Emotionally Aligned? Evaluating LLM Predictions in Affective Tasks*

Asutosh Hota[1,*], Jiayi Zhang[1] and Jussi P.P. Jokinen[1]

[1]*University of Jyvaskyla, Faculty of Information Technology, P.O. Box 35 (Agora), FI-40014 University of Jyvaskyla, Finland*

Abstract

Large Language Models (LLMs) are increasingly deployed in tasks that require sensitivity to human emotions, yet the scientific understanding of how these models process affective content remains limited. This gap poses challenges for developing emotionally aligned and trustworthy AI systems. In this paper, we propose a formal framework for evaluating emotional understanding in LLMs by combining controlled vignettes, human-annotated emotion ratings, and regression-based alignment analysis. Our approach allows for systematic testing of both emotion intensity and categorical recognition across diverse emotional contexts. Results show that while LLMs can predict high-level emotional patterns, they struggle with tasks involving subtle, aversive, or ambiguous emotions—areas where humans rely on deeper contextual and cognitive cues. We argue that augmenting LLMs with cognitive modeling, grounded in psychological theories of emotion, offers a promising path toward more nuanced and robust affective capabilities. This work contributes foundational tools for testing, interpreting, and improving emotional understanding in language-based AI systems.

Keywords

Generative Artificial Intelligence, Language Models, Appraisal Theory, Cognitive Modelling, Emotion Understanding

1. Introduction

As large language models (LLMs) and conversational AI systems become increasingly prevalent in everyday human interactions, their ability to detect, interpret, and respond to users' emotions is emerging as a critical aspect of human-AI alignment [1]. Recent industry developments emphasize emotional sensitivity as a key area of model development, with many commercial LLM and AI assistant announcements explicitly highlighting features such as "emotional intelligence," "emotional support," and "empathy" [2, 3, 4, 5]. These efforts reflect a broader trend in AI alignment that pays attention to emotionally conforming interaction and avoidance of insensitivity or manipulation. Accurate emotion recognition, prediction, and contextually appropriate responses are increasingly viewed as essential for developing aligned, user-centered AI systems.

However, current methods for incorporating emotional sensitivity into alignment practices remain limited. Existing approaches primarily rely on reinforcement learning from human feedback (RLHF), fine-tuning conversational tone, and aligning outputs with socially appropriate norms. While these techniques enable models to recognize basic emotional contexts and generate seemingly empathetic responses, they are largely based on statistical associations learned from large-scale text datasets. Empirical studies show that modern LLMs perform well on general affect recognition tasks, such as labeling emotions in vignettes or standardized assessments, occasionally exceeding average human performance [6, 7, 8, 9]. Despite these advances, significant limitations remain: models struggle to interpret context-specific emotional cues or individual differences, especially in emotionally complex or ambiguous situations.

In this paper, we argue that these limitations stem from the absence of a psychologically grounded theory of emotion in current LLMs. By contrast, affective computing (AC) has successfully incorporated established psychological theories, such as basic emotions, core affect, and appraisal theory, allowing

TKTP 2025: Annual Doctoral Symposium of Computer Science, 2.—3.6.2025 Helsinki, Finland

*Corresponding author.

asutosh.jyu.hota@jyu.fi (A. Hota); jiayi.eur.zhang@jyu.fi (J. Zhang); jussi.p.p.jokinen@jyu.fi (J. P.P. Jokinen)
0009-0002-0134-799X (A. Hota); 0000-0003-0235-9861 (J. Zhang); 0000-0002-3024-2209 (J. P.P. Jokinen)

© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-03-12
CEUR-WS.org/Vol-4181/paper05.pdf

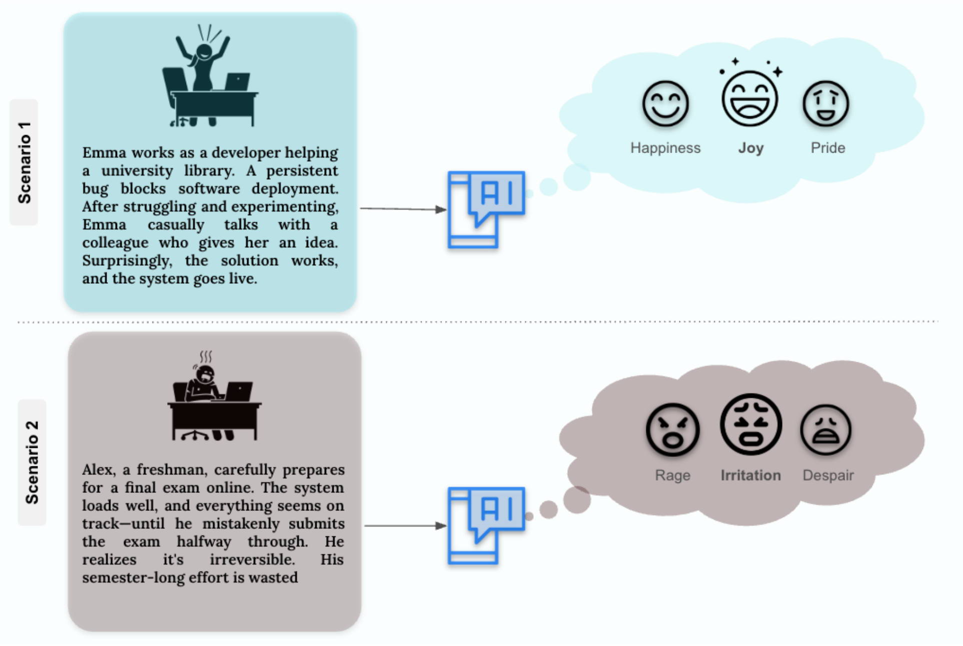
Scenario 1
Emma works as a developer helping a university library. A persistent bug blocks software deployment. After struggling and experimenting, Emma casually talks with a colleague who gives her an idea. Surprisingly, the solution works, and the system goes live.
AI
Happiness
Joy
Pride
Scenario 2
Alex, a freshman, carefully prepares for a final exam online. The system loads well, and everything seems on track—until he mistakenly submits the exam halfway through. He realizes it's irreversible. His semester-long effort is wasted
AI
Rage
Irritation
Despair

