# CAT-CoT: Instruction-Tuning LLMs via Cognitive Appraisal Theory-Inspired Chain-of-Thought Reasoning to Enhance Emotional Expressivity

by

Ashfaq Ahmad Saad
21301665
Asif Ahnaf Chowdhury
21301510
Fardeen Alam
21301116
Kazi Amzad Abid
21301750
A N M Jubair Tanvir
21301524

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
Brac University
June 2025.

# Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

**Student's Full Name & Signature:**

| | | |
|---|---|---|
| Ashfaq Ahmad Saad<br>21301665 | Asif Ahnaf Chowdhury<br>21301510 | Fardeen Alam<br>21301116 |

| | |
|---|---|
| Kazi Amzad Abid<br>21301750 | A N M Jubair Tanvir<br>21301524 |

# Approval

The thesis titled "CAT-CoT: Instruction-Tuning LLMs via Cognitive Appraisal Theory-Inspired Chain-of-Thought Reasoning to Enhance Emotional Expressivity" submitted by

1. Ashfaq Ahmad Saad (21301665)
2. Asif Ahnaf Chowdhury (21301510)
3. Fardeen Alam (21301116)
4. Kazi Amzad Abid (21301750)
5. A N M Jubair Tanvir (21301524)

Of Spring, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on June 20, 2025.

**Examining Committee:**

Supervisor:
(Member)

Dr. Jannatun Noor Mukta

Guest Associate Professor,
Department of Computer Science and Engineering,
Brac University.
Director, Data Science Program,
United International University.

Program Coordinator:
(Member)

Md. Golam Rabiul Alam, PhD

Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi

Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

# Ethics Statement

This statement affirms that the research presented in this paper upholds the principles of human rights, dignity, and well-being. The integrity and sincerity of the contributors have been preserved throughout the study, which draws upon thoroughly cited and credible sources. The data collection process was conducted without any form of bias or discrimination, and the findings have not been altered to serve any prejudiced agenda. We affirm that this work respects intellectual property and aspires to make a meaningful contribution to the long-term betterment of society.

# Abstract

Despite their conversational proficiency, large language models (LLMs) have a limited potential for true emotional comprehension, generally relying on surface-level indicators such as emotion keywords or sentiment heuristics. This gap becomes critical in domains like mental health support and human-robot interaction, where emotional misalignment can lead to user harm or reduced trust. To address this, we introduce CAT-CoT, a novel CoT framework that integrates Cognitive Appraisal Theory (CAT) with LLM via instruction tuning on a dataset named Appraisal-CoT. We created the Appraisal-CoT dataset, a synthetic corpus of 4,641 empathy-driven dialogues based on the EmpatheticDialogues benchmark and annotated with step-wise appraisal reasoning using GPT-4o-mini. Empirical evaluation demonstrates that our instruction-tuned models significantly outperform the untuned baselines on EmoBench, achieving the highest gains of 6.5% in Emotion Understanding (EU) and 4.0% in Emotion Application (EA) with Qwen3-4B-ACoT, compared to the base model. Also, the small models score matching or exceeding the performance of larger models (e.g., Qwen3 4B–ACoT rivals 8B variants). These findings demonstrate that employing psychologically grounded reasoning techniques can significantly enhance emotional expressivity in LLMs, marking a step toward safer, more empathetic AI systems.

**Keywords:** Cognitive Appraisal Theory, Instruction-Tuned LLMs, Emotion Generation, Empathetic AI, Emotional Intelligence (EI), Real-time Dialogue Systems.

# Acknowledgement

We thank and praise the Almighty Allah, who has granted us His blessings, guidance, and strength, allowing us to finalize our study on schedule. We are very grateful to our thesis supervisor, Dr. Jannatun Noor Mukta, who provided us with invaluable assistance, offering valuable suggestions, skills, and constant encouragement throughout every stage of the thesis journey. It is also our privilege to be guided and assisted by Dr. A. B. M. Alim Al Islam, Professor at BUET, whose inputs significantly enhanced our work. We would also like to thank Dr. Tanjir Rashid Soron for his insight into the psychological aspects of our research, as well as Dr. Abdullah Al Mamun and Abida Sultana for their contributions and expertise. We also greatly appreciate the hardware support given by BRAC University, which was crucial to obtain the results of our study. Lastly, we would like to extend our sincere appreciation to our parents, whose continuous support, love, and prayers have served as a supportive force and inspiration in this undertaking.

# Table of Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 Problem Statement

Large language models (LLMs), such as GPT-4, have demonstrated extraordinary conversational capabilities. However, their ability to capture emotional understanding and empathetic expression remains limited. Recent research suggests that although LLMs can provide answers that appear sympathetic when influenced by examples of context or fine-tuning, these glimpses of sympathy often conceal the absence of genuine emotional thinking [39]. In practice, LLM reactions depend primarily on the emotional phrases of pattern matching rather than modeling the basic cognitive processes that generate emotions [43]. LLMs cannot adapt to human emotional behavior and do not effectively combine compliance with persistent emotions [44]. LLM demonstrates some efficiency in emotional discourse; however, a comprehensive standard remains absent to assess their emotional intelligence in handling more complex feelings [39]. These results indicate an inequality between LLM emotion generation and human emotional feeling: the existing models can verify sympathy but lack real understanding or relevant emotional meaning.

A primary cause of this gap is the shallow quality of the current emotion modeling. Most models rely on clear emotional indicators (e.g., emotion symbols, emotion keywords) rather than standard, deeper reasoning. Many emotion benchmarks are primarily dependent on keyword-based emotional identity, which overlooks the necessary relevant factors required for a comprehensive understanding of emotions [61]. Many sympathetic conversation models only focus on superficial emotional data and ignore the ground causes of emotions and cognitive elements, resulting in answers that appear sympathetic but lack authentic emotional deficiency [42]. In practice, this means that sympathetic vocabulary or emotional adjectives in a large language model can be incorporated into its production, even if it does not analyze the underlying causes of the speaker's feelings (for example, what goals were thwarted or who is blamed). As a result, the current feature often captures clear signals and repeats the complex appraisal processes used by people to evaluate and respond to others.

In addition, modern LLM training structure rarely contains psychologically based emotional theories. Cognitive appraisal theory (CAT), a fundamental aspect of emotional psychology, is rarely incorporated into language models. Cognitive appraisal

theory suggests that emotions stem from a person's evaluation of factors such as goal relevance, responsibility, justice, and control [40]. These evaluations identify some emotions (e.g., anger emerges when one's goal is interrupted and the blame is assigned to another person). However, traditional NLP (natural language processing) models ignore this structure. LLM has difficulty correlating status and evaluation with special feelings and emphasizes the need for psychological theories in the teaching and evaluation of LLM on emotional arguments [70]. In other words, without a clear awareness of the appraisal dimensions, LLM lacks the cognitive structure required for authentic emotional understanding. Recent initiatives have begun to address this deficiency (for example, the "third-individual assessment agent" that utilizes cats to reduce emotions in discourse). However, these features are still limited [71]. Currently, no extensive LLM instruction or fine-tuning process systematically incorporates appraisal theory or other psychological models of emotion, resulting in a fundamental theory-practice gap.

This theoretical gap is exacerbated by a lack of training data that provides examples of cognitive-emotional reasoning. Most interactive spirit datasets (e.g., EmpatheticDialogues, IEMOCAP) provide only an emotional label for pronunciation without explaining the underlying causes of these feelings. "Existing ERC [Bhavnna recognition in conversation] data sets only offer the emotional pronunciation, devoid of widespread comments needed for emotional logic [71]." In short, trained models on this data learn to link references to the emotional labels and mainly neglect arbitrary appraisal steps. A limited number of special resources partially address this difference. For example one study compares a theory based on the theory-of-mind style dataset grounded in CAT (evaluating both context-to-emotion and emotion-to-context reasoning), while another study collects human assessments on status signals (the "EmotionBench" dataset) to examine LLM emotional alignment [44], [70]. However, they work forcibly as evaluation standards rather than extensive training corps and thus do not provide directly monitored teaching for evaluated response generation. Finally, there is no available interactive corpus that directly captures cognitive evaluation and conclusions that reduce emotions.

Overall, these observations underscore the importance of our problem. Current LLMs may appear emotionally expressive; however, they primarily operate based on subconscious success and superficial indicators. They lack a specific model for emotional evaluation and are trained on data that excludes the underlying cause of emotions. A recent study validates these deficiencies: While LLMs may respond well to fundamental emotional triggers, they cannot effectively normalize the "lack of adaptation with human emotional behavior" and analog conditions [44]. In other words, state-of-the-art LLMs show a superficial level of emotional intelligence. Improving this deficit requires the procurement or development of a dataset that promotes cognitive-emotion theories, especially CAT, in the architecture of large language models (LLM). With emphasis on these deficiencies and obstacles, we provide a basis for research focused on increasing LLM-intensive, psychologically informed emotional expression.

**Key Gaps:**

- Current models exhibit shallow empathy. Their outputs may sound sympa-

thetic but are generated through pattern matching rather than genuine emotional reasoning [42].

- Emotion modeling often relies on surface cues (keywords or sentiment heuristics), neglecting underlying cognitive factors [42].
- Large-scale psychological theories (e.g., CAT) are rarely integrated into training or prompting strategies, resulting in LLMs lacking structured appraisal-based reasoning [40].
- There is a scarcity of datasets annotated with appraisal or inference information; existing dialogue corpora only label emotions, not causal appraisals [71].

These issues highlight a pressing research problem: enhancing LLM emotional expressivity by grounding generation in cognitive appraisal processes and a richer theory of emotions rather than in disconnected surface keywords.

## 1.2 Motivation

Large language models are increasingly used in areas where it is necessary to understand and express human emotions accurately. In mental health support, LLM-operated chatbots (e.g., Woebot, Youper, ChatGPT-based services) promise scalable, on-demand help for those experiencing anxiety or depression. However, the main audit indicates that these systems are often limited: even LLMs are carefully and correctly set for emotional support, which is relatively backward compared to people [60]. Users and professionals have been disappointed by the LLM solutions for emergencies, as they often provide superficial responses. A user survey from CHI2024 has shown that LLM chatbots typically understand the complexities of LGBTQ+ customer problems and often provide platitudes or a superficial sense of compassion [50]. Generic, reference-sensitive reactions are not only unproductive but can also pose risks under high day conditions.

- **Generic, templated replies:** Chatbots often respond with safe but vacuous phrases ("I'm sorry that happened") that sound empathetic but offer no concrete guidance.
- **Context neglect:** Without an accurate appraisal, LLMs miss subtle user cues. They may overlook trauma or identity factors (as in the LGBTQ+ study), leading to advice that is misaligned with the user's situation [50].
- **Biases and inequity:** Evaluations show LLM empathy is not uniform; for example, GPT-4's replies to Black patients scored 2 to 15% lower in empathy than to White patients [50]. This demographic gap in emotional support highlights that current LLMs do not reliably align with all users' experiences.
- **Safety failures:** LLMs have already produced harmful suggestions (e.g., giving unsafe dieting tips to patients with eating disorders. In mental health,

under or over-reacting to suicide risk and dispensing shallow coping strategies can worsen outcomes. In summary, practical failures, from bland empathy to dangerous hallucinations, underscore the urgent need for more robust emotional reasoning in LLMs.

In the field of social robotics, the importance of emotional competence is equally paramount. Fellow and supportive robots, such as the therapeutic seal PARO, eldercare service robots, and social skill trainers for children, are designed to meet the emotional needs of individuals [35]. Users often attribute human-like properties to these robots and depend on them for companionship; therefore, any clear rebellion or malfunction may result in significant feelings of "deception" or despair [35]. Early efforts to integrate LLM with robotics show these dangers. A case study showed that a ChatGPT-powered Pepper Robot unlawfully assured a user that it would handle the necessary drug memory, a misleading impression of care. Such deception is acceptable in conditioned chatbots, but it is particularly concerning in healthcare robots, which can erode user confidence.

These operational deficiencies indicate the underlying theoretical deficiencies and lack of approach. Key deficiencies LLMs have with regard to this:

- **Shallow emotion cues:** Most current systems address emotions superficially. NLP involves highlighting sympathy when aligning the emotional mark rather than analyzing the underlying causes of the user's feelings [31]. For example, by identifying mood or expression and producing a reaction corresponding to Valens [31]. This overlooks individual and cultural variation in emotional expression.
- **Ignored appraisal processes:** Cognitive psychologists have consistently shown that human feelings arise from evaluating a situation [40]. By bypassing cognitive appraisal, LLMs lack insight into the meaning behind a user's words. They treat emotions as labels rather than the result of judgments (e.g., of goal conduciveness or fairness) that drive feelings [70].
- **Inconsistent alignment:** Without a theory of emotion, responses can be erratic and unpredictable. EmotionBench indicates that LLMs show inconsistent reactions to analog stimuli and are unable to forge connections between conditions similar to those of humans [38], [44].

In summary, beyond the domain of mental health and social robotics, modern LLMs depend on superficial spiritual modeling (emotional analysis, pre-written sympathy) that often proves to be inadequate. These deficiencies: vague, obvious help, prejudice or dangerous guidance, misleading, and occasional sympathy indicate a fundamental disagreement with human emotional thinking. This difference requires the integration of intensive psychological ideas, such as evaluation principles, in LLM training, enabling future systems to match their emotional reactions with human needs more accurately.

### 1.2.1 Overview of CAT-CoT

To cover superficial sympathy with authentic emotional argument, we introduce CAT-CoT (Cognitive Appraisal Theory-Chain-of-Thought). This instruction-tuning framework integrates psychologically informed appraisal processes directly into the training of large language models. In CAT-CoT, each conversation bend expands with a structured "appraisal chain" (primary appraisal → secondary appraisal → reappraisal) before the model that produces the final response. This process requires models:

1. Evaluate goal relevance, congruence, and ego involvement.
2. Assess coping potential, future expectancy, and perceived control.
3. Optionally reframe the situation; mirroring the automatic yet structured reasoning humans employ when examining their own and others' emotions.

Synthetic appraisal through CoT data, derived from authentic dialogues (such as EmpatheticDialogues) and informed by cognitive appraisal theory, equipped CAT-CoT LLM with its cognitive-emotional argument system. **Figure 1.1** showcases our models (Instruction tuned ACoT model) response versus the Baseline model response for the same user input.

### 1.2.2 Contributions

1. **CAT-Cot Framework:** We provide the first instructional setting approach that systematically incorporates cognitive Appraisal Theory in the large language model chain-of-thought region and offers models for performing sequential emotional evaluation before generation.
2. **Appraisal-CoT Dataset:** We have introduced a synthetic dataset of 4,641 single-turn conversations between a user and an assistant, which follows a noble Appraisal chain of thought structure.
3. **Acot Models:** Our instruction-tuned small models, ranging from 1B to 4B, achieved enhanced emotional intelligence. Also, 4B parameter models achieved optimal performance gains from Appraisal CoT instruction tuning, with some models (Qwen3 4B) matching the performance of their larger 8B variants.


Figure 1.1: Comparison of output responses generated by Qwen3 1.7B (baseline) and Qwen3 1.7B acot (instruction-tuned) models, illustrating the impact of Appraisal CoT on emotional reasoning and expression.

## 1.3 Thesis structure

**Chapter 1: Introduction**
This chapter establishes the central problem, the current inability of large language models (LLMs) to engage in deep emotional reasoning. This shortcoming is particularly critical in domains requiring emotional sensitivity, such as mental health and social robotics. The chapter introduces our novel framework, CAT-CoT, which leverages Cognitive Appraisal Theory (CAT) to enhance emotional expressivity in large language models (LLMs). It outlines the core contributions, including the development of a new framework, the construction of a specialized dataset, and the design of multifaceted evaluation methodologies. The chapter concludes with a roadmap that describes the thesis's structure.

**Chapter 2: Literature Review**
This chapter situates our work within three key research streams. First, we survey recent advancements in emotion generation and empathy modeling for LLMs, covering prompt-based methods, graph-based emotion decoders like E-CORE, and instruction-tuning approaches that leverage direct preference learning. Next, we review psychological theories of emotion, with emphasis on appraisal frameworks and their computational instantiations. Finally, we examine chain-of-thought (CoT) reasoning techniques in NLP and assess how existing CoT methods fail to incorporate affective cognition. Together, these surveys highlight the gap our CAT-CoT framework addresses.

**Chapter 3: Objectives**
Here, we articulate the research gap: Despite progress, no extant LLM training paradigm fully integrates a structured appraisal process into generation. We define three formal objectives:

1. **Theoretical Integration:** To operationalize Cognitive Appraisal Theory (CAT) as a parameterized reasoning schema for LLMs.

2. **Enhancing emotional intelligence of LLMs:** Instruction tuning LLMs to build enhanced emotionally intelligent LLM.

3. **Data Synthesis:** To construct a large-scale, CAT-grounded dataset in CoT format.

4. **Empirical Validation:** Conduct thorough empirical evaluations using the Emobench benchmark, along with automated and human-centered metrics.

These objectives guide our methodological design and evaluation criteria.

**Chapter 4: Methodology**
This chapter details the translation of Cognitive Appraisal Theory into a computational framework. This involves designing a psychologically grounded pipeline and developing an Appraisal Chain-of-Thought (CoT) algorithm that incorporates primary appraisal, secondary appraisal, and reappraisal mechanisms. The dataset construction process begins with the selection of EmpatheticDialogues as a seed corpus, followed by synthetic augmentation using GPT-4o-mini to generate appraisal chains. The chapter also describes the controlled rewriting and prompting strategies employed. Evaluation is multifaceted, combining automated metrics like BLEU, ROUGE, BERTScore, and Distinct-N with human assessments by domain experts. Inter-annotator agreement (IAA) is used for validation. The methodology concludes with instruction-tuning using Unsloth and details on the model configuration and training.

**Chapter 5: Dataset Description**
In this chapter, we present the Appraisal-CoT dataset. We first describe the EmpatheticDialogues seed corpus, its conversational structure, and its relevance to tasks involving emotional support. Next, we outline our preprocessing pipeline, which involves filtering to 4,641 conversations, normalizing the text, and ensuring diversity across emotion types. We then explicate the augmentation procedure, showing example prompts and generated appraisal CoT. Finally, we provide dataset statistics (turn counts and appraisal label distributions) and assess data quality via inter-annotator agreement scores and automated scores (BLEU, Distinct, ROUGE, and BERTScore).

**Chapter 6: Experimental Analysis**
This chapter reports our quantitative and qualitative evaluations. We first detail the training setups for A-CoT models, including the computing environment and hyperparameters. Performance is benchmarked using EmoBench, with a focus on emotional intelligence measures. Results from automated metrics, such as BLEU, Distinct, ROUGE, and BERTScore, are analyzed alongside findings from human evaluations that assess emotional understanding, empathy, appraisal reasoning, and overall helpfulness.

**Chapter 7: Discussion & Findings**
Here, we interpret the empirical results, illustrating how CAT-CoT improves emotional alignment and depth compared to baseline models. It features comparative performance analysis and includes case studies that highlight both the strengths and limitations of the model. The chapter also discusses the increased interpretability enabled by explicit appraisal reasoning steps.

**Chapter 8: Future Work & Limitations**
This chapter critically examines the limitations of our approach, including reliance on synthetic data and potential cultural biases in appraisal parameterization. We propose future extensions: (a) merging more psychological theories and (b) building a multi-turn dataset. We also discuss the need for larger-scale human studies to validate long-term user outcomes in mental health.

**Chapter 9: Conclusion**
In the final chapter, we summarize our contributions: the CAT-CoT framework, Appraisal-CoT dataset, and comprehensive evaluations, and reflect on their broader impact. We argue that embedding psychologically grounded reasoning into large language models (LLMs) marks a significant step toward truly empathetic AI. We conclude by presenting a vision for next-generation, emotionally intelligent systems that can understand, reason about, and respond to human emotions with cognitive fidelity.

# Chapter 2

# Literature Review

## 2.1 Emotion Generation and Empathy in LLMs

Modern research in natural language processing explores various methods to enable large language models to communicate understanding and empathy. A common theme is to guide LLMs via specialized prompts or fine-tuning, often informed by psychological theories. ChatGPT achieved better results than prior models on EmpatheticDialogues by just employing guidance similar to instructions [39]. By using a suitable prompt, LLMs can respond with greater empathy. A group of researchers developed an approach that ensures the background knowledge presented in the prompt is closely related to the question, thereby maintaining clear and precise answers [36]. Their approach performed better than standard ones in machine tests as well as in responses from people judging their empathy. Emotions are connected in their research through a graph, and the system then utilizes a new type of decoder to understand these relationships [37]. Models based on E-CORE react with more realistic emotions and provide better answers than models that only deal with one emotion. Such investigations demonstrate that increasing knowledge and fine-tuning how emotions are understood can help LLMs exhibit emotions more consistently.

### 2.1.1 Instruction Tuning and Preference Learning for Empathy

Experts have recently attempted to modify large language models to exhibit empathy by training them using instruction tuning and preference-based learning. Using instruction tuning or special datasets in recent research helps LLMs be more empathetic. Sotolar compiled a dataset centered on psychological research, in which each conversation offers a response that can be seen as better or worse than the one preceding it [53]. The better replies were selected by training the LLM with a technique named direct preference optimization. Thanks to this process, the model responded more effectively to questions about emotions while maintaining its overall knowledge, as indicated by diff-Epitome and BERTScore scores. Similarly, a training set was prepared specifically for counseling purposes [47]. Carefully written counseling prompts and advice from professional psychologists were given to LLaMA and GPT models. Related to empathy, relevance, support, and crisis management, the expert judges found that the trained models performed significantly better than those that

had received no training. The primary focus in both studies was ensuring the model adhered to clear guidelines to understand and manage emotions. We train our LLM by supplying psychology-based cognitive prompts, just like the other methods that use theories from psychology.

### 2.1.2 Chain-of-Thought and Psychological Grounding

A good approach is to show language models how to reason about emotions gradually, using well-structured examples from psychology. It was suggested Emotional Chain-of-Thought (ECoT) as a simple technique that tries to replicate Goleman's theory of emotional intelligence [48]. ECoT prompts us to notice our emotions, figure out the reason behind them, and decide how to respond, following this order: "First identify, then understand, and then act." It provides the model a chance to consider their emotions before responding. In the tests, ECoT was better at creating emotional captions and responses. The writers also created a combined Emotional Generation Score, powered by GPT-3.5, to assess the responses. Some other studies employed similar methods to probe cognitive processes related to emotions. For instance, experts developed a Cognition Chain (Stimulus–Evaluation–Reaction–Stress) that enables the model first to recognize the stress trigger and then label it, making the process more comprehensible [54]. The authors proposed the ECR-Chain (stimulus $\rightarrow$ appraisal $\rightarrow$ emotion) to aid models in determining the reasons behind certain emotions. The use of these steps enhanced the models' ability to detect emotions and also made their choices clearer. In brief, working with mental and emotional processes, such as our appraisal-chain method, enables models to respond with deeper and more human-like responses.

