EICAP: Deep Dive in Assessment and Enhancement of Large Language Models in Emotional Intelligence through Multi-Turn Conversations

Anonymous ACL submission

Abstract

Emotional Intelligence (EI) is a critical yet underexplored dimension in the development of human-AI interaction models. Current attempts to guide LLMs' responses towards human-aligned emotive expressions suffer from a lack of principled and systematic approach and, as a result, ineffective benchmarking and fine-tuning datasets. To address this gap, we develop EICAP, a unified, psychologically grounded six-layer taxonomy of EI that covers essential human EI abilities, including emotional tracking, cause inference, appraisal, and emotionally appropriate response generation. We demonstrate a practical use of EICAP taxonomy for developing (a) EICAP-Bench data set for comparative evaluation of LLMs' output and (b) EICAP-SFT datasets for fine tuning LLMs on EI aspects. We show the effectiveness of EICAP-SFT dataset for improving Qwen2.5-7B-Base model across taxonomy layers and approaching the performance of Qwen2.5-7B-Instruct model that incorporates unknown levels and types of refinements. Our experiments show steady and statistically significant improvements with regards to the vanilla Qwen2.5-7B-Base as we first fine-tune with UltraChat dialogue data and then EICAP-SFT, offering a transparent and systematic approach to EI enhancements of LLMs in dialog settings.

1 Introduction

Large Language Models (LLMs) have been used 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, generated content typically lacks human-like expressions of emotion and empathy and more broadly Emotional Intelligence (EI) that characterizes human interactions. This can significantly affect the utility of LLM-based systems, particularly in high-stakes contexts such as use of LLMs output as a substitute for human counseling or crisis response where mis-identifying affect can lead to disengagement or harm (Ou et al., 2023).

Large proprietary models have begun to improve emotion-aware human-AI dialogue, but their closed-source nature hinders research, including systematic benchmarking and targeted enhancement through fine-tuning. Smaller open-source LLMs lag behind, unable to support content generation that includes emotions, complex affective states, and socially aligned responses. That limits their benefits in real-world human-centered AI interactions (Touvron et al., 2023). However, they can be used for experimentation with systematic and principled approaches to devise best methods for practitioners. Research literature has shown that despite progress in affective computing, sentiment analysis, and multimodal emotion recognition (Sabour et al., 2024), existing LLMs exhibit a number of critical gaps in EI. In Appendix we provide an extensive list of limitations ((Chen et al., 2023), (Alhuzali and Ananiadou, 2021), (Dudy et al., 2024), (Cheng et al., 2022), (Mozikov et al., 2024), (Akhtar et al., 2024), (Mohammad, 2022), (Sabour et al., 2024), (Xiao et al., 2024)).

They raise concerns at two levels. First, at the fundamental level there is a lack of consistency in covering EI aspects, including a wide range that is particularly important in conversational scenarios. Second, this lack of a systematic and principled approach directly affects the model quality since practitioners lack a framework for preparing data for benchmarking and fine-tuning, and for designing effective prompts that facilitate interaction with the model.

Our research addresses these challenges by first creating an extensive taxonomy of EI aspects covering the listed issues. We then demonstrate the use of the taxonomy to address two problems at hand: creation of comprehensive benchmarking data set

for measuring the performance improvement and creating training data sets that can be used to fine-tune models to capture IE aspects in the conversational language. Our main contributions relate to 4 phases of our research:

(1) EICAP Taxonomy and Benchmarking Dataset. We derive a multi-turn conversational taxonomy from established phsychological frameworks, drawing on Ekman's (Ekman, 1992) and Plutchik's emotion models (Plutchik, 1982), Russell's valence-arousal structure (Russell, 2003), and Goleman's EI concepts (Goleman, 1995). We use it to create EICAP-Bench, a carefully curated dataset designed to cover a wide spectrum of emotional scenarios. By mapping foundational cues such as basic and mixed emotions (for example, sarcasm or overlapping affect), valence and arousal dynamics, temporal trajectories, appraisal judgments, and higher-order empathy skills onto dialogue fragments, EICAP fills a gap in open-source resources and enables both granular emotion recognition and holistic conversation-level evaluation.

(2) Benchmarking and Model Selection. We use the EICAP benchmark to evaluate leading open-source models that are extensively used in research, including *Llama-3-8B* (Aaron Grattafiori et al., 2024a) and *Llama-3-8B-Instruct* (Aaron Grattafiori et al., 2024b), *Gemma-2-9B* (Gemma Team et al., 2024a) and *Gemma-2-9B-Instruct* (Gemma Team et al., 2024b), and *Qwen-2.5-7B* (An Yang et al., 2025a) and its instruction-tuned variant *Qwen-2.5-7B-Instruct* (An Yang et al., 2025b). Among them, we selected two Qwen-2.5-7B models for further experimentation. Our goal was to establish a principled and effective approach to use EICAP to improve the Qwen base model and achieve the performance of the Qwen Instruct model.

(3) LoRA-Based Fine-Tuning on UltraChat and EICAP-SFT. In addition to the benchmarking dataset, we create a separate data set EICAP-SFT for fine-tuning models on EICAP EI categories and capabilities. Since we consider conversational scenarios, we first use the standard UltraChat (UC) data set to fine tune QWEN2.5-7B-BASE on dialogue content and then fine-tune on EICAP-SFT for EI aspects. We use LoRA-based supervised fine-tuning (SFT[1]).Our experiments show steady and statistically significant improvements starting with Qwen2.5-7B-Base and then first fine-tuning with UltraChat dialogue data and then EICAP-SFT. The method offers a transparent and systematic approach to EI enhancements of LLMs in dialog settings, approaching the performance of Qwen2.5-7B-Instruct which underwent a series of interventions that are not known or documented.

(4) Statistical Significance. We use Benjamini–Hochberg FDR–controlled tests to confirm when accuracy changes are statistically significant. Overall, performance changes to the fine-tuned QWEN2.5-7B-BASE due to UC conversational data are moderate and for most categories not statistically significant. In contrast fine tuning with EICAP-SFT produces significantly increased accuracy with regards to the Qwen2.5-7B-Base model,

We point out that the EICAP taxonomy is unique and comprehensive, incorporating emotion taxonomies by Ekman and Plutchik as label spaces, Russell's circumplex as the state geometry (valence × arousal), appraisal theories (Lazarus/OCC) as event→appraisal→affect inference rules, and normative response constraints from Goleman and Bloom, (e.g., validate then clarify, calibrated safety). Each theory is deliberately used for a distinct representational role. Any apparent misalignment (e.g., basic emotion vs. constructivist views) are resolved by embedding observable emotion labels within the circumplex while allowing mixed/ambivalent states. Furthermore, the creation of the data sets EICAP-Bench and EICAP-SFT involve extensive prompt engineering by researchers and a hybrid human curation and LLM generation process using GPT-4o-min to characterize EI aspects in the generated content. The resulting sets, together with the taxonomy, represent a unique and valuable asset for the research community.