Figure 1: Are language models and humans emotionally aligned? Here we illustrate the contrast between LLM predictions on basic versus aversive emotional scenarios. In Scenario 1, the model correctly infers positive emotions such as joy w.r.t. happiness, and pride following a surprising success, aligning closely with human interpretation. In Scenario 2, however, the model is presented with a more cognitively complex situation involving irreversible failure and high personal investment. While humans typically distinguish this as despair or regret, the LLM may conflate it with broader negative states such as rage or irritation. This highlights a core limitation: without structured appraisal reasoning, LLMs often misattribute or flatten nuanced emotional experiences into coarse affective categories, revealing a gap in their cognitive-affective understanding.

machine systems to more reliably interpret and predict human emotional states [10]. These models benefit from explicit theoretical frameworks that support emotional reasoning beyond pattern recognition. It remains an open question whether an LLM, trained solely on statistical associations from text, can internalize or replicate such psychologically informed capabilities. Without integration of explicit emotion theories, LLMs risk maintaining a superficial and stereotyped understanding of emotional phenomena.

To empirically demonstrate the lack of psychologically grounded emotional understanding in LLMs, we present experiments evaluating their ability to predict human emotional reactions to narrative vignettes explicitly designed using appraisal theory. These vignettes were systematically manipulated with subtle contextual changes intended to elicit nuanced emotional responses along appraisal dimensions such as conduciveness, urgency, and power. An example is given in Figure 1, displaying two short stories designed to evoke a particular emotion. Humans are able to distinguish the detailed emotional response, while LLMs resort to a more coarse analysis. The findings reported in this paper highlight a core limitation of current LLMs: without grounding in a psychologically valid theory, they lack sensitivity to the fine-grained appraisal dynamics that underpin human emotional interpretation.

To address the limitations identified in our study, we propose that LLMs can improve their emotional prediction capabilities by integrating an explicit computational cognitive model of human emotional processes, grounded in appraisal theory, into the model's internal representation of users. The effectiveness of such modeling has been demonstrated in systems that successfully simulate human emotional reactions across a variety of scenarios. By employing psychologically grounded cognitive simulations, LLMs could develop a more precise and theoretically coherent "theory of the user," enabling them to predict, explain, and adapt to users' nuanced emotional responses.

2. Background

Emotions have long been understood as arising from cognitive evaluations of events, an idea formalized in cognitive appraisal theory. According to this view, an individual's interpretation of a situation—its personal significance relative to one's goals and beliefs—triggers the emotional response [11]. For example, two people can appraise the same event very differently, leading to opposite emotions: if one sports fan's team loses a championship, he appraises it as an undesirable outcome and feels sadness, whereas another fan appraises the same outcome as positive (his team won) and feels happy [12]. Pioneering work in the 1960s–1990s cemented the idea that emotions are not direct results of events themselves but of our subjective evaluations [13, 14, 15, 16]. This appraisal framework accounts for individual variability in emotional reactions and emphasizes the role of cognition in emotion generation.

Several influential appraisal-based models have been proposed in psychology which provide a theoretical backbone for understanding how specific emotions are elicited by particular patterns of cognitive evaluation. One prominent example is the Ortony-Clore-Collins (OCC) model [17], a structural theory defining emotions in terms of evaluative categories of events, agents, and objects. The OCC model identifies different emotion types (e.g. hope, joy, fear, pride) by describing the cognitive antecedents that elicit each of them. For instance, hope arises from the appraisal of a future desirable event (an expectation of a good outcome). The OCC model provides a taxonomy of 22 emotion types stemming from evaluations such as an event's desirability with respect to one's goals, an agent's praiseworthiness with respect to social standards, or an object's attractiveness with respect to one's attitudes. Another foundational framework is Scherer's Component Process Model (CPM)[18], which conceives of emotion as a dynamic multicomponent process unfolding over time. In Scherer's model, a sequence of appraisal checks evaluates an event across criteria like novelty (suddenness), intrinsic pleasantness, goal relevance, goal conduciveness (does it help or harm one's goals), and coping potential (control, power).

Parallel to these psychological theories, the field of affective computing has focused on modeling emotion in computers both for emotion recognition (systems that perceive and interpret human emotions) and emotion synthesis (systems that simulate or exhibit emotions). Early computational models often drew directly from appraisal theory. For example, the OCC model was adopted in one of the first emotion reasoning systems, Elliott's Affective Reasoner [19], to endow software agents with rule-based emotional reactions. Subsequent frameworks expanded on these ideas, for example ALMA (A Layered Model of Affect) combined longer-term mood and personality with immediate OCC-based appraisals to drive an agent's emotions [20] and EMA model (Emotion and Adaptation), a notable appraisal-based simulation implemented the emotion process as a continuous interpretation of the agent–environment relationship [21]. Not only do such models enable machines to appear emotional, they also serve as tools to test theories of emotion. By implementing appraisal processes in software, researchers can simulate nuanced scenarios and examine if the system's emotion outputs match human data [22]. Building on these foundations there have been recent advancements focusing on cognitive modelling of human emotion. For example, [23] proposed a computational framework that combines cognitive appraisal theory with reinforcement learning to model human emotional responses in task-based scenarios. Their work validated the model using controlled narrative vignettes and human emotion ratings, demonstrating how specific appraisal patterns map onto distinct emotional experiences.

