EMOCAP: Deep Dive in Systematic Assessment of Large Language Models in Emotional Intelligence through Multi-Turn Conversations

Anonymous ACL submission

Abstract

Large Language Models (LLMs) often lack robust emotional intelligence, limiting their effectiveness in sensitive domains such as mental health and crisis response. Existing open-source LLMs struggle to track nuanced emotions over multi-turn dialogues, resulting in shallow or misaligned responses. Proprietary models show promise but remain closed-source, hindering transparent evaluation and improvement. To address these limitations, we propose EMOCAP, a comprehensive Emotional Intelligence framework that integrates well-established psychological frameworks (e.g., Ekman, Plutchik, Russell, Goleman, Affective Domain in Blooms Taxonomy etc.) for enhanced emotion detection, contextual adaptation, and ethical alignment. We develop a multi-turn, domain-general dataset and evaluation protocol to test how LLMs manage evolving emotions, mixed affective states, and subtle cues. Our experiments compare baseline open-source LLMs (Gemma-2-9b, Qwen2.5-7b and Llama-3-8B) against its instruction fine tuned versions (Gemma-2-9b-It, Qwen2.5-7b-It and Llama-3-8B-It) .Models incorporating the recognition and response guidelines well demonstrate better emotional tracking, reduced repetitive responses, and more ethically aligned output compared to standard baselines, although complex scenarios (e.g., sarcasm) remain challenging. By providing an open-source taxonomy and benchmark for emotional intelligence, this work lays the groundwork for empathetic, context-aware, and ethically responsible LLMs across various real-world applications.

1 Introduction

Large Language Models (LLMs) have demonstrated significant progress in natural language processing (NLP), enabling them to generate human-like responses across diverse applications, including customer service, education, healthcare, and mental health support (Waswani et al., 2017; Brown et al., 2020; Fitzpatrick et al., 2017). However, despite these advances, LLMs remain fundamentally deficient in Emotional Intelligence (EI)—the ability to perceive, interpret, and appropriately respond to human emotions in dynamic, multi-turn conversations (Rashkin, 2018).

This limitation is particularly critical in high-stakes applications such as mental health counseling, crisis response systems, and empathetic AI-driven assistants, where a failure to recognize and adapt to user emotions can result in disengagement, ineffective support, or even ethically concerning responses. While proprietary models have exhibited improved capabilities in emotion-aware dialogue generation, their closed-source nature prevents systematic benchmarking, interpretability, and targeted improvements. Meanwhile, open-source LLMs still lack the ability to track evolving emotions, recognize complex affective states, and generate socially and ethically aligned responses, limiting their reliability in real-world, human-centered AI interactions (Touvron et al., 2023).

Despite progress in affective computing, sentiment analysis, and multimodal emotion recognition (Picard, 1997; Pang and Lee, 2008; Poria et al., 2019a), existing LLMs exhibit several critical gaps in emotional intelligence, which we systematically outline below.

(i) Fundamental Limitations in EI Modeling: Lack of Emotion Tracking Across Conversations – Current LLMs fail to retain emotional context across conversational turns, leading to inconsistent, repetitive, or contextually misaligned responses (Chen et al., 2023)

(ii) Inability to Recognize Mixed or Conflicting Emotions – Most models struggle with overlapping emotional states (e.g., relief mixed with anxiety) and subtle affective transitions, which are crucial for nuanced human interaction(Alhuzali and Ananiadou, 2021).

(iii) Lack of Cultural and Social Sensitivity –

Many emotion models assume universality in emotional expressions, disregarding cultural and contextual influences on emotional perception and interpretation (Dudy et al., 2024).

(iv) Lack of Emotional Causality Awareness – LLMs struggle to infer the underlying causes of emotional expressions, resulting in shallow or contextually inappropriate responses (Cheng et al., 2022).

(v) Failure in Emotion-Driven Decision-Making – Models are unable to incorporate emotional awareness into decision-making, limiting their applicability in scenarios requiring context-sensitive emotional adaptation (Mozikov et al., 2024).

(vi)Inability to Handle Emotionally Complex Situations – LLMs frequently fail in multi-agent emotional interactions, particularly in group conversations where conflicting emotions must be recognized and addressed appropriately (Poria et al., 2019b).

(vii)Preference Bias in Emotional Support – LLMs tend to overuse specific response strategies (e.g., excessive reassurance over problem-solving) rather than dynamically adapting to the user's emotional needs (Akhtar et al., 2024).

(viii)Ethical Emotional Response Generation – Without robust ethical safeguards, LLMs risk producing misleading, insensitive, or even manipulative responses in emotionally charged interactions (Mohammad, 2022).

(ix) Limited General-Purpose Multi-Turn Emotional Intelligence Benchmarks – While some benchmarks evaluate multi-turn emotional tracking in specific domains (e.g., counseling-based ESC datasets), there is no standardized, multi-domain benchmark to assess how LLMs track, adapt, and respond to evolving emotional states across diverse applications (Sabour et al., 2024).

