	73.6	70.3	79.4	60.1	76.1
Llama-3.2-1B (MetaAI, 2024)	CoT	23.5	50.0	26.8	49.7	25.0	49.9
Llama-3.2-3B (MetaAI, 2024)		34.2	61.1	49.2	66.5	40.7	63.5
Llama-3.1-8B (Dubey et al., 2024)		47.5	69.4	61.3	75.4	53.5	72.0
Gemma-2-2B (Rivière et al., 2024)		29.6	57.4	40.1	58.6	34.2	57.9
Gemma-2-9B (Rivière et al., 2024)		48.1	69.8	56.5	71.0	51.8	70.4
s1-32B (Muennighoff et al., 2025)		50.6	71.6	55.1	71.4	52.2	71.1
Closed-source LLMs							
GPT-3.5-Turbo (OpenAI, 2023)	Base	39.7	65.7	65.8	75.7	51.0	70.0
GPT-4o (OpenAI, 2024a)		59.7	76.8	74.6	83.3	66.2	79.6
Claude 3.5 Sonnet (Anthropic, 2024)		63.6	79.5	74.9	84.7	68.5	81.8
o1 (OpenAI, 2024b)		65.8	81.6	77.4	85.5	70.8	83.3
GPT-3.5-Turbo (OpenAI, 2023)	CoT	42.3	65.7	59.9	72.7	49.9	68.8
GPT-4o (OpenAI, 2024a)		62.7	79.6	72.9	84.3	67.2	81.7
Claude 3.5 Sonnet (Anthropic, 2024)		61.4	78.9	69.2	81.9	64.8	80.2
o1 (OpenAI, 2024b)		66.2	81.9	79.1	87.3	71.8	84.3
Human	-	**86.3**	**93.1**	**95.7**	**97.1**	**90.7**	**95.0**

Table 2: The performances of LLMs on TRACE. For each prompting method (i.e., Base and CoT), the best results among LLMs are underlined.

we employ Llama 3.1 (8B) (Dubey et al., 2024), Llama 3.2 (1B, 3B) (MetaAI, 2024), Gemma 2 (2B, 9B) (Rivière et al., 2024), and s1-32B (Muennighoff et al., 2025). For closed-source LLMs, we experiment with GPT-3.5-Turbo (OpenAI, 2023), GPT-4o (OpenAI, 2024a), Claude 3.5 Sonnet (Anthropic, 2024), and o1 (OpenAI, 2024b). Lastly, we include a random choice baseline as a lower bound for comparison.

4.2 Prompting Methods

We adopt two prompting methods following the prior works (Sabour et al., 2024; Chen et al., 2024): a vanilla prompting (Base) and Chain-of-Thought (CoT) prompting. The Base prompting asks the model to predict the emotion without any intermediate reasoning, whereas the CoT prompting encourages step-by-step reasoning by appending "Let's think step by step." The prompt details are shown in Table 10 and 11.

4.3 Human Evaluation

Following the prior works (Zellers et al., 2019; Bisk et al., 2020), we calculate human performance using a majority vote. Five qualified crowdworkers answered each scenario, and the final answer was determined by majority vote. This aggregated human prediction is used as an upper-bound baseline. More detailed setups and results for the human evaluation are provided in Appendix G.

4.4 Evaluation Metrics

Following Fu et al. (2023), we use two evaluation metrics: pair-wise accuracy (Acc_p) and query-wise accuracy (Acc_q). Acc_p is counted only when a model correctly predicts emotions for both scenarios in the pair, making it a stricter metric. In contrast, Acc_q assigns credit for each correctly predicted scenario. Detailed formulations of these metrics are provided in Appendix D.

4.5 Main Results

We present our results in Table 2. The experimental results show that the recent LLMs significantly underperform compared to the human performance on TRACE. Even the best-performing model, o1, significantly underperforms humans. Specifically, it achieves 66.2% Acc_p and 81.9% Acc_q on the sensitivity test, falling short of human performance by 20.1% and 11.2%, respectively. On the robustness

Model	Time	Place	Agent	Avg.
Random	6.3	6.3	6.3	6.3
Open-source LLMs				
Llama-3.2-1B (2024)	2.9	10.7	6.5	6.5
Llama-3.2-3B (2024)	30.9	36.4	34.7	34.0
Llama-3.1-8B (2024)	36.0	54.5	45.2	44.9
Gemma-2-2B (2024)	9.4	19.0	9.0	11.8
Gemma-2-9B (2024)	46.8	65.3	55.8	55.6
s1-32B (2025)	48.2	53.7	53.3	52.3
Closed-source LLMs				
GPT-3.5-Turbo (2023)	35.3	41.3	41.7	39.7
GPT-4o (2024a)	51.1	66.1	61.8	59.7
Claude 3.5 Sonnet (2024)	60.4	67.8	63.3	63.6
o1 (2024b)	56.8	73.6	67.3	65.8
Human	**82.8**	**84.0**	**92.3**	**86.3**

Table 3: The performances of LLMs on the subsets of the sensitivity test, each representing a distinct context type. We report the results on the robustness counterpart in Table 6. The performance is measured by Acc_p.

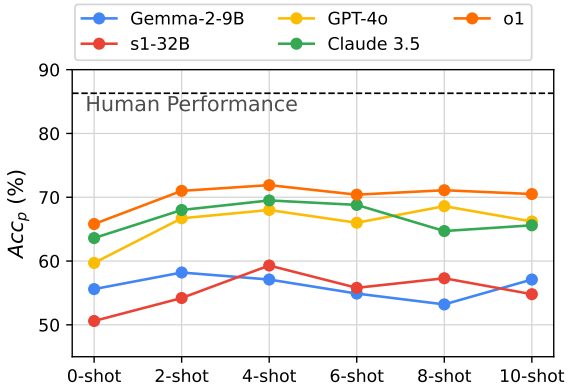


Figure 3: Results of few-shot in-context learning. We report the results on the robustness test in Figure 7. The performance is measured by Acc_p.

test, o1 reaches 79.1% Acc_p and 87.3% Acc_q, still lagging behind humans by 16.6% and 9.8%. These results indicate that while recent LLMs have shown notable performance on a wide range of NLP tasks, understanding the *emotion context sensitivity* of humans is still challenging. Additional results for open-source LLMs with larger sizes are shown in Appendix H.

4.6 Context Type Analysis

Our experimental results in Section 4.5 reveal that LLMs exhibit substantial difficulties in capturing the influence of context changes on emotional responses. However, it remains unclear whether these difficulties occur uniformly across all types of context or are particularly pronounced in a specific type. To investigate this further, we separately assess model performance across three distinct categories of context: time, place, and agent.

As shown in Table 3, capturing emotional shifts induced by changes in time context is most challenging for both humans and LLMs. Previous research suggests that temporal perception varies substantially across individuals and is influenced by personal factors such as age and sex (Hancock and Rausch, 2010). This individual variability in temporal construal may underlie the observed difficulty. Additionally, we observe a notable disparity between human and model performance with respect to agent context changes. This gap suggests that LLMs, unlike humans, struggle to capture the complex interplay between emotional responses and agent contexts. These observations suggest that, among various types of context, enhancing LLMs' ability to understand the influence of agent context on emotional responses may be especially important for developing a human-level understanding of *emotion context sensitivity*. Additional breakdown of the results across emotion labels can be found in Table 8.

4.7 Effect of Advanced Prompting Strategies

While vanilla prompting and CoT prompting in Section 4.5 establish a baseline for evaluating LLMs, recent studies indicate that advanced reasoning strategies can substantially influence the LLMs' performance. Thus, it remains unclear whether the observed performance gap between LLMs and humans arises from fundamental deficits in their emotion comprehension or from insufficient utilization of their capabilities due to suboptimal prompting strategies. To investigate this, we further examine whether the following advanced prompting strategies can mitigate the observed deficiencies.