While cognitively grounded models—such as [23], built on appraisal theory—offer interpretability through structured emotional reasoning, it remains unclear whether the same can be said for large language models (LLMs). Unlike appraisal-based systems, LLMs generate affective predictions from opaque, data-driven patterns, lacking explicit representations of core psychological variables like agency, control, and goal relevance. This raises a key question: can general-purpose LLMs capture the depth and nuance of human emotional understanding without structured appraisal mechanisms? LLMs such as GPT-4 are increasingly used in emotionally sensitive domains including mental health, education, and social interaction [24, 25]. These settings require more than surface-level emotion recognition—they demand contextual inference and empathy. Yet, the inner workings of LLMs in affective tasks remain poorly understood, raising concerns around reliability, explainability, and ethical deployment.

Recent research has evaluated the extent to which LLMs understand and reason about emotions through tasks such as emotion classification, empathetic response generation, and theory-of-mind reasoning [26, 27]. Approaches involving fine-tuning on affective dialogue datasets like EmpatheticDialogues [28] and reinforcement learning with human feedback (RLHF) have substantially enhanced emotional appropriateness and empathetic response quality in models like GPT-4 and Claude [24, 25]. Similarly, prompting techniques and chain-of-thought reasoning have improved LLM performance on theory-of-mind tasks, reflecting improved inference of human mental states and intentions [29]. However, despite these advancements, recent benchmarks such as EmoBench indicate that even the most advanced LLMs remain limited in their ability to capture nuanced emotions that depend heavily on contextual interpretation and cognitive appraisal mechanisms [27, 30].

This highlights a critical research gap: existing evaluations seldom explicitly test or incorporate cognitive appraisal frameworks into LLM emotional understanding. Our study directly addresses this gap by systematically evaluating LLMs against human-labeled emotional responses derived from cognitively grounded scenarios. By identifying specific appraisal dimensions where current LLMs show weaknesses, our results offer a clear pathway toward augmenting weaker models through explicit integration of cognitive appraisal structures. Such cognitive augmentation could provide weaker models with interpretable, structured emotional reasoning capabilities, helping them to better contextualize user emotions, improve affective responsiveness, and enhance overall user-model interaction quality in emotionally sensitive contexts.

3. Method

3.1. Human Participant Data (Previously Collected)

The human participant data used for comparison in this study was previously collected and analyzed in the original study by Zhang et al. [23]. That earlier work involved 106 human participants recruited via an online platform, ensuring diverse demographics with an average age of 37.2 years (SD = 11.6), including 89 women and 17 men, all native English speakers. Participants provided informed consent and were compensated for their involvement. For detailed procedures and analyses related to the human data, refer [23] or download the data here.

3.2. Materials

For consistency and direct comparability, we used the same 11 narrative vignettes developed in the original study [23]. These narratives were carefully designed based on established cognitive appraisal patterns from psychological literature to reliably elicit specific emotional responses. Seven of these vignettes targeted prototypical emotional states (Happiness, Joy, Pride, Boredom, Fear, Sadness, Shame), and the remaining four explored cognitively complex, socially nuanced negative emotions (Anxiety, Despair, Irritation, Rage). Each vignette was approximately 100-200 words, depicting realistic scenarios commonly encountered during human-computer interaction. The materials of this paper can be downloaded here.

3.3. Large Language Models (LLMs) Evaluation

Extending beyond human evaluations, this current study explicitly evaluates multiple state-of-the-art Large Language Models (LLMs) in affective understanding tasks. The LLMs evaluated include DeepSeek-R1, Gemma 3, LLaMA 3.2, Phi-4, GPT-4.5, and GPT-4o. The selection of models represents diverse architectures and varying scales of parameterization, reflecting cutting-edge capabilities across both commercial and open-source research-focused language models.

To ensure consistency with the human evaluation setup, all LLMs were prompted using identical narrative stimuli and experimental instructions as those presented to human participants. The prompts were carefully structured to direct the models either to rate emotion intensities on a continuous scale or to select discrete emotional labels, thus precisely replicating the task formats used in the original human experiments. Model generation parameters were held constant across evaluations—for example, the temperature setting was fixed at 0.5 for all runs. To enhance robustness and mitigate randomness in generation, each model was prompted ten times per story, and results were aggregated across these runs. This multi-run approach enabled the computation of average predictions and confidence intervals, providing a more reliable measure of each model's affective judgments.

Due to the significant computational cost and API constraints associated with closed-source commercial models, GPT-4.5 and GPT-4o were evaluated only in Experiment 3, which focuses on aversive emotions— an area where open-source models previously showed limited performance. Given the popularity and capabilities of proprietary models like GPT-4.5 and GPT-4o, our goal was to assess whether their advances in general language understanding translate into greater sensitivity to nuanced affective signals, particularly in complex emotional contexts.

3.4. Experimental Design

We conducted three distinct computational experiments with the LLMs, precisely mirroring the original experimental conditions from Zhang et al.'s (2024) human participant studies. Each experiment systematically assesses the capability of LLMs in approximating the depth and specificity of human emotional understanding demonstrated in prior human evaluations.

1. Experiment 1 (Emotion Intensity Prediction - Basic Emotions): Models rated the intensity of seven basic emotions (Happiness, Joy, Pride, Boredom, Fear, Sadness, Shame) elicited by the vignettes on a continuous scale from 0 (not at all) to 1 (extremely strong).
2. Experiment 2 (Discrete Emotion Classification - Basic Emotions): Models selected the most appropriate single emotion label from the same set of seven basic emotions, emulating forced-choice classification tasks previously performed by humans.
3. Experiment 3a (Emotion Intensity Prediction - Aversive Emotions): Models rated the intensity of four cognitively complex, nuanced negative emotions (Anxiety, Despair, Irritation, Rage) on the same continuous scale, reflecting tasks demanding subtle emotional differentiation.
4. Experiment 3b (Discrete Emotion Classification - Aversive Emotions): Models selected discrete labels from the set of aversive emotions, mirroring the forced-choice categorization approach.