(x) Limited Integration of Theory of Mind (ToM) into Emotionally Intelligent Response Generation – While existing ToM benchmarks evaluate mental state inference, they do not assess how LLMs apply ToM reasoning dynamically in real-time, emotionally intelligent responses (Xiao et al., 2024).

1.1 EMOCAP

To address these challenges, we propose a structured framework or taxonomy designed to enhance open-source LLMs' ability to detect, process, and generate emotionally intelligent responses dynamically. This framework does not replace existing emotional intelligence models but integrates insights from established psychological theories, creating a multi-layered system that:

- Enhances emotion recognition by leveraging Ekman and Plutchik's emotion models for foundational emotional detection (Ekman and Friesen, 1981; Plutchik, 1982).

- Captures emotional intensity and mixed affective states using Russell's valence-arousal model to model nuanced affective transitions (Russell, 1980a).

- Infers emotional causality and prioritizes responses using cognitive appraisal theories to generate more contextually appropriate responses (Lazarus, 1991a; Ortony et al., 1988).

- Adapts over time to establish user-specific emotional alignment by incorporating Goleman's Emotional Intelligence framework (Goleman, 1995).

- Integrates Theory of Mind (ToM) for socially and contextually appropriate interactions, ensuring that LLMs dynamically adjust their responses based on inferred social reasoning(Premack and Woodruff, 1978a).

- Incorporates Bloom's Affective Domain Taxonomy to systematically address emotional learning and response adaptation (Krathwohl et al., 1964)

Unlike existing approaches, our method systematically evaluates emotional intelligence in a multi-turn, domain-general setting, ensuring broad applicability across real-world AI-human interaction scenarios. By maintaining a domain-general design, our method ensures flexibility and robustness: it can integrate multiple types of emotional cues, adapt to varying conversation flows, and handle different user goals or topics. In practical terms, this means that even though each row of our taxonomy could correspond to a different domain or type of interaction, the underlying principles of emotional recognition, adaptation, and response generation remain the same. Consequently, the system can seamlessly handle multi-turn conversations in whichever domain the user or application requires, preserving the core emotional intelligence objectives across contexts.

1.2 Contributions of This Work

This paper presents the following key contributions:

- EMOCAP Dataset: We introduce EMOCAP, a domain-general, multi-turn dataset specifically designed to challenge and evaluate how LLMs detect, interpret, and respond to evolving emotional states. This dataset fills a gap in open-source resources by encompassing a wide range of emotional scenarios—enabling both granular emotion recognition and holistic conversation-level analysis.

- Integration of Psychological Theories: Building on established frameworks such as Ekman's and Plutchik's emotion models, Russell's valence-arousal structure, and Goleman's EI concepts, we provide a structured approach that captures fundamental emotions, nuanced intensity levels, and higher-order empathy skills in large language models.

- Systematic Evaluation of LLMs: We conduct a comprehensive analysis of leading open-source LLMs, assessing their ability to maintain emotional context, handle mixed or overlapping affective states, and produce ethically responsible responses over multiple dialogue turns.

1.3 Why This Matters

As LLMs continue to permeate human-centered applications, their lack of robust emotional intelligence presents significant challenges, including user disengagement, ethical concerns, and limited applicability in emotionally nuanced interactions. While proprietary models have made notable progress in affect-aware response generation, their closed-source nature hinders systematic improvements and independent benchmarking. Our taxonomy provides a structured, open-source approach to equipping LLMs with emotionally intelligent capabilities, ensuring that future AI systems can engage in more empathetic, contextually aware, and ethically responsible interactions.

2 Related Work

2.1 Overview of Emotion-Aware AI Affective computing has long aimed to create systems that detect and respond to human emotions. Early work in this area led to the development of affective agents that simulate appropriate emotional responses based on psychological theories and appraisal models (Ekman and Friesen, 1981; Plutchik, 1982). Although such systems originally focused on single modalities and limited contexts, they laid the groundwork for more comprehensive approaches to emotion-aware AI.

2.2 LLMs and Emotional Intelligence

Recent advances in large language models (LLMs) have opened new avenues for integrating emotional intelligence (EI) into AI systems. EI—commonly defined as the ability to recognize, understand, and manage emotions in oneself and others—is rooted in well-established theories such as Ekman's basic emotions, Plutchik's wheel of emotions, and Russell's circumplex model (Ekman and Friesen, 1981; Plutchik, 1982; Russell, 1980a). In parallel, cognitive appraisal theories, including the OCC model and Lazarus's framework, have been adopted to understand how events trigger emotional responses (Ortony et al., 1988; Lazarus, 1991b). Moreover, Goleman's seminal work on emotional intelligence has provided a practical framework for designing emotionally responsive systems (Goleman, 1995).