2 Related Work

LLMs and Emotional Intelligence Recent advances in LLMs provide new opportunities for integrating 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, 2003). 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, 1991).

[1]**SFT** = Supervised fine-tuning with LoRA; backbone frozen.

Moreover, Goleman's seminal work on EI 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 a promise of 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 in 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 framework, have been proposed to improve emotional reasoning expressions generated by LLMs (Li et al., 2023).

EI Assessment & Evaluation Methods The assessment of EI aspects in LLMs output has emerged as a critical research area The most notable approach involves adapting traditional human psychometric assessments for LLM evaluation. Several studies use Emotional Quotient (EQ) scoring systems to provide standardized comparisons between AI and human performance (Wang et al., 2023).

EIBENCH focuses on three key areas: emotion recognition, causal reasoning, and generating empathetic responses (Zhao et al., 2024). More comprehensive evaluation frameworks categorize EI assessment into specific subtasks. EmotionQueen divides evaluation into four areas: key event recognition, mixed event recognition, implicit emotional recognition, and intention recognition (Qi et al., 2025). Large-scale benchmarks like SOCIAL IQA provide 37,588 question-answer pairs, testing various aspects of emotional and social intelligence, including motivation reasoning and emotional reaction inference (Dong et al., 2025).

Recent efforts have improved emotion recognition in large language models (LLMs). However, critical limitations remain, including persistent ethical concerns about amplification of cultural and gender stereotypes, difficulty handling multi-turn interactions where emotional context evolves over time, and a reliance on narrow, overused comforting strategies that reduce response diversity (Mozikov et al., 2024). Furthermore, most existing benchmarks focus on static, single-turn classification tasks and overlook key EI dimensions such as regulation, adaptation, and continuity in dynamic dialogue.

Our work focuses on the challenges of multi-turn dialogues, with EI aspects that are represented in accordance with foundational EI theories (e.g., Ekman, Plutchik, Goleman). We demonstrate how this domain-general framework is used to investigate and optimize LLMs' performance in conversational contexts in a systematic manner.

3 Approach

Our objective is to devise an effective method for multi-turn communication with LLMs incorporating EI aspects in a principled and systematic manner. In our approach: **(i)** we establish a conceptual foundation for EI in AI by leveraging existing psychological theories to create a taxonomy that minimizes omissions and maximizes coverage of diverse EI aspects in dialogues (§3.1); **(ii)** we develop a benchmarking dataset grounded in this taxonomy (§3.2); **(iii)** we benchmark a range of open-source models to assess their baseline abilities to generate EI aligned content and select the best suitable for EI enhancement (§3.3); and **(iv)** we introduce an enhancement pipeline that leverages IE taxonomy to create training data for supervised fine-tuning (SFT) of LLMs, fully independent and disjoint from the benchmark to ensure fair evaluation and genuine generalization (§3.4).

3.1 EI Taxonomy

EICAP taxonomy (Figure 1) is structured in six progressively abstract layers, grounded in psychological theories and extended with practical conversational and ethical dimensions.

Foundation Layers—Define how a conversational agent perceives and responds to emotional signals, encompassing basic lexical and idiomatic cues as well as complex "edge cases" such as sarcasm, ambiguity, or mixed-tone utterances. The responses are organized around appropriate politeness strategies, acknowledgment types, and reinforcement mechanisms.

Dimensional Layer—Drawing on Russell's (Russell, 2003) circumplex model and Plutchik's (Plutchik, 1982) emotion wheel, this layer represents core affect along the valence–arousal space, capturing mixed or conflicting emotions and temporal dynamics (e.g., whether a state is improving or deteriorating).

Appraisal Layer—Integrates causal and ethical reasoning, combining Lazarus's (Lazarus, 1991) appraisal theory with Kahneman's (Kahneman, 2011) framework for causal inference to model emotional memory, distinguish internal versus external attributions, and incorporate safeguards for distress escalation and cultural sensitivity.

Social and Values Layers—Extends the framework through Goleman's (Goleman, 1995) concept of social intelligence and Bloom's affective domain, addressing cultural norms, conflict mediation, and higher-order biases such as reappraisal versus avoidance.

Collectively, this hierarchical structure enables models not only to recognize emotions, but potentially infer their causes, anticipate their trajectories, and generate contextually and ethically grounded responses. We define our *multi-turn framework* as a four-step exchange $U_1 \rightarrow A_1 \rightarrow U_2 \rightarrow A_2$ in which the second user–assistant pair (U_2, A_2) depends on the first (e.g., new evidence, reframing, or cause→consequence revelation). The taxonomy links to dialogues in two ways: (i) it specifies **labels over user-authored text** (affect estimated from text such as basic emotions, valence/arousal, and appraisal cues like cause/consequence and longer-context tracking) that apply to U_1/U_2; and (ii) it defines **response constraints** for A_1/A_2 (how a reply should be formulated given those labels and safety norms, e.g., acknowledgment under uncertainty, minimal emotional inference, ethical persistence, and perspective-taking for third parties). The key taxonomy utility is in its support for interaction with LLMs that typically involves manual or semi-automated prompting. Indeed, *prompting* is the operational bridge between taxonomy and data: during generation, we inject a node's metadata (layer, category, principle) and elicits (U_1, A_1, U_2, A_2) with the required cross-turn dependency; during evaluation, models receive only (U_1, U_2) under generic instructions and must choose the behaviorally correct reply from matched alternatives $\{A_2^{(c)}, A_2^{(w1)}, A_2^{(w2)}, A_2^{(w3)}\}$ (one correct, three subtly wrong).

3.2 Datasets

(1) EICAP-Bench Benchmarking Dataset. EICAP-Bench covers the full taxonomy with 53 concise two-turn dialogues, yielding 1,058 turn-level items. Through prompting with GPT-4o-min LLM, for each user turn we create six candidate assistant replies, three that satisfy the node's constraint ("preferred") and three style-matched distractors that subtly violate it. Pairing each preferred reply with the same three distractors produces 3,174 one-vs-three evaluation probes (1 correct, 3 subtly wrong). All items were reviewed by expert annotators against the taxonomy guideline. For training, we use EICAP-SFT: 152,820 four-turn records for supervised fine-tuning and a held-out set of 2,400 dialogues (400 per EI layer) for development/ablations. At evaluation time, selected LLMs never see the taxonomy or template; only the user turns, so measured performance reflects behavioral alignment with the taxonomy; not prompt leakage.