3.5. Evaluation Metrics

Consistent with the original evaluation methods employed by Zhang et al. (2024), the coefficient of determination (R^2) and the root mean squared error (RMSE) were used to quantitatively compare LLM predictions against human emotion ratings. The R^2 metric evaluates the degree of linear agreement between model predictions and human emotion intensity ratings, while RMSE measures the average magnitude of error between the predicted and actual human responses. For the discrete-choice experiments, model outputs were aggregated to compute the selection proportions for each emotion label. These proportions were then compared to the corresponding human response proportions to assess the alignment between human and model categorizations.

4. Results and Discussion

This section presents the alignment between LLM predictions and previously collected human emotion ratings [23] across 3 experiments.

4.1. Experiment 1: Emotion Intensity Prediction for Basic Emotions

In Experiment 1, we evaluated how well LLMs predict human emotion intensity ratings across seven prototypical emotions. The results, visualized in Figure 2, show that models such as DeepSeek and Phi-4 consistently tracked human ratings across most emotion categories. DeepSeek, in particular,

Model	Exp	R^2	RMSE
DeepSeek-R1	1	0.68	0.08
	2	0.70	0.24
	3a	0.20	0.03
	3b	0.01	0.26
Gemma 3	1	0.67	0.09
	2	0.00	0.41
	3a	0.18	0.03
	3b	0.00	0.22
LLaMA 3.2	1	0.60	0.09
	2	0.28	0.36
	3a	0.37	0.03
	3b	0.19	0.22
Phi-4	1	0.66	0.08
	2	0.43	0.31
	3a	0.32	0.03
	3b	0.55	0.18
GPT-4.5	3a	0.53	0.02
	3b	0.07	0.23
GPT-4o	3a	0.61	0.02
	3b	0.00	0.26

Table 1
Table summarizing the regression results comparing LLM predicted emotion scores to human annotated emotional responses across three experiments. These experiments range from relatively straightforward tasks (Experiment 1 & 2) to more subtle or aversive emotional contexts (Experiments 3a–3b). These findings validate that LLMs can approximate high level affective understanding, but they remain limited in contexts requiring nuanced or aversive emotional reasoning. While newer models like GPT-4o or Deepseek-R1 show clear gains in emotional intensity alignment, categorical emotion judgments particularly for negative affect remain a significant challenge

exhibited strong alignment for high intensity emotions such as happiness, pride, and fear. However, notable divergence appeared for more ambivalent states like boredom and shame, where models either overestimated or underestimated intensity compared to humans. Similarly, Gemma 3 performed well in this task with strong correlation to human data while LLaMA 3.2 tended to exaggerate ratings, especially for shame and sadness. The relatively high congruence in high salience categories indicates that LLMs capture surface level affective patterns reasonably well, though finer calibration is needed for subtler emotions Having assessed the models' ability to capture graded emotional salience through continuous intensity ratings, we next examined whether they could identify the single most dominant emotional category in a forced-choice setting which is an essential capability for applications that require discrete emotional labeling, such as emotion detection systems or empathetic response generation.

4.2. Experiment 2: Discrete Emotion Classification – Basic Emotions

In Experiment 2, models were tasked with selecting a single dominant emotion from a set of seven basic emotions per vignette. While all models showed moderate convergence with human majority choices as shown in figure 3, their ability to consistently match human judgments varied substantially. DeepSeek and Phi-4 performed relatively well in capturing dominant emotional interpretations in high salience scenarios such as joy and fear, with outputs often closely matching human label distributions. However, models like Gemma 3 and LLaMA 3.2 occasionally misclassified the dominant emotion, particularly for ambiguous stimuli such as those targeting shame or boredom. The overall trend suggests that LLMs can approximate surface level emotion selection reasonably well when emotional cues are strong, but they often struggle with single label disambiguation in emotionally layered narratives. While basic emotions provide a useful benchmark, they often lack the nuance of real world affective contexts; thus, we expanded our evaluation to more cognitively complex and socially grounded aversive emotions to test whether LLMs could handle subtler affective distinctions that depend on deeper appraisal reasoning.