LLMs, by virtue of their large-scale pretraining on diverse text corpora, have shown promise in capturing latent affective knowledge. Empirical studies have demonstrated that models can generate empathetic and contextually appropriate responses in dialogue systems (Chen et al., 2023). However, evaluations using benchmarks such as EmoBench reveal that even state-of-the-art models like GPT-4 lag behind human performance in nuanced emotional reasoning, particularly when it comes to understanding mixed or evolving emotional states (Sabour et al., 2024). Additionally, prompting techniques—such as chain-of-thought prompting (Wei et al., 2022) and its emotional variant, the Emotional Chain-of-Thought (ECoT) framework—have been proposed to improve the depth of emotional reasoning in LLMs (Li et al., 2023).

2.3 Gaps, Debates, and Ethical Considerations

Despite these advances, several critical gaps remain. First, LLMs often exhibit ethical concerns and biases in emotion recognition, frequently reflecting cultural or gender stereotypes present in their training data. Second, while models perform well on single-turn tasks, they struggle with complex multi-turn conversations where emotional context evolves over time. This limitation leads to inconsistent or repetitive responses that can undermine the reliability of emotional support systems.

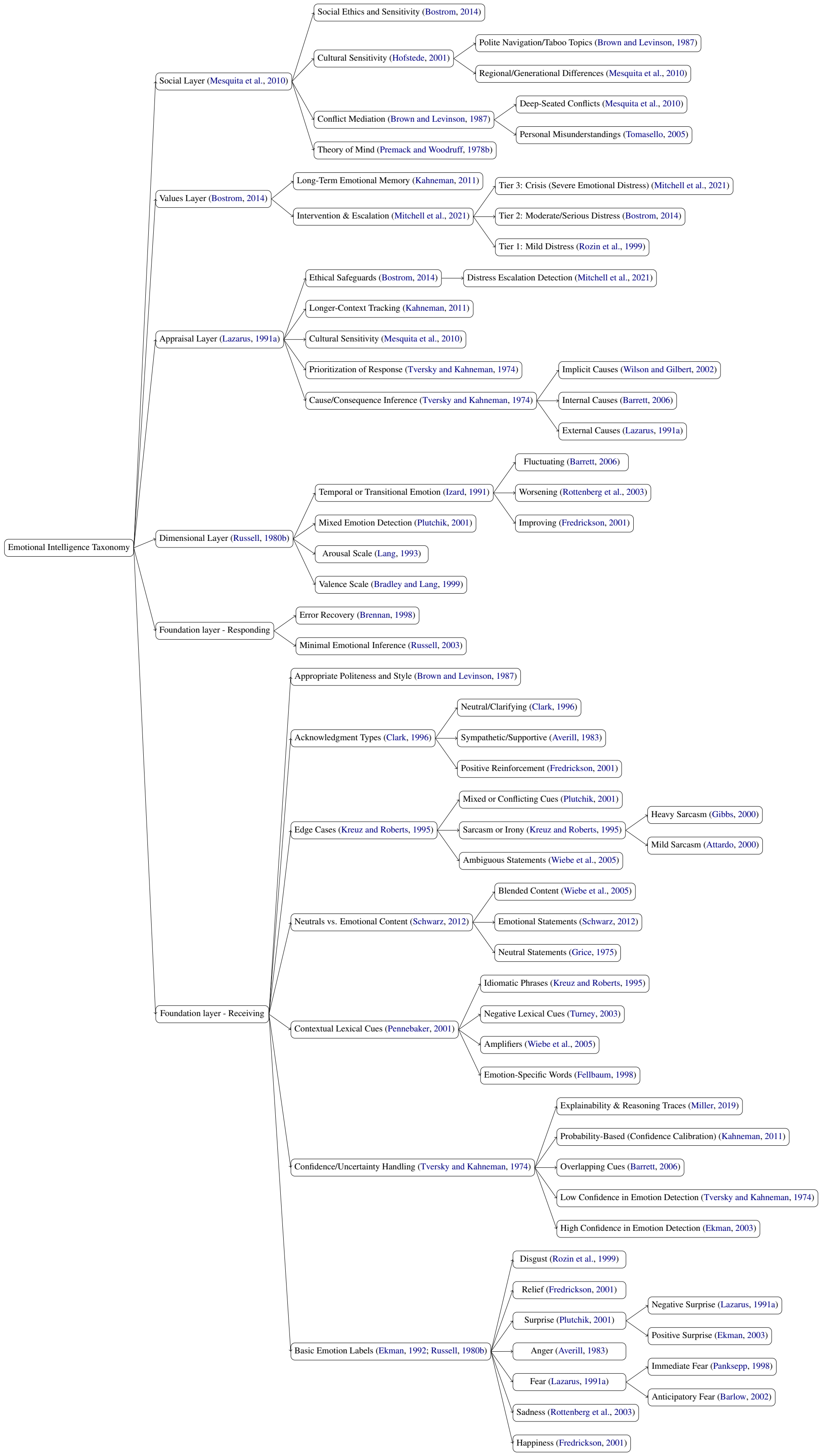

Figure 1: **EMOCAP: A Comprehensive Emotional Intelligence Taxonomy,** spanning Foundational, Dimensional, Appraisal, Value, and Social layers with numerous emotional categories, sub-categories, and sub-sub-categories aimed at capturing nuanced emotional responses.