In-context Learning. We first conduct few-shot in-context learning (Brown et al., 2020) experiments using 2-, 4-, 6-, 8-, and 10-shot settings. For each setting, we randomly sample scenario pairs from TRACE as a demonstration and evaluate the models on the remaining ones. We report median performance after repeating this process three times. As shown in Figure 3, few-shot in-context learning yields performance improvements across most models. However, these improvements remain marginal and significantly below human-level performance, suggesting that the observed limitations are not merely due to unfamiliarity with the task,

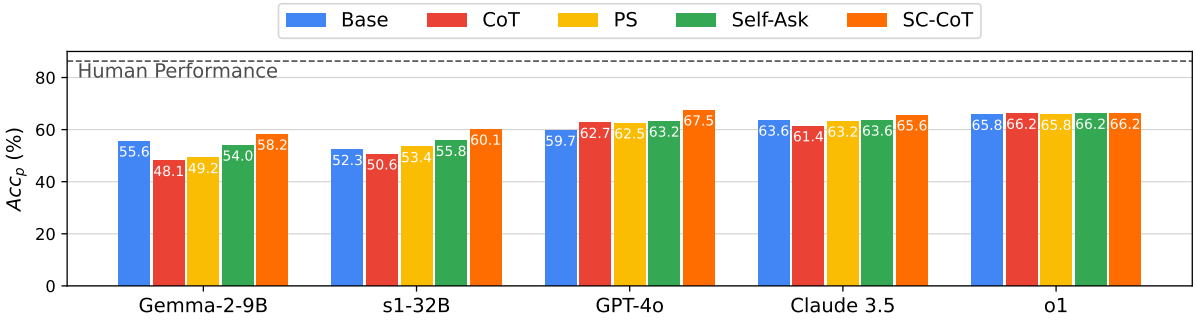


Figure 4: The performances of LLMs on the sensitivity test, with different prompting methods. We report the results on the robustness counterpart in Figure 6. The performance is measured by Acc_p.

but reflect an inherent deficiency in the models' ability to understand *emotion context sensitivity*.

CoT variants. We further apply the following CoT variants to both open-source models and closed-source models. 1) Plan-and-Solve (PS) (Wang et al., 2023a) prompts LLMs to plan to divide the task into smaller subtasks, and then carry out the subtasks according to the plan. 2) Self-Ask (Press et al., 2023) prompts LLMs to decompose complex questions into sub-questions and answer them before answering the main question. 3) Self-consistency CoT (SC-CoT) (Wang et al., 2023c) samples multiple reasoning paths and then selects the most consistent answer by marginalizing out these reasoning paths.

As shown in Figure 4, most structured reasoning methods, except for SC-CoT, result in minimal to no improvement, and sometimes even degrade performance, particularly in smaller models. While SC-CoT provides a slight gain, even the best-performing model using this method still trails human performance by nearly 20%. Taken together, these findings suggest that current LLMs exhibit the fundamental limitations in their capability for understanding human *emotion context sensitivity*.

4.8 Qualitative Analysis

To uncover latent patterns and gain deeper insight into LLM limitations, we conduct a qualitative analysis of incorrect predictions. This analysis consists of two parts: an error pattern analysis based on 50 sampled errors to identify recurring failure types, and a case study that highlights representative instances of these failures. Together, they reveal how and why models struggle to incorporate contextual cues when reasoning about emotional responses.

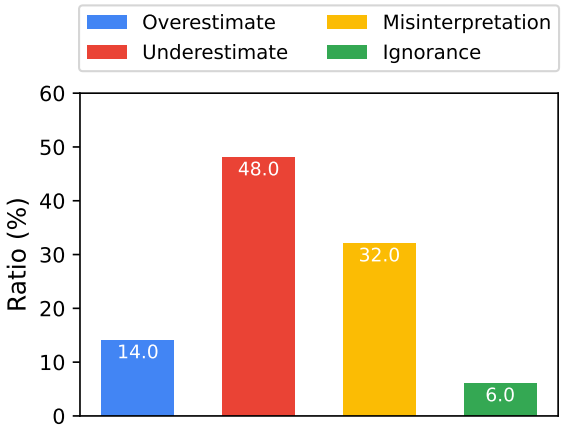


Figure 5: Pattern analysis of model errors in TRACE.

Error Pattern Analysis. To uncover recurring failure modes, we conducted a manual analysis of 50 incorrect predictions generated by the o1 model with CoT prompting. Each error was categorized by examining the model's reasoning path—specifically, how it incorporated (or failed to incorporate) contextual information when inferring emotional states. Based on this analysis, we defined four major error types: *Overestimate* (i.e., overinterpreting the emotional impact of context), *Underestimate* (i.e., underestimating or overlooking the emotional impact of context), *Misinterpretation* (i.e., misinterpreting the direction of emotional impact caused by contextual changes), and *Ignorance* (i.e., failing to reflect the effect of contextual change in the reasoning process).

The distribution of error types was as follows: 14% were classified as context overestimates, 48% as underestimates, 32% as misinterpretations, and 6% as context ignorance. These results suggest that current models tend to underestimate or misjudge how contextual changes affect emotional states.

	Example #1
Scenario 1	Sydney became sick the day *before an important job interview*.
Reasoning	Sydney became sick right before an important job ***interview***, which is typically a high-stake situation that a person would want to perform well in. "Sadness" could occur, but most would first worry about how the illness might affect their preparation and performance. "Awe" doesn't seem appropriate here, as there is no indication of wonder or amazement relating to the situation. "Relief" would only make sense if Sydney did not want to attend the interview. However, being sick right before a critical interview more likely escalates concern rather than providing comfort. "Anxiety" aligns strongly with the uncertainty about whether they can attend or do well in the interview while feeling unwell, reflecting the stress and worry Sydney would probably experience.
Prediction / GT	Anxiety / Anxiety ✓
Scenario 2	Sydney became sick the day *after an important job interview*.
Reasoning	Sydney may be worried about whether the ***interview*** went well and how her sudden illness might affect her upcoming opportunities or her ability to follow up on the interview. Doubts and concerns about her performance or the interviewer's decision could amplify her stress. Feeling sick right after such an important event often fuels uncertainty: she cannot be sure whether she presented herself in the best light, and now she might fear follow-up interviews while she's unwell. Hence, the predominant emotion is likely Anxiety.
Prediction / GT	Anxiety / Sadness ✗
	Example #2
Scenario 1	Kai got out of a muddy pond. Kai got the dirty water off their chest *during the mud festival*.
Reasoning	Kai participated in a mud ***festival***, which is typically a fun, playful environment. While getting muddy might be temporarily uncomfortable, the overall atmosphere is usually lighthearted. Kai successfully got the dirty water off, suggesting relief from discomfort. Mud ***festivals*** are generally associated with enjoyment and celebration rather than fear, frustration, or disapproval.
Prediction / GT	Joy / Joy ✓
Scenario 2	Kai got out of a muddy pond. Kai got the dirty water off their chest *before her wedding*.
Reasoning	Disapproval would imply a moral or ethical judgment, which doesn't necessarily fit Kai's situation. She likely wouldn't be disapproving of herself or her circumstances; the muddy pond was an accident. Frustration is possible initially because getting muddy right before a ***wedding*** can be aggravating. However, since she managed to clean up "just before" the ***wedding***, any lingering frustration would probably be overshadowed by relief. Fear might occur if she worried about ruining her dress or the ceremony, but once clean, fear would subside. Joy seems most likely the final emotion. Having successfully removed the mud in time, Kai would feel relieved and happy that her ***wedding*** could proceed smoothly, leading to an ultimate feeling of joy.
Prediction / GT	Joy / Frustration ✗

Table 4: Case study of o1's predictions. Scenario contexts are underlined, while salient words conventionally associated with emotions (e.g., festival-joy, interview-anxiety) are marked in ***bold italics and blue***. The reasoning associated with the salient words is highlighted in light blue for emphasis. "GT" denotes a ground-truth answer.