### 2.1.3 Commonsense and Perspective Modeling

There are strategies designed to expand the model's awareness of the situation and the feelings of the person speaking, thereby enhancing empathy. The Sibyl framework enabled the LLM to employ "visionary commonsense inference," which involves imagining the next steps or the underlying thoughts behind what the speaker says. Such insights generated form a mental chain within the model that helps it react more thoughtfully to future changes [68]. The recommendations were enhanced by incorporating real-life examples, which helped the system better understand user needs. Similarly, researchers compiled an extensive collection of therapy talks and provided emotional appraisals for each part of the conversation, as reported by both the therapist and the patient [55]. It is essential to align these ideas, they say, to effectively display empathy. Even though they do not offer solutions, they make it clear that considering emotions through appraisal theory is valuable in such circumstances. Likewise, our method incorporates appraisal as part of the model-building process. **Table 2.1** showcases Summary of recent methods and the comparison which includes baseline models, prompting techniques, graph-based approaches, and instruction-tuned frameworks, including our proposed method using Cognitive Appraisal CoT and the dataset.

| Study | Instruction Strategy | Dataset(s) | Evaluation Metrics | Key Findings |
|---|---|---|---|---|
| [25] | Fine-tuning on emotional dialogue | EmpatheticDialogues | BLEU, Distinct, BERTScore | Basic emotional quality; lacks deeper empathy reasoning |
| [39] | Zero/few-shot prompting with instructions | EmpatheticDialogues | Human empathy ratings | Chat-based LLMs surpass fine-tuned models in empathy |
| [36] | Adaptive commonsense selection in prompts | EmpatheticDialogues | Automatic scores, human ratings | Outperforms baselines in empathy and coherence |
| [37] | Emotion graph + correlation-aware decoder (E-CORE) | EmpatheticDialogues | Emotion prediction, response quality | Handles co-occurring emotions better; improves prediction |
| [53] | Instruction tuning via preference optimization (EmPO) | EmpatheticDialogues | diff-Epitome, BERTScore | Empathy improves without hurting general quality |
| [47] | Counseling-specific instruction tuning with expert feedback | Counseling dataset | Human ratings: empathy, support, crisis handling | Tuned models show better emotional and counseling ability |
| [48] | Emotional Chain-of-Thought (ECoT) | Emotion captioning/ dialogue tasks | Emotional Generation Score, human feedback | Better emotional articulation; more appropriate responses |
| [54] | Cognition Chain: Stimulus → Evaluation → Reaction → Stress | Stress detection | Label accuracy, explanation quality | Improves detection with cognitive step-wise reasoning |
| [44] | ECR-Chain: Stimulus → Appraisal → Emotion | Emotion cause datasets | Cause detection, explainability | Boosts performance and makes reasoning visible |
| [55] | Appraisal annotations in therapy dialogues (non-generative model) | Annotated therapy dataset | Alignment of appraisals between therapist and patient | Shows value of appraisal theory for empathy in dialogue |
| Our Method | Cognitive CoT-based instruction tuning | A noble Appraisal CoT Dataset | Emobench benchmark dataset, Automatic scores (BLEU, Distinct, BERTScore, ROUGE), Domain expert Psychiatrist human ratings | Performing well on the Benchmark Dataset, Performing well on Automated evaluations, Higher emotional expressiveness and reasoning than untuned models |

Table 2.1: Summary of recent methods in emotionally intelligent dialogue generation, highlighting instruction strategies, datasets, evaluation metrics, and key findings.

## 2.2 Psychological theories for emotion generation

### 2.2.1 The James-Lange theory

The James-Lange theory posits that an individual experiences an emotion in response to observing bodily changes [2]. The stimulus triggers specific physiological effects (e.g., the heart rate increases, the body starts to shake), and experiencing these bodily phenomena is the emotion. The strength of this theory lies in its ability to reverse the obvious logic. It explains why there is no reason to say, "I tremble because I am afraid," but instead, it must be said that because I tremble, then I am afraid. James and Lange employed introspective reasoning and observation. Initial experiments attempted to elicit emotions (e.g., fear, anger) through provocative stimuli, assessing physiological responses such as pulse rate and skin conductance. These attempts were made in search of individual visceral patterns that correspond

to emotional reports. The theory was critiqued by Cannon and Bard, who noted that physiological responses are too slow or not specific enough to respond to a single particular emotion [1].

### 2.2.2 Cannon–Bard Theory

Cannon and Bard advanced the idea that the brain processes emotional stimuli, simultaneously eliciting emotional experience and physiological arousal. This simultaneous activation refutes the conclusion that emotions are a result of bodily-level changes. They focused on subcortical brain structures and noted that the thalamus, in particular, plays a crucial role in emotion processing. Cannon used animal experiments to cut sensory nerve endings, thereby eliminating visceral feedback. Even after this, animals were found to exhibit emotional behaviors, which implies that emotions originate centrally, not peripherally. Another finding he made was the resemblance between various emotional and mental states in terms of their physiological reactions, which also reinforced the argument that bodily signals are not specific in their reactions to emotions [1].

### 2.2.3 Schachter–Singer (Two-Factor) Theory

According to the two-factor theory of emotions, as described by Schachter and Singer, emotions consist of two primary elements: cognitive identity and physical arousal [3]. Arousal is viewed as nonspecific, and it is put into context to evoke emotions. The novelty of this theory lay in its focus on misattribution: individuals could experience various emotions depending on how they interpreted the same physiological condition. In their pioneering research, participants were injected with epinephrine. The participants changed their emotional states based on how they were informed, misinformed, or uninformed about the effects when placed with euphoric or angry confederates. This argued in favor of the construct that cognition informs emotional apprehension of arousal.

From a comparative perspective, the Cognitive Appraisal Theory differs from the first three in its emphasized meaning. According to cognitive appraisal theory, cognition is not just a means but rather the decision that triggers an emotion. The thing that determines the emotion we are feeling is our goals and beliefs because we decide what emotions we are experiencing. It also depends on circumstances and personal meaning, which lies at the core. According to Lazarus, the rich range of different emotional states is attributed to various appraisals [9]. The theory of cognitive appraisal involves cognitive manipulations to alter interpretations, which consist of various techniques such as instructions, reappraisal tasks, vignettes, or diverse narrative contexts (e.g., films displayed under differing cover stories). Lazarus demonstrate that merely construing the situation differently (e.g., as less threatening) leads to a change in emotional outcome, which the James-Lange or Cannon-Bard theories cannot easily explain [4].

## 2.3 Cognitive Appraisal Theory in detail

### 2.3.1 Cognitive Appraisal Theory (CAT)

Cognitive appraisal theory is a psychological theory that explains how we perceive and evaluate a specific event rather than the event itself. According to Ellsworth, appraisal theories of emotion suggest that emotions are composed of systematic processes that evaluate one's relationship to the environment, as well as the corresponding behavioral preferences and physiological reactions [8]. For instance, someone will feel depressed if they have lost something they love, and it is believed that uncontrollable circumstances caused the loss. An individual evaluates the personal significance of an event even before the emotional reaction. The appraisal process determines the emotion and its intensity [9]. Ellsworth universalizes her concept, suggesting that, in general, similar types of appraisals will elicit similar emotions across cultures.

As emotions arise from evaluating a particular situation, Cognitive Appraisal Theory has set itself apart from other psychological theories [5], [9], [14], [19]. In terms of emotion and adaptation, emotions can be considered a cognitive system. According to Lazarus, emotions serve as mechanisms for understanding conditions of the universe that are relevant to individual well-being. The feelings are "about" the society. They are mental diagrams of particular types of realities. This is the reason they are "cognitive." Thus, "appraisal" must be a fundamental characteristic of an emotional reaction.

People are constantly evaluating what is happening in the world at large. Events judged meaningful to the individual are further evaluated. These judgment patterns, referred to as appraisal structures, produce emotions. Since emotions are linked to evaluations of events, the appraisal approach to emotions can be thought of as a "cognitive approach." However, to compare the method of evaluation to conscious reasoning would be incorrect. According to Roseman and Smith, no appraisal theory presupposes that evaluations have to be deliberate and regulated [13]. On the contrary, it is assumed that evaluation procedures are usually automatic and unconscious [16]. Because the environment might change quickly, appraisals must be completed swiftly [12].

Here are the main components of cognitive appraisal theory:

1. **Primary Appraisal:** This is the initial assessment of an event, where we determine whether it is irrelevant, benign-positive (something good or pleasant), or harmful/threatening. For example, if you receive an unexpected email, you first assess whether it is something that could affect you positively, negatively, or not at all.

2. **Secondary Appraisal:** After assessing whether an event poses a threat, we proceed to evaluate whether we have the necessary resources or abilities to cope with the situation. This could involve assessing the extent of our control or the strategies available to manage it.

3. **Reappraisal:** This is an ongoing process where our evaluation of a situation might change as we gain more information or as the situation unfolds. Emotions can evolve depending on how we update our appraisals.

Lazarus believed that emotions arise from this dynamic and subjective process of appraisal, meaning that the same event might cause different emotional reactions in different people, depending on how they perceive and evaluate it.

### 2.3.2 Historical Use of Cognitive Appraisal Theory Across Disciplines

CAT has been applied across diverse domains to explain and predict emotional experiences:

- **Emotion Differentiation:** Ellsworth empirically mapped specific emotions to distinct patterns of appraisal [8]. For instance, fear is often associated with low control and high uncertainty, while anger arises from appraising an event as unjust but controllable. This validated the theory's predictive strength across a range of emotions.
- **Art and Aesthetics:** Silvia extended CAT to the domain of aesthetic emotions, particularly interest [15]. He demonstrated that people feel interested when they appraise art as novel but understandable; a balance between complexity and comprehensibility. This showed how CAT could predict emotional responses beyond survival or interpersonal contexts.
- **Stress and Coping:** In Lazarus and Folkman's classic work, CAT was foundational to understanding how people experience stress [4]. Stress was conceptualized as an emotional outcome of appraising a situation as threatening and beyond one's coping abilities.
- **Cross-cultural Analysis:** Shweder critically analyzed CAT's assumption of universal emotional responses to certain appraisals [10]. He argued that while some patterns may appear consistent, cultural variations in appraisal structures suggest that emotions and appraisals are co-constructed within cultural frameworks.
- **Emotion Prediction in AI:** More recently, CAT has been influential in designing emotionally aware artificial agents, where appraisal models are used to simulate emotional responses in machines based on input contexts and goal relevance [22].

### 2.3.3 Relation of CAT to Cognitive Decision-Making

CAT shares deep ties with cognitive decision-making processes because both involve evaluating relevance, goals, consequences, and control. While decision-making traditionally focuses on rational analysis, CAT highlights how emotions emerge from those same evaluations, thereby influencing future decisions.

According to Lazarus, the appraisal process is inherently evaluative, it determines not only what emotion is felt but guides the selection of coping strategies, much like

how cognitive decision-making involves weighing options and outcomes [9]. Emotions act as feedback mechanisms during decision-making, signaling progress toward goals [8].

Furthermore, secondary appraisals in CAT (blame, coping potential, and future expectations) are nearly identical to cognitive evaluations made during complex decision making. This positions CAT as a bridge between the affective and cognitive sciences, making it foundational for fields such as emotional intelligence, behavioral economics, and human-centered AI.

## 2.4 Chain of Thought and Emotional Intelligence

### 2.4.1 Chain-of-Thought (CoT) Prompting for Emotional Reasoning:

Chain-of-Thought (CoT) prompting is a technique that aids large language models (LLMs) in better reasoning by presenting their chains of thought in a structured, step-by-step manner [33]. Rather than merely arriving at the answer, the model is designed to reason out how it arrived at that answer by tracing a chain of reasoning. This method has been demonstrated to significantly enhance performance on tasks that involve math, logic, and common sense. Subsequent work has applied CoT to a wide variety of domains, confirming that scale can enable large models to learn multiple steps of thinking when they are trained or given similar examples.

Emotional reasoning has also been explored recently through CoT, with models being trained to understand and express emotions in a similar step-by-step manner.

**Emotional CoT (ECoT):**
Researchers proposed the Emotional Chain-of-Thought (ECoT) [48]. This straightforward yet efficient prompting algorithm enables language models to perceive and react to emotions in a more human-like manner. ECoT operates in understandable steps: the model initially detects emotional indicators in the user's words, followed by an explanation of how the user is likely to feel, and produces a thoughtful reply that is empathetic. It adheres to professional principles to ensure every action is emotional intelligence. This approach significantly enhanced the model's performance in providing emotionally adequate answers by revealing the reasoning process involved.

**Cognition Chains:**
The authors were inspired by the cognitive appraisal theory to develop a Cognition Chain format for detecting stress [59]. This is accomplished by having the model's reasoning follow a series: Stimulus $\rightarrow$ Evaluation $\rightarrow$ Reaction $\rightarrow$ Stress State, a condition similar to the human feelings experienced when mentally processing a stressful situation. The model recognizes the occurrence of what has happened, assesses the meaning of what has happened, responds emotionally, and, lastly, reckons the degree of stress. With the help of such a format, they were able to guide a language model through artificial learning on synthetic examples, producing good results in stress detection and making reasoning more intuitive in the model.

**Emotion–Cause Reasoning Chain (ECR-Chain):**
The idea of Chain-of-Thought prompting received further development with experts introducing, theoretically after the appraisal, the Emotion Cause Reasoning Chain (ECR-Chain): It appears as follows: "theme → reaction → appraisal → stimulus [60]." Here, the theme is the overall situation, the reaction is the emotional response, the appraisal is the evaluation of the situation, and the stimulus is the emotional event. They triggered language models with ECR-Chains and constructed annotated datasets to train language models to think about the reasons why something happened step-by-step about explaining the occurrence of an emotion. The technique helped bring about very positive shifts in activities, which were directed toward identifying the origin of feelings.

**Emotional-CoT (for Support):**
A study utilized the ESConv benchmark to conduct Chain-of-Thought prompting on emotional support discussions [45]. They propose their method, known as "Emotional-CoT," which directs the model to comprehend the user's emotional state first, e.g., what they feel or what they are worried about, and only then decide on a supportive strategy to use in responding to them. This implies that the model explicitly considers how the user feels and uses the same to influence a more understanding and responsive reaction. The improved emotional support and more meaningful conversations were the results of this two-step mental mechanism.

All of these CoT-based methods seek to make the process of emotional reasoning in the model more transparent. However, they vary in structure and theory, as they are dependent upon the theories they are based upon. Emotional CoT and Emotional-CoT do not decompose emotions into specific appraisal steps but primarily approach CoT as a guided prompt to aid the model in demonstrating empathy. Cognition Chains and ECR-Chain, by contrast, do contain appraisal theory ideas, albeit typically in only a single general step, such as evaluation or appraisal. Our Appraisal-CoT approach is more detailed in its explicit inclusion of several psychological appraisal dimensions in its chain of thought. Rather than a single general evaluative step of appraisal, the model is processed through specific cognitive determinations related to appraisal theory, including the relevance of goals, congruence of the goals, controllability, and coping potential.

This forms a more theoretical and comprehensive framework: whereas Cognition Chains follow the pattern Stimulus → Evaluation → Reaction, and ECR-Chain follow the sequence Theme → Reaction → Appraisal → Stimulus, Appraisal-CoT shows and distinguishes all the significant types of the appraisal elements of an incident before determining the ensuing emotion hypothesis. Appraisal-CoT will help provide a fundamental understanding of how humans react to feelings by connecting each step with the psychological dimensions that researchers such as Arnold, Lazarus, and Scherer have postulated. In brief, the emotional reasoning chains provided by previous CoT models were presented; however, the new appraisal cot is much more representative, as it is not only fully structured but also organized around the cognitive appraisal framework, which has not been achieved in existing instruction-tuning strategies.

### 2.4.2 Emotional Intelligence (EI) in LLMs

In psychology, Emotional Intelligence (EI) refers to the capacity to identify, comprehend, manage, and apply emotions to everyday life. Salovey and Mayer initially described it as the ability to observe one's own emotions and those of others, distinguish one emotion from another, and base thinking and actions on this emotional awareness. Their perception encompasses four main domains: the perception of emotions, the utilization of emotions in facilitating thinking, perceiving the meaning of emotions, and managing emotions. The review of EI was later popularized by Goleman, who emphasized practical skills such as self-awareness, self-control, empathy, and social interaction [11]. In general, EI entails not only being aware of emotions but also contemplating them clearly and applying their knowledge to act accordingly.

Emotional Intelligence (EI) is estimated in large language models (LLMs) through linguistic activities associated with emotion analysis and utilization. These are typically the ability to identify emotions in the text (e.g., by recognizing sadness or anger), the capacity to formulate emotionally considerate responses (e.g., by offering comfort), and the ability to apply emotional knowledge in tasks such as providing advice or managing emotions. According to research, LLMs are highly competent in detecting emotions. For instance, it was noted that models such as ChatGPT and LLaMA excel in emotion detection tasks [47]. Although basic classification is helpful, the real EI entails having context awareness and responding to it appropriately. It has also been observed in studies that LLMs can exhibit at least cognitive empathy, as they can understand the feelings of others and respond with feelings of compassion. GPT-4 and Gemini 1.5, as well as other models, were found by researchers capable of obtaining a performance of approximately 81% in addressing emotional intelligence test items, which is substantially above the human average (56%) [67]. Indeed, such models can formulate new questions that are valid and meaningful. As empathetic care or emotionally sensitive communication is relevant, especially in real-life applications such as healthcare, ChatGPT was ranked to be more empathetic than human physicians in certain instances, demonstrating the increasing potential of LLMs in the realm of emotional care.

Current studies of Emotional Intelligence (EI) within the context of NLP focus on two aspects. One direction is toward better empathy in dialogues itself; e.g., a transition model (LLaMA-based dialogue model) named SoulChat was trained with a patient support application in mind [57]. It scored higher on both automatic measures, ROUGE and BLEU, as well as on human-rated empathy scores (approximately 1.9 versus 1.6 on the 0-2 scale). A separate area tests the level of understanding and reasoning of models using emotions through Chain-of-Thought or appraisal techniques. To measure these benchmarks, 400 well-structured tasks in English and Chinese have been developed, assessing "Emotional Understanding" (i.e., understanding emotions and their causes) and "Emotional Application" (i.e., selecting the appropriate responses). Even GPT-4 cannot surpass the average human in these categories, indicating that modern models have yet to achieve significant emotional intelligence. Equally, EmoBench-M further expands this analysis to multimodal models in 13 conditions, including basic emotion recognition, conversation comprehension, and socially complex scenarios [63].

**Benchmarking and metrics:** A set of approaches is used to measure emotional intelligence in large language models (LLMs). Other benchmarks, including EmoBench, also evaluate the accuracy of models on multiple-choice questions that assess understanding and reasoning of emotional expressions. Some studies may compare LLMs to standardized EI tests, such as the one in Nature Communications Psychology, which is analogous to the MSCEIT and reports a total percentile score. In tasks involving models that generate responses, measures such as BLEU or ROUGE do not provide a reliable assessment of emotional quality. Consequently, researchers have developed new automatic measures or employed human evaluation. For example, a proposal was made that the Emotional Generation Score (EGS), an automated metric based on the psychological theory of EI, which rates answers according to dimensions such as self-awareness, emotion management, empathy, and relationship skills [48]. EGS closely matches the human rating of emotional quality. Finally, human assessment cannot be neglected; the results are frequently graded by experts or crowd workers in terms of empathy, relevance, and suitability. Patients ranked the medical responses of ChatGPT in comparison with those of doctors, evaluating them as more empathetic and helpful. This suggests that it is possible to perceive LLMs as more emotionally intelligent in real-life scenarios.

**Recognition vs. Generation vs. Regulation:** There are three primary aspects through which emotional intelligence can be interpreted in language models: recognition, generation, and regulation. Recognition deals with the model's capacity to detect or identify emotions in text, such as whether a person feels happy, sad, or angry, aspects that are already within the scope of LLMs, as the model can support sentiment or emotion classification [56]. Generation also enables responding to the model appropriately with empathy, e.g., through emotional storytelling, supportive responses, or advice-giving, which frequently occur in empathetic response generation tasks [48]. The hardest thing to do is regulation, which is assistance in controlling or transforming their emotional states, which has been partially addressed by providing coping mechanisms or emotional advice, which has been discussed in part via the use of dialogue systems that act as a substitute of therapeutic conversations [60]. Although the majority of benchmarks focus on recognition and generation, regulation is an equally important aspect of actual emotional intelligence, which is not yet mature in the field of LLM research [11], [52]. To achieve human-like emotional intelligence, the models will not only need to recognize and communicate emotional states but also assist the user in regulating those emotions with context-informed, psychologically informed rationality.

# Chapter 3

# Objective

## 3.1 Research Gap: Psychological Theory in LLM Emotion Models

Despite recent progress, current large language models (LLMs) primarily utilize expressions such as emotional words rather than more sophisticated theories of emotions. To illustrate, Yeo and Jaidka note that LLMs typically notice easily observable details but struggle to perceive things from another person's perspective [54], [70]. add that the general purpose of most LLMs does not include familiarity with specific psychological theories on emotional processing. Currently, no major language model directly utilizes Cognitive Appraisal Theory (CAT), which posits that we develop emotions based on how we judge events as they are perceived or experienced. Recently, scientists have employed new techniques to investigate this issue. For example, APTNESS addresses the selection of responses using appraisal theory, and ECR-Chain utilizes appraisal to determine what led to a particular emotion. Still, these are just the beginning. All in all, LLMs do not truly learn how people emotionally respond based on critical thinking, which occurs naturally in people.

- **Missing CAT focus:** Almost all LLMs can only analyze emotions or copy someone's style without grasping their meaning.
- **Shallow empathy:** Most explanations do not explain the reasons for a person's feelings after certain events. It is common for emotional understanding to be regarded as a basic code rather than a methodical, reason-based step.
- **Limited use of psychology:** Not many models (for instance, APTNESS and ECR-Chain) focus on appraisal, demonstrating that this area is underdeveloped in leading NLP research

### 3.1.1 Motivation: Failures in Emotionally Sensitive Contexts

As a result of these gaps, society faces actual consequences. LLMs may respond inappropriately to people going through mental health challenges as they do not understand emotions like humans. It has been shown in one study that ChatGPT

tends to provide scores indicating a lower risk of suicide than professional experts, which could lead to poor results in emergencies. It was also found that ChatGPT's mental health advice is short and usually focuses on basic methods like CBT. When a situation becomes complicated, it becomes difficult to feel empathy [40].

The feedback from real users also supports this. Even though ChatGPT responds well to various discussions, its reaction to answers about suicidal thinking and trauma was not satisfactory, according to the study. A study further highlighted bias, revealing that in some cases, GPT-4 expressed more empathy for women than for men, and it often failed to show empathy in happy moments, reflecting how it can perpetuate human prejudices [54]. Even some therapy bots available on the market may not be helpful. Research demonstrates that the functionality of Woebot is not superior to basic approaches such as self-journaling, lessons about mental health, or a decades-old chatbot (ELIZA) [51].