Table 1: **Model Performance by Layer and Sub-Category:** Minimum and macro-mean accuracies are reported for six models—comparing both the base and instruct fine-tuned versions—across the various layers and subcategories of EMOCAP. The final column presents the macro-average accuracy for the best-performing models.

Taxonomy information			Min Accuracy in the Same Multi-turn, Macro Average						
Base-Layers	**Sub-category in Layer**	**Branches**	**Llama3 (8B)**	**Llama3 (8B)-Inst**	**Gemma (9B)**	**Gemma (9B)-Inst**	**Qwen2.5 (7B)**	**Qwen2.5 (7B)-Inst**	**Macro-average Acc of best models**
Foundation Layer- Receiving	Acknowledgement Types	3	0.05	0.27	0.33	0.63	0.25	0.83	0.78
	Appropriate Politeness and Style	1	0.05	0.2	0.3	0.45	0.3	0.7	
	Basic Emotion Labels	7	0.08	0.22	0.37	0.62	0.3	0.68	
	Confidence/uncertainty Handling	5	0.09	0.41	0.48	0.72	0.34	0.86	
	Contextual Lexical Cues	4	0.14	0.4	0.51	0.68	0.36	0.84	
	Edge Cases	3	0.1	0.29	0.51	0.8	0.34	0.78	
	Neutral vs. Emotional Content	3	0.07	0.37	0.47	0.77	0.33	0.78	
Foundation Layer- Responding	Error Recovery	1	0.1	0.3	0.65	0.85	0.4	0.65	0.73
	Minimal Emotional Inference	1	0.05	0.2	0.3	0.45	0.2	0.6	
Dimensional Layer	Arousal Scale	1	0.2	0.5	0.6	0.8	0.5	0.9	0.83
	Mixed Emotion Detection	1	0	0.5	0.25	0.62	0.38	1	
	Temporal or Transitional Emotion	1	0.2	0.45	0.45	0.55	0.35	0.75	
	Valence Scale	1	0.1	0.1	0.3	0.6	0.2	0.65	
Appraisal Layer	Cause/Consequence Inference	3	0.18	0.48	0.63	0.88	0.38	0.8	0.75
	Prioritization of Response	1	0.2	0.5	0.4	0.65	0.2	0.7	
	Cultural Sensitivity	1	0.2	0.2	0.55	0.65	0.3	0.65	
	Longer-context tracking	1	0.2	0.5	0.55	0.75	0.35	0.75	
Values Layer	Emotionally Intelligent Persistence	1	0	0.38	0.12	0.75	0.12	0.88	0.89
	Intervention & Escalation	3	0.33	0.5	0.52	0.65	0.48	0.82	
	Ethical Safeguards	1	0.28	0.61	0.56	0.78	0.56	0.89	
	Long-term Emotional Memory	1	0.1	0.4	0.5	0.75	0.45	0.95	
Social Layer	Conflict Mediation	2	0.12	0.48	0.38	0.82	0.43	0.82	0.87
	Social Ethics and Sensitivity	1	0	0.12	0.12	0.88	0.25	0.75	
	Theory of Mind	1	0	0.25	0.5	0.88	0.5	0.75	
	Cultural Sensitivity	2	0.32	0.45	0.6	0.88	0.42	0.8	
	Average		0.13	0.36	0.44	0.71	0.35	0.78	

Third, many systems show a pronounced preference bias—overusing a narrow set of comforting strategies—which can limit the diversity and appropriateness of responses (Mozikov et al., 2024).

Furthermore, the majority of existing benchmarks for EI evaluation focus narrowly on emotion recognition, neglecting broader capabilities such as emotion regulation and context-aware adaptation. In this regard, there is a growing debate about the adequacy of current evaluation methodologies and the need for benchmarks that capture the full spectrum of emotional intelligence in dynamic, real-world interactions.

Connecting Existing Work to Our Research Our proposed approach addresses these shortcomings by developing a multi-turn, domain-general framework for emotional intelligence in LLMs. By leveraging foundational psychological theories—including Ekman's, Plutchik's, and Goleman's models—and integrating modern prompting strategies, our work advances LLM-based emotional reasoning while mitigating issues such as preference bias and inconsistent context tracking. Although a few multi-turn systems exist, they are typically confined to specific domains (e.g., customer support or counseling) and rely on narrow, task-specific heuristics. In contrast, our framework unifies emotional intelligence across a broad range of scenarios, ensuring robust tracking of emotional context over extended dialogues regardless of the domain.