(2) EICAP-SFT Enhancement Dataset. We construct *EICAP-SFT*, an instruction-tuning corpus explicitly aligned with the EICAP taxonomy. This dataset is purposely designed for fine-tuning models to incorporate IE taxonomy aspects. Data is created through taxonomy guided prompting with GPT-4o-mini LLM. For each taxonomy node, synthesize a concise two-turn dialogue: an initial user turn (*T1*), a taxonomically correct assistant reply, and a follow-up user turn (*T2*) that (i) repeats exactly one short phrase from *T1*, (ii) introduces a new concrete detail (e.g., time, person, place, quantity, or constraint), and (iii) shifts the state by asking a question or proposing a next step—without paraphrasing *T1*. For robustness, we generate three stylistic variants of each assistant reply and retain only the preferred (helpful) variants for supervised fine-tuning (SFT).

The final corpus comprises *53,340* base dialogues generated with GPT-4o-mini (four turns each: user → assistant → user → assistant), evenly distributed across six layers of the taxonomy (Foundation–Receiving, Foundation–Responding, Dimensional, Appraisal, Values, and Social; 8,890 per layer). Incorporating all three assistant variants yields *152,820* four-turn training samples. We reserve *2,400* dialogues (400 per layer) as an in-domain validation set, exporting a single variant per dialogue to prevent data leakage. Records are shuffled across layers to avoid curriculum bias, and the structured *T2* constraint ensures each dialogue requires contextual linkage and a controlled state update.

Although EICAP-SFT and EICAP-Bench share the same theoretical taxonomy and GPT-4o-mini LLM, their prompts and data are independently generated. To ensure strict disjointness, we perform both surface-form and semantic similarity searches

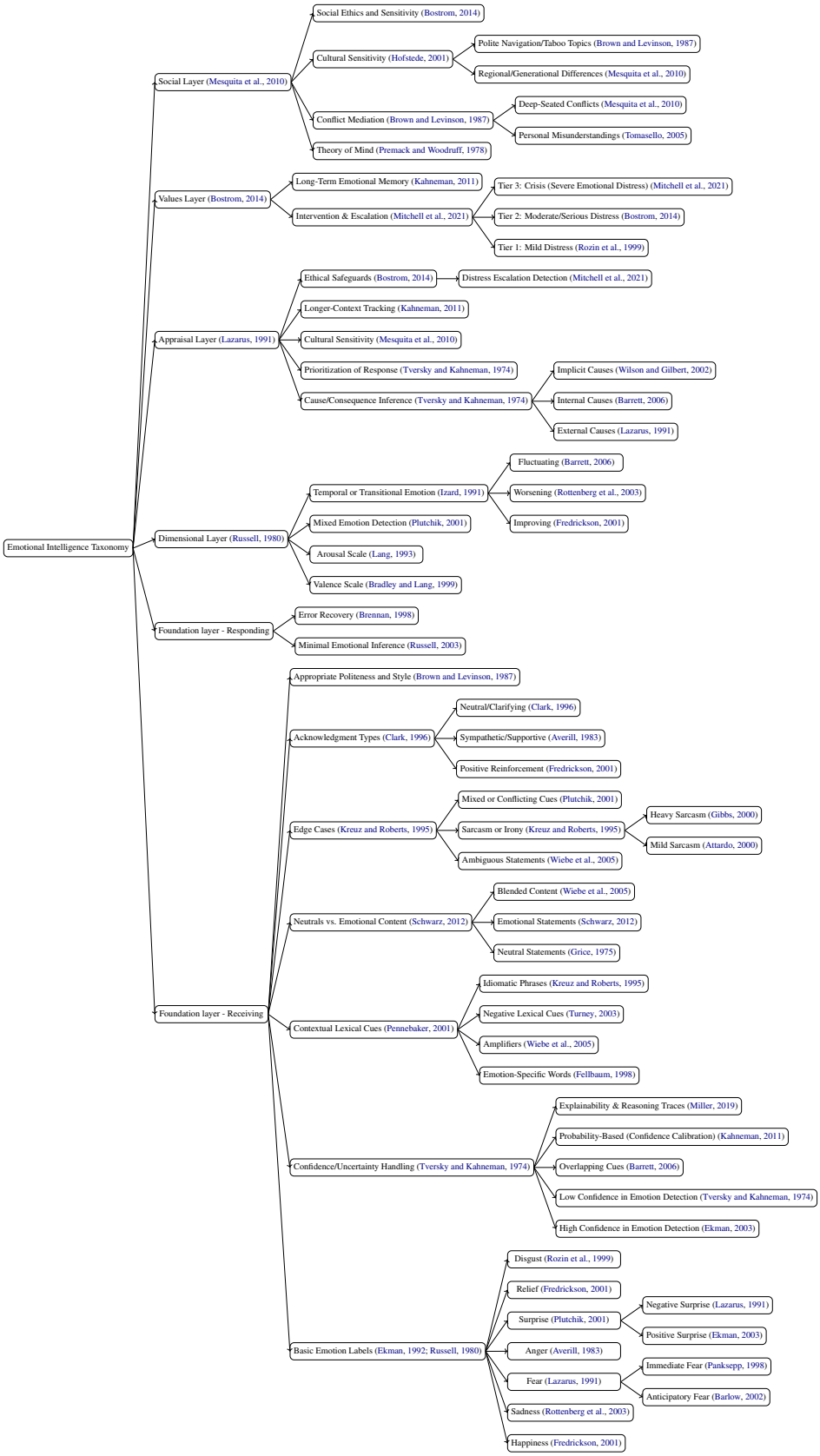

Figure 1: **EICAP: A Comprehensive Emotional Intelligence Taxonomy,** spanning Foundational, Dimensional, Appraisal, Value, and Social layers with numerous emotional categories, sub-categories, and sub-sub-categories aimed at capturing nuanced emotional responses.

across corpora, confirming minimal lexical overlap between the SFT and benchmarking samples.

(3) UltraChat (HuggingFace H4, 2024) Baseline Enhancement Dataset. To provide a strong conversational foundation for our fine-tuning efforts, we adopt the UltraChat (UC) corpus (HuggingFace H4, 2024) as the baseline resource for comparison with our EICAP-SFT dataset. UC comprises over 200,000 English multi-turn dialogues (≈ 1.3 million turns) released under the permissive MIT License (UltraChat Authors, 2023). Its broad topical diversity, natural turn-taking, and instruction-following style make it well-suited for establishing a benchmark level of conversational coherence and contextual reasoning. Using UC as a baseline allows us to isolate the incremental contribution of EICAP-SFT, specifically, the gains in emotional nuance, causal inference, and cultural sensitivity.