Case Study. While our error pattern analysis categorizes common types of failure based on model predictions, it does not fully reveal the internal reasoning processes that lead to these errors. To better understand how models arrive at their predictions, we manually and thoroughly analyze the CoT reasoning trajectories. CoT explicitly elicits intermediate reasoning steps, enabling us to assess whether models follow contextually appropriate reasoning paths or exhibit systematic biases.

A recurring issue we observe is that LLMs tend to over-rely on emotionally salient keywords. Rather than performing structured reasoning based on the full context, models often latch onto emotionally salient words—such as *interview*, *festival*, or *wedding*—and generate reasoning paths that reflect conventional emotional associations (e.g., *interview-anxiety*, *festival-joy*). This leads to correct predictions when the emotional connotation aligns with the actual context, but fails when it does not.

For instance, as shown in Table 4, scenario 1 of each example shows that the model correctly reasons about emotional responses when the context supports the conventional association. In contrast, in scenario 2, the model ignores conflicting contextual cues and instead generates explanations that reflect stereotypical associations, resulting in incorrect predictions.

5 Conclusion

We present the first systematic study to assess the extent to which LLMs understand *emotion context sensitivity*—a key component of human emotional intelligence. Grounded in behavioral theory and emotion psychology, we define time, place, and agent as three core contextual dimensions, and use them to construct TRACE—a controlled benchmark that isolates each type of context change to assess whether LLMs can track emotional responses. Our

analysis on TRACE results in three key findings. First, current LLMs exhibit a large performance gap relative to humans in understanding *emotion context sensitivity*. Second, even with advanced prompting techniques, models show limited improvement, suggesting an inherent limitation in processing how contextual cues shape emotional responses. Third, LLMs struggle to account for individual differences in emotional experience, as evidenced by their pronounced performance gap in agent-modified contexts. We hope that TRACE serves as a challenging benchmark for future research in the emotional intelligence of AI and facilitates the development of models capable of context-aware emotional reasoning.

6 Limitations

While TRACE provides a systematic evaluation of the LLMs' capability in understanding *emotion context sensitivity* of humans, it has several limitations that future research should address.

First, TRACE reflects individual differences in only a subset of scenarios. To clearly attribute emotional changes to each context dimension, we designed the benchmark such that only one type of context is modified at a time. However, even the same contextual changes in time or place can lead to different emotional responses depending on personal factors, such as cultural background (Greenaway et al., 2018). Therefore, this design choice can make the character's emotional state ambiguous in scenarios without agent-specific context. This issue is reflected in lower human performance for non-agent context types compared to agent-modified scenarios.

Second, TRACE is limited to the textual modality. We adopted this design to isolate failures in understanding of *emotion context sensitivity* from those arising due to difficulties in processing non-textual modalities (Kervadec et al., 2021; Lao et al., 2023; Tong et al., 2024). However, this comes at the cost of overlooking the inherently multimodal nature of contextual cues, which are often conveyed through visual, auditory, and other sensory signals in real-world scenarios.

Third, our benchmark focuses solely on affective experience—that is, whether LLMs can infer emotional states given a specific context—while other important components of emotion context sensitivity, such as emotional expression and regulation, remain unaddressed. We intentionally limit our scope to affective experience as a first step; however, it inevitably restricts a more comprehensive evaluation of LLMs' understanding of *emotion context sensitivity*.

Lastly, we acknowledge that the benchmark skews towards US residents. As such, the collected scenarios may reflect emotional experiences more common to this group. Therefore, future work should prioritize collecting scenarios from under-represented populations in order to assess LLMs' understanding of the emotion context sensitivity of humans in these groups. This, however, poses a notable challenge, given the current demographic distribution of workers on crowdsourcing platforms and the additional complexity of designing data collection forms in languages other than English.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.RS-2025-00517221 and No.RS-2024-00415812) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2024-00439328, Karma: Towards Knowledge Augmentation for Complex Reasoning (SW Starlab), No.RS-2024-00457882, AI Research Hub Project, and No.RS-2019-II190079, Artificial Intelligence Graduate School Program (Korea University)).

References

Anthropic. 2024. Claude 3.5 sonnet.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: reasoning about physical commonsense in natural language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 7432–7439. AAAI Press.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Nancy J Burke, Galen Joseph, Rena J Pasick, and Judith C Barker. 2009. Theorizing social context: Rethinking behavioral theory. *Health Education & Behavior*, 36(5_suppl):55S–70S.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. 2024. Tombench: Benchmarking theory of mind in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024*, pages 15959–15983. Association for Computational Linguistics.

Karin G Coifman and George A Bonanno. 2009. Emotion context sensitivity in adaptation and recovery. *Emotion regulation and psychotherapy*, pages 157–173.

Karin G Coifman and George A Bonanno. 2010. When distress does not become depression: emotion context sensitivity and adjustment to bereavement. *Journal of abnormal psychology*, 119(3):479.

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

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. 2023. MME: A comprehensive evaluation benchmark for multimodal large language models. *CoRR*, abs/2306.13394.

Deepanway Ghosal, Siqi Shen, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. 2022. CICERO: A dataset for contextualized commonsense inference in dialogues. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, ACL 2022*, pages 5010–5028. Association for Computational Linguistics.

K Elise Goubet and Evangelia G Chrysikou. 2019. Emotion regulation flexibility: Gender differences in context sensitivity and repertoire. *Frontiers in psychology*, 10:935.

Katharine H Greenaway, Elise K Kalokerinos, and Lisa A Williams. 2018. Context is everything (in emotion research). *Social and Personality Psychology Compass*, 12(6):e12393.

Seungju Han, Beomsu Kim, Jin Yong Yoo, Seokjun Seo, Sangbum Kim, Enkhbayar Erdenee, and Buru Chang. 2022. Meet your favorite character: Open-domain chatbot mimicking fictional characters with only a few utterances. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022*, pages 5114–5132. Association for Computational Linguistics.

Peter A Hancock and Robert Rausch. 2010. The effects of sex, age, and interval duration on the perception of time. *Acta psychologica*, 133(2):170–179.

Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao K. Huang, and Lun-Wei Ku. 2018. Emotionlines: An emotion corpus of multi-party conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018*. European Language Resources Association.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *CoRR*, abs/2401.04088.

Molly Joy and Asha Mathew. 2018. Emotional maturity and general well-being of adolescents. *IOSR Journal of Pharmacy*, 8(5):01–06.

Corentin Kervadec, Grigory Antipov, Moez Baccouche, and Christian Wolf. 2021. Roses are red, violets are blue... but should VQA expect them to? In *IEEE*

Conference on Computer Vision and Pattern Recognition, CVPR 2021, pages 2776–2785. Computer Vision Foundation / IEEE.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2023. SODA: million-scale dialogue distillation with social commonsense contextualization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023*, pages 12930–12949. Association for Computational Linguistics.

Mingrui Lao, Nan Pu, Yu Liu, Kai He, Erwin M. Bakker, and Michael S. Lew. 2023. COCA: collaborative causal regularization for audio-visual question answering. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023*, pages 12995–13003. AAAI Press.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017*, pages 986–995. Asian Federation of Natural Language Processing.

Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. 2020. Towards conversational recommendation over multi-type dialogs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, pages 1036–1049. Association for Computational Linguistics.

MetaAI. 2024. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. *CoRR*, abs/2501.19393.

OpenAI. 2023. Gpt-3.5 turbo fine-tuning and api updates.

OpenAI. 2024a. Hello gpt-4o.

OpenAI. 2024b. Learning to reason with llms.

Samuel J. Paech. 2023. Eq-bench: An emotional intelligence benchmark for large language models. *CoRR*, abs/2312.06281.