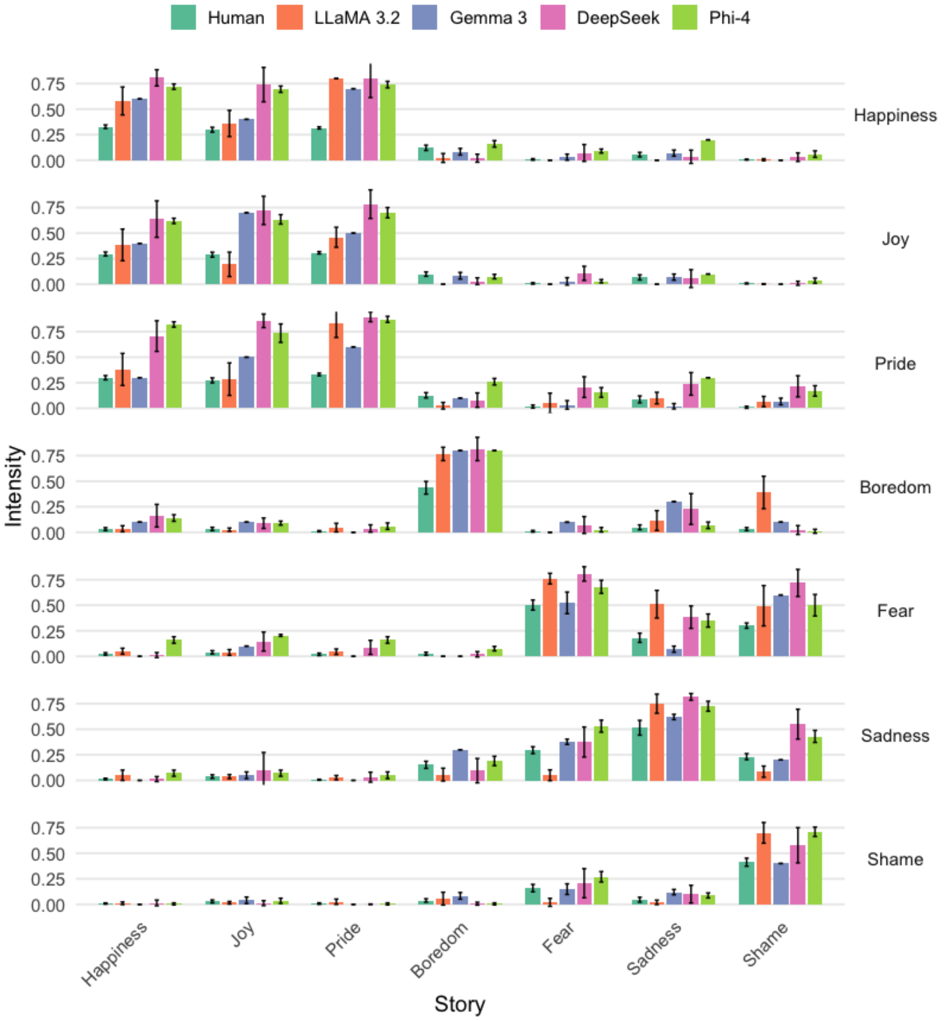

Figure 2: Emotion intensity ratings for seven basic emotions across vignettes in Experiment 1. Human ratings are shown alongside predictions from four LLMs. Error bars represent standard errors of the mean.

4.3. Experiment 3a: Emotion Intensity Prediction – Aversive Emotions

Experiment 3a required models to rate the intensity of four cognitively complex negative emotions: anxiety, despair, irritation, and rage. Figure 4 illustrates a consistent drop in performance relative to basic emotion prediction (Experiment 1). We included GPT-family models in experiment 3a and 3b specifically to compare their performance against the open-source models. Given the popularity and reported capabilities of proprietary models like GPT-4.5 and GPT-4o, our goal was to assess whether their improvements in general language understanding extend to nuanced affective reasoning. By targeting aversive emotional contexts where open-source models previously struggled, we aimed to probe the limits of even state of the art commercial LLMs. Human ratings showed moderate variation across stories, but model predictions tended to be more uniform, with several models (especially GPT-4.5 and GPT-4o) overestimating intensity across all stories. While GPT-4o achieved relatively high alignment with human ratings in terms of correlation (R^2 = 0.61), the results suggest that even high performing models rely on overly generalized mappings of aversive emotional cues. DeepSeek and Phi-4 also performed reasonably but exhibited limited nuance, frequently assigning mid to high intensity scores across all stimuli regardless of specific narrative content. These results indicate that while LLMs can reflect broad affective tone, their differentiation across nuanced negative emotions remains shallow.

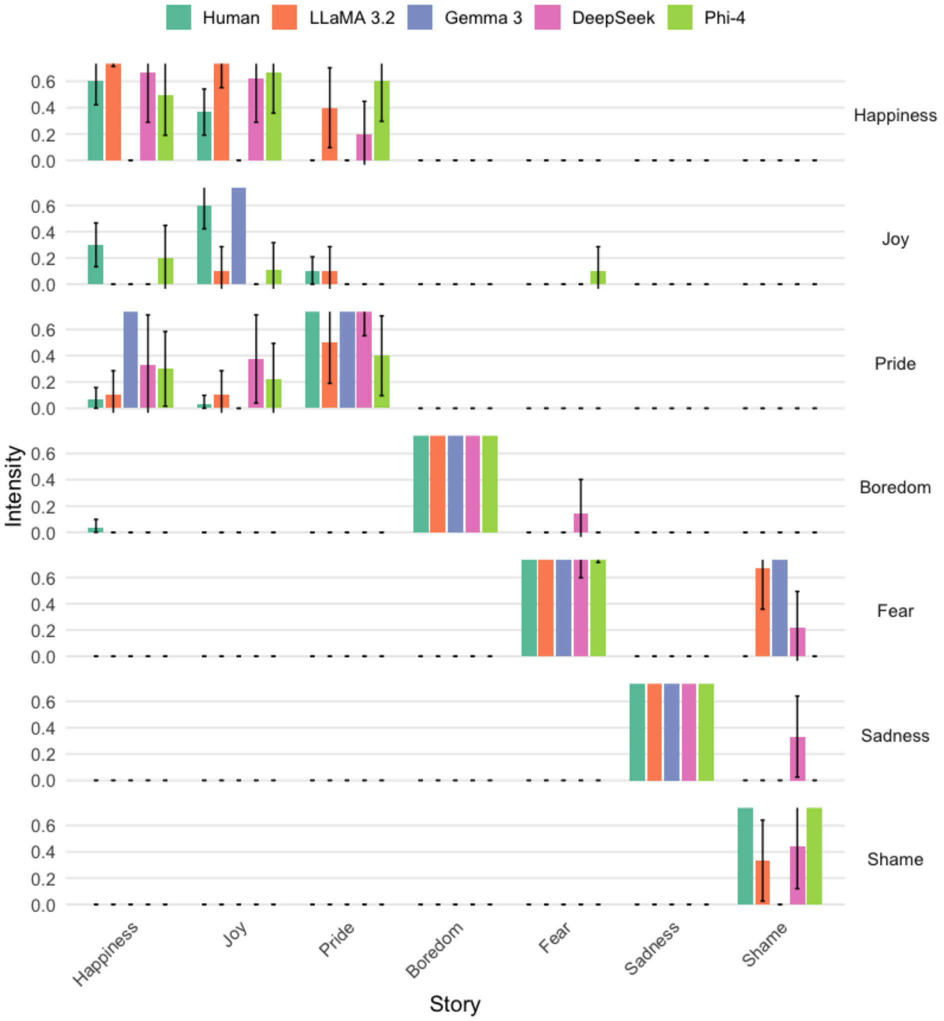

Figure 3: Experiment 2: Discrete emotion label predictions for each story targeting basic emotions. Bars show the proportion of label selection for humans and LLMs across trials.