- **Underreacting to crises:** ChatGPT did not notice as many suicide risks as experts did, which might have ignored the urgency of the danger.
- **Generic, context-insensitive answers:** According to clinical reviewers, ChatGPT typically suggests brief solutions for everyone (such as CBT or mindfulness), disregarding details from the user's background. "It does not take into account everything about the patient" and lacks warmth and empathy.
- **Insensitive to trauma/loneliness:** Many times, when the subject of self-harm or loss comes up, the bot does not give the expected supportive response. It suggests that LLMs may not accurately detect subtle signs of someone being in distress.
- **Empathy gaps and biases:** During user studies, GPT-4 demonstrated a lack of empathy in happy situations and exhibited gender differences, unlike humans.
- **Limited benefit over trivial methods:** AI counselors created specifically to help people have not proven more effective than simple interventions, which illustrates how LLM-driven support lacks basic standards.

Such real-life situations highlight the fact that LLMs may sometimes respond with inappropriate and harmful words without the aid of psychology.

## 3.2 Objective: Appraisal-Based Instruction Tuning for Empathetic LLMs

Our research focuses on enhancing the utility of LLMs by fine-tuning them with cognitive appraisal and chain-of-thought data. As a result, the model can generate responses that are both impactful and consistent with psychological concepts. We create training cases that require the LLM to go through the entire process of appraisal, from noticing a stimulus to reacting, before giving a response. This reflects

how people deal with emotions when events happen to them. The model learns to respond appropriately based on genuine feelings rather than just the immediate tone used in the text.

1. **Theoretical Integration:** To operationalize Cognitive Appraisal Theory (CAT) as a parameterized reasoning schema for large language models (LLMs), our primary objective is to develop a CAT-CoT framework.
2. **Building Synthetic Dataset on psychological grounding:** Applying Cognitive Appraisal Theory (CAT) directly to the training set makes the model better understand emotions. Previous studies have employed appraisal theory in only one or two ways, such as labeling data or prompting. In contrast, we take it a step further by training the model using entire sequences of the appraisal framework. This is based on the belief that making LLMs use step-by-step thinking (like in chain-of-thought reasoning) helps them do more complex tasks. The idea is to ensure the model describes its feelings and actions using appraisal logic, enabling it to react with empathy.
3. **Enhancing emotional intelligence of LLMs:** Instruction tuning LLMs to build enhanced emotionally intelligent LLMs.
4. **Empirical Validation:** Conducting thorough empirical evaluations using Emobench benchmark, automated, and human-centered metrics.

To summarize, we will utilize examples of appraisal chain of thought to train the LLMs. This approach is based on new findings in psychology and neurolinguistic programming (NLP). We believe it will enable the model to respond with greater empathy and relevance when responses from typical LLMs are insufficient. Using cognitive appraisal theory (as many related studies suggest) helps us address the shortcomings in emotional reasoning found in current models.

# Chapter 4

# Methodology

## 4.1 Task Formulation: Conceptualizing the Workflow and Research Design

In our paper, we have integrated Cognitive Appraisal Theory (CAT) into large language models (LLMs) to enhance their capacity for emotion generation and understanding. Firstly, we have parameterized cognitive appraisal theory in a way that allows its essence to be integrated into large language models. As a result, we have these parameters: Goal Relevance, Goal Congruence, Ego Involvement, Coping Potential, Future Expectancy, and Perceived Control. These parameters successfully capture the idea of CAT. Translating the theorem into our algorithmic framework proved to be a challenging process. We held multiple meetings with professional psychiatrists whose insights were instrumental in deepening our understanding of the theorem.

This dataset needed to be formed in a way that allowed the model to learn to reason about a given scenario using CAT. Initially, we chose the Empathetic Dialogues dataset to collect our data. The current dataset, the EMPATHETICDIALOGUES, was sequentially assembled in two crowdsourcing tasks on Amazon Mechanical Turk. During the first stage, the workers were individually presented with one out of 32 fine-grained emotion labels and asked to compose a brief (1-3 sentence) personal situation in which they had truly felt that emotion. In the second stage, one scenario involved the author (the Speaker) being matched with another worker (the Listener). The Speaker engaged in a brief (4-8 utterances) one-on-one conversation based entirely on their account of that scene, neither the Speaker nor the Listener ever saw the original emotion label or description [25].

This project was headed by Hannah Rashkin of the University of Washington, in collaboration with the Paul G. Allen School of Computer Science & Engineering, along with Eric Michael Smith, Margaret Li, and Y-Lan Boureau of Facebook AI Research. To stimulate replication and subsequent research, the authors made the full dataset publicly available, consisting of nearly 24,850 dialogues and all experimental code [25].

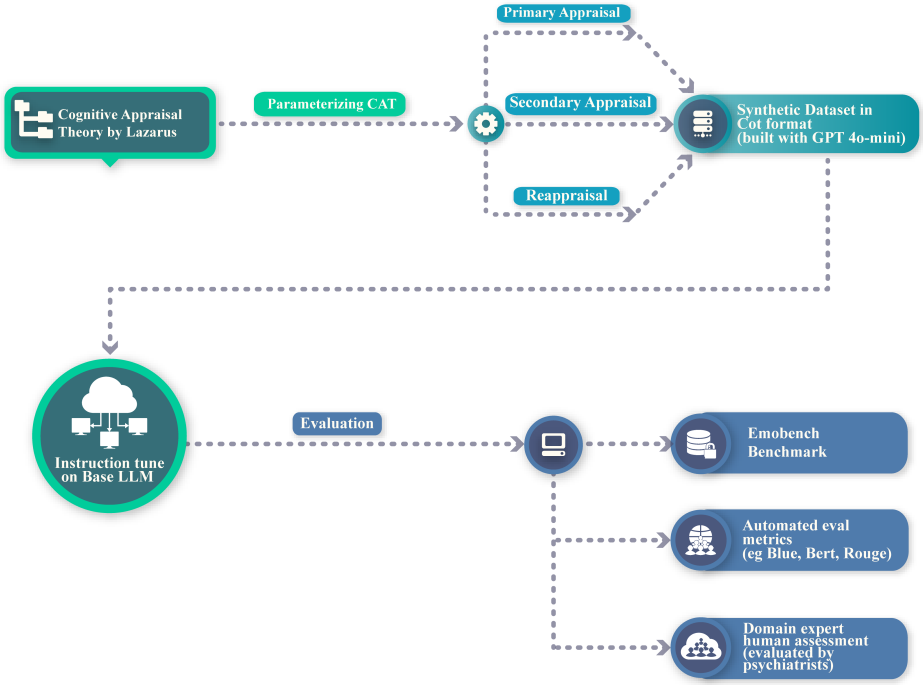


Figure 4.1: End-to-end diagram of the proposed system architecture and experimental pipeline.

The dialogs consist of a pair of Speakers (the worker authoring a situation) and a Listener (another worker), engaging in chit-chat in an open domain and an empathetic style. Listeners also pay attention to recognizing and confirming the emotions of the speaker rather than achieving informational or transactional objectives. Conversations, on average, last 4.31 turns, and each utterance has an average of 15.2 words [25].

The set consisted of all the information gathered through the platform ParlAI at Facebook AI Research, which comprises 810 Mechanical Turk workers located in the U.S. The collection protocol was dynamically directed to cause contributors to select the least frequently chosen labels in follow-up tasks, resulting in an even proportion of positive and negative emotions in personal, between-individual scenarios [25].

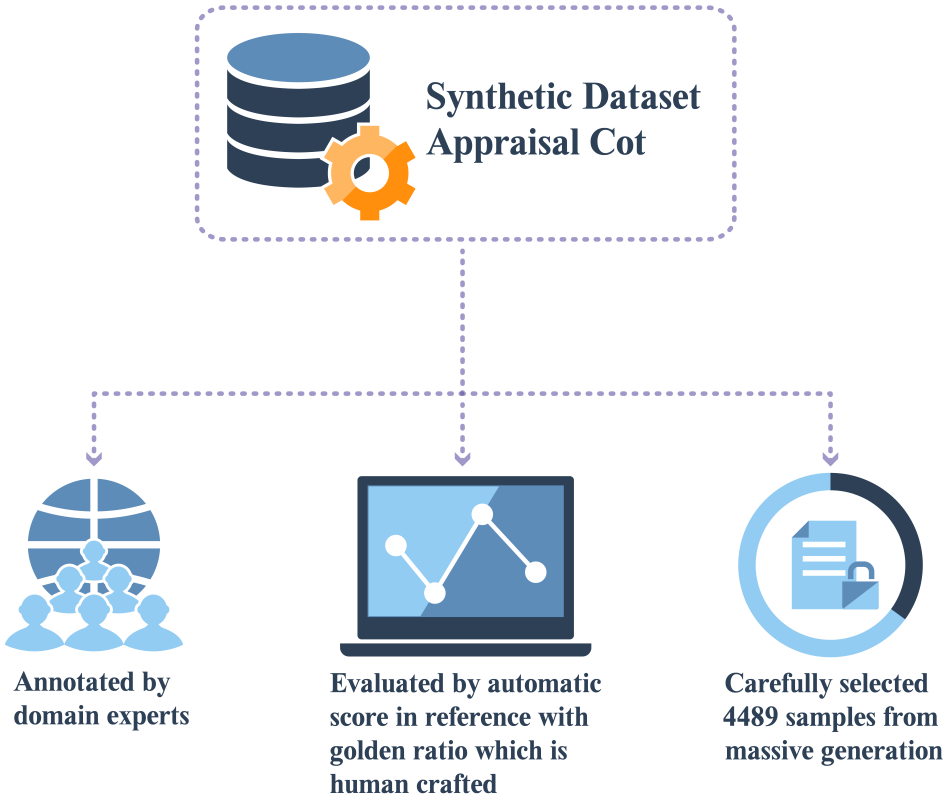


Figure 4.2: Illustration of the synthetic dataset validation process, detailing the steps used to assess the quality and reliability of generated data.

Now, from the ED dataset, we have collected the initial user message. We have only selected the initial user scenario because we wanted to generate the assistant response part synthetically. The reason for ignoring the actual human response and generating the assistant part synthetically was to demonstrate the impact of Cognitive Appraisal Theory on emotion generation. In real life, when a human responds, they utilize an unknown number of algorithms in their brain. So, we cannot definitively say what triggered the response. Thus, we needed to create an ideal environment where only CAT is available to the assistant. This is only possible if a dataset is created in a way that allows the model to use a chain-of-thought process inspired by CAT to reach a viable solution. That is why we have not taken human responses into account. Moreover, observe how it responds to users' queries in a single turn after learning about Appraisal CoT. After that, we developed a synthetic dataset using GPT-4o-mini in the form of chain-of-thought reasoning while using the parameters of CAT.

The reason behind choosing Gpt-4o-mini is its cost-effectiveness and emotion generation capability. A recent study, Sentiment Analysis in Software Engineering, evaluated generatively pre-trained transformers, including GPT-4o-mini, on senti-

ment and emotion detection tasks. It found that fine-tuned GPT-4o-mini achieved macro-F1 scores of 0.93 and 0.98 on structured datasets (GitHub, Jira) and 85.3% accuracy on complex, imbalanced inputs; outperforming full-scale GPT-4's emotion recognition in some cases (Benchmarking Zero-Shot Facial Emotion Annotation With Large Language Models: A Multi-Class and Multi-Frame Approach in DailyLife, n.d.) [64]. Another work, "ELECTRA and GPT-4o: Cost-Effective Partners for Sentiment Analysis," places GPT-4o-mini near full GPT -4 performance—achieving 86.77 macro-F1 after fine-tuning—while maintaining substantially lower token costs [41]. GPT-4o-mini offers a highly competitive pricing model: $0.15 per million input tokens and $0.60 per million output tokens, making it over 60% cheaper than GPT-3.5 Turbo without sacrificing performance. OpenAI's official launch highlighted its superior benchmark scores (e.g., 82% MMLU) at a significantly reduced cost.

After building the dataset, we have validated our dataset in these three ways:

1. Automatic evaluation using metrics such as Blue, Bert, Rouge, and distinct
2. 3 Domain expert psychiatrists annotated a small portion of our dataset. Which then has been compared with another set of annotations by non-domain expert annotators
3. IAA score has been calculated for both of the annotated datasets

**Figure 4.2** is depicting the validation methodology applied to the synthetic dataset.

We now have instruction-tuned LLMs using this Appraisal CoT dataset, thereby equipping them with a cognitively grounded emotional reasoning process that enhances their emotional expressivity and alignment with human affective states. This instruction tuned LLM then eventually was evaluated in 3 different ways:

1. The Emobench Benchmark dataset was chosen to test our model's emotional intelligence.
2. Automated Evaluation metrics (Rouge-1, Rouge-2, BLEU-1, 2, 3, 4, BERTScore, Distinct-1, 2) were chosen to judge the ability of model generation capability compared to a handcrafted golden dataset.
3. Human Evaluation metrics (Emotional Understanding, Empathy, Appraisal Reasoning, and Helpfulness) were captured from three domain expert psychiatrists to evaluate the final model response against human expectations and assess the quality of the response structure.

**figure 4.1** is demonstrating comprehensive representation of the research plan, including data preparation, model training, and evaluation stages.

## 4.2 The way of incorporating Appraisal CoT

### 4.2.1 Parameterizing Appraisal CoT

Appraisal Chain of thought is a model that parametrizes Lazarus's Cognitive appraisal theory of stress and coping. This model helps one to evaluate and perceive their environment. According to Lazarus and Folkman, Cognitive appraisal can be categorized as primary appraisal and secondary appraisal [4].

Primary appraisal can be defined using the points below:

1. **Irrelevant:** An interaction with the environment is considered unimportant if it has no impact on an individual's well-being. The individual has no stake in the potential results, which is another way of expressing that there is no need, value, or commitment involved; there is nothing to gain or lose from the deal.

2. **Benign-positive:** Benign-positive assessments occur when an interaction is perceived as beneficial, preserving or improving well-being, or promising to do so.

3. **Stressful:** Harm, loss, threat, and challenge are all components of stress evaluations. Harm or loss occurs when a person has already suffered damage, such as an injury or disease, loss of self-esteem, or the loss of a loved one. Losing significant commitments is one of the most painful life events.

On the other hand, Secondary appraisal mainly focuses on the individual's coping mechanism in response to their stress or environment. He uses the term outcome expectancy to refer to the person's evaluation that a given behavior will lead to certain outcomes and efficacy expectation to refer to the person's conviction that they can successfully execute the behavior required to produce the outcomes [4]. A study found that coping strategies differed depending on the situation (primary appraisal) and the available options (secondary appraisal) [6]. For example, when people's self-esteem was threatened, they exhibited more confrontational coping, self-control, escape-avoidance, and acceptance of responsibility compared to when it was not threatened.

Later, Lazarus and Folkman also introduced the concept of reappraisal. Reappraisal is the process of continuously reevaluating a situation in light of fresh information. This step of reappraisal occurs throughout the process and has the potential to alter an individual's perception of the circumstance [4].

In our paper, we have attempted to parametrize the Cognitive Appraisal Theory. We attempted to incorporate it by building a framework that embodies the essence of the theory. Below are the parameters chosen by us, which reflect the ideas of Lazarus and Folkman:

**Primary Appraisal:**

- Goal Relevance: (Is this situation relevant to the listener's goals?)

- Goal Congruence: (Is this situation consistent with the listener's goals, or does it obstruct them?)
- Ego Involvement: (Does this situation impact the listener's self-esteem, values, or social identity?)

**Secondary Appraisal:**

- Coping Potential: (What coping options does the listener believe they have available? Do they feel they can influence or manage the situation?)
- Future Expectancy: (What does the listener expect will happen in the future as a result of this situation? Do they feel hopeful or pessimistic?)
- Perceived Control: (To what extent does the listener feel they have control over the situation and its outcome?)

To represent the degree of each Primary appraisal parameter, we have introduced three categories: High, Moderate, and Low. On the other hand, to represent the degree of each Secondary appraisal parameter, we have used Positive, Uncertain, and Negative.

## 4.3 Detailed Explanation of Appraisal CoT Parameters and Levels

This portion explicitly describes the levels of CAT.

### 4.3.1 Primary Appraisal: "What is at stake?"

This is the initial assessment of a situation to determine its significance and meaning for one's well-being.

1. **goal_relevance:**

   **Meaning:** Does this situation matter to my personal goals, values, or well-being? Is it relevant to what I care about?

   **High:** The situation is critical and directly impacts one or more significant personal goals, values, or aspects of well-being. For example, a final exam for a student aiming to graduate or a health diagnosis.

   **Moderate:** The situation has some connection to personal goals or well-being, but it is not critical or central to them. It might be a minor inconvenience or a small opportunity. For example, a slight change in work routine that does not drastically affect outcomes.

   **Low:** The situation has little or no bearing on personal goals, values, or well-being. It is perceived as irrelevant or trivial. For example, news about a minor event in a distant country that has no personal connection.

2. **goal_congruence (or incongruence):**

   **Meaning:** Does this situation help or hinder the achievement of my goals? Is it consistent or inconsistent with what I want?

   **High (Congruence):** The situation is highly consistent with or actively helps achieve personal goals. It is seen as beneficial or positive. For example, receiving an unexpected promotion that aligns with career ambitions.

   **Moderate (Congruence/Incongruence):** The situation has a mixed impact or a mild positive or negative impact on goals. It may facilitate or impede goal attainment. For example, a new project that is both interesting and adds to the workload.

   **Low (Congruence) / High (Incongruence):** (As in your example, "Low" indicates low congruence, meaning it obstructs goals). The situation is highly inconsistent with, thwarts, or obstructs personal goals. It is seen as harmful, threatening, or a loss. For example, failing an important exam or losing a job.

3. **ego_involvement:**

   **Meaning:** How much does this situation relate to my sense of self, identity, self-esteem, values, or commitments?

   **High:** The situation deeply involves core aspects of the self, such as self-esteem, moral values, personal identity, or significant commitments. The outcome has substantial implications for how the person sees. For example, public criticism of one's work or a situation that challenges one's deeply held beliefs is deeply distressing.

   **Moderate:** The situation has some relevance to the self or personal values, but it is neither a profound challenge nor a significant affirmation. For example, receiving feedback on a task that is not central to one's identity.

   **Low:** The situation has little to no connection to one's self-concept, self-esteem, or personal values. It is perceived as impersonal—for example, a technical issue with a piece of equipment that does not reflect personal competence.

### 4.3.2 Secondary Appraisal: "What can be done?"

This follows primary appraisal (if the situation is deemed relevant and stressful) and involves evaluating one's coping resources and options.

1. **coping_potential:**

   **Meaning:** What are my options for addressing this situation, and how effective do I think they will be? Do I have the necessary resources (internal, such

as skills, and external, such as support) to manage this?

**Positive (High coping potential):** The individual believes they have sufficient resources and practical strategies to manage or alter the stressful situation. They feel capable and resourceful. For example, facing a challenging project but feeling confident in their skills and having a supportive team.

**Moderate (coping potential):** The individual believes they have some resources or options but may be unsure of their effectiveness or sufficiency. There is a mixed sense of capability and doubt. For example, facing a difficult task with some relevant skills but lacking complete confidence or support.

**Negative (Low coping potential):** The individual believes they lack the necessary resources, skills, or options to deal effectively with the stressor. They feel overwhelmed, helpless, or that their efforts will be futile. For example, facing a significant financial crisis with no savings or job prospects.

2. **future_expectancy:**

   **Meaning:** What is my outlook on how this situation will unfold or change in the future, considering my coping efforts or the nature of the situation itself?

   **Positive:** The individual expects the situation to improve, for their coping efforts to be successful, or for positive outcomes to emerge. There is an optimistic outlook. For example, believing that hard work will eventually lead to overcoming a current challenge.

   **Moderate:** The future outlook is uncertain or mixed. Things could improve or worsen, or the situation might persist without a clear resolution. For example, being unsure if a new treatment for an illness will be effective.

   **Negative:** The individual expects the situation to worsen, persist in a negative manner, or for their coping efforts to be ineffective, resulting in unfavorable outcomes. There is a pessimistic outlook. For example, believing that a conflict will escalate no matter what they do.

3. **perceived_control:**

   **Meaning:** To what extent do I believe I can influence or change the situation itself or my emotional reaction to it?

   **Positive (High perceived control):** The individual believes they have a significant degree of control or influence over the stressor or their response to it. They feel empowered to make changes or manage their reactions effectively. For example, feeling they can directly address a misunderstanding with a colleague.

**Moderate (perceived control):** The individual believes they have some, but not complete, control over the situation or their response. They can influence certain aspects but not others. For example, being able to manage their stress about an upcoming exam, but not the difficulty of the exam itself.

**Negative (Low perceived control):** The individual believes they have little or no control over the situation or their response to it. They feel subject to external forces or their unmanageable reactions. For example, feeling helpless in the face of a natural disaster or a company-wide layoff.

### 4.3.3 Reappraisal:

Given the listener's initial emotional response and the conversation so far, is there an opportunity for reappraisal? (Can the listener reframe the situation in a more positive light or develop more effective coping strategies?). If yes, describe the potential reappraisal process. If no, state, "No opportunity for reappraisal in this situation."

## 4.4 Core Principles Behind Synthetic Dataset Generation

### 4.4.1 Motivation Behind Choosing the Empathetic Dialogues(ED) Dataset

Researchers introduced a benchmark dataset to improve empathetic dialogue generation [25]. We have chosen their Empathetic Dialogues dataset as a seed dataset for building our Appraisal-cot dataset.

We have selected a total of 4641 full conversations between customers and Agents from the ED dataset. We have selected conversations that have a maximum of 4 turns.

Primarily the ED dataset looks like the **figure 4.3**.