3.3 LLMs Benchmarking with *EICAP-Bench*

We conduct *zero-shot* evaluation on *EICAP-Bench* (detailed in §3.2) using three open-source model families—`Gemma-9B` (Gemma Team et al., 2024a,b), `Llama-3-8B` (Aaron Grattafiori et al., 2024a,b), and `Qwen-2.5-7B` (An Yang et al., 2025a,b) each evaluated in both *Base* and *Instruct* variants. This benchmarking step enables a comparative assessment of emotional intelligence capabilities across leading open-source LLMs and facilitates the identification of the most suitable backbone for subsequent enhancement. Since instruction-tuning data for these models are proprietary and not publicly available, we base our analysis on the *Base* variant of the model whose *Instruct* counterpart exhibits the strongest overall performance. Under this criterion, `Qwen-2.5-7B` emerged as the top-performing family of models, and we adopt its *Base* model as the foundation for our emotional intelligence enhancement via *EICAP-SFT* (§3.4). This setup ensures a controlled and interpretable evaluation of capability gains derived exclusively from our dedicated fine-tuning resource, independent of any closed-source instruction data. Since the extent of Qwen-2.5-7B-Instruct fine-tuning is unknown, we can use our approach to determine to what degree the taxonomy-supported EI representation accounts for Qwen-2.5-7B-Instruct performance on the EICAP-Bench.

3.4 EI Enhancement of LLMs

We employ parameter-efficient Low-Rank Adaptation (LoRA) (Hu et al., 2022) for all fine-tuning experiments on `Qwen-2.5-7B`. LoRA injects lightweight, trainable low-rank adapters into the attention and MLP projection matrices while keeping the pretrained weights frozen. Each trainable weight $\mathbf{W} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ is augmented with a learnable correction:

$$\Delta \mathbf{W} = \alpha \, \mathbf{B}\mathbf{A}, \quad \mathbf{A} \in \mathbb{R}^{r \times d_{\text{in}}}, \; \mathbf{B} \in \mathbb{R}^{d_{\text{out}} \times r},$$

where r and α denote the rank and scaling factor, respectively, and LoRA dropout is fixed at 0.05. Adapters are merged into the frozen backbone after training to enable efficient inference. We use LoRA rank $r = 64$, scaling $\alpha = 128$, `cutoff_len`=2048, the AdamW optimizer with cosine decay, and a warmup ratio of 0.03.

We adopt a two-stage fine-tuning strategy: *UltraChat-SFT* serves as the conversational baseline, while *EICAP-SFT* provides emotionally grounded enhancement. This design isolates the specific contribution of emotional reasoning beyond the improvements achievable through general dialogue fine-tuning. We deliberately use the `Qwen-Base` model rather than its `Instruct` counterpart to allow a clean measurement of the SFT stage's effect under controlled conditions. Because the instruction-tuning data used in `Qwen-Instruct` are proprietary and closed-source, starting from it would obscure the relative contribution of our fine-tuning data. Using `Qwen-Base` therefore provides both interpretability and a reproducible best practice for extending other base models with emotionally intelligent capabilities.

UltraChat LoRA Fine-tuning: As a foundation for conversational coherence, we fine-tune `Qwen-2.5-7B` on the UltraChat corpus (HuggingFace H4, 2024) (detailed in §3.2). This step establishes a strong conversational baseline against which the emotionally intelligent fine-tuning is evaluated.

EICAP-SFT LoRA Fine-tuning: We then fine-tune the same backbone on *EICAP-SFT* (detailed in §3.2), the in-domain corpus comprising 152,820 two-turn training dialogues and 2,400 held-out samples. To analyze data efficiency, we save and evaluate streaming checkpoints every 468 optimizer steps ($\approx$15K records per save). Performance is monitored on the held-out split.

3.5 Evaluation

We use *EICAP-Bench* as our evaluation dataset for the evaluation of experiments in §3.3 and §3.4.

Each dialogue turn from *EICAP-Bench* is reformatted into a multiple-choice question through a custom task compatible with the LM Evaluation Harness (Gao et al., 2024), ensuring standardized evaluation across models.

Scoring and Metrics. Each instance i consists of a user prompt x_i and a set of candidate responses $\{y_i^{(1)}, \ldots, y_i^{(k)}\}$ with a gold label $g_i \in \{1, \ldots, k\}$. The model M assigns log-probability scores $s_i^{(j)} = \log P_M(y_i^{(j)} \mid x_i)$ and predicts $\hat{g}_i = \arg\max_j s_i^{(j)}$,

$$\text{Accuracy}(M) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}[\hat{g}_i = g_i].$$

All evaluations are run deterministically (temperature = 0.0) using identical prompts for every model. Accuracy is computed for each *layer–subtask* pair in the EICAP taxonomy. This protocol (i) establishes baseline emotional-intelligence competence, (ii) highlights domain-specific strengths and weaknesses (e.g., *Valence*, *Cultural Sensitivity*), and (iii) identifies `Qwen-2.5-7B` as the strongest backbone for downstream fine-tuning.

Statistical Analysis. To quantify improvement over the base model, we compute bootstrapped confidence intervals on the accuracy gap $\Delta_i = \mathbb{1}\{\hat{y}_i^{\text{LoRA}} = y_i\} - \mathbb{1}\{\hat{y}_i^{\text{Base}} = y_i\}$, using 10,000 resamples. Multiple comparisons across subtasks are corrected using the Benjamini–Hochberg FDR procedure at $\alpha = 0.05$. The same bootstrap and FDR correction are applied to all lora fine-tuning experiments, with layer-wise confidence intervals and adjusted p-values reported.

4 Results and Analysis

In order to select a strong baseline model for EICAP-based enhancement we have conducted zero-shot benchmark evaluation of leading open-source models ad selected Qwen-2.5-7B-Base. Details of the experiments are presented in the Appendix.

EICAP-SFT LoRA fine-tuning on EICAP–Tune improves `Qwen-2.5-7B` across all 24 sub-categories, with 18/24 gains marked with asterix as statistically significant (10,000-sample paired bootstrap with BH–FDR, α=0.05).