4.4. Experiment 3b: Discrete Emotion Classification – Aversive Emotions

Experiment 3b further tested the granularity of LLM emotional understanding by asking for single label classification among the four aversive emotions. As shown in Figure 5, performance across models was inconsistent and often misaligned with human majority responses. GPT-based models (particularly GPT-4.5 and GPT-4o) were most frequently aligned with human selections on stories targeting anxiety and despair, but misclassifications were common in more ambiguous narratives involving irritation and rage. Notably, models like LLaMA 3.2 and DeepSeek showed wide variance in label distribution, often spreading selections across multiple categories with low confidence. Human responses were also more concentrated, suggesting a shared emotional interpretation that LLMs failed to converge on. These findings underscore that even with high text understanding capabilities, LLMs lack the cognitive and emotional granularity to reliably disambiguate similarly valenced emotions with different appraisal underpinnings. Having tested models' ability to rate the intensity of complex emotions (in 3a), this experiment further challenged LLMs to perform discrete classification among these nuanced states, mirroring real world scenarios where systems must make categorical emotional judgments despite overlapping valence and ambiguous context.

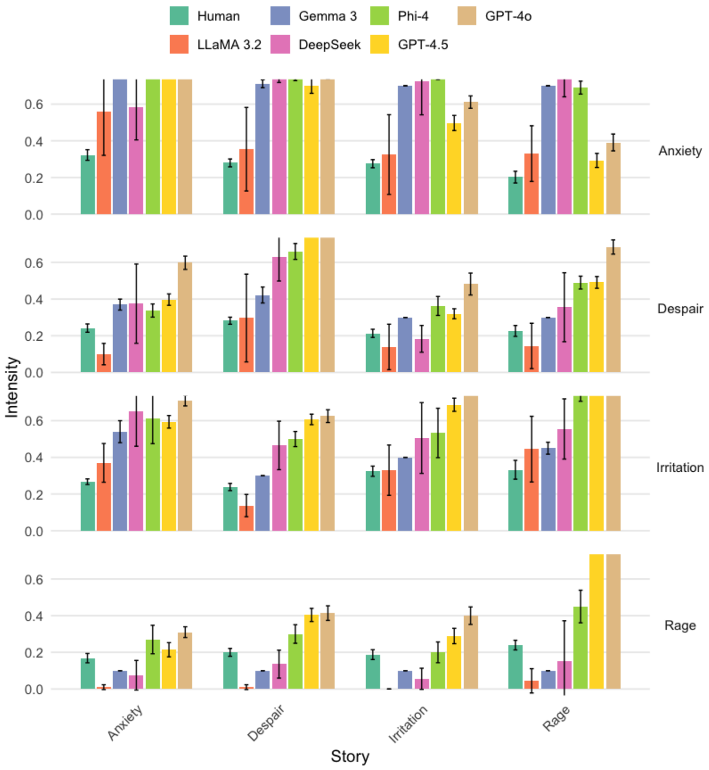

Figure 4: Experiment 3a: Emotion intensity predictions for aversive emotions. Human and LLM ratings are shown for four emotions across different stories.

4.5. Discussions

The results indicate that current LLMs predict basic emotional understanding reasonably well but consistently falter when tasks involve nuanced, cognitively complex emotions requiring detailed appraisal reasoning. While models demonstrate surface level alignment with human affective responses particularly in scenarios involving clear and high salience emotions, their performance deteriorates significantly when tasked with distinguishing between subtly different emotions that rely on context, social norms, or internal psychological states. Furthermore, LLMs over-estimate their emotional understanding in most cases. Despite their linguistic fluency and statistical mimicry of human language, LLMs operate without an understanding of the cognitive mechanisms behind human emotions.

Explicit understanding of emotion is critical for AI systems that aim to interact meaningfully, ethically, and empathetically with humans. Emotions are not merely labels or affective tones; they are deeply intertwined with cognition, intentions, beliefs, and social interpretation. For instance, differentiating between fear and anxiety requires recognizing not just the presence of threat but the degree of uncertainty, control, and expected timing, all of which are grounded in cognitive appraisals. Without internal models of these appraisal dynamics, LLMs risk generating emotionally inappropriate, tone-deaf, or even harmful outputs in sensitive domains such as mental health, education, or customer support.

To address this gap, cognitive models of emotion grounded in psychological theory such as the CPM or the OCC model offer a principled way forward. These models formalize how specific emotions arise from evaluations of events in terms of novelty, goal relevance, agency, norm violation, and coping potential. By incorporating such structured appraisal representations, LLMs could gain interpretable internal states that support contextual emotional reasoning. Rather than relying solely on statistical associations between words and affect labels, a cognitively augmented model could simulate appraisal processes to infer why an emotion is appropriate, what social or motivational context underlies it, and how it might evolve over time.