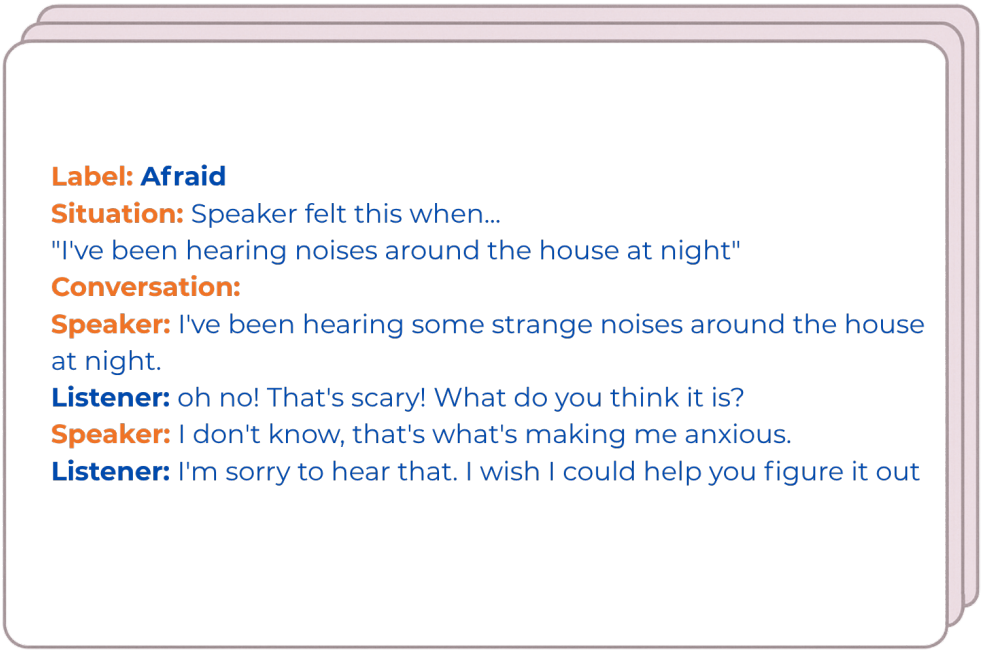


Figure 4.3: Sample entry from the EmpatheticDialogues (ED) dataset, illustrating the structure and emotional context of a dialogue instance.

### 4.4.2 Procedure for generating Appraisal-cot dataset turn-by-turn using cognitive appraisal theory and controlled rewriting

The ED dataset consisted of multiturn conversations between speaker and listener. We need to build the synthetic dataset from this seed dataset. So, we have chosen to follow a noble method for building this Appraisal-cot dataset. Here algorithm 1 is showcasing how the Appraisal CoT should be structured. Furthermore, algorithm 2 is demonstrating the procedure of generating the synthetic dataset.

The algorithm that we followed:

---

**Algorithm 1: Appraisal Chain of Thought (ACot)**

---

**Input:**
Customer: <Here is the Customer's dialogue from the conversation>
Assistant : Performing ACot
Step 1: **Describe the content of the conversation.**
Step 2: **Identify the Customer's situation.**
Step 3: **Primary Appraisal:**

- Goal Relevance: High/Moderate/Low - (Is this situation relevant to the customer's goals?)
- Goal Congruence: High/Moderate/Low - (Is this situation consistent with the customer's goals, or does it obstruct them?)
- Ego Involvement: High/Moderate/Low - (Does this situation impact the customer's self-esteem, values, or social identity?)

Step 4: **Secondary Appraisal:**

- Coping Potential: Positive/Moderate/Negative - (What coping options does the customer believe they have available? Do they feel they can influence or manage the situation?)
- Future Expectancy: Positive/Moderate/Negative - (What does the customer expect will happen in the future as a result of this situation? Do they feel hopeful or pessimistic?)
- Perceived Control: Positive/Moderate/Negative - (To what extent does the customer feel they have control over the situation and its outcome?)

Step 5: **Based on the primary and secondary appraisals, identify the Customer's emotions and explain why.** (Connect appraisal patterns to expected emotions)
Step 6: **Reappraisal (Optional):** Given the Customer's initial emotional response and the conversation so far, is there an opportunity for reappraisal? (Can the Customer reframe the situation in a more positive light or develop more effective coping strategies?). If yes, describe the potential reappraisal process. If no, state, "No opportunity for reappraisal in this situation."
Step 7: **You are the "Assistant"; consider how to respond to the "Customer" with empathy based on the appraisal and their likely emotional state. Consider the potential impact of your reply on "Customer" and generate an empathic response that acknowledges their appraisal.**
**Response:** Combine the above thoughts and give your response to "Customer". Your response should acknowledge their emotions and possibly guide them toward reappraisal if appropriate.

**Algorithm 2: Procedure for Single Turn Empathetic Dialogue Generation Using Cognitive Appraisall**

**Input**
A dialogue sample with four turns: (Customer_t1, Assistant_t1, Customer_t2, Assistant_t2)
**Output:**
Generated empathetic dialogue sample with intermediate cognitive appraisals and rewrites

**foreach** dialogue sample do

1. **Parse Dialogue:**
   Extract individual turns:
   **Extract individual turns:**
   Customer_t1, Assistant_t1, Customer_t2, Assistant_t2 ←
   ParseTurns(dialogue_sample)

2. **Generate Assistant Turn 1 (Appraisal + Response):**
   Input: Customer_t1
   Perform cognitive appraisal and generate response:
   (Appraisal_1, Response_1) ← LLM(Customer_t1)

**end**

Firstly, we have collected the first turn of Speaker conversation from the ED dataset. This was done to keep the contexts grounded in real-life events.
Then, in the first assistant turn (Assistant_t1) generation phase, the LLM is prompted using only the first customer utterance to perform the Appraisal-cot algorithm and generate a response.

### 4.4.3 Prompt that was given for Dataset Generation

The prompt structure for generating the Assistant response to the customer's input :

**Prompt given to ChatGpt 4-o-mini**

**Variables:** System Prompt = *"You are an expert in emotional psychology specializing in cognitive appraisal theory. Your task is to analyze the customer's statement(s), perform a cognitive appraisal from the customer's perspective, and generate an empathetic response."*

customer_t1_text = Extracted user's first utterance from the ED dataset

Appraisal Chain of thought = Perform the following Cognitive Appraisal Steps:

- Step 1: Describe the content of the conversation.
  (Briefly summarize what the customer said).

- Step 2: Identify the listener's situation.
  (What specific event or state is the customer describing?).
- Step 3: Primary Appraisal:
  - Goal Relevance (Is this situation relevant to the listener's goals?)
  - Goal Congruence (Is this situation consistent with the listener's goals, or does it obstruct them?)
  - Ego Involvement (Does this situation impact the listener's self-esteem, values, or social identity?)
  - High/Moderate/Low with explanation.
- Step 4: Secondary Appraisal:
  - Coping Potential (What coping options does the listener believe they have available? Do they feel they can influence or manage the situation?)
  - Future Expectancy (What does the listener expect will happen in the future as a result of this situation? Do they feel hopeful or pessimistic?)
  - Perceived Control (To what extent does the listener feel they have control over the situation and its outcome?)
  - Positive/Moderate/Negative with explanation.
- Step 5: Based on the primary and secondary appraisals, identify the listener's likely emotions and explain why.
  (Connect appraisal patterns to expected emotions).
- Step 6: Reappraisal (Optional):
  Is there an opportunity for reappraisal? If yes, describe potential process. If no, state "No clear opportunity...".
- Step 7: Speaker's thought process:
  Briefly outline how to reply empathetically based on the appraisal.
- Response:
  Generate the empathetic response text for the assistant.

**Output Format = "**
Provide your entire output strictly as a JSON object with two keys:

- "appraisal_cot": containing steps 1–7 as nested objects/strings
- "response": containing the generated assistant text "

**Example Output = "**
{
"appraisal_cot": {
"step1": "The customer expresses feelings of being overwhelmed...",
"step2": "They likely feel anxious, exhausted...",
"step3": {
"primary_appraisal": {

"goal_relevance": "High - Work performance and well-being are key goals.",
"goal_congruence": "Low - The current situation obstructs the goal of feeling capable and managing work.",
"ego_involvement": "High - The situation impacts their sense of competence and self-esteem."
}
},
"step4": {
"secondary_appraisal":
"coping_potential": "Moderate - Currently feels low, but potential options might exist (like suggested breaks).",
"future_expectancy": "Negative - Without change, stress is likely to persist or worsen.",
"perceived_control": "Negative - Feels overwhelmed by external demands and pressure."
}
},
"step5": "The customer's mental state is one of distress...",
"step6": { "reappraisal": "Encouraging small, actionable coping strategies..."
},
"step7": "Acknowledge their struggle, validate emotions, suggest steps..."
},
"response": "It sounds like you're facing a lot of pressure..."
}
"

---

***{System Prompt}***

Customer: [The actual first customer utterance,
'customer_t1_text ', would be inserted here.]

***{Appraisal Chain of thought}***

***{Output format}***

Example Output JSON structure (content will vary based on input):

***{Example Output}***

---

## 4.5 Dataset Evaluation: Justification and Suitability Assessment

### 4.5.1 Quantitative Validation via Automatic Evaluation Metrics

#### BLEU and ROUGE

Automated measures of evaluating generated dialogues can provide quantitative assessments of how well a dialogue is calibrated on a lexical, semantic, and emotional level compared to an ideal set of dialogues. BLEU was developed which, measures n-gram precision between candidate text and reference [17]. BLEU (1-4) scores indicate the number of unigram, bigram, trigram, and 4-gram sequences of lexemes in the candidate data that are equal to the reference data. Moreover, ROUGE-1 and ROUGE-2 [18] compute the matching recall of unigrams and bigrams. The value represents the normalized count of the content in the reference data that covers the candidate information and addresses charges of omission. BLEU and ROUGE are frequently used to quantify improvements in research on dialogue evaluations, often by combining the BLEU and ROUGE metrics in lexical similarity. Where BLEU gives more stress to precision (the n-grams present in the candidate text discovered in the reference text), and ROUGE also gives more stress to recall (the n-grams present in the reference text discovered in the candidate text). Emotional dialogue evaluation is also extensively employed, utilizing BLEU and ROUGE. For instance, these methods are used in research aimed at achieving emotionally suitable responses, such as the Emotional Chatting Machine and the EmpatheticDialogues dataset [25], [26]. Moreover it was found that BLEU correlates reasonably with human judgments of conversational response quality during their research [23]. Therefore, a dataset with high scores on these metrics can enhance a trained model's ability to represent an ideal emotional style and dialogue.

#### BERTScore

Lexical overlap checking often lacks semantic match, particularly in conversation-type datasets, as emotional conversations commonly contain some degree of paraphrasing. Hence, checking similarity by its semantic meanings is crucial for making a correct judgment. BERTScore utilizes cosine similarity on the BERT embeddings to make comparisons of tokens between two documents. It will penalize the use of words instead of rewarding semantic paraphrases [27]. Additionally it was found that BERTScore yields higher scores when evaluating grounded emotions and applies it to instances of emotion response development, preserving the original meaning [39]. When BERTScore is high, it indicates that the candidate data and its related meaning share a similar emotional background to that of the ideal set. Being more closely correlated with human judgments compared to BLEU in text-generation tasks, BERTScore can be an appropriate measure to evaluate the semantic quality of generated utterances in dialogue [27].

#### Distinct-N

In research on emotional generation, the Emotional Chatting Machine revealed an inverse relationship between distinct-n scores and less expressive, mixed emotional responses [26]. Distinct-1 and Distinct-2 calculate lexical diversity, which is part of the unique unigrams/bigrams in the text. In particular, distinct-1 = (number of distinct unigrams) / (number of tokens) and likewise distinct-2 with bigrams. This ensures that the metric improvement is not a result of the parroting of some safe answers. However, the distinction ratio of a single response may be insignificant, and so in most cases, Distinct is calculated among a group of outputs to measure lexical variety. For our case, the purpose of instruction-tuning models with a dataset that achieves better distinct-1/2 scores means that the model utilizes more extensive vocabulary resources and is less monotonous. Li demonstrated that neural conversational models are affected by low diversity and introduced the distinct-n to quantify this problem [23]. In such a way, the higher the evaluation score in distinct-1/2, the more the linguistic diversity in sets.

In short, BLEU and ROUGE count the number of overlapping tokens (precision and recall) between the candidate and ideal datasets. BERTScore determines semantic similarity by way of contextual embeddings, whereas Distinct-1/2 answers the question of the inherent lexical richness of the candidate data. The synthesis of these measures forms a complete portrait of quality and model evaluation for a dataset, providing well-balanced and quantitatively reasonable reasoning. These metrics have also been used in Automated model evaluation to showcase how well the model has grasped the essence of Appraisal CoT.

### 4.5.2 Qualitative Validation through Expert Human Assessment

The nature of our dataset requires human assessments that are primarily subjective. To ensure the most valid assessment, we have followed combinations of well-practiced human evaluation methodologies. In our case, understanding human psychology is a definitive way to make a quality assessment. So, we arranged a team of three psychiatrists as domain experts. Psychiatrists possess a deep understanding of emotion theories and appraisal constructs; they regularly assess human emotions and responses. As our domain experts, their judgments on our synthetic dataset are crucial to understanding its capability. Studies also show that agreement between expert annotators is more frequent than that between non-expert annotators in subjective questionnaires [21]. However, we arranged a team of data annotators to evaluate the dataset on a larger scale. Before evaluating, these annotators were well-trained to understand the topic, the method of evaluation, and the integrity of the evaluation.

**Human Evaluation Method:**

To all the teams, we presented a brief refresher on cognitive appraisal theory and its application to LLMs. We presented sample data and a questionnaire that will provide parameters for evaluating the dataset. The questionnaire provided us with the following parameters:

- On cognitive appraisal theory's procedure:
    1. Primary Appraisal: How psychologically plausible and accurate is it?
    2. Secondary Appraisal: How psychologically plausible and accurate is it?
    3. Coherence & Logic: How logical, coherent, and easy to follow is the step-by-step reasoning?
    4. Empathy: How well does the response convey empathy, understanding, and validation of the user's situation and likely emotions?
- On overall conversation:
    1. Coherence: How coherent is the overall conversation?

All annotation teams quantized these parameters for each assigned data. Scherer and Wallbott employed a 5-point Likert scale to collect scores for cognitive appraisal tasks [7]. Likert scales are the standard method for rating subjective parameters and their correlation measurements [24]. Hence, our annotators scored each data set on parameters using a 5-point Likert scale (1 = Very Poor to 5 = Excellent).

For the non-expert annotators' team, we have employed a methodology to enhance evaluation quality compared to the experts. As non-experts are more prone to have disagreements, a study proposed a systematic procedure to increase quality, which encourages discussions among annotators to justify and clarify their annotations upon disagreement [29]. This was formulated based on previous research into the Interrater Agreement, also known as the Inter Annotator Agreement (IAA). Where IAA discarded disagreements, Oortwijn's procedure prevented loss of information systematically and increased reliability [29]. Taking into account the justification for their procedure, we have prepared the following system (see **Figure 4.4**), inspired by the study, for non-expert annotators.

- **Step-1:** Two independent annotator teams (Team A and Team B) are formed. Each team comprised two annotators.
- **Step-2:** Team members will discuss amongst themselves to annotate each data.
- **Step-3:** If both members of a team cannot agree, another annotator from outside of both teams will vote to finalise a score to annotate.
- Step 2 and 3 will loop until all selected data are annotated.

Finally, when both teams are done, we have two sheets of 500 data annotations

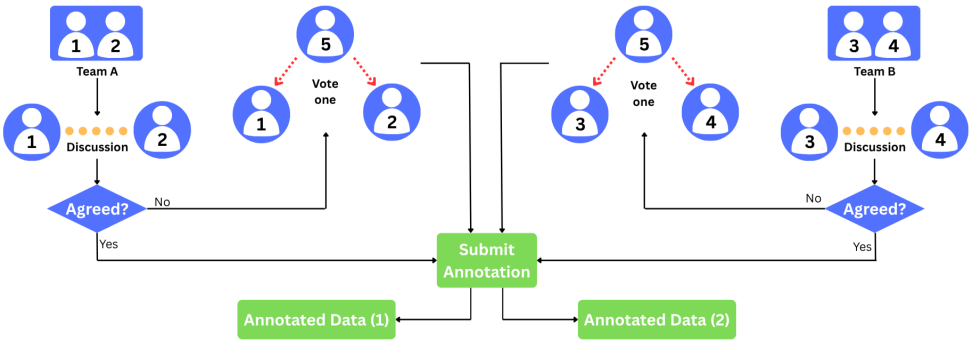


Figure 4.4: Non-expert Data annotation method for each data

A study proposed a system where a "third annotator" adjudicator makes one authoritative annotation from two clinical annotations [66]. Complimentarily, we encourage discussion and tie-breaking decisions by an additional annotator prior to annotation to increase efficiency and reliability.

### 4.5.3 Inter-Annotator Agreement as a Reliability Justification

The Inter Annotator Agreement demonstrates the consistency and reliability of all data annotations for the same dataset. It is measured using standard methods, such as Cohen's kappa and Krippendorff's alpha [20]. Although we must discard the idea of clearing disagreements, we need them to ensure the quality of the annotations by analyzing the produced scores. Krippendorff's alpha is a reliability-checking coefficient that supports our use of multiple annotators per instance and ordinal type of data. Furthermore, to measure ordinal data, such as emotions and sentiments, Spearman's correlation is commonly used to quantify agreement. It will measure the consistency of annotations and complement Krippendorff's Alpha for reliability check. However, scholars argue that correlation checks only reveal variations in variables rather than actual similarity. For example, even if an annotator is biased to score high and another is biased to score low, the correlation score will be high, but not the actual agreement. We will also use Intraclass Correlation (ICC), which measures agreement. It compares total variance and variations between annotators' annotations and gives a rating of agreement [46]. Using these measurements, IAA ensures the reliability and unbiased scoring of annotations [20].

## 4.6 Justification of Model Evaluation Approach

### 4.6.1 Benchmarking with Emobench for Emotional Intelligence Assessment

EmoBench is a novel benchmark crafted with theoretical grounds to mainly evaluate the emotional intelligence (EI) of large language models (LLMs). It contains 400 carefully designed multiple-choice questions (MCQs) printed in English and Chinese,

with an equal split across two fundamental parameters: Emotional Understanding (EU) and Emotional Application (EA). On the one hand, in the EU task, each scenario, usually corresponding to one or three persons, provides questions such as: What is the emotion of Person A? What is the reason for it? This way, the model will be forced to presuppose the implicit emotional states and the potential causes [52]. The EA task would then present a social dilemma characterized by the type of relationship and the problem situation, which the model should subsequently select to provide the most suitable response or action towards the problem situation. In addition, both tasks utilize the selection of well-chosen distractors. They are presented in a multiple-choice question (MCQ) format, making it clear that adequate performance cannot be achieved through pattern recognition but rather through profound contextual thinking. Lastly, the answers are obtained by repeatedly prompting zero-shot prompting, both with and without chain-of-thought reasoning, and then consolidated through a process of majority voting and random answer ordering [52]. The expectation from this benchmark is to understand how well the instruction-tuned model is responding with emotional intelligence.

**Theoretical Foundations of EmoBench:**
In the broader context of affective computing, the goal is to equip computers with the ability to detect and simulate human emotions. However, older datasets of emotion classification often tend to foster simple pattern recognition, where one can associate the notion of losing with sadness. On the other hand, unlike other benchmarks, which [52]. emphasize are more inclined toward emotion recognition rather than incorporating key elements of emotional intelligence (EI), namely emotion regulation and strategic response, EmoBench helps to fill critical gaps [52]. In this way, through a combination of recognition (label and cause) and application (best response), EmoBench will more accurately reflect the complex nature of human emotional intelligence.

**Justification for Using EmoBench:**
EmoBench is also an invaluable evaluating tool for instruction-tuned LLMs with synthetic emotional datasets. One of the latest surveys of affective computing in LLMs notes that emotional ability is barely examined, and EmoBench is an all-inclusive EI benchmark at both the comprehension and utilization levels [59]. Additionally, high-quality data may considerably improve the production of empathetic responses, which are commonly assessed on dialogue assessment measures only (e.g., EmpatheticDialogues) [49]. Comprehensively, EmoBench checks the internal emotional intelligence of models: it verifies whether a model not only produces empathetic language but also genuinely understands emotional situations and applies that knowledge when relevant. Thus, training and assessing with EmoBench will ensure that the improvements in synthetic-data training are considered accurate compared to emotional-reasoning tasks, as described by experts, rather than superficial fluency. Besides, at an academic level, rigorous evaluation criteria are desirable, as conventional evaluation datasets, such as BLEU and ROUGE, and even human ratings, only measure a partial aspect of emotion-related quality. EmoBench, on the other hand, offers objective, theory-driven evaluation identical in nature to psychological EI tests [48]. This means that EmoBench has been used to present a rigorously multi-dimensional justification of emotional intelligence in an instruction-tuned model.

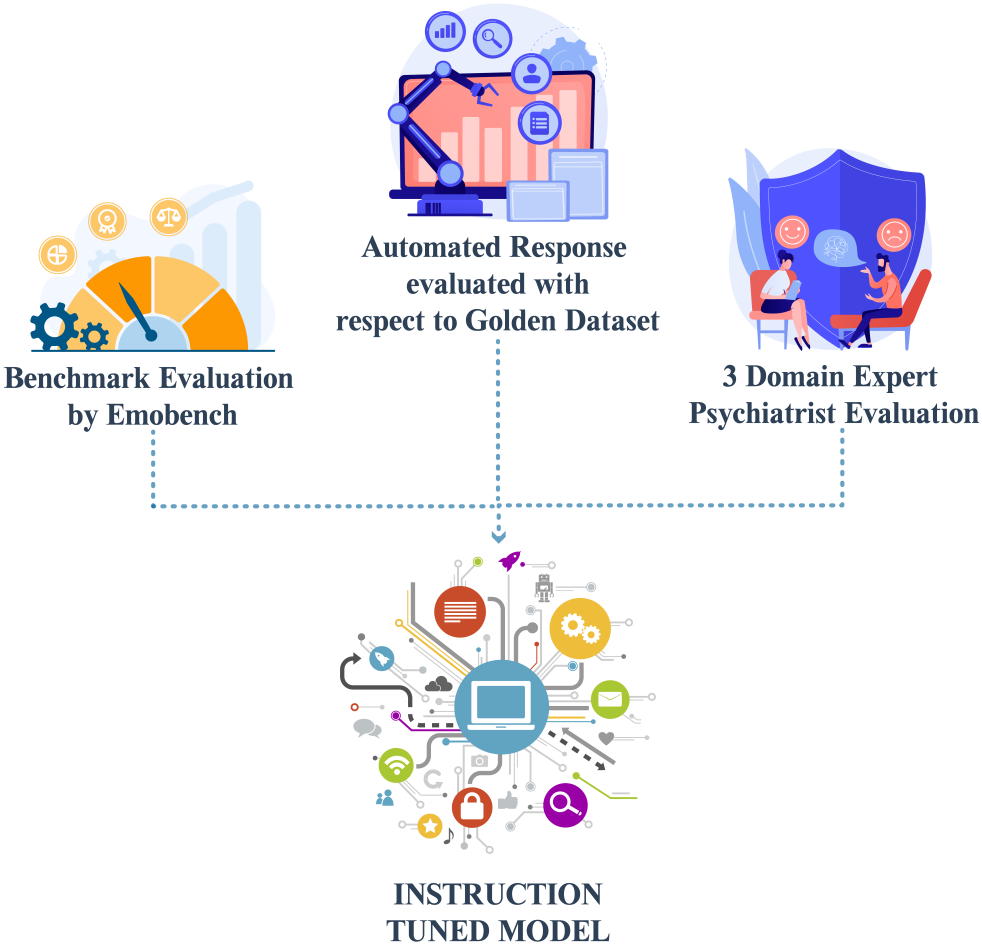


Figure 4.5: Evaluation framework for the instruction-tuned model. The model's performance is assessed through three parallel approaches: (i) Emobench benchmark evaluation, (ii) automated scoring against a golden reference dataset, and (iii) expert human evaluation by three domain-specialist psychiatrists.