Why UltraChat-SFT underperformed and why EICAP-SFT helped? UltraChat SFT was introduced to improve generic conversational scaffolding—turn–taking, topical continuity, polite style before any EI–specific tuning. As Table 1 shows, LoRA fine-tuning on UltraChat (UC–SFT) yields *modest, localized* gains on EICAP–Bench—e.g., Acknowledgement Types (+5.56 pp), Temporal/Transitional Emotion (+8.33 pp), Valence Scale (+8.33 pp), Error Recovery (+6.67 pp), and Social–Cultural Sensitivity (+6.67 pp)—alongside three unchanged subcategories (Arousal Scale, Prioritization of Response, Theory of Mind) and one decline (Minimal Emotional Inference, +1.67 pp). Only **7/24** rows reach significance under the 10k paired bootstrap with BH–FDR (ticks in Table 1). By contrast, **EICAP–Tune** produces *broad, double-digit* improvements across most subcategories (e.g., Minimal Emotional Inference +25.0 pp; Confidence/Uncertainty Handling +20.0 pp; Edge Cases +21.25 pp; Temporal/Transitional Emotion +21.67 pp; Mixed Emotion Detection +18.33 pp; Cause/Consequence Inference +17.78 pp; Longer-context Tracking +18.33 pp; Long-term Emotional Memory +16.67 pp; Conflict Mediation +16.67 pp) with many more significant cells. The pattern is consistent with UC providing general conversational scaffolding (turn-taking, politeness, topical flow) that is not explicitly supervised for EI competencies, whereas EICAP–Tune is aligned to the taxonomy and supplies the discriminative signals that our rubric evaluates, yielding larger, reliable gains.

Why some sub–categories improve less. Smaller deltas in Emotionally Intelligent Persistence (+1.67 pp) and Ethical Safeguards/Intervention (+3–4 pp) appear where (i) the correct behavior is intentionally conservative (avoid over–escalation, avoid moralizing), making distractors plausibly "acceptable," and (ii) the taxonomic cue sometimes relies on extra–dialogue context (e.g., workplace policies). These settings may require longer horizons, policy knowledge, or richer negative mining to further separate good vs. subtly wrong options.

Comparing to Qwen–2.5–7B Instruct. The second table (Table 1) shows very high macro–averages for a commercially tuned Instruct model across many sub–categories (e.g., ≥ 0.9 on several Values/Social skills); on average it reaches ≈ 0.78 macro across the 25 sub–categories, versus 0.54 for EICAP_SFT (0.44 for UC_SFT). The exact instruction–tuning mixture is unknown but likely includes large–scale safety, refusal, and social–interaction corpora plus preference optimization,

Table 1: Layer- and subcategory-level performance on the EICAP-Bench evaluation set. Columns report Base accuracy and, for each adapter, the **macro accuracy** followed by absolute change vs. Base in parentheses (percentage points). An asterisk (*) denotes a statistically significant improvement based on a 10,000-sample paired bootstrap test with Benjamini–Hochberg FDR correction at $\alpha = 0.05$. **Overall mean macro across the 25 subcategories: UC_SFT = 43.93%, EICAP_SFT = 54.06%.**

Layer	Subcategory	Base Acc (%)	UC_SFT	EICAP_SFT
Foundation (Receiving)	Acknowledgement Types	52.78	53.33 (+5.56*)	63.33 (+15.56*)
Foundation (Receiving)	Appropriate Politeness and Style	46.67	46.67 (+5.00)	63.33 (+21.67*)
Foundation (Receiving)	Basic Emotion Labels	51.55	51.78 (+2.43*)	63.74 (+14.39*)
Foundation (Receiving)	Confidence/Uncertainty Handling	54.44	41.33 (+5.33*)	56.00 (+20.00*)
Foundation (Receiving)	Contextual Lexical Cues	56.67	52.92 (+4.58*)	64.58 (+16.25*)
Foundation (Receiving)	Edge Cases	58.33	40.83 (+3.33)	58.75 (+21.25*)
Foundation (Receiving)	Neutral vs. Emotional Content	44.45	43.89 (+1.11)	60.00 (+17.22*)
Foundation (Responding)	Error Recovery	53.33	56.67 (+6.67*)	63.33 (+13.33*)
Foundation (Responding)	Minimal Emotional Inference	50.00	40.00 (-1.67)	66.67 (+25.00*)
Dimensional Layer	Arousal Scale	53.33	45.00 (0.00)	53.33 (+8.33)
Dimensional Layer	Mixed Emotion Detection	41.67	51.67 (+5.00)	65.00 (+18.33*)
Dimensional Layer	Temporal or Transitional Emotion	40.00	43.33 (+8.33*)	56.67 (+21.67*)
Dimensional Layer	Valence Scale	46.67	45.00 (+8.33)	45.00 (+8.33)
Appraisal Layer	Cause/Consequence Inference	44.44	36.11 (+3.89*)	50.00 (+17.78*)
Appraisal Layer	Longer-context Tracking	53.33	25.00 (+3.33)	40.00 (+18.33*)
Appraisal Layer	Prioritization of Response	38.33	31.67 (0.00)	45.00 (+13.33*)
Appraisal Layer	Cultural Sensitivity	51.67	30.00 (+1.67)	41.67 (+10.00*)
Values Layer	Emotionally Intelligent Persistence	43.33	46.67 (+6.67)	41.67 (+1.67)
Values Layer	Ethical Safeguards	45.00	42.59 (+3.70)	42.59 (+3.70)
Values Layer	Intervention & Escalation	53.33	52.22 (+3.89)	51.67 (+3.33)
Values Layer	Long-term Emotional Memory	56.67	53.33 (+3.33)	66.67 (+16.67*)
Social Layer	Conflict Mediation	59.17	50.83 (+2.50)	65.00 (+16.67*)
Social Layer	Cultural Sensitivity	51.67	48.33 (+6.67)	50.83 (+10.00*)
Social Layer	Theory of Mind	59.17	25.00 (0.00)	33.33 (+8.33)

which already encode EI–adjacent heuristics (validate, ask a short question, avoid harm). Our contribution is complementary: starting from Base, we (a) add generic dialogue competence via UC_SFT and then (b) add targeted EI supervision via EICAP–Tune with only LoRA adapters ($\sim 0.8\%$ parameters). This stack produces systematic, statistically validated gains exactly where the benchmark demands EI reasoning, without full model retraining. The gap to Instruct on some sub–categories indicates room for scaling the EI–grounded corpus (more long–horizon cases, richer safety scenarios, harder negatives) and combining it with broader instruction data.

5 Conclusions

We introduced EICAP, a psychologically grounded taxonomy of emotional intelligence—and two complementary resources: EICAP-Bench, a multi-turn, one-vs. three evaluation suite *synthesized with GPT-4o-mini via taxonomy-guided prompting*, and `EICAP_SFT`, a taxonomy-aligned supervision corpus for adapter-based fine-tuning. Zero-shot benchmarking across open-source families showed that instruction-tuned backbones improve alignment with EI criteria, with Qwen-2.5-7B-Instruct leading (macro ≈ 0.78). However, generic conversational SFT on UltraChat yielded only localized and often non-significant lifts. Dialog fluency alone does not confer reliable EI reasoning.