Integrating these models into LLMs could also improve generalization and robustness. For example, if

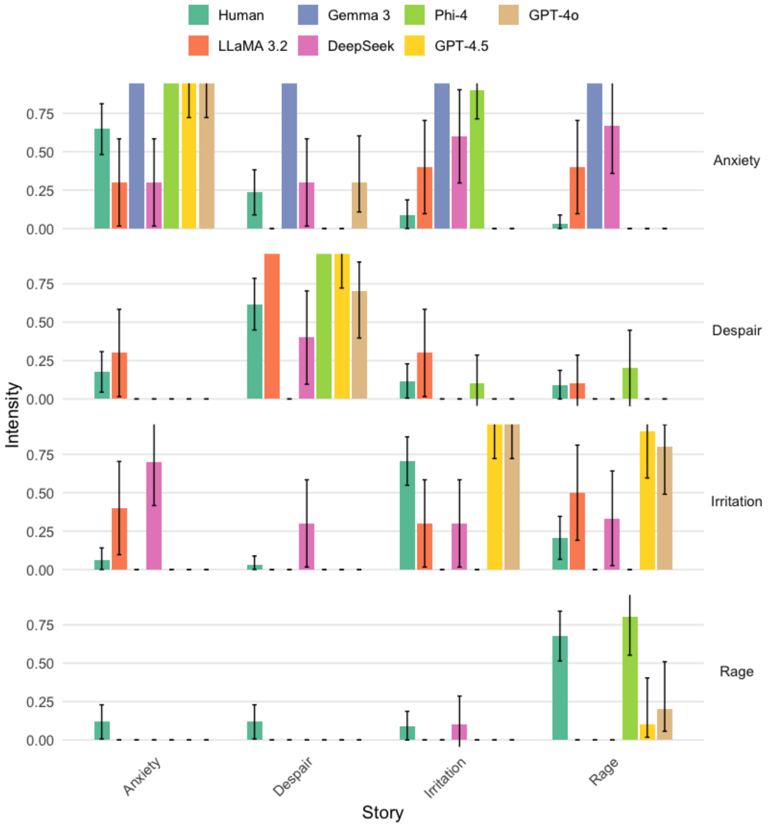

Figure 5: Experiment 3b: Proportions of discrete emotion label predictions for aversive emotional stories. Model and human responses shown as bar plots across all vignettes.

a model understands that despair arises when a negative outcome is both uncontrollable and irrevocable, it could more accurately generate or classify emotional responses even in novel situations. Moreover, embedding appraisal reasoning into LLM architectures may support counterfactual and dynamic emotion prediction—understanding how emotions might change under different circumstances or how people regulate emotions in social interactions. Bridging the gap between statistical language understanding and structured emotional cognition is essential for building emotionally aligned AI. We highlight the limitations of current LLMs and motivate future research on hybrid architectures that combine the flexibility of language models with the interpretability and theoretical grounding of cognitive emotion frameworks.

5. Future Work

Integrating Appraisal Models into Language Models Future research should develop hybrid architectures that integrate cognitive appraisal theories (CCPM or OCC) into LLMs. Embedding appraisal dimensions (novelty, goal relevance, agency, norm compatibility) into intermediate representations would let models simulate evaluative processes rather than merely correlate emotion labels with linguistic patterns. Achieving this may require architectural changes to support structured, interpretable appraisal representations or integration with cognitive models to enhance text generation.

Learning Structured Emotional Representations A major technical challenge involves operationalizing appraisal variables in formats compatible with neural network learning. Future work should explore multitask and supervised learning strategies using annotated corpora that capture appraisal based emotion dimensions. Alternatively, incorporating neuro-symbolic approaches may support the integration of appraisal rules with distributed representations, yielding systems capable of both flexible generalization and interpretable emotional reasoning. Embedding such structures into LLMs may also support compositional emotion understanding and enable emotion reasoning in novel contexts.

Modeling Dynamic and Temporal Emotional Processes Human emotions are temporally dynamic and context sensitive, evolving over the course of events, social interactions, and internal cognitive regulation. To capture this, future models should incorporate mechanisms for emotion updating, regulation, and counterfactual reasoning. For instance, models could track shifts in goal relevance or perceived control to simulate how emotions like hope, frustration, or despair emerge and evolve. Incorporating temporal emotion dynamics may also enhance LLMs' performance in dialog systems, story understanding, and affective forecasting.

Developing Richer Evaluation Benchmarks Current benchmarks for emotion understanding primarily rely on classification tasks with discrete emotion labels, which fail to capture the subtlety and context dependence of emotional appraisals. Future benchmarks should include datasets annotated with multidimensional appraisal features, causal emotion chains, and social context cues. Narrative and conversational datasets, in particular, can provide testbeds for evaluating whether models understand why an emotion is appropriate and how it might shift under changing conditions.

Interdisciplinary Collaboration and Ethical Implications Progress in emotionally aware AI will benefit from interdisciplinary collaboration across affective computing, cognitive science, social psychology, and human computer interaction. Integrating insights from these fields can help guide model development, ensure psychological validity, and establish ethical boundaries for emotionally responsive AI. Furthermore, as emotional understanding of systems gets better and are deployed in sensitive domains such as education, therapy, or customer service, it is critical to evaluate not only their accuracy but also their potential to build trust, avoid harm, and support user well-being.

6. Conclusion

The results demonstrate a clear limitation in current LLMs' emotional reasoning capabilities, especially concerning subtle or socially nuanced emotions. While basic emotional states are predicted adequately, nuanced differentiation critical in authentic human interaction is lacking. Our findings suggest the necessity of integrating more explicit cognitive appraisal mechanisms, as demonstrated by previous computational models grounded in psychological theory. One limitation of the present study is its non interactive, vignette based design; future work will focus on testing the utility of augmenting cognitive models of user emotion with LLMs within interactive tasks, where context, feedback, and user specific adaptation are critical. Integrating psychological theories, such as cognitive appraisal, into LLM architectures could significantly enhance their ability to handle emotional subtleties, moving towards genuinely emotionally aligned AI systems.