### 4.6.2 Qualitative Model Validation through Expert Human Assessment

For human evaluation, we expected that the model would be judged based on its response to a structured appraisal Chain of thought, along with emotional understanding. Also, if the final response is empathetic and helpful. As practiced in research, to assess the emotional expressivity and clinical commensurateness of our instruction-tuned model, we conducted a comparative human assessment with psychiatrists, who are domain experts [62], [65]. To be more specific, we introduced a carefully selected appraisal form to a panel of three domain experts, which included seven situations taken from reality. The separate scenarios contained user input, matched responses of two models (Model A: instruction-tuned; Model B: base), and appraisal reasoning about the same. Each of the responses was rated independently by the experts on three parameters: Emotional Understanding, Empathy, and Helpfulness, as well as appraisal reasoning quality for the reasoning steps. A 5-point Likert scale was used to assign ratings.

**Figure 4.5** is showcasing both qualitative and quantitative evaluation approaches.

## 4.7 Bias-Oriented Dataset Inclusion: Ethical and Analytical Considerations

Bias was introduced in the case of instruction tuning smaller models, which means training our acot and base models with a dataset that is formatted like the Emobench benchmark dataset. This was only necessary in cases of Qwen3 0.6B and Qwen3 1.7B base models, as well as the tuned models. The reason behind this was that the model's parameters were so small that after training it with our dataset, their behavior changed. As a result, they were unable to answer the MCQs on the emobench. Another reason was that even the base models were not able to answer the MCQs. Thus making it hard to evaluate their capability. Thus, they needed to be tuned in a way so that the emobench could evaluate it. So, Qwen3 0.6B, Qwen3 0.6B, Qwen3 1.7B, and Qwen3 1.7B were instruction-tuned on the Demo Emobench dataset.

# Chapter 5

# Dataset Description

## 5.1 Source Dataset: EmpatheticDialogues (ED)

The benchmark dataset we used is Empathetic Dialogues, which was collected on Amazon Mechanical Turk (AMT) [25]. Initially, it contained 24,850 turns of conversations, each between two crowd workers (a speaker and a listener). It is an open domain and has conversations of any context that express emotions. Among many other emotion-based datasets, this benchmark dataset construction focused on how real-world emotional conversations occur. Speakers discussed complex scenarios from their daily lives or any emotionally charged event. Such conversational behavior and complexity are only seen in real-world events. The dataset provides conversations of 32 unique emotion labels. They are, 'sentimental', 'afraid', 'proud', 'faithful', 'terrified', 'joyful', 'angry', 'sad', 'jealous', 'grateful', 'prepared', 'embarrassed', 'excited', 'annoyed', 'lonely', 'ashamed', 'guilty', 'surprised', 'nostalgic', 'confident', 'furious', 'disappointed', 'caring', 'trusting', 'disgusted', 'anticipating', 'anxious', 'hopeful', 'content', 'impressed', 'apprehensive', and 'devastated' (see distribution in **figure 5.1**).

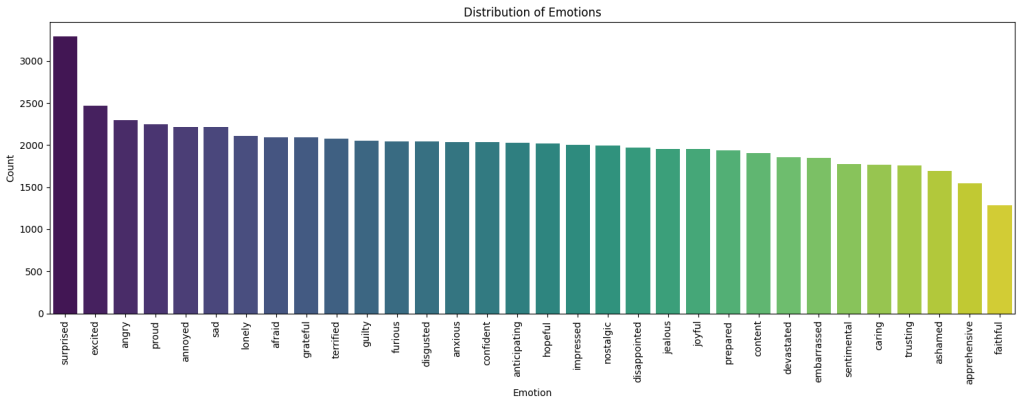


Figure 5.1: Data Preprocessing: Cleaning and Selection of Conversations

## 5.2 Data Preprocessing: Cleaning and Selection of Conversations

Each row of the original dataset contained two turns from each person. Upon observing, we found some significant problems. In the dataset, many messages from the speaker were labeled under the listener's label and vice versa. This also led to repeated messages from each side. We have cleaned the issue from the dataset by running it through a function. Furthermore, the dataset had some unknown characters, which we have cleaned. At this point, the dataset was ready for subsequent processing in the correct format.

Now, the conversation is in the following format:
Human: ...
Assistant: ...
Human: ...
Assistant: ...

We have renamed the 'Speaker' to 'Human' and 'Listener' to 'Assistant,' the role played by our model. After formatting, the dataset consisted of 16,789 dialogues (see in **figure 5.1**). with contexts, each having two turns and one of 32 types of emotion. For further procedures, we encountered limitations regarding the use of commercial model usage tokens. Hence, we randomly selected approximately 150 dialogues from each emotion, totaling 4,640 dialogues. Additionally, in addition to the previously selected dialogues, we have randomly selected 675 dialogues for evaluation purposes.

## 5.3 Synthetic Dataset Construction: Empathetic Response Generation and Augmentation

We have selected the ChatGPT 4o mini for our synthetic data generation. Firstly, we have built the dataset using the prompt structure outlined in Section 4.3.3 of the Methodology. However, prior to model training and analysis, the raw dataset underwent a crucial preprocessing stage to address significant structural inconsistencies and ensure data integrity. The primary issue identified was the varied and often malformed structure of the step-6 field within the appraisal_cot object of assistant-generated conversation turns. While the expected format was a dictionary containing a reappraisal key (e.g., 'reappraisal': '...'), numerous entries contained a raw string, were null, or were missing the field entirely. To rectify this, a programmatic cleaning script was executed. This script systematically iterated through each entry, transforming any string-based step6 values into the correct dictionary format. For entries where the step6 field or its mandatory reappraisal sub-key was absent, a default dictionary was inserted to ensure its presence across all records. This standardization process ensured a uniform data structure, which was crucial for reliable downstream processing and analysis, ultimately yielding a clean and corrected dataset.

## 5.4 Development of Bias and Evaluation Datasets

### 5.4.1 Construction of the Bias Dataset Using Emobench

The Bias dataset was a demo dataset that mimicked the output style of the Emobench benchmark dataset. **Figure 5.2**, **Figure 5.3** and **Figure 5.4** is demonstrating the prompts which were giving to Gpt 4o mini for building the bias dataset for Emobench.

```
You are an AI assistant helping to create diverse training data.
Your task is to generate a list of unique and diverse situation themes.

Theme Type: Emotion Understanding (EU) tasks. These themes
should describe situations that involve complex, nuanced, hidden,
mixed, or seemingly contradictory emotions, requiring deeper
insight into 'what' emotions are felt and 'why'.

Instructions:
1. Generate exactly 75 unique themes.
2. Each theme should be a short phrase (typically 3-10 words)
describing a relatable scenario or event.
3. Ensure the themes are diverse and cover a range of common
human experiences relevant to the theme type.
4. The themes should be suitable for constructing EmoBench-style
scenarios.
5. Do NOT repeat themes from the examples below or make them
too similar. Be creative.

Example themes (for guidance on style and topic, do not copy
these):
["someone laughing hysterically after a series of unfortunate
personal events", "a person crying tears of joy at a surprise reunion",
"an individual becoming unusually quiet and withdrawn after
receiving seemingly good news", "a character expressing anger
when they are actually feeling deep fear or hurt", "observing
someone trying to subtly hide their disappointment behind a facade
of indifference"]

Output your response strictly as a single JSON object containing a
key "themes" whose value is a list of strings.
For example:
{
  "themes": [
    "theme example 1",
    "another theme example",
    "yet another distinct theme"
  ]
}
Ensure the output is only this JSON object and nothing else.
```

Figure 5.2: Prompt template used for generating the Bias Dataset, targeting emotional variability across controlled scenarios.

Generate a single EmoBench-style Emotion Appraisal (EA) task.
The task must be a valid JSON object.

**Instructions for generation:**
1. **Overall Structure:** The JSON must have a "messages" key, which is a list of 3 dictionaries:
    * Message 1: role "system", content: "In this task, you are presented with a scenario, a question, and multiple choices. Carefully analyze the scenario and take the perspective of the individual involved. Then, select the option that best reflects their perspective or emotional response."
    * Message 2: role "user", content includes Scenario, Question, and Choices.
    * Message 3: role "assistant", content is a JSON string like '{"answer": "X"}' where X is A, B, or C.
2. **Scenario:**
    * Create a brief, relatable scenario involving the protagonist '{protagonist_name}'.
    * The scenario should revolve around the theme: '{theme}'.
    * It should clearly set up an emotional situation where the protagonist or someone interacting with them needs to decide on an effective or empathetic course of action.
3. **Question:**
    * Pose a question asking for the most effective, empathetic, or constructive response or perspective for '{protagonist_name}' or someone interacting with them in that scenario.
    * Example question format: "In this scenario, what is the most effective response for {protagonist_name}?" or "How might {protagonist_name} best approach this situation?"
4. **Choices (A, B, C):**
    * Provide three distinct choices labeled A), B), and C).
    * One choice should be clearly the "best" or most emotionally intelligent/constructive option.
    * The other two choices should be plausible but less ideal (e.g., common but unhelpful reactions, avoidant behaviors, overly negative or aggressive responses).
    * Ensure choices are concise.
5. **Correct Answer:**
    * The "assistant" message content must specify the correct choice (A, B, or C).

**Example of desired output structure (values will differ based on your generated scenario):**
{{
 "messages": [
  {{
   "role": "system",
   "content": "In this task, you are presented with a scenario, a question, and multiple choices. Carefully analyze the scenario and take the perspective of the individual involved. Then, select the option that best reflects their perspective or emotional response."
  }},
  {{
   "role": "user",
   "content": "Scenario: {protagonist_name} worked hard on a report, but their boss pointed out several errors publicly in a meeting. \\nQuestion: In this scenario, what is the most constructive way for {protagonist_name} to react immediately after the meeting? \\nChoices:\\nA) Argue with the boss about the feedback.\\nB) Acknowledge the feedback and ask for a private discussion to understand better.\\nC) Complain about the boss to colleagues."
  }},
  {{
   "role": "assistant",
   "content": "{{\\"answer\\": \\"B\\"}}"
  }}
 ]
}}

Protagonist: {protagonist_name}
Theme for current scenario: {theme}

Output *only* the JSON object. Do not include any other text or explanations.

Figure 5.3: Instructional prompt used to generate outputs for the Emotion Application (EA) task.

Generate a single EmoBench-style Emotion Understanding (EU) task.
The task must be a valid JSON object.

**Instructions for generation:**
1. **Overall Structure:** The JSON must have a "messages" key, which is a list of 3 dictionaries:
    * Message 1: role "system", content: "In this task, you are presented with a scenario, a question, and multiple choices. Carefully analyze the scenario and take the perspective of the individual involved. Then, select the option that best reflects their perspective or emotional response."
    * Message 2: role "user", content includes Scenario, Question 1, Choices for Q1, Question 2, and Choices for Q2.
    * Message 3: role "assistant", content is a JSON string like '{"answer_q1": "X", "answer_q2": "Y"}' where X and Y are A, B, or C.
2. **Scenario:**
    * Create a brief scenario involving the protagonist '{protagonist_name}'.
    * The scenario should revolve around the theme: '{theme}'. This theme should ideally hint at a complex, nuanced, hidden, or seemingly contradictory emotional state.
    * The scenario should allow for probing into *what* emotion is felt and *why*.
3. **Question 1:**
    * Typically asks to identify the primary or most likely (possibly complex or mixed) emotion(s) '{protagonist_name}' is experiencing.
    * Example: "What primary emotion(s) is {protagonist_name} likely feeling in this situation?"
4. **Choices for Question 1 (A, B, C):**
    * Provide three distinct emotional states or combinations.
    * One should be the most accurate based on the scenario's nuance. The others should be plausible but less fitting.
5. **Question 2:**
    * Typically asks for the reason behind the emotion identified in Q1, or a deeper understanding of '{protagonist_name}'s perspective or the situation's underlying psychological dynamics.
    * Example: "Why might {protagonist_name} be feeling this way?" or "What underlying factor contributes most to {protagonist_name}'s reaction?"

```
6.  **Choices for Question 2 (A, B, C):**
    *   Provide three distinct explanations or factors.
    *   One should be the most accurate and insightful. The others plausible
distractors.
7.  **Correct Answers:**
    *   The "assistant" message content must specify the correct choices for
both Q1 and Q2.

**Example of desired output structure (values will differ based on your
generated scenario):**
{{
  "messages": [
    {{
      "role": "system",
      "content": "In this task, you are presented with a scenario, a question,
and multiple choices. Carefully analyze the scenario and take the perspective
of the individual involved. Then, select the option that best reflects their
perspective or emotional response."
    }},
    {{
      "role": "user",
      "content": "Scenario: {protagonist_name} received a promotion they'd
worked towards for years, but on the same day, their close friend announced
they were moving far away.\\nQuestion 1: What conflicting emotions might
{protagonist_name} be primarily experiencing?\\nChoices for
Question 1:\\nA) Pure joy and excitement\\nB) Joy mixed with
sadness/loss\\nC) Anger and frustration\\nQuestion 2:
Why would {protagonist_name} feel this mix of emotions?\\nChoices for
Question 2:\\nA) Because the promotion wasn't what they expected.\\nB)
Because personal news is overshadowing professional success.\\nC)
Because significant positive and negative life events are occurring
simultaneously."
    }},
    {{
      "role": "assistant",
      "content": "{{\\"answer_q1\\": \\"B\\", \\"answer_q2\\": \\"C\\"}}"
    }}
  ]
}}

Protagonist: {protagonist_name}
Theme for current scenario: {theme}

Output *only* the JSON object. Do not include any other text or
explanations.
```

Figure 5.4: Prompt design employed for generating data samples for the Emotion Understanding (EU) task.

### 5.4.2 Creation of the Evaluation Reference Dataset

Our acot model and the Baseline model were given the same instruction prompt. Furthermore, a final response (in **Figure 5.5**) was requested in 80 words so that the evaluation becomes valid.

```
instruction_ prompt = (
    "You are an expert in emotional psychology and
you can accurately assess people's emotional states and
help them process difficult situations using cognitive
appraisal theory. Understand the listener's situation,
follow the listener's point of view, and respond
empathetically using primary appraisal, secondary
appraisal, and reappraisal processes. Avoid dismissing
their feelings, and instead, help them see the situation
from a growth-oriented perspective. Give a Final Response
at the end in 80 words. Give the response like this -
Final Response: { I'm so sorry to hear about your
grandma. Losing someone we love is can feel utterly
overwhelming, and it's completely natural to feel lost right
now. Please take all the time you need to grieve and find
a way to remember her that feels right for you. Talking
about it might help. You're not alone in this. } "
)
```

Figure 5.5: Evaluation prompt template utilized to create reference responses for the Evaluation Dataset.

### 5.4.3 Human Assessment Dataset for Model Output Evaluation

Our model was prompted with the instruction on which it was trained. The baseline model was just given the user input. **Figure 5.6** is showcasing the instruction prompt which was given to both the Baseline models and instruction tuned (acot) models to generate response for same user input.

instruction_prompt = ( You are an expert in emotional psychology and you can accurately assess people's emotional states and help them process difficult situations using cognitive appraisal theory. Understand the listener's situation, follow the listener's point of view, and respond empathetically using primary appraisal, secondary appraisal, and reappraisal processes. Avoid dismissing their feelings, and instead, help them see the situation from a growth-oriented perspective. )

Figure 5.6: Specialized prompt exclusively applied to Appraisal CoT–tuned models for context-aware emotional reasoning.

## 5.5 Automated Evaluation of the Synthetic Appraisal CoT Dataset

We sampled 32 user inputs as a subset of the gold-standard set that we had already used to conduct synthetic augmentation. For each of these inputs, we generated optimal assistant response and appraisal steps that specifically corresponded to the structure of our simulated data. Subsequently, we automatically evaluated our produced data against that of the cohort of these thirty-two hand-crafted responses. Specifically, we quantified the match between every synthetic reply and the ideal reply, thereby assessing the completeness of the synthetic data based on the quality of the content in the professionally created examples.

**BLEU Score Analysis**

| Metric | Score |
|---|---|
| BLEU-1 | 0.5375 |
| BLEU-2 | 0.3950 |
| BLEU-3 | 0.3174 |
| BLEU-4 | 0.2605 |

Table 5.1: Sentence-Level Average BLEU Scores

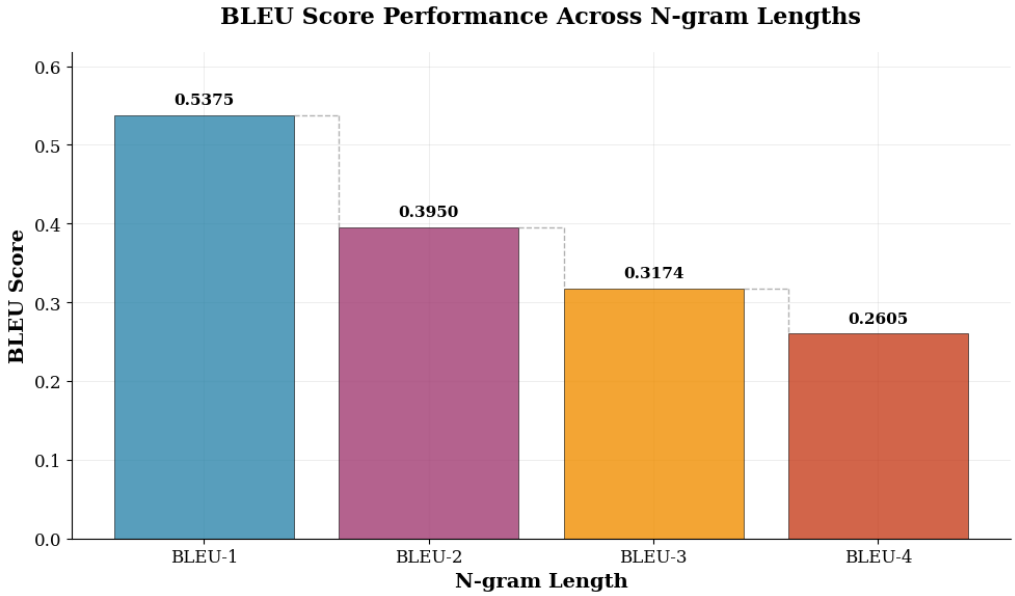


Figure 5.7: BLEU score performance showing characteristic degradation pattern

As expected in text generation evaluation, the degradation characteristic is exhibited by the BLEU scores at **Figure 5.7**, which show a monotonic increase in the length of n-grams. From **Table 5.1** BLEU-1 (0.5375) indicates moderate unigram overlap between the synthetic dataset and the gold-standard reference, suggesting that the realistic match of individual words in the two is approximately 54%. This gradual weakening by BLEU-2 (0.3950), BLEU-3 (0.3174), and BLEU-4 (0.2605) shows how the mental capacity of the text in keeping the longer n-grams is at stake. BLEU-4 (0.2605) is rather low yet does not reject the premises of highly involving reasoning processes when the matching of phrases becomes more challenging by the complexity in the expression of cognitive appraisal processes.

### BERTScore Analysis

| Metric | Score |
|---|---|
| Precision | 0.7645 ± 0.0383 |
| Recall | 0.7444 ± 0.0356 |
| F1-Score | 0.7542 ± 0.0355 |

Table 5.2: BERTScore Results

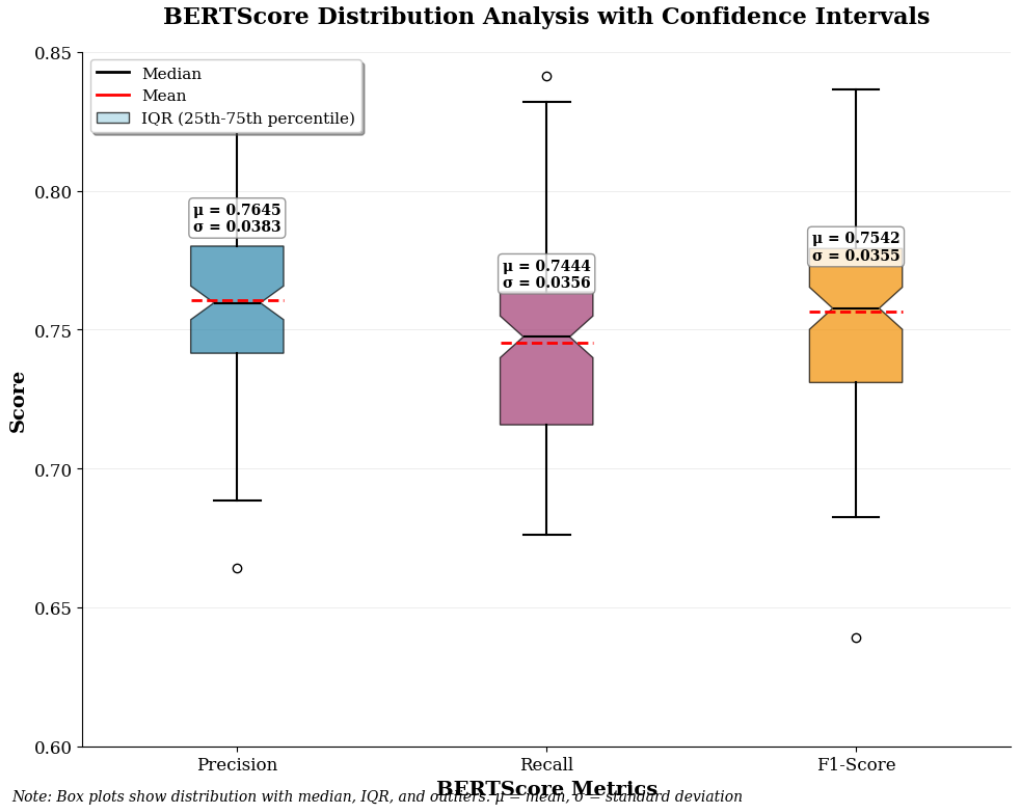


Figure 5.8: BERTScore evaluation results demonstrating semantic similarity performance. $\mu$ = mean, $\sigma$ = standard deviation

The BERTScore metric provides a more exacting analysis of semantically similar performances, yielding promising results across all scores. From the **Table 5.2**, the precision of 0.7645 ($\pm$0.0383) implies that in the synthetic dataset the semantic appropriateness between the reference cognitive appraisal patterns and the synthetic one is very large. The recall of 0.7444 ($\pm$0.0356) indicates that the idea of handling conceptual material found in the gold-standard data was well covered. The F1-score of 0.7542 ($\pm$0.0355) reflects the balanced achievement of both precision and recall aspects, implying that synthetic data effectively captures the meta-linguistic content of cognitive appraisal forms of reasoning, while also reflecting their corresponding coherent language. Moreover, the standard deviations (see **Figure 5.8**) of all the BERTScore metrics are fairly small, indicating good quality across the entire synthetic dataset.