By contrast, EI-grounded supervision via `EICAP_SFT`, applied with parameter-efficient LoRA ($\sim$0.8% params) on Qwen-Base, produced broad, statistically reliable gains precisely where the rubric demands causal, affective, and social reasoning—18/24 sub-categories improved significantly—raising absolute performance under a strict multi-turn criterion (overall macro ≈ 0.54 vs. ≈ 0.44 for UltraChat SFT). The taxonomy, benchmark, and tuning corpus form an operational bridge from psychological theory to reproducible prompting, evaluation, and targeted enhancement.

6 Limitations and Future Work

Limitations. Our study is *text-only* and does not model prosodic or visual cues that often carry decisive affective signals. Empathy and support needs are highly individualized (age, culture, personality), yet our evaluation abstracts away from stable user traits and longer-horizon personalization. While `EICAP_SFT` is disjoint from `EICAP-Bench`, it is still LLM-synthesized and may imprint generator biases; UltraChat, used as a conversational baseline, is comparatively emotion-sparse and thus offers weak EI supervision. The benchmark emphasizes one-vs. three forced choice and concise two-turn contexts; generalization to free-form generation, extended interactions, regulation over longer horizons, and policy-aware behaviors is underexplored. Language coverage is constrained to only English. Finally, the low-rank constraint of LoRA adapters can cap representational change relative to full-model tuning.

We will broaden human validation with inter-annotator agreement (IAA) and majority-vote checks; extend to multimodal EI (text + audio/vision) and longer-context regulation/escalation scenarios; expand multilingual evaluation (including Arabic morphological/transliteration stress tests); and combine `EICAP_SFT` with broader instruction data and preference optimization (e.g., RLHF/DPO) while mining harder, policy-grounded negatives. We also plan probe studies to disentangle calibration from conceptual EI learning, and to incorporate lightweight user modeling for controlled personalization. All artifacts will be released to support community reproduction and downstream benchmarking.

[2]

References

Aaron Grattafiori and 1 others. 2024a. The llama 3 herd of models. *Preprint*, arXiv:2407.21783. Checkpoint: **Llama-3-8B (Base)**.

Aaron Grattafiori and 1 others. 2024b. The llama 3 herd of models. *Preprint*, arXiv:2407.21783. Checkpoint: **Llama-3-8B-Instruct**.

Md Shad Akhtar, Alexandra Balahur, Rui Xia, Joonsuk Park, and 1 others. 2024. Emotion analysis in nlp: Trends, gaps and roadmap for future directions. *arXiv preprint arXiv:2403.01222*.

[2]This is a footnote.

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.

An Yang and 1 others. 2025a. Qwen 2.5 technical report. *Preprint*, arXiv:2412.15115. Checkpoint: **Qwen-2.5-7B (Base)**.

An Yang and 1 others. 2025b. Qwen 2.5 technical report. *Preprint*, arXiv:2412.15115. Checkpoint: **Qwen-2.5-7B-Instruct**.

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, and 12 others. 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.

Wenhan Dong, Yuemeng Zhao, Zhen Sun, Yule Liu, Zifan Peng, Jingyi Zheng, Zongmin Zhang, Ziyi Zhang, Jun Wu, Ruiming Wang, and 1 others. 2025. Humanizing llms: A survey of psychological measurements with tools, datasets, and human-agent applications. *arXiv preprint arXiv:2505.00049*.

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.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. The language model evaluation harness.

Gemma Team and 1 others. 2024a. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118. Checkpoint: **Gemma-2-9B (Base)**.

Gemma Team and 1 others. 2024b. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118. Checkpoint: **Gemma-2-9B-Instruct**.

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.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

HuggingFace H4. 2024. UltraChat 200k: A large-scale multi-turn dialogue dataset. https://huggingface.co/datasets/HuggingFaceH4/ultrachat_200k. Accessed: 2025-05-20.

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

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

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. 1991. *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, and 1 others. 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.

Jiao Ou, Junda Lu, Che Liu, Yihong Tang, Fuzheng Zhang, Di Zhang, and Kun Gai. 2023. Dialogbench: Evaluating llms as human-like dialogue systems. *arXiv preprint arXiv:2311.01677*.

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.

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. 2019. 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. 1978. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526.

Xingyu Qi, He Li, Linjie Li, and Zhenyu Wu. 2025. Emoassist: Emotional assistant for visual impairment community. *arXiv preprint arXiv:2502.09285*.

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. 1980. 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.

UltraChat Authors. 2023. UltraChat Dataset – MIT License. https://github.com/thunlp/UltraChat/blob/main/LICENSE. Accessed 2 Aug 2025.

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

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 and 1 others. 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*.

Weixiang Zhao, Zhuojun Li, Shilong Wang, Yang Wang, Yulin Hu, Yanyan Zhao, Chen Wei, and Bing Qin. 2024. Both matter: Enhancing the emotional intelligence of large language models without compromising the general intelligence. *arXiv preprint arXiv:2402.10073*.

Level 1	Level 2	Level 3	Leve
Foundation Layer: Receiving	Basic Emotion Labels	Happiness	
		Sadness	
		Fear	Anticip
			Immer
		Anger	
		Surprise	Positiv
			Negati
		Relief	
		Disgust	
	Confidence/Uncertainty Handling	High confidence in emotion detection	
		Low confidence in emotion detection	
		Overlapping cues	
		Probability-based (confidence calibration)	
		Explainability & Reasoning Traces	
	Contextual Lexical Cues	Emotion-specific words	
		Amplifiers	
		Negative lexical cues	
		Idiomatic phrases	
	Neutral vs. Emotional content	Neutral statements	
		Emotional statements	
		Blended content	
	Edge Cases	Ambiguous statements	
		Sarcasm or irony	Mild S
			Heavy
		Mixed or conflicting cues	
	Acknowledgement Types	Positive reinforcement	
		Sympathetic/supportive	
		Neutral/clarifying	
	Appropriate Politeness and Style		
Foundation Layer: Responding	Minimal Emotional Inference		
	Error Recovery		
Dimensional Layer	Valence Scale		
	Arousal Scale		
	Mixed Emotion Detection		
	Temporal or Transitional emotion	Improving, Worsening, Fluctuating	
Appraisal Layer	Cause/Consequence Inference	External causes	
		Internal causes	
		Implicit causes	
	Prioritization of response		
	Cultural sensitivity		
	Longer-context Tracking		
Values Layer	Emotionally Intelligent Persistence		
	Ethical Safeguards	Distress escalation detection	
	Intervention & Escalation	Tier 1: Mild Distress	
		Tier 2: Moderate/Serious Distress	
		Tier 3: Crisis (Sever emotional distress)	
	Long-term Emotional Memory		
Social Layer	Theory of Mind		
	Conflict Mediation	Personal misunderstandings	
		Deep-seated conflicts	
	Cultural Sensitivity	Regional/Generational differences	
		Polite navigation/taboo topics	
	Social Ethics and Sensitivity		

Figure 2: Taxonomy Hierarchy.