R Plutchik. 1982. A psycho evolutionary theory of emotions. *Social Science Information*, 21(4-5):529–553.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019*, pages 527–536. Association for Computational Linguistics.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711. Association for Computational Linguistics.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019*, pages 5370–5381. Association for Computational Linguistics.

Morgane Rivière, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozinska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucinska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju-yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjösund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, and Lilly McNealus. 2024. Gemma 2: Improving open language models at a practical size. *CoRR*, abs/2408.00118.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June M. Liu, Jinfeng Zhou, Alvionna S. Sunaryo, Tatia M. C. Lee, Rada Mihalcea, and Minlie Huang. 2024. Emobench: Evaluating the emotional intelligence of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024*, pages 5986–6004. Association for Computational Linguistics.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social iqa: Commonsense reasoning about social interactions. In *Proceed-*

ings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, pages 4462–4472. Association for Computational Linguistics.

Thierry Steimer. 2002. The biology of fear-and anxiety-related behaviors. *Dialogues in clinical neuroscience*, 4(3):231–249.

Carlo Strapparava and Alessandro Valitutti. 2004. Wordnet affect: an affective extension of wordnet. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004*. European Language Resources Association.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024*, pages 9568–9578. IEEE.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023a. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, ACL 2023*, pages 2609–2634. Association for Computational Linguistics.

Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Jia Liu. 2023b. Emotional intelligence of large language models. *Journal of Pacific Rim Psychology*, 17:18344909231213958.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023c. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023*. OpenReview.net.

Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. 2024. Opentom: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024*, pages 8593–8623. Association for Computational Linguistics.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019*, pages 6720–6731. Computer Vision Foundation / IEEE.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, pages 3361–3367. Association for Computational Linguistics.

Kun Zhou, Xiaolei Wang, Yuanhang Zhou, Chenzhan Shang, Yuan Cheng, Wayne Xin Zhao, Yaliang Li, and Ji-Rong Wen. 2021. Crslab: An open-source toolkit for building conversational recommender system. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL 2021 - System Demonstrations*, pages 185–193. Association for Computational Linguistics.

Appendix

A Human Annotation Quality Control

To ensure high-quality annotations, we employed workers on MTurk with a qualification test. 19% of the candidates passed the qualification test and proceeded to the data collection pipeline. Furthermore, we continuously monitored the quality of submitted work, revoking qualifications for workers who repeatedly submitted low-quality data. Figure 12 shows the excerpt from HIT instructions of the qualification test, and Table 5 summarizes the demographic information of workers who passed the qualification test. We include only individuals over the age of 20 to ensure emotional maturity, the ability to stabilize emotions (Joy and Mathew, 2018). We informed all the workers of the purpose of our study, and the workers provided their consent to participate in the project. Each task was compensated at $9.03 per hour, which is above the U.S. federal minimum wage.

B Details of Emotion Taxonomy Construction

In this section, we detail the construction of the emotion labeling taxonomy used in data collection. Our taxonomy is grounded in *Plutchik's Wheel of Emotions* (Plutchik, 1982), a widely used framework in psychological literature. Specifically, *Plutchik's Wheel* proposes a total of 32 emotions, comprising eight basic emotions, eight stronger and eight milder variants of those basic emotions, and eight compound emotions (i.e., dyads) formed by combining pairs of adjacent basic emotions. While this model provides a structured foundation, it lacks intensity variations for the eight compound emotions. This absence may introduce ambiguity during annotation, as annotators have no clear reference points for differentiating intensity levels in these cases.

To fill this gap, we expanded the taxonomy using 280 emotion words from WordNet-Affect (Strapparava and Valitutti, 2004), a lexical resource that offers broad coverage of affective vocabulary. Based on this pool, we consulted experts to identify high- and low-intensity variants for the eight compound emotions, resulting in 16 new emotion labels that introduce intensity distinctions previously missing from the original model. If multiple candidate words were available for a given emotion, we selected the most frequently used term. In addition, during the qualification test, annotators occasionally labeled emotions not present in the taxonomy. Among these, several terms appeared consistently across multiple annotators. We incorporated the most commonly mentioned of these into the taxonomy, grouping them into two supplementary categories—Extreme Emotions and Other Emotions—which brought the total number of emotion labels to 64. Figure 11 shows the complete taxonomy of emotions used to construct TRACE.

C Observer-based Evaluation in TRACE

In this study, we adopt an observer-based evaluation framework for TRACE, where LLMs are tasked with analyzing scenarios from a third-person perspective during evaluation. A natural question may arise: why choose this approach over alternatives such as role-playing frameworks?

We adopt the observer-based framework for two primary reasons. First, we believe this approach more accurately reflects the typical setting in which current LLMs operate: they are generally expected to analyze and interpret situations from an external perspective, rather than simulate internal emotional states. Second, prompts framed from an observer's viewpoint tend to be less ambiguous and easier for LLMs to process. In contrast, role-playing tasks introduce additional challenges, such as defining fictional characters with only minimal contextual information (Han et al., 2022). Nevertheless, we consider role-playing a promising direction for future research, as it may offer complementary insights into LLMs' capacity for contextual emotion understanding.

D Definition of Evaluation Metrics

In this section, we formally define the two evaluation metrics used in our experiments: pair-wise accuracy (Acc_p) and query-based accuracy (Acc_q). Let o_i $(i = 1, 2)$ denote the emotion predicted by the LLM for the pivot character p in scenario s_i, and let y_i be the correct label among the four emotion options. The metrics are computed as follows:

$$\text{Acc}_p = \frac{1}{n_p}\sum_{j=1}^{n_p}(\mathbb{1}[o_{1j} = y_{1j}] \times \mathbb{1}[o_{2j} = y_{2j}]),$$

$$\text{Acc}_q = \frac{1}{2n_p}\sum_{j=1}^{n_p}(\mathbb{1}[o_{1j} = y_{1j}] + \mathbb{1}[o_{2j} = y_{2j}]), \quad (1)$$

Age		Gender		Education		Economic class		Location	
20-29	34.3%	Female	47.8%	High-school or equivalent	6.0%	Lower	9.0%	US	97.0%
30-39	26.9%	Male	52.2%	Some college (no degree)	6.0%	Working	43.3%	non-US	3.0%
40-49	23.9%			Bachelor's degree	67.2%	Middle	35.8%		
50-59	11.9%			Graduate degree	20.8%	Upper-middle	11.9%		
60-	3.0%								

Table 5: Breakdown of crowdworker demographics by age, gender, education, economic class, and location.

Model	Time	Place	Agent	Avg.
Random	6.3	6.3	6.3	6.3
Open-source LLMs				
Llama-3.2-1B (2024)	5.4	10.7	9.1	8.2
Llama-3.2-3B (2024)	56.2	58.3	48.8	54.2
Llama-3.1-8B (2024)	66.2	67.0	66.1	66.4
Gemma-2-2B (2024)	11.5	14.6	12.4	12.7
Gemma-2-9B (2024)	65.4	65.0	69.4	66.7
s1-32B (2025)	71.5	68.0	68.6	69.5
Closed-source LLMs				
GPT-3.5-Turbo (2023)	67.7	67.0	62.8	65.8
GPT-4o (2024a)	83.1	67.0	71.9	74.6
Claude 3.5 Sonnet (2024)	<u>83.1</u>	73.8	66.9	74.9
o1 (2024b)	75.4	<u>83.5</u>	<u>74.4</u>	<u>77.4</u>
Human	**95.2**	**96.2**	**95.7**	**95.7**

Table 6: The performances of LLMs on the subsets of the robustness test, each representing a distinct context type. Performance is measured by Acc_p. Second-best results are <u>underlined</u>.

where n_p denotes the number of scenario pairs.