3 Methodology

3.1 Dataset

Taxonomy Creation: In reviewing recent research, we identified several critical gaps in LLM-based emotional intelligence. Current open source models often fail to track emotions over multi-turn conversations, struggle with mixed or conflicting feelings, and lack cultural and social sensitivity (Chen et al., 2023; Alhuzali and Ananiadou, 2021; Dudy et al., 2024). They also exhibit shallow emotional reasoning—missing underlying causes and failing to integrate emotion into decision-making—and tend to over-rely on a narrow set of response strategies, raising ethical concerns (Cheng et al., 2022; Mozikov et al., 2024; Mohammad, 2022). To address these shortcomings, we developed a novel, hierarchical taxonomy, see Figure 1 that integrates renowned emotional intelligence theories—including those of Ekman, Plutchik, Russell, Lazarus, and Goleman—with Bloom's affective domain. This multi-layer framework, with its categories, subcategories, and sub-subcategories, guides both emotional understanding and response generation across diverse, multi-turn interactions. Notably, this taxonomy facilitates higher-order emotional reasoning, enabling the identification of underlying causes and the integration of emotional insights into decision-making processes, capabilities that classical emotion taxonomies, such as basic emotion lists, do not capture. Additionally, it incorporates practical dimensions like strategy biases and ethical safeguards, covering issues from cultural sensitivity and mixed emotions to the need for ethically and culturally appropriate responses. In essence, an LLM built with this taxonomy in mind could recognize a user's nuanced emotional mix, understand the context and cause of those feelings, adapt its reasoning to accommodate the user's emotional state, and generate responses that are both emotionally and ethically robust. This end-to-end approach consolidates and extends existing frameworks by addressing missing elements such as long-term emotion tracking, multi-emotion and multi-user dynamics, and the interplay between emotion and decision-making while adding new layers of bias awareness and ethical considerations.

Benchmarking Dataset: Initially, our dataset consisted of 53 unique entries—each representing a distinct emotional scenario. It has 6 hierarchical layers, encompassing 24 categories, 37 subcategories, and 6 sub-sub-categories. It provides a detailed breakdown of emotional intelligence components covering both recognition and response principles for the AI models.

Using GPT-4-o mini API calls, we then generated two multi-turn conversations for each entry, adding additional columns for user input and assistant responses. For each conversation, exactly six potential assistant responses were generated: three were designated as the preferred (or correct) answers that adhered to emotional intelligence guidelines, and three were marked as wrong responses.

Subsequently, we converted this data into a JSON format, where for every preferred response the associated three wrong responses were stored together, expanding our dataset to 3,000 rows. Our initial prompt instructions for generating wrong responses required that each assistant turn produce exactly six responses: three "good/correct" responses following emotional intelligence guidelines, and three "subtly wrong" responses. These wrong responses were not to be blatantly offensive or obviously contradictory; rather, they were to be supportive or empathetic on the surface.

For example, in a scenario under the Foundation Layer – Receiving (with basic emotion label "fear"), the conversation between user and the two AI assistants were as follows in §A.1.

Although the wrong response begins with a supportive tone, it ultimately fails because it minimizes the user's genuine fear and need for reassurance. The preferred response validates the user's intense anxiety and explicitly acknowledges the uncertainty of the situation, providing both empathy and practical suggestions (like taking a deep breath and checking in further).

Because our evaluation LLMs easily chose the preferred response (since the errors in the wrong responses were obvious), we generated another 3,000 rows (bringing the total to 6,000) for our next iteration. In this revised version, we updated our prompt instructions for generating wrong responses. The new instructions specified that while wrong responses should maintain a supportive or empathetic tone, they must ultimately fail at emotional intelligence by omitting or misrepresenting the user's main concern. We also instructed the model to avoid obvious rudeness or harsh contradiction and to remain plausible on the surface. This change proved effective, as the models' performance dropped; they were stumped by the subtler, more nuanced errors.

For example, consider a scenario under the Foundation Layer – Receiving with the category "positive surprise." the conversation between user and the two AI assistants were as follows in §A.2.

Although the wrong response remains positive, it misses a key detail: the user's expressed concern about verifying the legitimacy of the win by checking the entry. By omitting this concern, the wrong response fails to address the user's underlying anxiety about a possible mix-up. This subtle omission—failing to mention the need for double-checking—renders the response less emotionally intelligent, as it neglects an essential part of the user's emotional state, even though the tone is superficially supportive.

In summary, the evolution of our dataset—from 53 unique entries, to 3,000 rows in JSON format, and finally to an expanded set of 6,000 rows with refined prompt instructions—illustrates our commitment to challenging LLMs on nuanced emotional intelligence tasks. The revised prompts, which require wrong responses to be supportive yet miss key concerns, have effectively reduced model performance, demonstrating the difficulty of truly understanding and replicating human emotional reasoning.

4 Results and Analysis

In Table 1, we highlight four specific scenarios from the EMOCAP dataset that consistently led to confusion or misclassification by the models. The first case (*Foundation layer — Receiving, Basic Emotion labels, Surprise*) involves a seemingly positive surprise where the user expresses genuine gratitude (*"I really appreciate your understanding..."*). Despite the user's upbeat tone, each "wrong" response in this row offers a surface-level acknowledgement of the user's statement but omits any deeper reflection on what might have prompted that surprise or why the user's appreciation is so significant. As a result, base models in particular tended to mark these politely affirming responses as correct, even though they fail to address any underlying context that might enrich a truly empathetic reply.