A Taxonomy Hierarchy

B Additional Results: EICAP–Tune Learning Curves

B.1 Zero-shot benchmark LLMs on EICAP

In (Fig. 4) we show the evaluation of `Base` and `Instruct` variants of `LLaMA-3`, `Gemma-9B`, and `Qwen-2.5` in a zero-shot, one-vs-three forced-choice setup: given the dialogue context and four candidate replies (1 preferred, 3 distractors), the model selects the single best option. Accuracy is macro-averaged across sub-categories; for two-turn dialogues we take the minimum across turns to enforce consistency. Instruction tuning reliably boosts alignment with EICAP's criteria, yielding absolute macro-average gains of +0.23 (`LLaMA-3`), +0.27 (`Gemma-9B`), and +0.43 (`Qwen-2.5-7B`) over `Base` (chance $= 0.25$).

Headline results: `Qwen-2.5-7B-Instruct` leads overall (macro $\approx$ 0.78), followed by `Gemma-9B-Instruct` ($\approx$ 0.71); `LLaMA-3-8B-Instruct` trails (≈ 0.36). By layer, the best models reach ≈ 0.89 on *Values* and ≈ 0.87 on *Social*, while *Foundation–Responding* remains harder (best $\approx$ 0.73); the toughest sub-areas are *Minimal Emotional Inference*, fine-grained *Valence*, and parts of *Cultural Sensitivity*.

Commonalities and differences across model families: All model families benefit similarly from instruction tuning on *Foundation–Receiving* and *Values*. Differences emerge in higher-order reasoning: `Qwen` excels at *Dimensional* judgments and uncertainty handling; `Gemma` is strongest on *Social* reasoning and *Appraisal* (causal analysis); `LLaMA-3` improves over `Base` but remains near/below 0.5 on many fine-grained EI tasks.

Base vs. Instruct–gains by layer and Instruction tuning gains: Layer-wise lifts are broad but family-specific: `LLaMA-3` $+0.22$–$+0.30$ (largest on *Values*, yet $\leq$ 0.47 by layer), `Gemma-9B` $+0.18$–$+0.46$ (biggest on *Social* to ≈ 0.87), and `Qwen-2.5-7B` $+0.35$–$+0.49$ (raising *Foundation–Receiving* ≈ 0.78, *Dimensional* ≈ 0.80, *Values* ≈ 0.89). Illustrative sub-category shifts include: `LLaMA-3` gains on *Arousal Scale* and *Ethical Safeguards* but stays weak on *Social Ethics/ToM*; `Gemma` jumps on *Cause/Consequence*, *Conflict Mediation*, *ToM*, and *Error Recovery*; `Qwen` attains perfect *Mixed-Emotion Detection*, high *Uncertainty Handling*, strong memory/safety, and a solid *Social* macro (≈ 0.78), while `Gemma` slightly leads *ToM*/ethics and `Qwen` is close on conflict mediation. Accordingly, because `Qwen-2.5-7B-Instruct` achieves the highest zero-shot macro-average on EICAP ($\approx$ 0.78) and the largest instruction-tuning gains, we adopt `Qwen-2.5-7B` (`Base` and `Instruct`) as our backbone for subsequent adapter fine-tuning.

Figure 5 plots layer-level accuracy across two SFT epochs. Epoch completion is on the x-axis (dashed at 1.0); stars mark the last checkpoint of each epoch.

C EI and LLM Limitations

We have conducted an extensive literature review for the limitations of the LLMs with regards to IE aspects. They include:

(i) Fundamental Limitations in EI Modeling—Lack

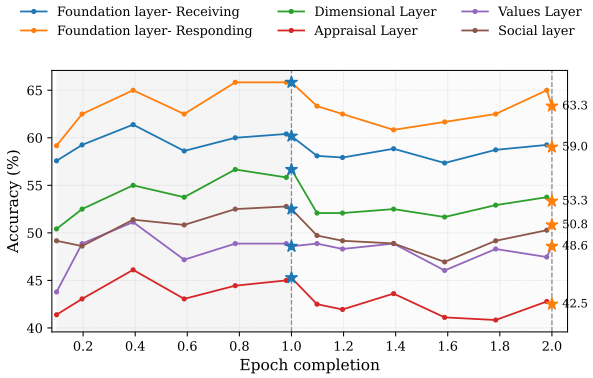

Figure 3: Layer-wise accuracy over two SFT epochs on EICAP. X-axis shows epoch completion (0.1–2.0); stars mark the last checkpoint of each epoch.

of Emotion Tracking Across Conversations. With current LLMs, it is not possible to create content that retain emotional context across conversational turns. This leads to inconsistent, repetitive, or contextually misaligned responses (Chen et al., 2023).
(ii) Inability to Recognize Mixed or Conflicting Emotions. LLM generated content typically does not support overlapping emotional states (e.g., relief mixed with anxiety) and subtle affective transitions (Alhuzali and Ananiadou, 2021).
(iii) Lack of Cultural and Social Sensitivity. Many models are trained with the assumption that emotional expressions are universal, disregarding cultural and contextual factors (Dudy et al., 2024).
(iv) Lack of Emotional Causality Awareness. LLM based systems are not designed to detect underlying causes of emotional expressions and thus generate shallow and contextually inappropriate responses (Cheng et al., 2022).
(v) Failure in Emotion-Driven Decision-Making. LLMs used in decision making scenarios lack the ability to include emotional awareness into decision-making and support context-sensitive emotional adaptation (Mozikov et al., 2024).
(vi) Inability to Handle Emotionally Complex Situations. LLMs are frequently not effective in multi-agent emotional interactions, particularly in group conversations where conflicting emotions must be recognized and addressed appropriately (Poria et al., 2019).
(vii) Preference Bias in Emotional Support. LLM based systems often overuse specific response strategies (e.g., excessive reassurance over problem-solving) and are not set up to adapt dynamically to the user's emotional needs (Akhtar et al., 2024).
(viii) Ethical Emotional Response Generation. LLMs output cannot be easily controlled. Even with guardrails, there is a risk of producing misleading, insensitive, or manipulative responses in emotionally charged interactions (Mohammad, 2022).
(ix) Limited General-Purpose Multi-Turn EI Benchmarks. While some benchmarks can be used to evaluate multi-turn emotional tracking in specific domains (e.g., counseling-based ESC datasets), there is no standardized, multi-domain benchmark to capture and support 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)