### ROUGE Score Analysis

| Metric | ROUGE-1 | ROUGE-2 |
|---|---|---|
| Precision | 0.5959 | 0.2624 |
| Recall | 0.5386 | 0.2369 |
| F1-Score | 0.5628 | 0.2477 |

Table 5.3: ROUGE-1 and ROUGE-2 Results

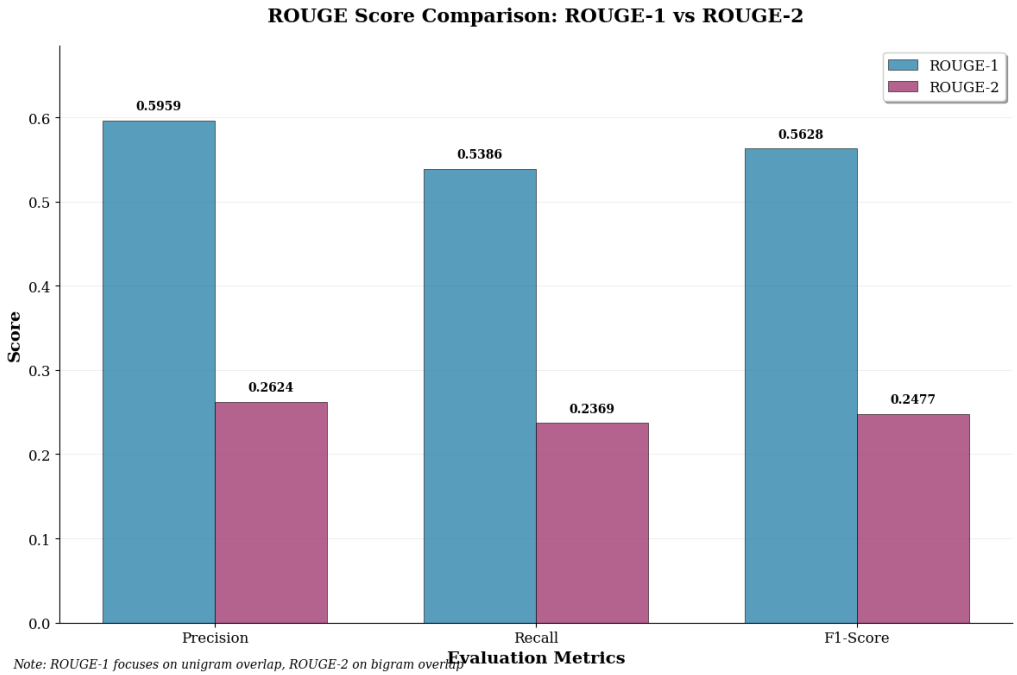


Figure 5.9: Comparison of ROUGE-1 and ROUGE-2 for lexical overlap assessment.

The ROUGE-1 scores reveal that there is moderate lexical overlap with precision of 0.5959, recall of 0.5386 and F1-score of 0.5628 (see **Table 5.3**). These scores suggest that, although the synthetic dataset covers the basic vocabulary linked to the theory of cognitive appraisal, lexical matching still needs improvement. The scores obtained by ROUGE-2 are significantly worse (precision: 0.2624, recall: 0.2369, F1-score: 0.2477), as there is a specific emphasis on bigram overlap. This trend (see **Figure 5.9**) suggests that although individual terms closely match the reference set, the individual sequence matchings of terms are more varied, which can be advantageous in training a variety of reasoning skills for reasoning tasks.

### Diversity Analysis (Distinct-N)

| Metric | Score |
|---|---|
| Distinct-1 | 0.1315 |
| Distinct-2 | 0.4769 |

Table 5.4: Distinct-N Results

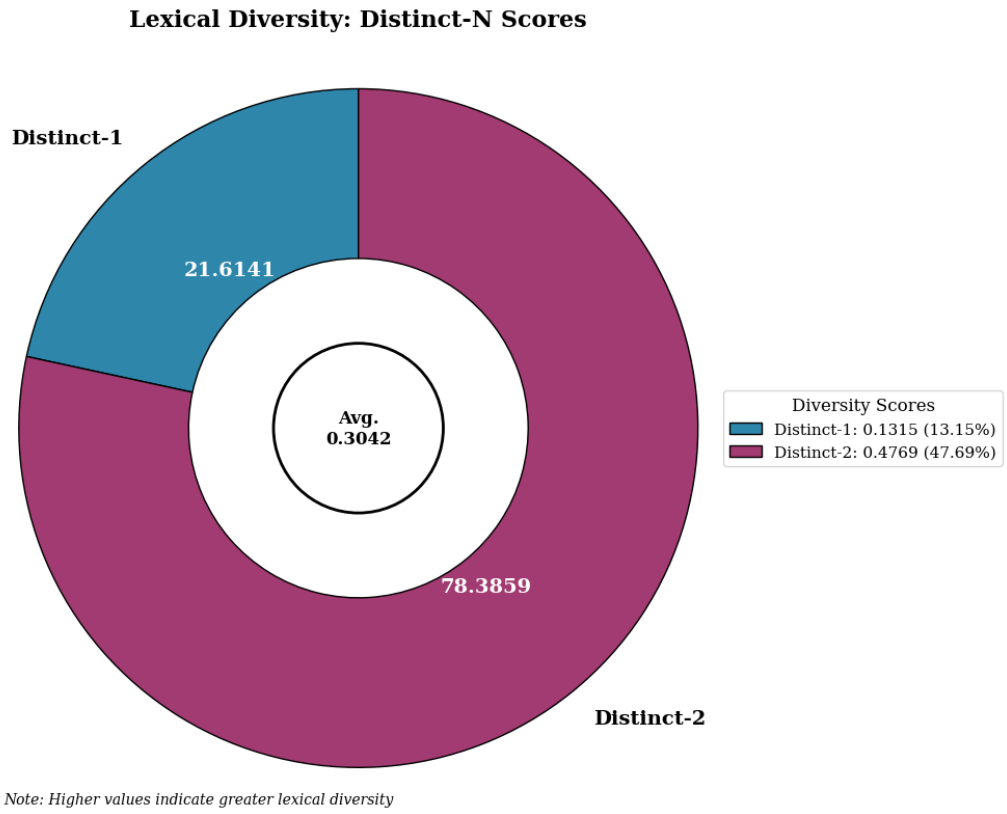


Figure 5.10: Distinct-N score showing lexical variety at unigram and bigram level

The Distinct-N outcomes provide significant information regarding the lexical variety of the manufactured data set (see **Figure 5.10**). From the **Table 5.4**, the Distinct-1 score of 0.1315 implies that approximately 13% of unigrams are unique, indicating that the vocabulary is narrow and may be adapted to the specific field of cognitive appraisal theory. The much greater Distinct-2 score of 0.4769 shows that there is great diversity in bigrams, and almost 48 percent of word pairs are unique. The given pattern is especially well-suited for instruction tuning, as it indicates that the synthetic dataset offers diversity in terms of expression while describing patterns, all while remaining consistent in domain-specific terminology.

A combination of all the evaluation measures shows that the synthetic dataset is not only reasonable in terms of its fidelity to the gold-standard reference but also manages to describe sufficient diversity to allow effective instruction tuning.

## 5.6 Human Evaluation of the Appraisal CoT Dataset

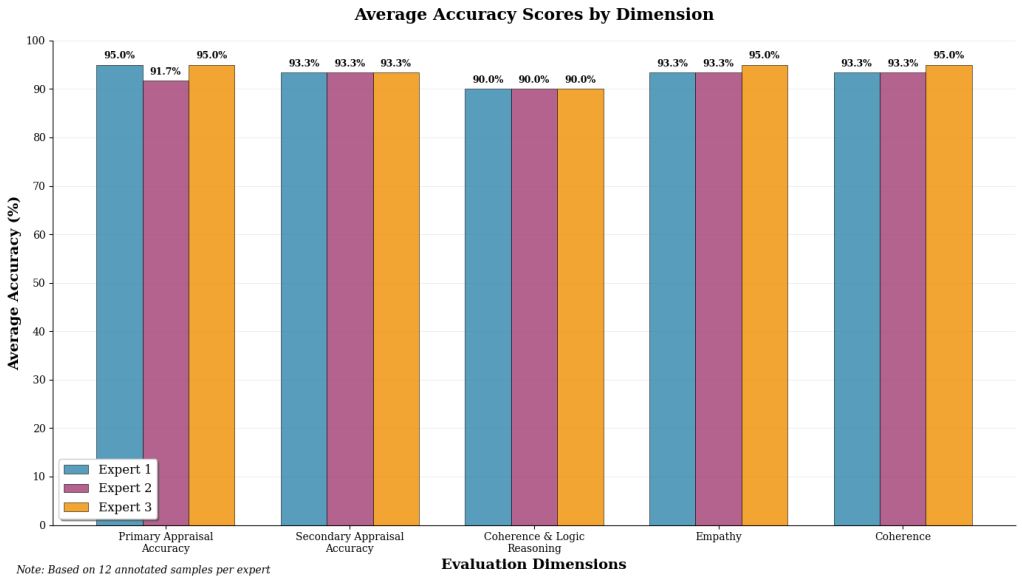


Figure 5.11: Average accuracy of A-CoT dataset in each parameter by three domain experts

The domain experts showed remarkable levels of agreement, sharing an average standard deviation of only 0.63 percent across dimensions. In the **Figure 5.11**, the averages ranged from 92.33% to 93.67%. Coherence and Primary Appraisal Accuracy were found to be the most robust dimensions, with scores of 95%. These findings demonstrate that the A-CoT dataset meets a high quality standard, as judged by domain experts, and can capture the complex nature of cognitive appraisal reasoning that requires expert judgment.

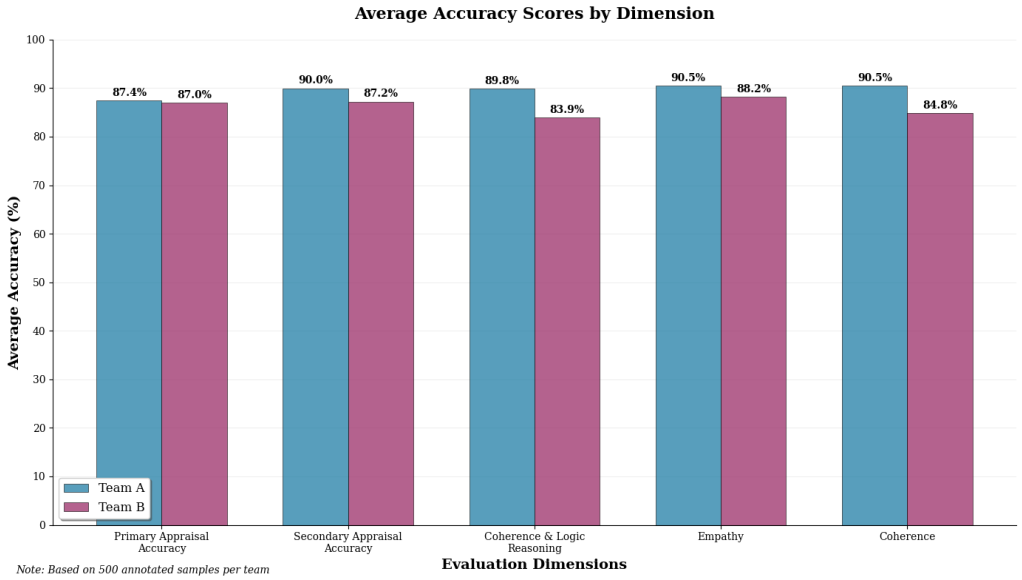


Figure 5.12: Average accuracy of A-CoT dataset in each parameter by two annotation teams

The assessment showed an excellent performance in all five dimensions. The accuracy of primary appraisal had the closest similarity among teams (0.40% point difference), whereas, coherence and logic reasoning had the widest difference (5.92% points). The dimensions of empathy and general coherence scored the highest in absolute terms, with Team A achieving more than 90% in this approach (See **Figure 5.12**). These findings comprise a study that reveals, in both objective and subjective evaluation scales, that the A-CoT dataset is of high quality.

## 5.7 Inter-Annotator Agreement (IAA) Analysis

**Between Domain Experts:**

| Dimension | Krippendorff Alpha | ICC | Spearman $\rho$ |
|---|---|---|---|
| Primary Appraisal | 0.7917 | 0.6106 | 0.8024 |
| Secondary Appraisal | 1 | 1 | 1 |
| Coherence & Logic | 1 | 1 | 1 |
| Empathy | 0.9911 | 0.9209 | 0.9939 |
| Coherence | 0.8727 | 0.8791 | 0.8777 |

Table 5.5: Inter-Annotator Agreement metric results of domain experts.

The IAA of the 3 domain expert annotation outcomes proves to display very high levels of consistency and reliability, which is a synonym for a high level of agreement when assessing the A-CoT dataset. From the **Table 5.5**, Krippendorff's Alpha = 0.931: This surpasses the generally agreed-upon value of 0.800, indicating strong agreement, and suggests that the evaluators were in good consensus in terms of observing similar patterns in the data. This high coefficient indicates the transparency and coherence of the information applied in evaluation. Intraclass Correlation Coefficient (ICC) = 0.882: This value is considered good reliability and is close to the excellent point (>0.900), implying that the absolute agreement between raters is strong. Spearman's Rank Correlation Coefficient ($\rho$) = 0.935: The rank-order correlation coefficient value of 0.89 proves the point that raters had been very consistent in their rank ordering of sample quality, supporting the criterion of reliability of ordinal judgments against different evaluations. In short, the annotation process and the quality of the A-CoT dataset are well-reinforced by these IAA results of domain experts. The high agreement is a sign of overall methodological reliability, which effectively eliminates concerns about subjective bias. This high IAA provides significant evidence of the dataset's quality for use in instruction tuning of large language models (LLMs).

**Between Non-Domain Expert Annotators:**

| Dimension | Krippendorff Alpha | ICC | Spearman $\rho$ |
|---|---|---|---|
| Primary Appraisal | 0.5228 | 0.6323 | 0.5220 |
| Secondary Appraisal | 0.6090 | 0.7184 | 0.6205 |
| Coherence & Logic | 0.4722 | 0.6107 | 0.5307 |
| Empathy | 0.5852 | 0.7488 | 0.6016 |
| Coherence | 0.4790 | 0.6383 | 0.5466 |

Table 5.6: Inter-Annotator Agreement metric results of non-domain experts.

High or medium levels of inter-annotator agreement (IAA) observed in non-expert annotators (see **Table 5.6**) are not only expected but also methodologically adequate, given the task's specifics and the study design. The evaluated dataset's dimensions, such as empathy and coherence, are subjective. Hence, opinions regarding emotional tasks may vary even among experts. Furthermore, as discussed above, non-domain expert annotators tend to have more disagreements. Therefore, according to the **Table 5.6**, moderate agreement (Krippendorff Alpha = 0.53) corresponds to the real challenge of implementing the technical appraisal standards without domain expertise. Furthermore, an ICC of 0.60-0.75 has often been seen as an indicator of moderate to good agreement, showing that the raters reliably employed the criteria and generated the same ratings when working through hundreds of cases. Similarly, a Spearman's Rho greater than 0.50 indicates a significant, positive rank-order correlation between the judgments of the annotators, demonstrating that most of them agree on the relative importance of sample quality. Collectively, these measurements confirm that non-expert annotators consistently recorded clear and reliable judgments, thereby creating a solid dataset upon which downstream analyses and model development can occur.

# Chapter 6

# Experimental Analysis

## 6.1 Experimental Setup: Dataset Selection, Fine-Tuning Framework, and Training Configuration

### 6.1.1 Fine tuning

To bring pre-trained LLMs to a particular project, they are often fine-tuned by training on labeled data specific to the target task. For instance, OpenAI launched GPT-3, which demonstrated that it can be fine-tuned on a substantial amount of task-specific information (or prompts) to achieve high scores on a range of NLP tests [28]. It was demonstrated that by fine-tuning a 137B-parameter LLM on a combination of 60 varied NLP tasks (whose description was provided in the form of natural-language instructions) - an effort that they term FLAN - they were able to significantly improve the zero-shot results of the model on unseen tasks [30]. Despite such achievements, in the service of a given user, conventional fine-tuning may nonetheless leave a gap between the internal training target of a model and the user's natural language target. Specifically, it is observed that pre-trained LLMs provide token predictions, whereas most end users assume the model takes instructions conveyed using natural language.

### 6.1.2 Instruction Tuning

Instruction tuning is a finer version of fine-tuning to fill this gap. With instruction tuning, a pre-trained large language model (LLM) is retrained (supervised) using a set of instruction-output pairs, each consisting of a natural language prompt or instruction and the corresponding desired response. According to research, this paradigm is an effective method for LLMs to enhance their capabilities and controllability, enabling them to comply with human-style instructions and become more similar to human brain processing [34], [58]. The point is that the model will learn to understand and implement the user orders more precisely with the help of training on examples formulated in the form of instructions. Empirically, it was discovered that the effect on generalization was strong: A group of researchers compared their instruction-tuned model (FLAN) to a significantly stronger GPT-3

model and demonstrated that, on most held-out tasks, FLAN outperformed GPT-3 by a substantial margin [30]. The same can be said about InstructGPT, which fine-tunes GPT-3 using instruction-response pairs written by humans and noticed significant improvements in the quality and safety of model generations [32]. In simplified terms, the instruction tuning is fine-tuning of human-format prompts that can make the model more initiated to assistance, truthfulness, and user intentions.

Chain-of-thought (CoT) is an instructional method that deliberately encourages the model to present some intermediate steps in their reasoning before providing an output. A study demonstrated that adding examples of CoT to prompts resulted in a significant increase in performance on math, common sense, and logical reasoning tasks. CoT-prompted instruction-tuned models can be more open to the CoT prompts, and recent work has shown that including CoT-style explanations in the fine-tuning data, in addition, provides even greater improvements in reasoning [33]. Indeed, by cloning the cues of CoT reasoning in training examples, even on smaller language models (LMs), it is possible to train step-by-step thinking and, to some extent, reduce the capabilities of more advanced models to those of smaller ones. With chain-of-thought instructions integrated into fine-tuning, instruction-tuned LLMs learn to perform multi-step reasoning and produce structured outputs to a greater extent.

### 6.1.3 Preprocess Dataset for Instruction-Tuning

We have proposed a noble synthetic dataset named Appraisal CoT. This dataset had a total 4,489 single turn conversations between user and assistant along with the system prompt. Figure **Figure 6.1** showcasing our dataset format. Now , our target was to instruction tune the basic chat models of Qwen and Gemma to enhance their emotion generation capability. As a result the dataset needed to be formatted in a way the model is expecting.

```
"messages:" : [
    {
    "content": "You are an expert in emotional psychology and you can accurately assess people's emotional states and help them process difficult situations using cognitive appraisal theory. Understand the listener's situation, follow the listener's point of view, and respond empathetically using primary appraisal, secondary appraisal, and reappraisal processes. Avoid dismissing their feelings, and instead, help them see the situation from a growth-oriented perspective.",
    "role": "system"
    },
    {
    "content": "we tried for a long time to get pregnant.",
    "role": "user"
    },
    {
    "content": "Performing appraisal_cot:\n\nStep 1: The customer shares their experience of trying for a long time to get pregnant.\n\nStep 2: They are describing a significant and emotionally charged journey related to conception, which likely involves feelings of hope, anticipation, and possibly frustration or disappointment.\n\nStep 3 (Primary Appraisal):\n- Goal relevance: High - The situation is directly related to their personal and family goals of starting a family.\n- Goal congruence: Low - The prolonged struggle to conceive is obstructing their goal of becoming parents, leading to feelings of inadequacy.\n- Ego involvement: High - This situation deeply impacts their self-esteem, values related to family, and social identity as potential parents.\n\nStep 4 (Secondary Appraisal):\n- Coping potential: Moderate - They may feel some options are available, such as seeking medical advice or support, but may also feel overwhelmed by the situation.\n- Future expectancy: Negative - They might feel uncertain or pessimistic about their chances of conceiving in the future, given the length of time they've been trying.\n- Perceived control: Negative - They likely feel a lack of control over their ability to conceive, which can lead to feelings of helplessness\n\nStep 5: The customer's mental state is likely one of frustration, sadness, and anxiety due to the prolonged effort to conceive without success. These emotions stem from the perceived blockage of their personal goals and the impact on their self-identity.\n\nStep 6 (Reappraisal):\n- There may be an opportunity for reappraisal by exploring new avenues for support or treatment, as well as reframing the situation to focus on what they can control and the support they can seek.\n\nStep 7: Acknowledge the difficulty of their journey, validate their emotions, and suggest exploring options for support or new approaches to their situation.\n\nResponse:\nI can only imagine how challenging and emotionally draining this journey has been for you. Trying for a long time to get pregnant can lead to feelings of frustration and sadness, especially when it seems like your goal is just out of reach. It's completely understandable to feel overwhelmed. Please know that you're not alone in this, and there are options and support available to help you navigate this difficult time. Exploring new avenues or seeking support can sometimes provide a renewed sense of hope. I'm here to listen and support you in any way you need.",
    "role": "assistant"
    },
```


Figure 6.1: Appraisal CoT dataset format

For Qwen3 models, `<think> </think>` token is introduced to enhance its reasoning responses. Also, the final response that the model gives is given in a `\boxed{}` format [69].

Qwen3 also allows it to be finetuned on non-reasoning dataset. Since our dataset is a CoT dataset for explicit instruction following we have chosen to format the dataset as a reasoning dataset.

---

***chat template for Qwen3***

---

```
<|im_start|>{system Prompt} <|im_end|>
<|im_start|>{User Input} <|im_end|>
<|im_start|><think>{Assistant response}</think><|im_end|>
boxed{Final Response}
```

Now, for Gemma3 models we had to format the dataset differently. The `<start_of_turn>` token is used as the beginning of a user/assistant turn, and `<end_of_turn>` is used as an ending of the token.