E Results on the Robustness Test

Here, we present results on the robustness test, including an analysis of performance across different context types and the effect of advanced prompting strategies.

Context Type Analysis. In Table 6, we observe a similar pattern of model vulnerability: LLMs exhibit the lowest performance in scenarios involving changes in time and agent context. Interestingly, human annotators achieve near-ceiling performance (above 95%) across all context types, suggesting that context changes unrelated to emotion are generally unambiguous to humans. These results underscore the importance of enhancing models' robustness to misleading yet emotionally irrelevant context shifts.

Advanced Prompting Strategies. We evaluated the effects of advanced prompting strategies—including few-shot in-context learning and CoT variants—on the robustness test. As shown in Figure 6 and Figure 7, both types of prompting tend to degrade performance across most models. This trend suggests that increased reasoning, rather than helping models ignore irrelevant context changes, often leads them to overestimate these changes and produce incorrect emotional predictions. In particular, smaller models are more prone to this overestimation behavior. Notably, o1 is the only model that benefits from these prompting strategies, showing consistent performance improvements. While the precise reason remains unclear, future investigation of this phenomenon may help reveal the reasoning capabilities required for understanding human *emotion context sensitivity*.

Overall, the findings suggest that emotional context robustness requires the ability to selectively reason about emotionally relevant cues—an ability that current LLMs often lack.

F Subcategory Analysis of Contexts

In this section, we provide additional details on three context types used in TRACE: time, place, and agent. While these high-level categories were designed based on psychological literature (Burke et al., 2009; Greenaway et al., 2018), their broad scope raises a natural question: does the dataset sufficiently capture meaning diversity within each context type?

As a response, we analyzed the collected data using subcategories identified in prior research. For the time context, we adopted four temporal dimensions from (Zhou et al., 2019): duration (how long an event takes), temporal ordering (temporal order of events), typical time (when an event happens), and frequency (how often an event occurs). For the place context, we focused on concrete physical settings such as parks, offices, and airports. For the agent context, inspired by Greenaway et al. (2018), we encouraged workers to include details related to demographics, personality traits, situational appraisals, and social relationships.

Figure 9 shows the distribution of these subcategories within the time and agent contexts. The re-

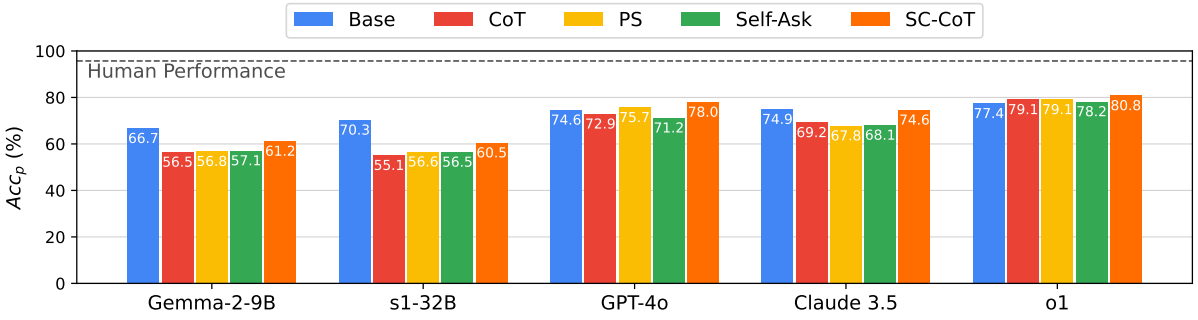


Figure 6: The performances of LLMs on the robustness test, with different prompting methods. The performance is measured by Acc_p.

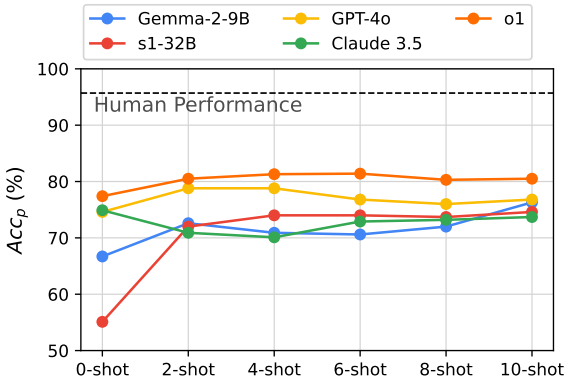


Figure 7: Results of LLMs with few-shot in-context learning on the robustness test. Each setting is evaluated three times, and the median performance is reported.

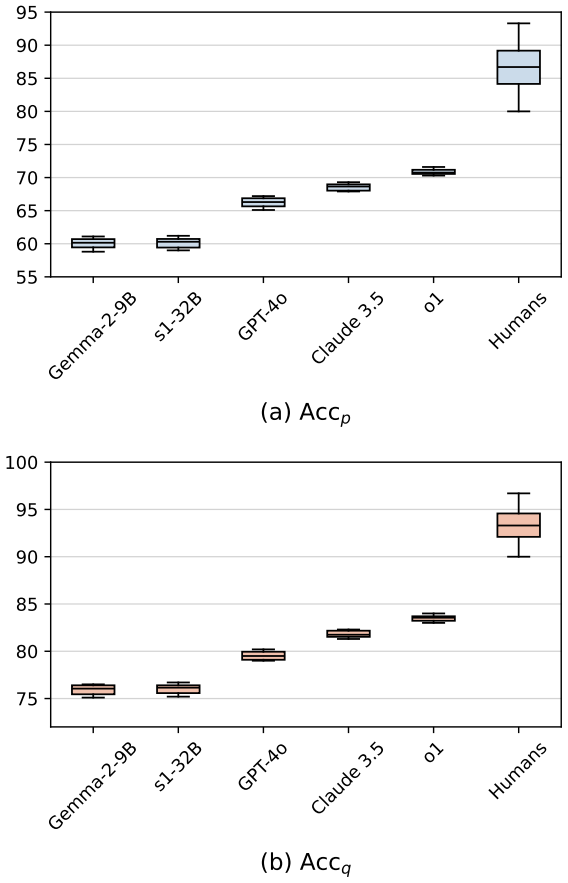


Figure 8: Distribution of the overall performances of LLMs and humans. The results for LLMs are based on 10 different runs. We use Acc_p and Acc_q as evaluation metrics for subplots (a) and (b), respectively.

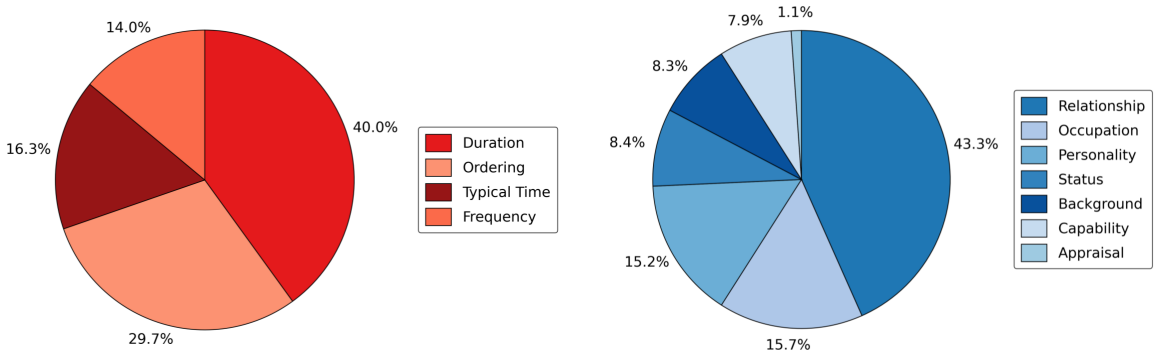
sults confirm that our dataset reflects a wide range of psychologically meaningful variations within each context type, addressing concerns about over-generalization in our design.

G Human Performance Measurement

We report human performance measured under different settings in the main text and the appendix, respectively.