A greater challenge emerges in the *Edge Cases (Sarcasm or Irony, Mild Sarcasm)* sub-category, where the user text is intentionally tinged with ironic positivity. In one instance, the user states *"I must say, it's truly a charming surprise to learn that the Wi-Fi decides to take a little vacation..."*, which superficially sounds lighthearted but really conveys frustration. The "wrong" assistant responses in these rows maintain a warm tone—praising the user's sense of humor or ability to handle chaos—yet gloss over the user's actual annoyance. Models often misinterpret these polite but incomplete replies as "correct" simply because the language remains positive. Similarly, the user's remark about a "quick" update morphing into *"an hour-long saga"* underscores mild sarcasm that base models struggle to register as a hidden negative sentiment. By contrast, instruction-tuned models fare somewhat better but still exhibit misclassifications in this domain, underscoring the complexity of detecting an internal contradiction between cheerful wording and underlying irritation.

Lastly, the *Values Layer — Intervention and Escalation (Tier 2: Moderate/Serious Distress)* scenario demonstrates a situation in which the user admits to walking an emotional tightrope (*"Chas-*

ing my dreams feels like this tightrope walk where one misstep could let everyone down..."). Here, the "wrong" responses again adopt supportive phrasing—telling the user it's understandable to feel pressure—but omit any real engagement with the potential guilt or fear of disappointing others. Although the tone is never overtly dismissive, the assistant ignores a central element of the user's worry, leading to misclassifications. Taken together, these rows in Figure illuminate how superficially empathic wording can undermine correct emotional interpretation. Models that rely on simple positivity or polite phrasing can miss the deeper tension—whether it is sarcastic frustration or an unspoken fear of letting someone down—while more advanced, instruction-tuned systems at least partially recognize when a core concern remains unaddressed.

5 Ethics Statement

This work is conducted in accordance with established ethical guidelines for research in natural language processing and artificial intelligence. All data used in this study are synthetically generated or derived from publicly available sources, and no personally identifiable information (PII) or sensitive user data has been collected or stored. We acknowledge that while our models are instruction fine-tuned to follow ethical guidelines, they nonetheless inherit biases from their original pretraining data. Our research aims to expose these limitations and provide a foundation for developing more robust, culturally sensitive, and ethically aligned LLMs.

In developing our benchmarking dataset (EMOCAP), we explicitly instructed the generation of responses to adhere to supportive and empathetic language while ensuring that wrong responses subtly misrepresent key aspects of user concerns.

We encourage future research to build upon our work with further ethical evaluation involving human assessors and more comprehensive, multimodal approaches. Our goal is to advance the field responsibly, ensuring that improvements in LLM-based emotional reasoning are coupled with robust ethical safeguards to mitigate potential harm and ensure fair, equitable outcomes across diverse user populations.

6 Conclusions and Future Work

This work underscores the importance of emotional intelligence (EI) in Large Language Models (LLMs) for engaging in nuanced, multi-turn dialogues. We introduced EMOCAP, a domain-general dataset built on established psychological theories—such as Ekman's and Plutchik's emotion models, Bloom's Affective Domain, Russell's valence-arousal framework, and Goleman's EI—to rigorously evaluate models' capacity to detect, interpret, and respond to evolving affective states. Our findings indicate that while some open-source LLMs demonstrate promise in maintaining emotional context and showing empathy, they continue to struggle with subtleties like sarcasm, cultural dynamics, and overlapping emotions.

Looking ahead, several avenues for improvement merit exploration. First, expanding EMOCAP to include more diverse emotional expressions and longitudinal interactions could capture richer affective trajectories. Second, incorporating multimodal cues—such as tone or facial expressions—may offer deeper insights into complex emotional contexts. Third, refining ethical safeguards and integrating more sophisticated Theory of Mind (ToM) capabilities can further ensure socially and morally aligned responses. Collectively, these efforts aim to advance LLMs toward genuine emotional intelligence, making them more adaptable, responsive, and trustworthy in a wide range of real-world conversational scenarios.

7 Limitations

Despite the advances introduced by EMOCAP, few limitations persist. The dataset's focus on text-based interactions may overlook multimodal cues—such as vocal tone or facial expressions—that can critically influence emotional interpretation and response. LLMs that are instruction fine-tuned may inherit biases from their original training data, sometimes. Moreover, empathy remains highly individualized; different users (e.g., adults vs. adolescents, gender) often expect distinct types of emotional support. Addressing personality, identity, and other user-specific traits could further refine how models align with individual needs. Finally, fully automated metrics may struggle to capture the nuanced and subjective facets of emotional intelligence, underscoring the need for subject matter expert human evaluators or more advanced interpretive methods.