---

***chat template for gemma3***

---

```
<bos><start_of_turn>{system_prompt}{user}
<end_of_turn>
<start_of_turn>model<end_of_turn>
```

In the case of **Qwen2** series, we see a different chat template. Here, `<|im_start|>` token is used at the beginning of user/Assistant?system turn, and `<|im_end|>` at

the end.

---

***chat template for Qwen2***

---

```
<|im_start|>{system Prompt} <|im_end|>
<|im_start|>{User Input} <|im_end|>
<|im_start|>{Assistant response}<|im_end|>
```

After formatting these texts, the users and system prompt is masked so that Only the assistant's part of the conversation is used for loss computation, i.e., gradients are calculated only from the assistant's responses.This prevents the model from learning from the user input (so it doesn't try to predict user messages). Eventually, it is sent to the trainer for training.

### 6.1.4 Implementation Details

#### 6.1.4.1 Baseline Model Selection

For our experiment we have chosen different variants of Qwen and Gemma. The reason behind is that, eventually the Emobench benchmark we were going to test on suggested that among open source models Qwen outperformed every other model [52]. And Gemma was brought on as a comparison with other models. So as a baseline, we have chosen Qwen3 0.6B, Qwen3 1.7B , Qwen3 4B , Qwen3 8B, Qwen3 14B, Qwen2.5 7B instruct , Qwen2.5 14B instruct , Qwen2 7B instruct and Gemma3 4B it. Also, all of these models are instruct models or also known as chat models . The reason behind choosing such models is our format of experiment. As our goal was to train models to follow instruction in a certain way , thus the instruct models became the best options.

#### 6.1.4.2 Fine-Tuning with Unsloth: Framework, Hyperparameter Configuration, and Training Environment

All of our experiment were ran on Nvidia RTX A6000. Although smaller GPUs can also be used to run these experiments. We fine-tuned the all models using parameter-efficient training through Low-Rank Adaptation (LoRA) implemented via the Unsloth framework. The base model was loaded with 4-bit quantization to optimize memory usage, with a maximum sequence length of 2,048 tokens. LoRA was applied with a rank (r) of 32 and alpha value of 32, targeting the attention projection layers (q_proj, k_proj, v_proj, o_proj) and feed-forward network projections (gate_proj, up_proj, down_proj) of the transformer architecture. No dropout was applied to LoRA layers, and bias terms were excluded for computational efficiency.

The training configuration employed a per-device batch size of 2 with 4 gradient accumulation steps, yielding an effective batch size of 8. The model was trained for 3 epochs using the AdamW optimizer with 8-bit precision, a learning rate of $2 \times 10^{-5}$, and linear learning rate scheduling with 5 warm up steps. Weight decay was set to

0.01 to prevent overfitting. Gradient checkpointing with Unsloth's optimized implementation was enabled to further reduce memory consumption during training. The random seed was fixed at 3407 to ensure reproducibility across experimental runs.

In case of Qwen3 0.6B acot LoRA was applied with a rank (r) of 64 and alpha value of 64. And the model was trained for 4 epochs. Rest of the hyperparameters were the same as above.

Qwen3 0.6B acot, Qwen3 0.6B, Qwen3 1.7B acot, and Qwen3 1.7B were instruction tuned on the Demo Emobench dataset. The hyperparameters were same except - LoRA was applied with a rank (r) of 16 and alpha value of 16. And the model was trained for 2 epochs. Rest of the hyperparameters were the same as above.

After tuning all of our models, we have received these models - Qwen3 0.6B acot Emobench, Qwen3 1.7B acot Emobench, Qwen3 4B acot, Qwen3 8B acot, Qwen3 14B acot, Qwen2.5 7B instruct acot, Qwen2.5 14B instruct acot, Qwen2 7B instruct acot, and Gemma3 4B it acot.

## 6.2 Emobench Benchmark Results

The Emobench dataset by Sabour and his team was developed in both the English and Chinese languages. However, in the case of performing these experiments, we have only chosen the results of the English (en) language as our models were only turned on an English dataset. We have not used any cot for benchmarking because, in the original paper, the cot has little to no difference in model performance [52].

Increasing parameters resulted in enhanced performance, similar to earlier LLM scaling studies [28]. LLMs found emotional understanding more difficult than application. This is for several reasons. Unlike the EA task, the EU samples demand LLMs to correctly answer two questions (the emotion and its origin), which presents a greater challenge [52].

**Table 6.1** presents the values of the EU for various task segments (Complex Emotions, Emotional Cues, Personal Beliefs and Experiences and Perspective Taking) performed on different Chat models and instruction tuned on Appraisal CoT (acot) dataset models. Instruction tuning Chat models (instruct models) with our proposed Appraisal CoT has significantly increased the EU and EA of Qwen3 4B, Qwen3 8B, Qwen2.5 7B instruct, Qwen 2 7B instruct, and Gemma3-4B-it models. Among them, Qwen3 4b acot has outperformed its base instruct model (Qwen3 4B) in both tasks.

Qwen3 4B's absolute jump from 17.0% to 23.5% (+6.5%) is the single most significant EU gain. Qwen3 4b acot showcases an increase in all segments of MCQs from the EU. Like Qwen3 4B acot, Gemma3 4B acot gains (+3.0%), suggesting that the combination of moderate scale plus an appraisal CoT is a general strategy for lifting EU. This underscores that mid-tier capacity models benefit most from our dataset.

Furthermore, consistent improvements can be observed in Qwen3 8B acot (+2.5%) and Qwen2.5 7B acot (+2.0%), though less dramatically. Qwen3 8B acot showcases improvement in PBE and PT segments. On the other hand, the Qwen2.5 7B acot improves in EC and PT. On another note, Gemma3 shows improvement in CE and PBE. It suggests that the Appraisal CoT dataset does not necessarily improve a specific segment of a model's Emotional Intelligence. Rather, it improves the segments which require improvement. These medium-high capacity models have adequate plasticity to refine their emotion attributions.

Additionally, a slight decline in the EU can be observed in both Qwen3 14B (-2.5%) and Qwen2.5 14B (-2.0%). Their powerful pre-training may already capture most of the EU, and any fine-tuning without more concentrated data risks overwriting proper priors. **Figure 6.2** is showcasing EU for both the baseline and the acot models.

| | Model | Complex Emotions (CE) | Emotional Cues (EC) | Personal Beliefs and Experiences (PBE) | Perspective Taking (PT) | Overall |
|---|---|---|---|---|---|---|
| Our (+bias) | **Qwen3 0.6B acot emobench** | 10.20 | 0 | 7.14 | 10.44 | 8 |
| | **Qwen 3 0.6B Emobench** | 16.32 | 7.14 | 14.28 | 10.44 | 12.5 |
| Our (+bias) | **Qwen_3_1.7B_acot_Emobench** | 22.44 | 21.42 | 19.64 | 8.95 | 17 |
| | **Qwen3 - 1.7B-emobench** | 22.44 | 25 | 21.42 | 8.95 | 18 |
| Our | **Qwen 3 4b acot** | 32.65 ↑ | 35.71 ↑ | 17.85 ↑ | 16.41 ↑ | 23.5 ↑ |
| | **Qwen 3 4b** | 26.53 | 28.57 | 10.71 | 10.44 | 17 |
| Our | **Qwen3 8b acot** | 32.65 | 50 | 26.78 ↑ | 19.40 ↑ | 29 ↑ |
| | **Qwen3 8b :** | 32.65 | 50 | 19.64 | 17.91 | 26.5 |
| Our | **Qwen3 14B acot** | 55.10 | 50 | 37.5 | 38.80 | 44 |
| | **Q**wen 3 14 B | 57.14 | 53.57 | 41.07 | 40.29 | 46.5 |
| Our | **Qwen 2.5 7B instruct acot** | 36.73 | 35.71 ↑ | 26.78 | 28.35 ↑ | 31 ↑ |
| | **Qwen 2.5 7B instruct** | 40.81 | 32.14 | 26.78 | 20.89 | 29 |
| Our | **Qwen 2.5_14B_Instruct acot** | 63.26 | 42.85 | 28.57 | 32.83 | 40.5 |
| | **Qwen 2.5_14B_Instruct** | 61.22 | 46.42 | 35.71 | 32.83 | 42.5 |
| Our | **Qwen 2 7b instruct acot** | 48.97 | 39.28 | 23.21 | 17.91 | 30 |
| | **Qwen2 7B instruct** | 46.94 | 39.29 | 26.79 | 23.88 | 34.22 |
| Our | **Gemma 3 - 4b - it -acot** | 36.73 ↑ | 39.28 | 21.42 ↑ | 17.91 | 26.5 ↑ |
| | **Gemma 3 -4b-it** | 30.61 | 39.28 | 16.07 | 17.91 | 23.5 |

Table 6.1: Emotional Understanding (EU) Accuracy (%)

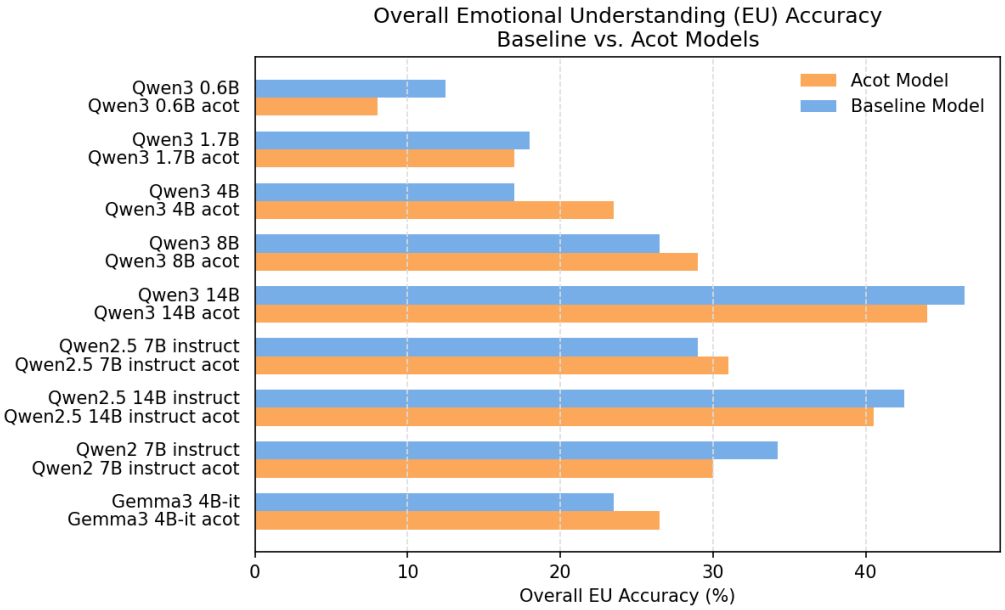


Figure 6.2: Emotional Understanding Accuracy ( % )

Table 6.2 is showcasing the values of EA for different segments of tasks (Personal-Self, Personal-Other,Social-Self and Social-Other and Perspective Taking) performed on different Chat models and instruction tuned on Appraical CoT (ACoT) dataset models. In the case of EA, both Qwen3 1.7B acot Emobench (bias) and Qwen3 4B acot exhibit the most significant gains (+4.5% and +4.0%). This indicates that instruction-tuning with the Appraisal CoT dataset is most effective on mid-sized models, where there is sufficient capacity to internalize the new patterns without overfitting or disregarding the tuning signal.

The Qwen2 7B instruct acot model shows a slight improvement (+2.5%). Instruction tuning improves EA, but not as much as mid-scale configurations. Qwen2.5 7B instruct acot and Qwen2.5 14B instruct acot see neutral or negative changes. Moreover, Qwen3 8B and 14B show slight drops (-5.0% and -0.5%) in EA. Their stronger pre-trained priors may resist additional adaptation on EA, implying that Instruction-tuning with Appraisal CoT delivers minimal benefit once a certain parameter barrier is crossed.

| | Model | Personal-Self | Personal-Other | Social-Self | Social-Other | Overall |
|---|---|---|---|---|---|---|
| Our (+bias ) | **Qwen3 0.6B acot emobench** | 54 | 40 | 52 | 46 | 48 |
| | **Qwen 3 0.6B Emobench** | 52 | 38 | 56 | 46 | 48 |
| Our (+bias) | **Qwen_3_1.7B_acot_ Emobench** | 68 ↑ | 46 ↑ | 58 ↑ | 52 ↑ | 56 ↑ |
| | **Qwen3 - 1.7B-emobench** | 64 | 40 | 56 | 46 | 51.5 |
| Our | **Qwen 3 4b acot** | 70 | 58 ↑ | 58 ↑ | 56 | 60.5 ↑ |
| | **Qwen 3 4b** | 78 | 30 | 56 | 62 | 56.5 |
| Our | **Qwen3 8b acot** | 74 | 52 | 64 | 56 | 61.5 |
| | **Qwen3 8b** | 80 | 58 | 72 | 56 | 66.5 |
| Our | **Qwen3 14B acot** | 78 | 66 | 66 | 68 | 69.5 |
| | **Qwen 3 14 B** | 74 | 68 | 74 | 64 | 70 |
| Our | **Qwen 2.5 7B instruct acot** | 74 | 56 | 70 | 62 | 65.5 |
| | **Qwen 2.5 7B instruct** | 78 | 62 | 74 | 60 | 68.5 |
| Our | **Qwen 2.5_14B_Instruct acot** | 76 | 60 | 70 | 70 | 69 |
| | **Qwen 2.5_14B_Instruct** | 74 | 62 | 66 | 74 | 69 |
| Our | **Qwen 2 7b instruct acot** | 74 | 60 | 68 | 62 ↑ | 66 ↑ |
| | **Qwen2 7B instruct** | 74 | 62 | 68 | 5 | 63.5 |
| Our | **Gemma 3 - 4b - it -acot** | 76 | 50 | 58 | 50 | 58.5 |
| | **Gemma 3 -4b-it** | 72 | 52 | 66 | 58 | 62 |

Table 6.2: Emotional Application (EA) Accuracy (%)

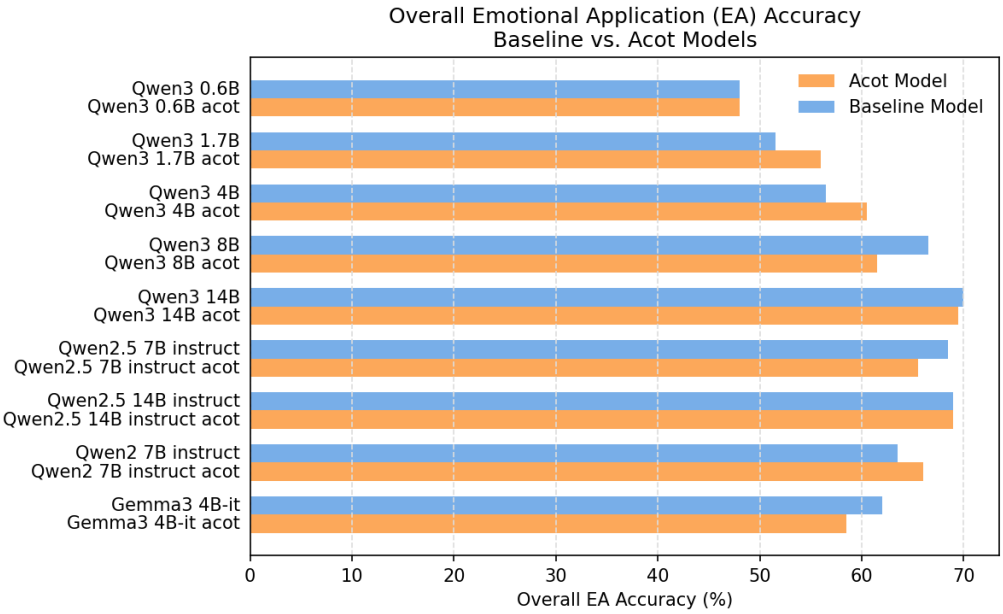


Figure 6.3: Emotional Application Accuracy (%)

**Figure 6.3** is showcasing EA accuracy score for both the baseline and the ACoT models.

## 6.3 Automated Evaluation Metrics: ROUGE, BLEU, and Distinct Scores

The reason for using automated evaluation metrics is to determine how closely the ultimate response aligns with our evaluation dataset. Which effectively describes how empathetic our models have become after Instruction tuning.

For evaluation, we have chosen our best-performing models. Models that have performed well in the EU or EA are Instruction-tuned versions of Qwen3-1.7B, Qwen3-4B, Qwen3-8B, Qwen2-7B-instruct, Qwen2.5-7B-instruct, and Gemma3-4B-it. We will compare these models with the vanilla models.

We have synthetically generated an evaluation dataset that perfectly catches the essence of CAT. We have created the evaluation dataset in the same manner as the original Acot dataset. The evaluation dataset consists of 675 items. Now, as we will be comparing our acot model's performance in those metrics with the Base model, we have modified our evaluation dataset. We have retained only the final response from the ACOT dataset to ensure fairness for the Base models, as they will not be able to generate the ACOT reasoning part as well, which might create unfair results. **Figure 6.4** is displaying the evaluation dataset format.

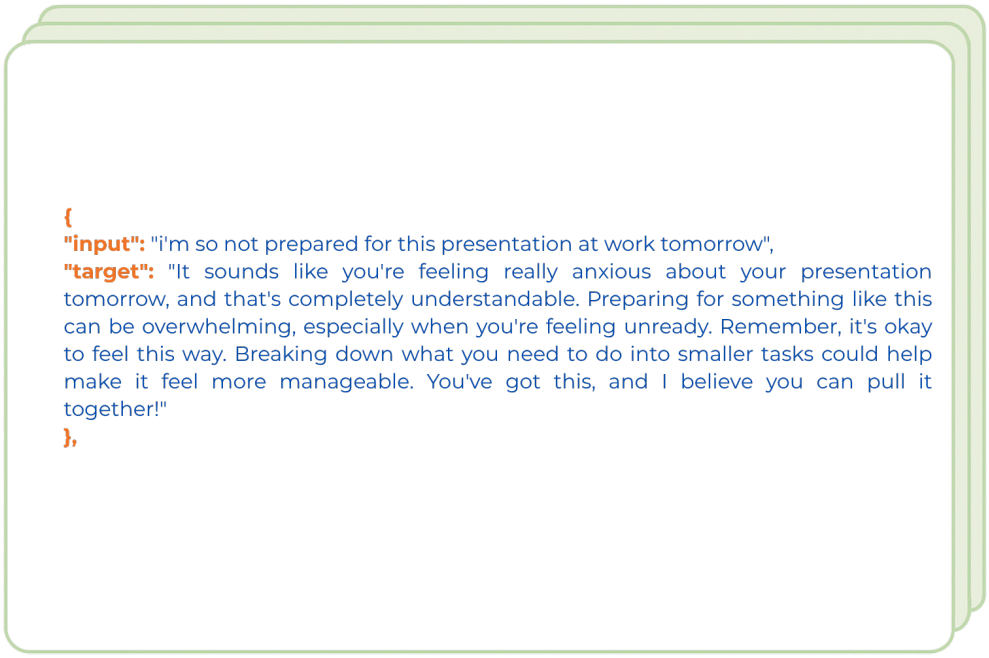


Figure 6.4: A sample entry from Evaluation dataset

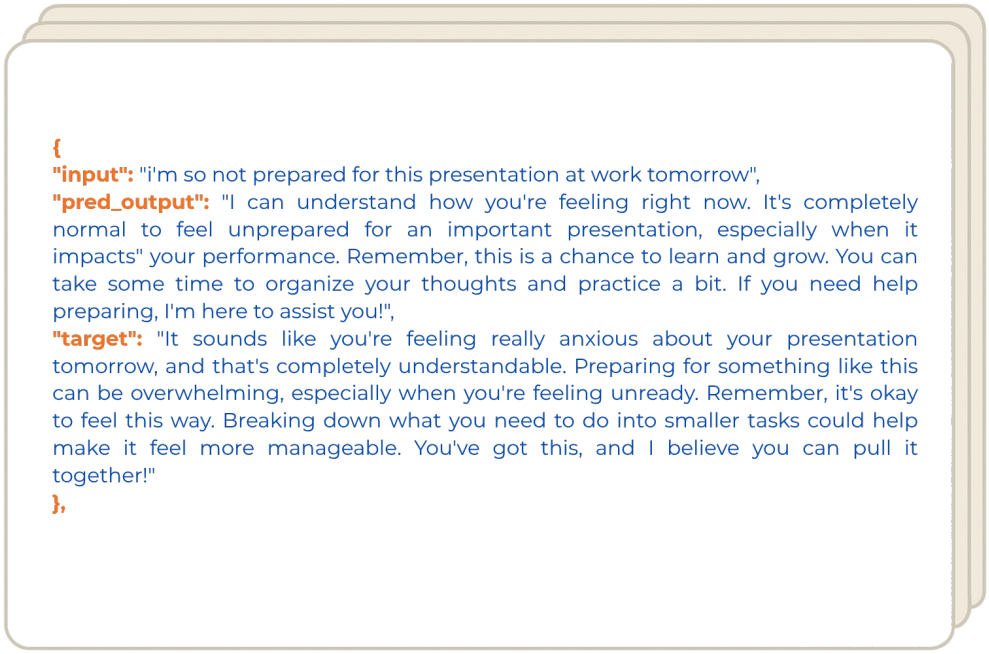


Figure 6.5: Example entry from the model evaluation dataset generated by Qwen3 4B-ACoT, illustrating emotionally tuned output aligned with the Appraisal CoT framework.

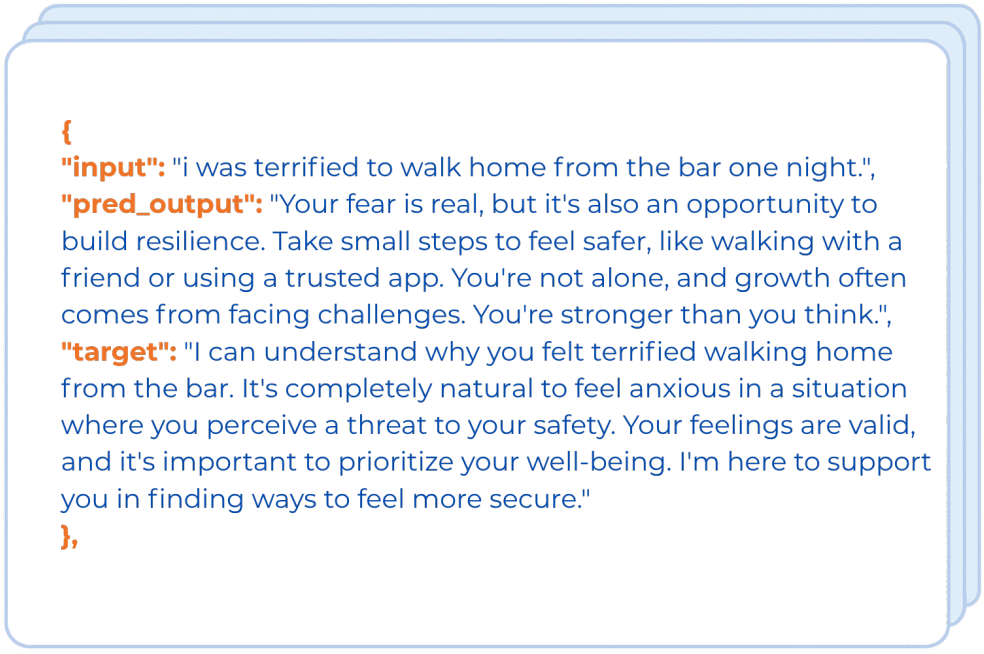


Figure 6.6: Corresponding example from the baseline Qwen3 4B model, demonstrating output prior to Appraisal CoT instruction tuning.