Details of Human Performance Measurement in Main Text In the first experiment, we randomly assigned five different crowdworkers to each of the 300 scenarios sampled from our benchmark. Following the setup commonly used in prior work (Zellers et al., 2019; Bisk et al., 2020), final answers were aggregated via majority vote, and the resulting accuracy was used as the reference for human performance in Section 4.

Individual Variability in Human Performance To examine individual variability in human performance, we conducted an additional experiment focusing on differences across crowdworkers. In this experiment, 10 crowdworkers from the initial human performance measurement were each asked to independently complete the same set of 60 randomly selected scenarios from TRACE. This setup allowed us to measure the variance in human performance and whether model performance is statisti-



(a) Time (b) Agent

Figure 9: Distribution of subcategories in the (a) time and (b) agent contexts.

cally distinguishable from that of humans. Figure 8 summarizes individual human performances. Acc_p scores ranged from 80.0% to 93.3% (SD: 4.1%), and Acc_q scores ranged from 90.0% to 96.7% (SD: 2.0%), indicating notable variability across crowdworkers. This variation reflects individual differences in emotional understanding ability, as reported in a prior study (Wang et al., 2023b). Despite this, crowdworkers consistently outperform LLMs. A two-sample t-test confirms that the performance gap is statistically significant ($p < 0.05$).

Figure 13 shows an excerpt from the MTurk HIT instructions used in the human evaluation. Each task was compensated at $9.03 per hour, which is above the U.S. federal minimum wage.

H Additional Results for Larger Open-source LLMs

To further investigate how state-of-the-art LLMs understand *emotion context sensitivity* of humans, we present the additional results for the larger open-source models, Llama-3-70B (MetaAI, 2024) and Mixtral 8x7B (Jiang et al., 2024), in Table 7. As shown, both models perform notably below human levels in both sensitivity and robustness tests, consistent with the findings in Table 2. These results show that understanding the *emotion context sensitivity* of humans remains a substantial challenge even for larger open-source models.

I Statistical Analysis of Collected Data

To better understand the characteristics of our dataset, we conduct a statistical analysis on the distribution of emotions and contextual dimensions. Figure 10 presents the overall distribution of emotions in our dataset, showing the relative frequency of each emotion category.

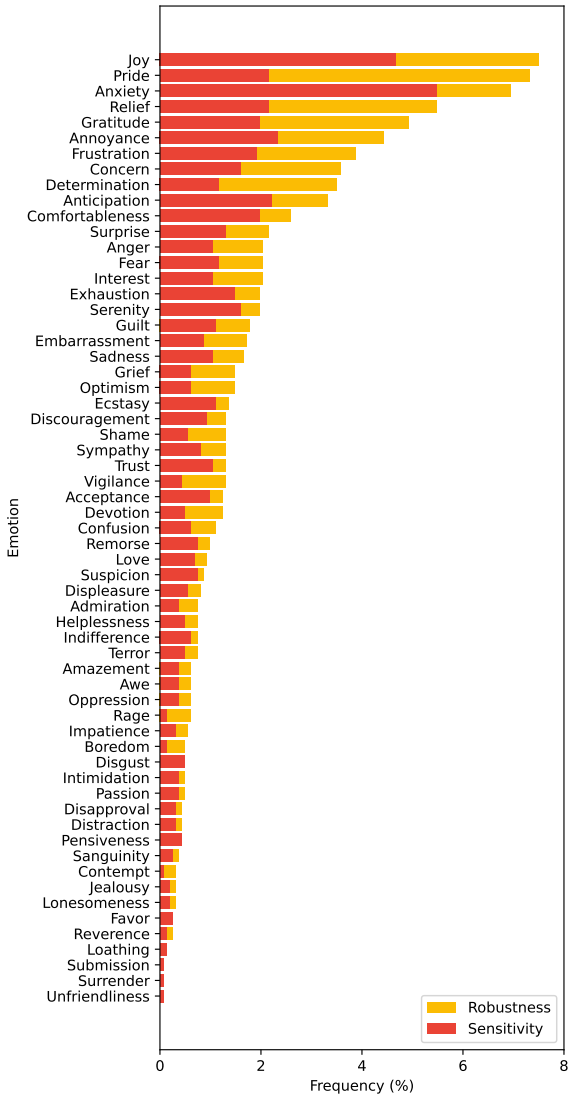


Figure 10: Emotion distribution in TRACE.

J Crowdsourcing for Data Collection

To collect high-quality annotations for our dataset, we conducted a crowdsourcing task using MTurk.

Model	Prompt	Sensitivity		Robustness		Avg.	
		Acc_p	Acc_q	Acc_p	Acc_q	Acc_p	Acc_q
Llama-3-70B (MetaAI, 2024)	Base	60.6	78.3	72.6	82.1	65.8	80.0
Mixtral 8x7B (Jiang et al., 2024)		40.7	64.8	57.1	70.6	47.8	67.3
Llama-3-70B (MetaAI, 2024)	CoT	61.4	79.5	74.0	83.7	66.9	81.3
Mixtral 8x7B (Jiang et al., 2024)		42.0	65.4	59.1	71.1	49.4	67.9

Table 7: The performances of additional open-source LLMs on TRACE.

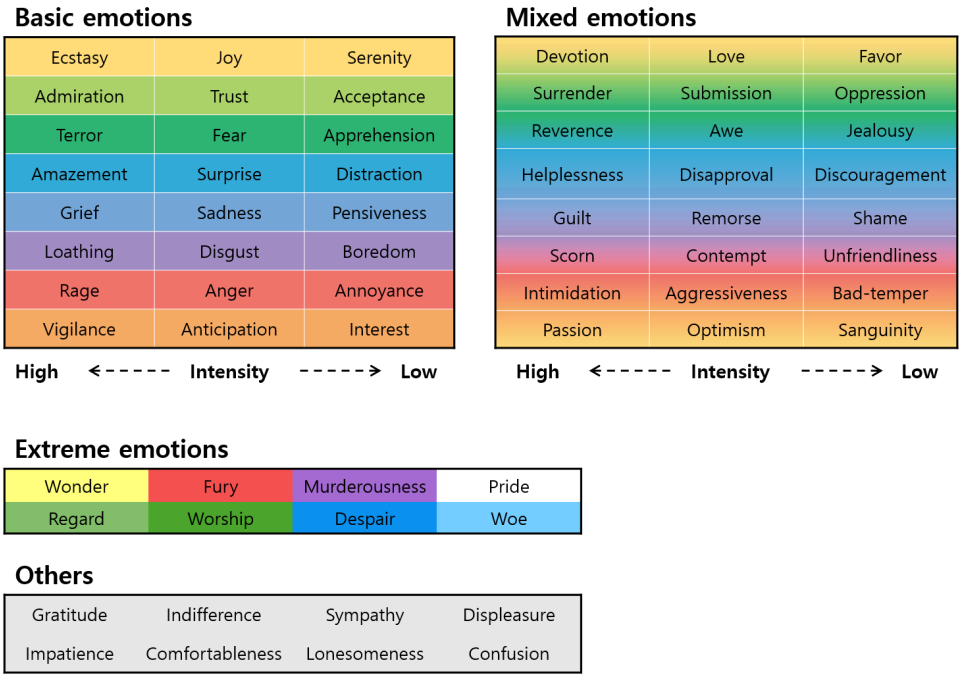


Figure 11: Emotion taxonomy in TRACE.

Examples of HITs are shown in Figure 12, Figure 13, and Figure 14. The full crowdsourcing interface will be made available on GitHub.

K Experiment Prompts

Table 10 and Table 11 present the prompts used in our experiments. We designed these prompts to effectively assess how well LLMs understand *emotion context sensitivity* of humans across different contextual variations. Each prompt was carefully crafted to maintain consistency while ensuring that models receive sufficient information to make context-aware emotional predictions.