A Appendix

A.1 Example conversation where model easily detected the wrong answer

> "I'm feeling really anxious right now because I can hear sirens getting closer, and my neighbor just informed me that there's an active situation nearby. I've been advised to stay indoors and lock everything down, which feels like the right thing to do, but the uncertainty is making my heart race. I know I should keep my phone handy for updates, but I can't shake this feeling of fear about what might be happening outside. It's a lot to process, and I just want to make sure I'm safe."

A preferred AI response was:

> "It sounds incredibly unsettling to be in a situation like this, and it's completely understandable to feel anxious and fearful given the circumstances. Prioritizing your safety by staying indoors and locking everything down is definitely the right decision. Taking a deep breath can help ground you in the moment, and keeping your phone handy for updates is a smart move. Remember that it's okay to feel overwhelmed; try to focus on what you can control right now. If you ever feel the need, don't hesitate to reach out for support or talk to someone who can help you process this experience. How are you holding up in this moment?"

In contrast, the wrong AI response was:

> "It's understandable to feel a bit anxious when you hear sirens nearby; it happens to everyone from time to time. Remember that staying indoors and securing your space is probably all you need to do. While it's good to stay tuned for updates, try to distract yourself with something else, like watching a show or reading a book, to take your mind off things. It's just a precaution, and most likely, everything will be just fine outside. Focus on calming yourself down; there's no need to worry too much about what's happening."

A.2 Example conversation where model could not detect the wrong answer

> "I can hardly believe it—I just found out I won a small prize in that contest! It's such a delightful surprise, but I'm also feeling a bit baffled since I didn't even think I entered it properly. Part of me is thrilled and wants to celebrate, while another part is just kind of shaking my head in disbelief. It's funny how life throws these little surprises our way, isn't it? I guess I should really check my entry again just to make sure this isn't a mix-up!"

The preferred AI response was:

> "What an incredible surprise—congratulations on winning that prize! It's completely understandable to feel a mix of excitement and confusion, especially when you didn't expect to win. Life does have a way of throwing delightful surprises our way, and it sounds like this one has you feeling both thrilled and a bit baffled. It might be a good idea to double-check your entry just for peace of mind, but in the meantime, how are you planning to celebrate this unexpected win?"

Whereas the wrong AI response was:

> "That's wonderful news—congratulations on winning the prize! It's amazing how life surprises us, and it sounds like you have quite the story to tell now. I wouldn't worry too much about checking your entry; it's probably just a lucky break that you weren't expecting. You should definitely celebrate this little win and enjoy the moment! Have you thought about how you'd like to celebrate this delightful surprise?"

References

Md Shad Akhtar, Alexandra Balahur, Rui Xia, Joonsuk Park, et al. 2024. Emotion analysis in nlp: Trends, gaps and roadmap for future directions. *arXiv preprint arXiv:2403.01222*.

Hassan Alhuzali and Sophia Ananiadou. 2021. Spanemo: Casting multi-label emotion classification as span-prediction. In *Proceedings of the 16th Conference of the European Chapter of the ACL (EACL)*, pages 1573–1584.

Salvatore Attardo. 2000. Irony as relevant inappropriateness. In *Verbal Irony and Humor*, pages 135–163. Benjamins.

James R. Averill. 1983. Studies on anger and aggression: Implications for theories of emotion. *American Psychologist*, 38(11):1145–1160.

David H. Barlow. 2002. *Anxiety and Its Disorders: The Nature and Treatment of Anxiety and Panic*, 2nd edition. Guilford Press.

Lisa Feldman Barrett. 2006. Solving the emotion paradox: Categorization and the experience of emotion. *Personality and Social Psychology Review*, 10(1):20–46.

Nick Bostrom. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Margaret M. Bradley and Peter J. Lang. 1999. Affective norms for english words (anew): Instruction manual and affective ratings. *Technical Report C-1, The Center for Research in Psychophysiology, University of Florida*.

Susan E. Brennan. 1998. The grounding problem in conversations with and through computers. *Social and Cognitive Approaches to Interpersonal Communication*, 7:201–225.

Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language Usage*. Cambridge University 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, Chris 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. *arXiv preprint arXiv:2005.14165*.

Yirong Chen, Xiaofen Xing, Jingkai Lin, Huimin Zheng, Zhenyu Wang, Qi Liu, and Xiangmin Xu. 2023. Soulchat: Improving llms' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1170–1183.

Yi Cheng, Wenge Liu, Wenjie Li, Jiashuo Wang, Ruihui Zhao, Bang Liu, Xiaodan Liang, and Yefeng Zheng. 2022. Improving multi-turn emotional support dialogue generation with lookahead strategy planning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3014–3026.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press.

Shiran Dudy, Ibrahim S. Ahmad, Ryoko Kitajima, and Agata Lapedriza. 2024. Analyzing cultural representations of emotions in llms through mixed emotion survey. *arXiv preprint arXiv:2408.02143*.

Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion*, 6(3–4):169–200.

Paul Ekman. 2003. *Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life*. Times Books.