7. Declaration on Generative AI

The author(s) acknowledge the use of GenAI tools (specifically, OpenAI's ChatGPT 4.1) in the preparation of this manuscript. These tools were employed solely for formatting assistance, language polishing, and other editorial tasks (e.g., improving clarity, correcting grammar, and ensuring consistent style). All substantive ideas, analyses, conceptual contributions, and interpretations presented in this paper are the original work of the authors, who bear full responsibility for its content. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] S. Russell, Human compatible: AI and the problem of control, Penguin Uk, 2019.

[2] Anthropic, Hume ai creates emotionally intelligent voice interactions with claude, https://www.anthropic.com/customers/hume, 2025. Accessed: 2025-04-01.

[3] Google, Emotion analyzer, https://ai.google.dev/competition/projects/emotion-analyzer, 2025. Accessed: 2025-04-01.

[4] OpenAI, Introducing gpt-4.5, https://openai.com/index/introducing-gpt-4-5/, 2025. Accessed: 2025-04-01.

[5] B. Nguyen, A new ai chatbot called pi is designed to serve as your personal assistant, https://www.businessinsider.com/ai-chatbot-pi-offers-personal-assistant-advice-how-it-works-2023-5, 2023. Accessed: 2025-04-01.

[6] Z. Elyoseph, D. Hadar-Shoval, K. Asraf, M. Lvovsky, Chatgpt outperforms humans in emotional awareness evaluations, Frontiers in psychology 14 (2023) 1199058.

[7] J.-t. Huang, M. H. Lam, E. J. Li, S. Ren, W. Wang, W. Jiao, Z. Tu, M. R. Lyu, Apathetic or empathetic? evaluating llms' emotional alignments with humans, Advances in Neural Information Processing Systems 37 (2024) 97053–97087.

[8] A. N. Tak, J. Gratch, Gpt-4 emulates average-human emotional cognition from a third-person perspective, arXiv preprint arXiv:2408.13718 (2024).

[9] G. D. Vzorinab, A. M. Bukinichac, A. V. Sedykha, I. I. Vetrovab, E. A. Sergienkob, The emotional intelligence of the gpt-4 large language model, Psychology in Russia: State of the art 17 (2024) 85–99.

[10] K. R. Scherer, Towards a prediction and data driven computational process model of emotion, IEEE Transactions on Affective Computing 12 (2019) 279–292.

[11] K. R. Scherer, A. Schorr, T. Johnstone, Appraisal processes in emotion: Theory, methods, research, Oxford University Press, 2001.

[12] P. A. Jaques, R. M. Viccari, Infering emotions and applying affective tactics for a better learning, in: Agent-based tutoring systems by cognitive and affective modeling, IGI Global, 2008, pp. 135–155.

[13] C. A. Smith, R. S. Lazarus, et al., Emotion and adaptation, Handbook of personality: Theory and research 21 (1990) 609–637.

[14] S. Folkman, Stress: appraisal and coping, in: Encyclopedia of behavioral medicine, Springer, 2020, pp. 2177–2179.

[15] N. H. Frijda, The laws of emotion, Psychology Press, 2017.

[16] K. R. Scherer, Emotion as a process: Function, origin and regulation, 1982.

[17] A. Ortony, G. L. Clore, A. Collins, The cognitive structure of emotions, Cambridge university press, 2022.

[18] K. R. Scherer, The dynamic architecture of emotion: Evidence for the component process model, Cognition and emotion 23 (2009) 1307–1351.

[19] C. D. Elliott, The affective reasoner: a process model of emotions in a multiagent system, Northwestern University, 1992.

[20] P. Gebhard, Alma: a layered model of affect, in: Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems, 2005, pp. 29–36.

[21] S. C. Marsella, J. Gratch, Ema: A process model of appraisal dynamics, Cognitive Systems Research 10 (2009) 70–90.

[22] S. Marsella, J. Gratch, Computationally modeling human emotion, Communications of the ACM 57 (2014) 56–67.

[23] J. E. Zhang, J. Broekens, J. Jokinen, Modeling cognitive-affective processes with appraisal and reinforcement learning, IEEE Transactions on Affective Computing (2024).

[24] Z. Guo, A. Lai, J. H. Thygesen, J. Farrington, T. Keen, K. Li, et al., Large language models for mental health applications: Systematic review, JMIR mental health 11 (2024) e57400.

[25] J. Yang, H. Jin, R. Tang, X. Han, Q. Feng, H. Jiang, S. Zhong, B. Yin, X. Hu, Harnessing the power of llms in practice: A survey on chatgpt and beyond, ACM Transactions on Knowledge Discovery

from Data 18 (2024) 1–32.

[26] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, S. Ravi, Goemotions: A dataset of fine-grained emotions, arXiv preprint arXiv:2005.00547 (2020).

[27] S. Sabour, S. Liu, Z. Zhang, J. M. Liu, J. Zhou, A. S. Sunaryo, J. Li, T. Lee, R. Mihalcea, M. Huang, Emobench: Evaluating the emotional intelligence of large language models, arXiv preprint arXiv:2402.12071 (2024).

[28] H. Rashkin, E. M. Smith, M. Li, Y.-L. Boureau, Towards empathetic open-domain conversation models: A new benchmark and dataset, arXiv preprint arXiv:1811.00207 (2018).

[29] M. Kosinski, Theory of mind may have spontaneously emerged in large language models, arXiv preprint arXiv:2302.02083 4 (2023) 169.

[30] Y. Chen, H. Wang, S. Yan, S. Liu, Y. Li, Y. Zhao, Y. Xiao, Emotionqueen: A benchmark for evaluating empathy of large language models, arXiv preprint arXiv:2409.13359 (2024).