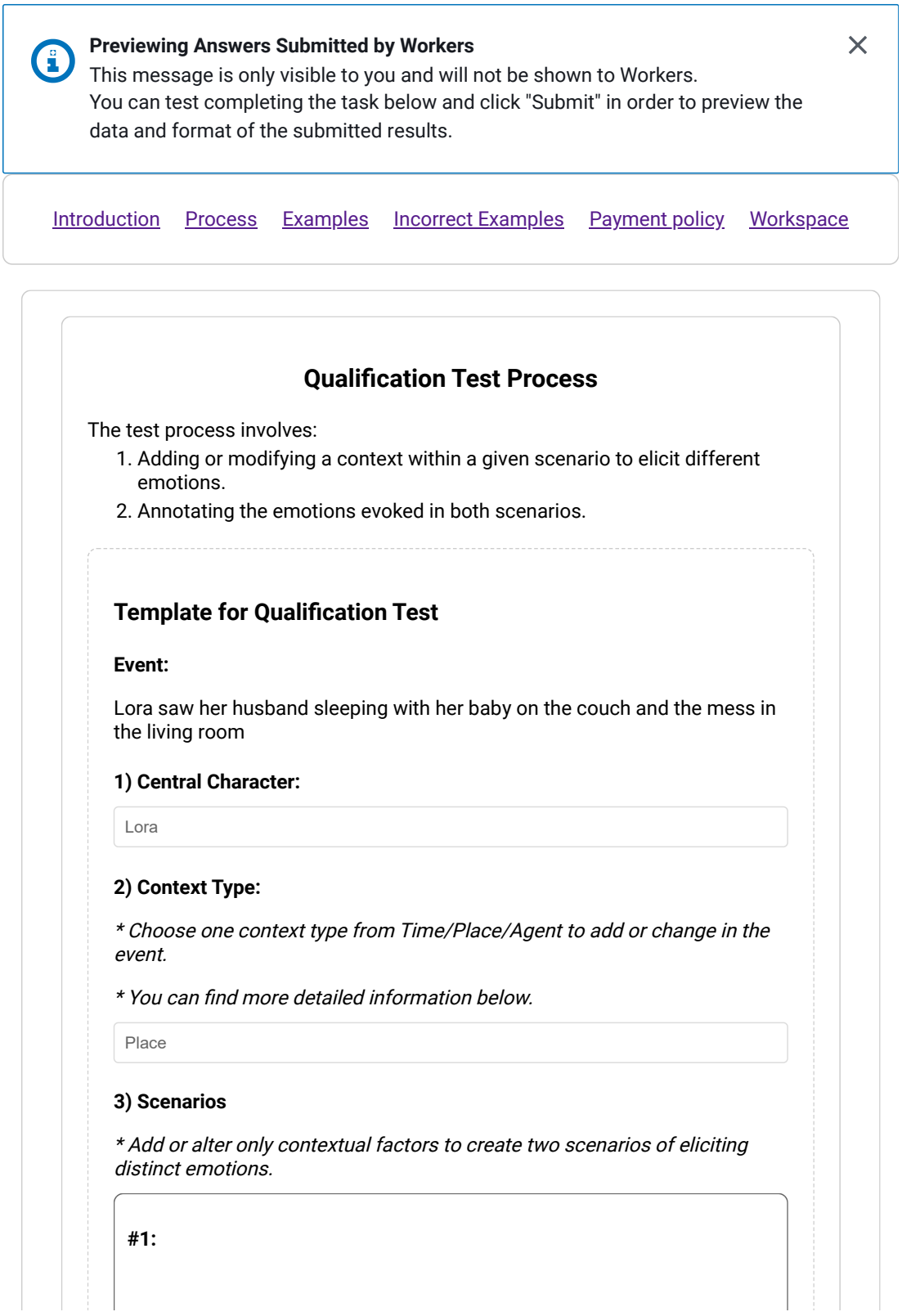


Figure 12: Excerpt from MTurk HIT instructions: Qualification test.

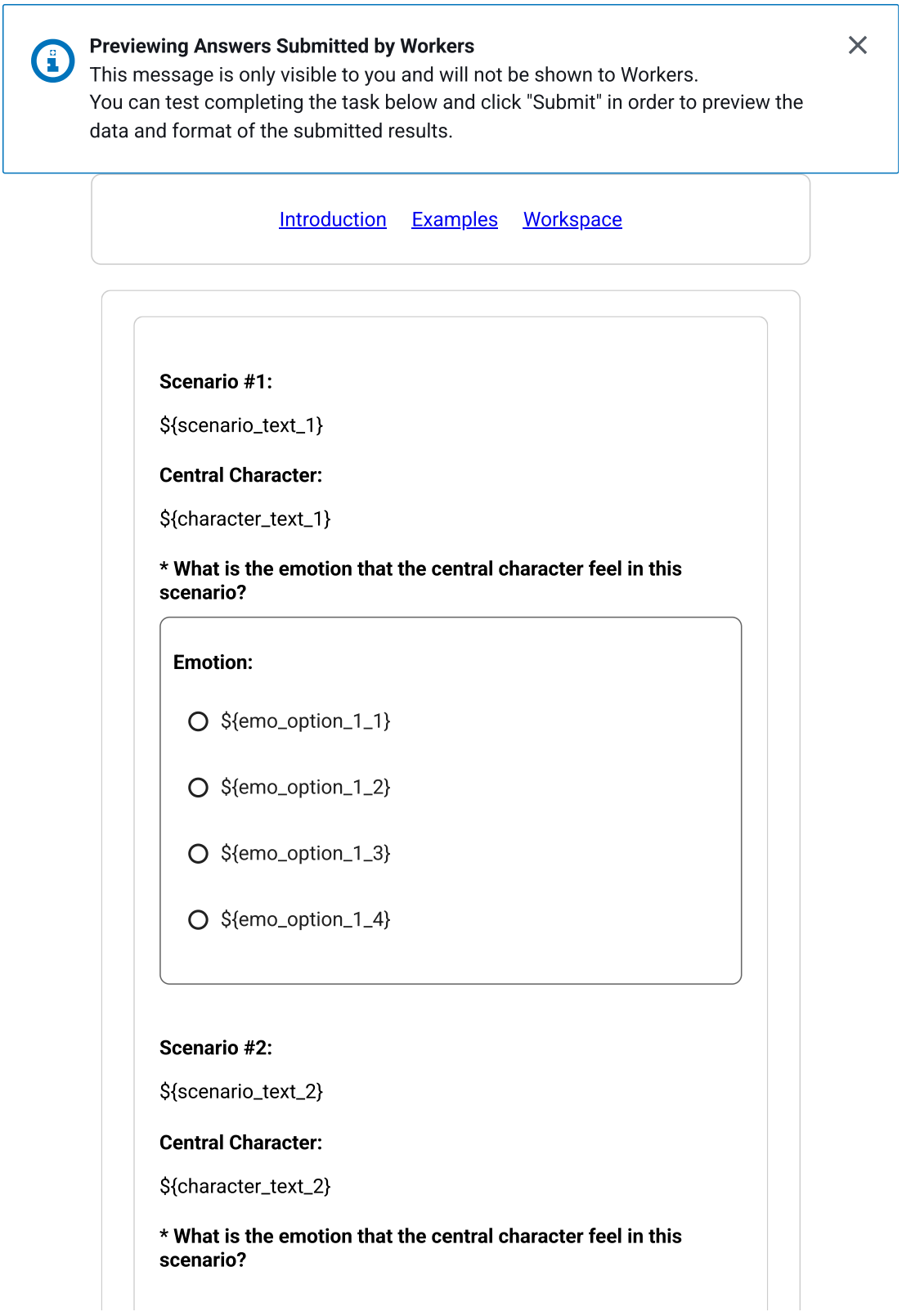


Figure 13: Excerpt from MTurk HIT instructions: Human evaluation.