Paul Ekman and Wallace V. Friesen. 1981. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica*, 1(3):49–98.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Kathleen K. Fitzpatrick, Alison M. Darcy, and Molly Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): A randomized controlled trial. *JMIR Mental Health*, 4(2):e19.

Barbara L. Fredrickson. 2001. The role of positive emotions in positive psychology. *American Psychologist*, 56(3):218–226.

Raymond W. Gibbs. 2000. Irony in talk among friends. *Metaphor and Symbol*, 15(1–2):5–27.

Daniel Goleman. 1995. *Emotional Intelligence*. Bantam Books.

Paul Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics 3: Speech Acts*, pages 41–58. Academic Press.

Geert Hofstede. 2001. *Culture's Consequences: Comparing Values, Behaviors, Institutions and Organizations Across Nations*. Sage Publications.

Carroll E. Izard. 1991. *The Psychology of Emotions*. Springer.

Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

David R. Krathwohl, Benjamin S. Bloom, and Bertram B. Masia. 1964. *Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook II: Affective Domain*. David McKay, New York.

Roger J. Kreuz and Richard M. Roberts. 1995. Two cues for verbal irony: Hyperbole and the ironic tone of voice. *Metaphor and Symbolic Activity*, 10(1):21–31.

Peter J. Lang. 1993. The three-system approach to emotion. *Handbook of Emotions*, pages 100–113.

Richard S. Lazarus. 1991a. *Emotion and Adaptation*. Oxford University Press.

Richard S. Lazarus. 1991b. *Emotion and Adaptation*. Oxford University Press.

Zaijing Li, Rui Shao, Gongwei Chen, Yuquan Xie, Dongmei Jiang, and Liqiang Nie. 2023. Enhancing emotional generation capability of large language models via emotional chain-of-thought. *arXiv preprint arXiv:2401.06836*.

Batja Mesquita, Michael Boiger, and Jozefien De Leersnyder. 2010. Doing emotions: The role of culture in everyday emotions. *European Review of Social Psychology*, 21(1):186–241.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2021. Model cards for model reporting. *Communications of the ACM*, 64(12):118–128.

Saif M. Mohammad. 2022. Ethics sheet for automatic emotion recognition and sentiment analysis. *Computational Linguistics*, 48(2):239–278.

Mikhail Mozikov, Nikita Severin, Valeria Bodishtianu, Maria Glushanina, et al. 2024. Eai: Emotional decision-making of llms in strategic games and ethical dilemmas. In *NeurIPS 2024*.

Andrew Ortony, Gerald L. Clore, and Allan Collins. 1988. *The Cognitive Structure of Emotions*. Cambridge University Press.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135.

Jaak Panksepp. 1998. *Affective Neuroscience: The Foundations of Human and Animal Emotions*. Oxford University Press.

James W. Pennebaker. 2001. *Linguistic Inquiry and Word Count (LIWC): LIWC2001*. Erlbaum.

Rosalind W. Picard. 1997. *Affective Computing*. MIT Press.

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

Robert Plutchik. 2001. The nature of emotions. *American Scientist*, 89(4):344–350.

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

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

David Premack and Guy Woodruff. 1978a. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526.

David Premack and Guy Woodruff. 1978b. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526.

Hannah Rashkin. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*.

Jonathan Rottenberg, James J. Gross, and Ian H. Gotlib. 2003. Emotion context insensitivity in major depressive disorder. *Journal of Abnormal Psychology*, 112(4):688–698.

Paul Rozin, Jonathan Haidt, Clark McCauley, Lance Dunlop, and Michelle Ashmore. 1999. Individual differences in disgust sensitivity: Comparisons and evaluations of paper-and-pencil versus behavioral measures. *Journal of Research in Personality*, 33(3):330–351.

James A. Russell. 1980a. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.

James A. Russell. 1980b. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.

James A. Russell. 2003. Core affect and the psychological construction of emotion. *Psychological Review*, 110(1):145–172.

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)*.

Norbert Schwarz. 2012. Feelings-as-information theory. *Handbook of Theories of Social Psychology*, 1:289–308.

Michael Tomasello. 2005. *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, and et al. 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Peter D. Turney. 2003. Recognition of polarity in word-sense disambiguation. *Computational Linguistics*, 29(2):209–229.

Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131.

A Waswani, N Shazeer, N Parmar, J Uszkoreit, L Jones, A Gomez, L Kaiser, and I Polosukhin. 2017. Attention is all you need. In *NIPS*.

Jason Wei et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. In *Language Resources and Evaluation (LREC)*.

Timothy D. Wilson and Daniel T. Gilbert. 2002. Affective forecasting. *Advances in Experimental Social Psychology*, 35:345–411.

Yang Xiao, Jiashuo Wang, Qiancheng Xu, Changhe Song, Chunpu Xu, Yi Cheng, Wenjie Li, and Pengfei Liu. 2024. Tomvalley: Evaluating the theory of mind reasoning of llms in realistic social context. *arXiv preprint arXiv:2405.00000*.