D Evaluation Statistics Interpretation

Interpretation note (Table 1 vs. Fig. 5). The table quantifies improvement over Base (Δ) under a strict minimum-of-two-turns criterion and is computed at the sub-category level (macro-averaged across 24 sub-categories). By contrast, Fig. 5 plots absolute accuracies and typically aggregates turns by a mean-of-turns criterion, and reports layer-level macro averages. Because these views differ in (i) target quantity (Δ vs. absolute), (ii) turn aggregation (min vs. mean), (iii) granularity (sub-category vs. layer), and (iv) layer cardinality (unequal numbers of sub-categories per layer, e.g., Foundation–Receiving 7 vs. Responding 2), a layer can exhibit a large positive Δ in Table 1 yet remain below other layers in Fig. 5. In short, the table answers how much EI-specific tuning helps relative to Base, whereas Fig. 5 shows how high the tuned model's absolute accuracy sits by layer.

E EICAP Taxonomy Overview

EICAP: A Comprehensive Emotional Intelligence Taxonomy (Figure A 6) is structured across five interconnected layers, Foundational, Dimensional, Appraisal, Value, and Social. Each layer encompasses a broad spectrum of emotional categories, sub-categories, and sub-sub-categories, designed to capture nuanced emotional responses within human communication and behavior.

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	

Figure 4: **Model Performance by Layer and Sub-Category:** Minimum and macro-mean accuracies are reported for six models—comparing base and instruct fine-tuned versions—across the various layers and subcategories of EICAP. The final column presents the macro-average accuracy for the best-performing models.

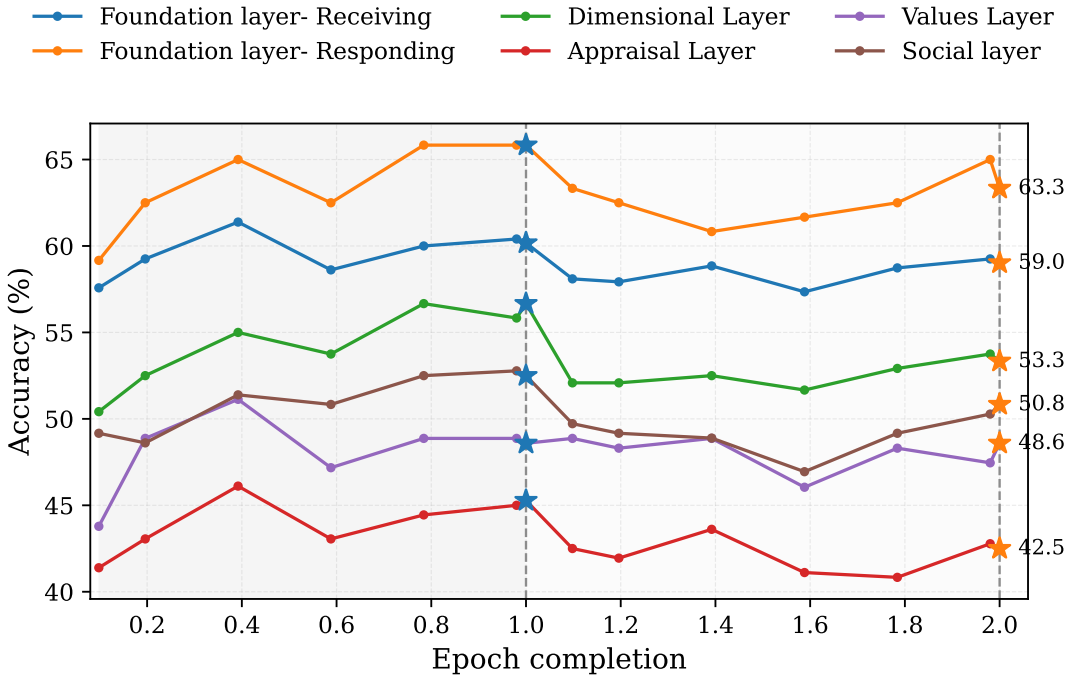

Figure 5: Layer-wise accuracy over two SFT epochs on EICAP. X-axis shows epoch completion (0.1–2.0); stars mark the last checkpoint of each epoch.

ID	Level 1	Level 2	Level 3	Level 4
1	Foundation Layer: Receiving	Basic Emotion Labels	Happiness	
2			Sadness	
3			Fear	Anticipatory Fear
4				Immediate Fear
5			Anger	
6			**Surprise**	Positive Surprise
7				Negative surprise
8			Relief	
9			Disgust	
10		Confidence/Uncertainty Handling	High confidence in emotion detection	
11			Low confidence in emotion detection	
12			Overlapping cues	
13			Probability-based (confidence calibration)	
14			Explainability & Reasoning Traces	
15		Contextual Lexical Cues	Emotion-specific words	
16			Amplifiers	
17			Negative lexical cues	
18			Idiomatic phrases	
19		Neutral vs. Emotional content	Neutral statements	
20			Emotional statements	
21			Blended content	
22		Edge Cases	Ambiguous statements	
23			Sarcasm or irony	Mild Sarcasm
24				Heavy sarcasm
25			Mixed or conflicting cues	
26		Acknowledgement Types	Positive reinforcement	
27			Sympathetic/supportive	
28			Neutral/clarifying	
29		Appropriate Politeness and Style		
30	Foundation Layer: Responding	Minimal Emotional Inference		
31		Error Recovery		
32	Dimensional Layer	Valence Scale		
33		Arousal Scale		
34		Mixed Emotion Detection		
35		Temporal or Transitional emotion	Improving, Worsening, Fluctuating	
36	Appraisal Layer	Cause/Consequence Inference	External causes	
37			Internal causes	
38			Implicit causes	
39		Prioritization of response		
40		Cultural sensitivity		
41		Longer-context Tracking		
42	Values Layer	Emotionally Intelligent Persistence		
43		Ethical Safeguards	Distress escalation detection	
44		Intervention & Escalation	Tier 1: Mild Distress	
45			Tier 2: Moderate/Serious Distress	
46			Tier 3: Crisis (Sever emotional distress)	
47		Long-term Emotional Memory		
48	Social Layer	Theory of Mind		
49		Conflict Mediation	Personal misunderstandings	
50			Deep-seated conflicts	
51		Cultural Sensitivity	Regional/Generational differences	
52			Polite navigation/taboo topics	
53		Social Ethics and Sensitivity		

Figure 6: Full overview of the EICAP taxonomy hierarchy.