Previewing Answers Submitted by Workers
This message is only visible to you and will not be shown to Workers.
You can test completing the task below and click "Submit" in order to preview the data and format of the submitted results.

Introduction Process Examples Payment policy Workspace

Emotion Types

Basic emotions

Ecstasy	Joy	Serenity
Admiration	Trust	Acceptance
Terror	Fear	Apprehension
Amazement	Surprise	Distraction
Grief	Sadness	Pensiveness
Loathing	Disgust	Boredom
Rage	Anger	Annoyance
Vigilance	Anticipation	Interest

High <----- Intensity -----> Low

Mixed emotions

Devotion	Love	Favor
Surrender	Submission	Oppression
Reverence	Awe	Jealousy
Helplessness	Disapproval	Discouragement
Guilt	Remorse	Shame
Scorn	Contempt	Unfriendliness
Intimidation	Aggressiveness	Bad-temper
Passion	Optimism	Sanguinity

High <----- Intensity -----> Low

Extreme emotions

Wonder	Fury	Murderousness	Pride
Regard	Worship	Despair	Woe

Others

Gratitude	Indifference	Sympathy	Displeasure
Impatience	Comfortableness	Lonesomeness	Confusion

You have to choose one emotion within the above list.

* **Basic emotions** represent various intensities of feelings derived from eight core emotions. (e.g., joy, anger)

* **Mixed emotions** occur when a person experiences two or more emotions simultaneously, often of opposite valence (e.g., one positive and one negative).

Workspace

Event:

${event_text}

1) Central Character:

Figure 14: Excerpt from MTurk HIT instructions: Data collection.

Emotion	Gemma-2-9B	s1-32B	GPT-4o	Claude 3.5	o1
Joy	91.0	88.5	98.4	96.7	95.1
Pride	79.0	83.2	88.2	89.1	87.4
Anxiety	85.8	59.3	74.3	85.0	88.5
Relief	87.6	91.0	93.3	83.1	93.3
Gratitude	88.8	97.5	92.5	93.8	95.0
Annoyance	79.2	68.1	81.9	88.9	86.1
Frustration	95.2	90.5	82.5	84.1	87.3
Concern	72.4	60.3	70.7	74.1	77.6
Determination	47.4	50.9	59.6	64.9	68.4
Anticipation	83.3	81.5	79.6	83.3	81.5
Comfortableness	81.0	83.3	88.1	88.1	90.5
Surprise	45.7	45.7	71.4	51.4	62.9
Anger	75.8	75.8	90.9	81.8	90.9
Fear	75.8	60.6	66.7	84.8	84.8
Interest	57.6	60.6	72.7	69.7	63.6
Exhaustion	68.8	84.4	90.6	78.1	93.8
Serenity	75.0	96.9	87.5	93.8	87.5
Guilt	75.0	89.3	78.6	92.9	82.1
Embarrassment	67.9	67.9	64.3	71.4	89.3
Sadness	88.9	85.2	88.9	88.9	85.2
Grief	87.5	83.3	87.5	83.3	83.3
Optimism	87.5	87.5	83.3	83.3	79.2
Ecstasy	81.8	95.5	77.3	90.9	90.9
Discouragement	81.0	85.7	66.7	81.0	81.0
Shame	71.4	66.7	66.7	71.4	85.7
Sympathy	52.4	66.7	81.0	90.5	81.0
Trust	47.6	76.2	52.4	85.7	66.7
Vigilance	52.4	57.1	66.7	61.9	57.1
Acceptance	65.0	90.0	55.0	85.0	75.0
Devotion	65.0	70.0	60.0	60.0	50.0
Confusion	61.1	50.0	72.2	88.9	77.8
Remorse	75.0	87.5	68.8	81.3	68.8
Love	86.7	93.3	93.3	86.7	93.3
Suspicion	42.9	42.9	78.6	78.6	85.7
Displeasure	92.3	92.3	84.6	92.3	84.6
Admiration	58.3	75.0	91.7	58.3	91.7
Helplessness	66.7	58.3	58.3	58.3	83.3
Indifference	50.0	33.3	75.0	33.3	50.0
Terror	83.3	83.3	91.7	75.0	91.7
Amazement	70.0	50.0	50.0	40.0	70.0
Awe	60.0	70.0	40.0	70.0	70.0
Oppression	70.0	60.0	80.0	90.0	40.0
Rage	50.0	50.0	50.0	60.0	80.0
Impatience	44.4	44.4	44.4	55.6	33.3
Boredom	87.5	100.0	87.5	87.5	100.0
Disgust	87.5	50.0	87.5	100.0	87.5
Intimidation	75.0	87.5	62.5	75.0	87.5
Passion	62.5	62.5	75.0	62.5	50.0
Disapproval	57.1	85.7	100.0	100.0	100.0
Distraction	71.4	57.1	71.4	85.7	85.7
Pensiveness	71.4	42.9	57.1	28.6	42.9
Sanguinity	83.3	83.3	66.7	83.3	66.7
Contempt	20.0	60.0	60.0	60.0	60.0
Jealousy	60.0	60.0	80.0	80.0	100.0
Lonesomeness	40.0	60.0	40.0	80.0	80.0
Favor	100.0	100.0	100.0	75.0	100.0
Reverence	75.0	100.0	100.0	75.0	100.0
Loathing	100.0	100.0	50.0	50.0	100.0
Submission	0.0	0.0	100.0	100.0	100.0
Surrender	100.0	100.0	100.0	100.0	100.0
Unfriendliness	0.0	0.0	0.0	0.0	0.0

Table 8: Breakdown of results on TRACE across emotion labels. We use Acc_q as an evaluation metric because Acc_p for each emotion label is not defined.

System Prompt
You are a helpful assistant.
User Prompt
Scenario: [scenario] **Question**: What emotions would [subject] ultimately feel in this situation? **Choices**: [choices] Directly choose the top 3 emotions that the individual is most likely to feel within the choices.

Table 9: Prompts we used in distractor generation.

System Prompt
****Instructions**** In this task, you are presented with a scenario, a question, and multiple choices. Please carefully analyze the scenario and take the perspective of the individual involved. ****Note**** Provide only one single correct answer to the question and respond only with the corresponding letter. Do not provide explanations for your response.
User Prompt
Scenario: [scenario] **Question**: What emotion(s) would [subject] ultimately feel in this situation? **Choices**: [choices]
Answer
Answer (Only reply with the corresponding letter numbering):

Table 10: Prompts for vanilla prompting we used in our experiments.

System Prompt
****Instructions**** **1. **Reason****: Read the scenario carefully, paying close attention to the emotions, intentions, and perspectives of the individuals involved. Then, using reason step by step by exploring each option's potential impact on the individual(s) in question. Consider their emotions, previous experiences mentioned in the scenario, and the possible outcomes of each choice. **2. **Conclude**** by selecting the option that best reflects the individual's perspective or emotional response. Your final response should be the letter of the option you predict they would choose, based on your reasoning. ****Note**** The last line of your reply should only contain the letter numbering of your final choice.
User Prompt
Scenario: [scenario] **Question**: What emotion(s) would [subject] ultimately feel in this situation? **Choices**: [choices]
Answer
CoT → Answer: Let's think step by step. **PS** → Answer: Let's first understand the problem and devise a plan to solve the problem. Then, let's carry out the plan to solve the problem step by step. **Self-Ask** → Answer: Break the original question into sub-questions. Explicitly state the follow-up questions, and the answers to the follow-up questions. Aggregate the answers to the follow-up questions and write the answer in the end as "Final Answer: [answer]".

Table 11: Prompts for advanced prompting strategies we used in our experiments. For Self-Ask, we follow the prompt design in Xu et al. (2024).