The pred_output has the value of the model giving output for that user input. Moreover, the target is the output from the evaluation dataset, specifically the final response value corresponding to that user input. **Figure 6.5** is showing the evaluation dataset sample entry.

To evaluate these parameters, we have modified the dataset by taking only the user input and the Final response as output. The entire thought process of Acot was not taken into consideration, as it would not be fair to the base model. Our acot model and the Baseline model were given the same instruction prompt. Furthermore, a final response was requested in 80 words. So that the evaluation becomes valid. **Figure 6.6** is showing sample entry from evaluation dataset of base model.

### 6.3.1 Inference Configuration

The configuration while generating the outputs of both the Baseline models and the acot models are:

#### 6.3.1.1 Inference Configuration for Qwen3 Models

For model inference, input messages were formatted using the tokenizer's chat template with generation prompts enabled and thinking mode activated to leverage the model's reasoning capabilities. Text generation was performed with a maximum output length of 1024 new tokens. Sampling parameters were configured with a temperature of 0.6 to balance creativity and coherence, top-p nucleus sampling of 0.95 to focus on the most probable token distributions, and top-k sampling limited to the 20 most likely tokens at each step. All inference was conducted on CUDA-enabled GPUs with real-time text streaming to monitor generation progress.

#### 6.3.1.2 Inference Configuration for Qwen2.5 Models

For model inference, input messages were formatted using the tokenizer's chat template with generation prompts enabled. Text generation was performed with a maximum output length of 1024 new tokens in sampling mode, using a temperature of 1.5 to promote diverse and creative outputs and top-p nucleus sampling of 0.9 to maintain response quality while allowing broader token selection. Gradient computation was turned off during inference for computational efficiency, and model caching was enabled to accelerate sequential token generation. The padding token was set to the end-of-sequence token to ensure proper sequence termination.

#### 6.3.1.3 Inference Configuration for For Gemma3 models

For model inference, input messages were formatted using the tokenizer's chat template with generation prompts enabled. Text generation was performed with a maximum output length of 800 new tokens under sampling mode with a temperature of 1.0 for balanced randomness, top-p nucleus sampling of 0.95 to include most probable tokens while maintaining diversity, and top-k sampling limited to the 64 most likely tokens at each generation step. Model caching was enabled to accelerate sequential token generation, and the padding token was set to the end-of-sequence token to ensure proper sequence termination.

Now, we want to test our models using BLEU metrics. BLEU-1, BLEU-2, BLEU-3, BLEU-4, ROUGE-1, ROUGE-2, Distinct-1, Distinct-2.

Table-6.3 is demonstrating the BLEU score (B-1, B-2, B-3, B-4) and the Distinct score (D-1, D-2). Moreover, the model column contains our instruction-tuned model and the Baseline model. While evaluating BLEU (n-grams), we calculate the literal matching of word sequences (n-grams) with the reference. It describes how closely our model's output matches the responses in the evaluation dataset. Similarly, BLEU (n-grams) is being calculated for the Baseline models. Thus, giving us an analogy about how well our model is generating compared to the Evaluation dataset.

Additionally, for baseline models, it displays how those models would have initially responded in comparison with the ideal response (evaluation dataset). Its scale is 0-1. The higher score indicates that the model frequently generates responses consistent with those found in the references. All of our models, instruction tuned on the Appraisal CoT dataset, showcased dramatically higher performance than the Baseline models. For example, the Qwen3-8B-acot achieves 0.5316, and Qwen3-8B achieves 0.3222 in the BLEU-1 score, displaying a 65% improvement in BLEU-1. Consistent improvements across all n-gram levels indicate learning at both word and phrase levels. This means that our instruction-tuned models produce replies that are significantly more similar to our intended emotional responses, implying that they learned to produce the exact language patterns required for empathic communication. Its limitation is that it penalizes valid rephrasings; utilizing synonyms or changing word order may result in a lower score, even if the meaning is preserved.

The values of Distinct-1 (D-1) and Distinct-2 (D-2) can also be observed in **Table 6.3**. Distinct (n-grams) usually express the ratio of unique n-grams to total n-grams. It depicts vocabulary diversity and creativity. It helps to detect repetitive or overly formulaic responses. It cannot be too low as it would indicate repetitive words being used.

On the other hand, it also cannot be too high as it would suggest incoherence in response. That is why a balance is needed. The limitation of Distinct is that the model can generate random words and achieve high diversity scores. Distinct does not guarantee relevance. Our instruction-tuned ACoT models have slightly lower Distinct-1 values but mixed Distinct-2 results compared to baseline models. This implies that they utilize more targeted, suitable terminology in emotional settings rather than random variation. This is particularly beneficial because emotional support often requires consistent, warm language rather than creative diversity. **Figure 6.7** and **Figure 6.8** is showcasing the BLEU and Distinct scores.

| | Model | B-1 | B-2 | B-3 | B-4 | D-1 | D-2 |
|---|---|---|---|---|---|---|---|
| Our | **Qwen3-1.7B-acot** | 0.4750 | 0.3127 | 0.2274 | 0.1682 | 0.0606 | 0.3112 |
| | **Qwen3-1.7B** | 0.2491 | 0.1073 | 0.0551 | 0.0300 | 0.0726 | 0.3597 |
| Our | **Qwen3-4B-acot** | 0.4905 | 0.3352 | 0.2527 | 0.1939 | 0.0612 | 0.3255 |
| | **Qwen3-4B** | 0.2563 | 0.1104 | 0.0554 | 0.0292 | 0.0673 | 0.3779 |
| Our | **Qwen3-8B-acot** | 0.5316 | 0.3847 | 0.3038 | 0.2437 | 0.0633 | 0.3394 |
| | **Qwen3-8B** | 0.3222 | 0.1556 | 0.0882 | 0.0519 | 0.0811 | 0.4237 |
| Our | **Qwen2.5-7B-instruct-acot** | 0.4242 | 0.2570 | 0.1761 | 0.1234 | 0.0764 | 0.4258 |
| | **Qwen2.5-7B-instruct** | 0.3011 | 0.1269 | 0.0666 | 0.0369 | 0.0810 | 0.4616 |
| Our | **Qwen2-7B-instruct-acot** | 0.4466 | 0.2897 | 0.2103 | 0.1556 | 0.0760 | 0.4018 |
| | **Qwen2-7B-instruct** | 0.2664 | 0.1276 | 0.0712 | 0.0405 | 0.0733 | 0.5040 |
| Our | **Gemma3-4B-it-acot** | 0.3238 | 0.2227 | 0.1692 | 0.1291 | 0.0479 | 0.3188 |
| | **Gemma3-4B-it** | 0.1948 | 0.0811 | 0.0393 | 0.0201 | 0.0419 | 0.2422 |

Table 6.3: Comparison of BLEU and Distinct Scores Between Instruction-Tuned and Baseline Models

Moving forward to **Table 6.4**, which presents the BERTScore of our instruction-tuned models and baseline models. BERTScore demonstrates the semantic similarity between the output and reference using deep contextual embeddings. It computes precision, recall, and F1 by comparing the similarity of word embeddings. Its scale is 0-1 (higher is better). As BERTScore understands synonyms and paraphrases, not just exact word matches, it is more sophisticated than BLEU. Strong improvements across all models (e.g., Qwen3-8B-acot F1: 0.7555 vs. base: 0.6507) indicate that our models not only memorize exact phrases but also understand the underlying emotional concepts. They can express empathy using varied language while maintaining semantic accuracy. Its limitation is that it depends on the embedding quality and can be fooled by surface-level similarity. **Figure 6.9** is showcasing the BERTScore.

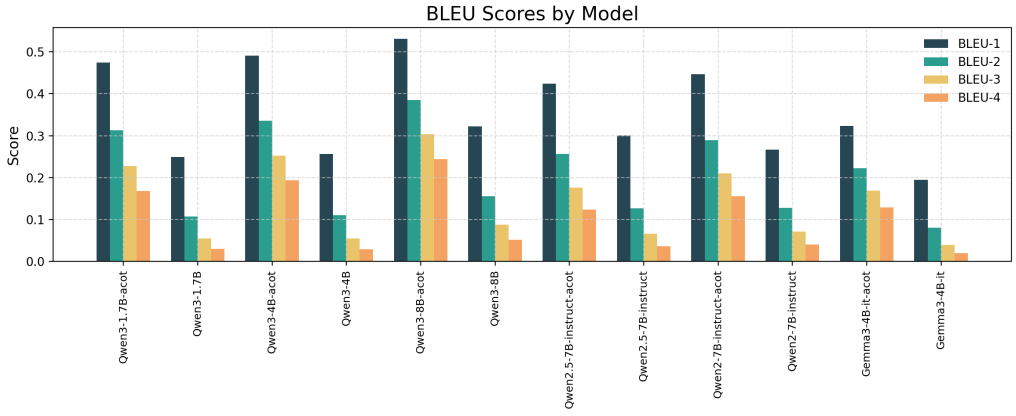


Figure 6.7: BLEU Scores (B-1 to B-4) Comparing output quality with respect to Evaluation dataset Instruction-Tuned and Baseline Models

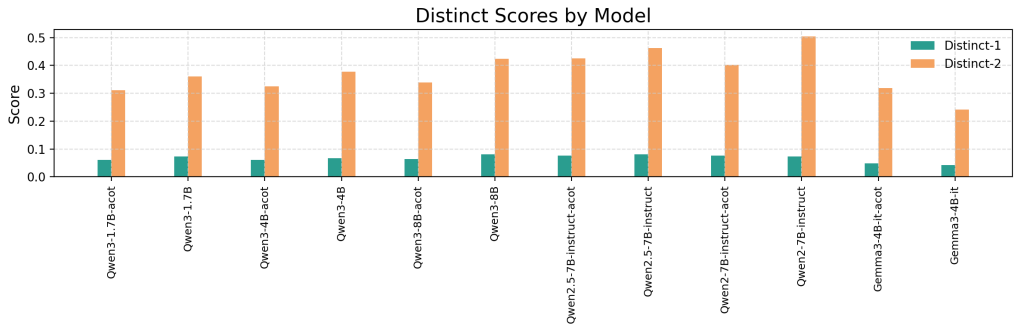


Figure 6.8: Distinct Scores (D-1, D-2) Reflecting Diversity in Model Responses

| | Model | Precision | Recall | F1-Score |
|---|---|---|---|---|
| Our | **Qwen3-1.7B-acot** | 0.7211 ± 0.0572 | 0.7207 ± 0.0537 | 0.7202 ± 0.0515 |
| | **Qwen3-1.7B** | 0.6431 ± 0.0512 | 0.6178 ± 0.0472 | 0.6293 ± 0.0447 |
| Our | **Qwen3-4B-acot** | 0.7356 ± 0.0579 | 0.7371 ± 0.0579 | 0.7357 ± 0.0515 |
| | **Qwen3-4B** | 0.6062 ± 0.0627 | 0.6197 ± 0.0417 | 0.6114 ± 0.0464 |
| Our | **Qwen3-8B-acot** | 0.7553 ± 0.0632 | 0.7572 ± 0.0568 | 0.7555 ± 0.0559 |
| | **Qwen3-8B** | 0.6350 ± 0.0446 | 0.6686 ± 0.0479 | 0.6507 ± 0.0434 |
| Our | **Qwen2.5-7B-instruct-acot** | 0.7032 ± 0.0536 | 0.7020 ± 0.0477 | 0.7017 ± 0.0447 |
| | **Qwen2.5-7B-instruct** | 0.6132 ± 0.0521 | 0.6197 ± 0.0567 | 0.6138 ± 0.0373 |
| Our | **Qwen2-7B-instruct-acot** | 0.7294 ± 0.0545 | 0.7223 ± 0.0509 | 0.7251 ± 0.0476 |
| | **Qwen2-7B-instruct** | 0.6127 ± 0.0430 | 0.6747 ± 0.0418 | 0.6413 ± 0.0360 |
| Our | **Gemma3-4B-it-acot** | 0.6150 ± 0.0810 | 0.7172 ± 0.0482 | 0.6602 ± 0.0610 |
| | **Gemma3-4B-it** | 0.5534 ± 0.0571 | 0.5933 ± 0.0508 | 0.5714 ± 0.0486 |

Table 6.4: BERTScore Evaluation Comparing Semantic Similarity Between Model Outputs and References

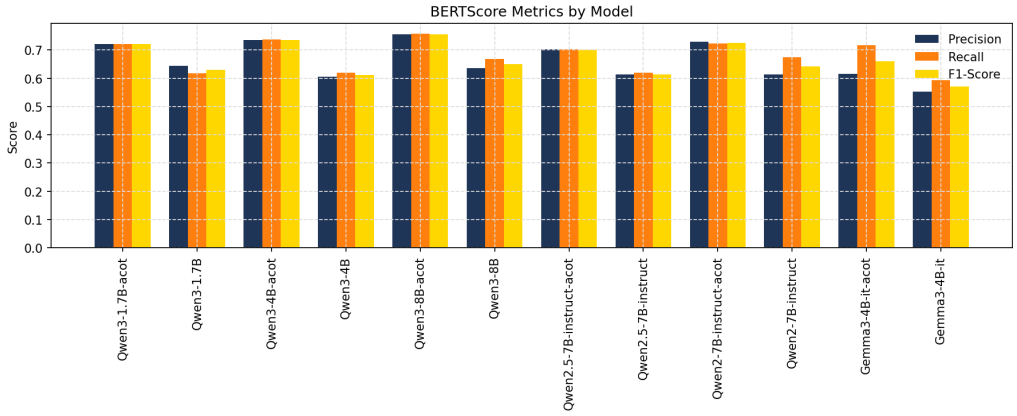


Figure 6.9: Visualization of BERTScore Across Instruction-Tuned and Baseline Models

**Figure 6.5** demonstrates the scores of ROUGE-1 and ROUGE-2. The ROUGE-1 metric captures the ratio of how many important words are captured. In comparison, the ROUGE-2 captures the important phrase coverage. It provides precision,

Recall, and F-1 score. The scale ranges from 0 to 1 (the higher, the better). Massive gains in ROUGE scores (e.g., Qwen3-8B-acot ROUGE-2 F1: 0.2960 versus base: 0.0838) indicate that our models produce more thorough emotional reactions. Dramatic improvements can be seen in content coverage (300-400% improvements in some cases). ROUGE-2 improvements are powerful, indicating better phrase-level emotional expression. They do not just match words; they also incorporate essential emotional concepts and contextual components to make responses more full and valuable. **Figure 6.10** and **Figure 6.11** showcasing the ROUGE-1 and ROUGE-2 scores.

| | Model | Precision | | Recall | | F1-score | |
|---|---|---|---|---|---|---|---|
| | | R-1 | R-2 | R-1 | R-2 | R-1 | R-2 |
| Our | **Qwen3-1.7B-acot** | 0.5037 | 0.2202 | 0.5125 | 0.2248 | 0.4971 | 0.2176 |
| | **Qwen3-1.7B** | 0.3636 | 0.0618 | 0.2596 | 0.0470 | 0.2884 | 0.0507 |
| Our | **Qwen3-4B-acot** | 0.5266 | 0.2489 | 0.5382 | 0.2540 | 0.5210 | 0.2462 |
| | **Qwen3-4B** | 0.3246 | 0.0598 | 0.3011 | 0.0578 | 0.2843 | 0.0526 |
| Our | **Qwen3-8B-acot** | 0.5614 | 0.3002 | 0.5677 | 0.3040 | 0.5532 | 0.2960 |
| | **Qwen3-8B** | 0.3339 | 0.0782 | 0.4000 | 0.0949 | 0.3553 | 0.0838 |
| Our | **Qwen2.5-7B-instruct-acot** | 0.4829 | 0.1834 | 0.4381 | 0.1662 | 0.4452 | 0.1685 |
| | **Qwen2.5-7B-instruct** | 0.3366 | 0.0607 | 0.3109 | 0.0603 | 0.2893 | 0.0542 |
| Our | **Qwen2-7B-instruct-acot** | 0.5283 | 0.2280 | 0.4670 | 0.2017 | 0.4839 | 0.2085 |
| | **Qwen2-7B-instruct** | 0.2834 | 0.0602 | 0.4690 | 0.1056 | 0.3391 | 0.0738 |
| Our | **Gemma3-4B-it-acot** | 0.3589 | 0.1674 | 0.6649 | 0.3188 | 0.4452 | 0.2088 |
| | **Gemma3-4B-it** | 0.2473 | 0.0390 | 0.3670 | 0.0607 | 0.2743 | 0.0435 |

Table 6.5: ROUGE-1 and ROUGE-2 Scores Measuring Overlap with Reference Responses

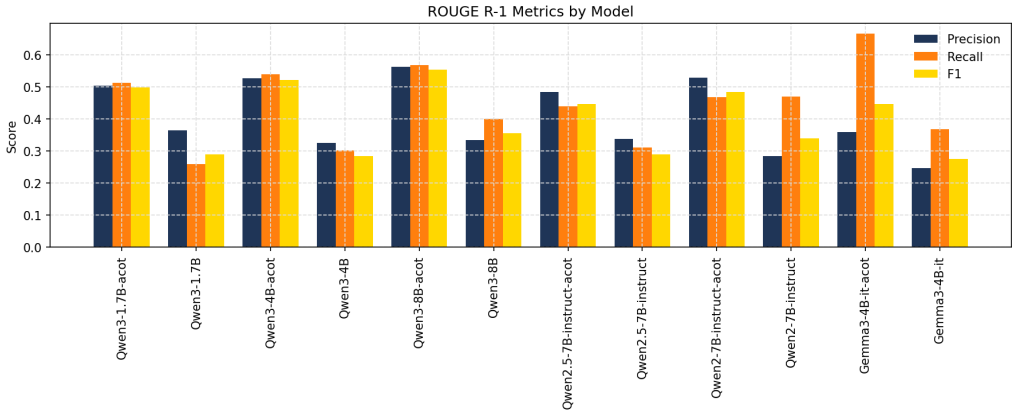


Figure 6.10: ROUGE-1 Score Comparison

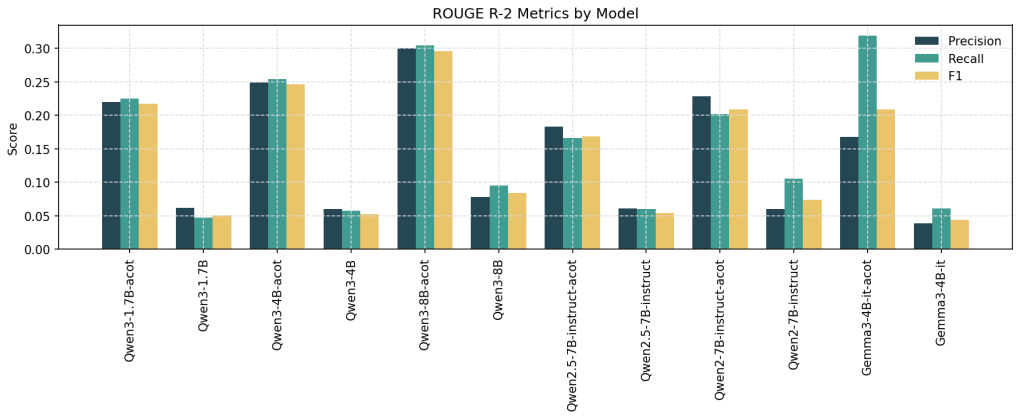


Figure 6.11: ROUGE-2 Score Comparison Between Instruction-Tuned and Baseline Models

## 6.4 Human Evaluation Metrics: Expert Ratings of Emotional Intelligence

The goal of human assessment is to judge the model's output based on emotional understanding, empathy, and helpfulness. Additionally, the structure of the final output needed to be verified through this process.

We have chosen the Qwen3 1.7B acot, not the Emobench version, because we do not need the output in that format to evaluate now. Additionally, for the baseline, only the Qwen3 1.7B was used.

### Inference Configuration

For model inference, input messages were formatted using the tokenizer's chat template with generation prompts and thinking mode enabled to leverage the model's reasoning capabilities. Text generation was performed with a maximum output length of 1024 new tokens using a temperature of 0.7 for moderate randomness, top-p nucleus sampling of 0.8, and top-k sampling limited to the 20 most likely tokens at each step. Real-time text streaming was implemented to monitor the progress of generation, with the original input prompt excluded from the streamed output. All inference operations were conducted on CUDA-enabled GPU hardware.

| | Emotional Understanding A | Empathy A | Helpfulness A | Emotional Understanding B | Empathy B | Helpfulness B | Appraisal Reasoning Of A |
|---|---|---|---|---|---|---|---|
| Psychiatrist 1 | 100 | 94.28 | 100 | 100 | 97.14 | 97.14 | 100 |
| Psychiatrist 2 | 91.42 | 94.28 | 94.28 | 88.57 | 91.42 | 97.14 | 94.28 |
| Psychiatrist 3 | 85.71 | 97.14 | 88.57 | 85.71 | 91.42 | 88.57 | 88.57 |

Table 6.6: Comparison of Responses of Qwen3 1.7B acot vs Responses of Qwen3 1.7B based on Human Evaluation metrics. Here model A is our acot model and model B is our Baseline model (Accuracy %)

Results in **Table 6.6** looking at the human evaluation results comparing the Qwen3 1.7B ACoT model (Model A) against the baseline Qwen3 1.7B model (Model B), several clear patterns emerge across the three key evaluation metrics.

**Emotional Understanding** shows mixed results across evaluators. Expert 1 rated both models equally at 100%, suggesting that for this evaluator, both models demonstrated excellent emotional comprehension. However, Expert 2 favored Model A (91.42%) over Model B (88.57%), while Expert 3 found no difference between the models (both at 85.71%). This variation suggests that improvements in emotional understanding may be context-dependent or evaluator-specific.

**Empathy** demonstrates the most consistent advantage of the ACoT model. Model A achieved scores of 94.28%, 94.28%, and 97.14% from the three experts, respectively, while Model B scored 97.14%, 91.42%, and 91.42%. Notably, Model A maintained more consistent empathy scores across evaluators (standard deviation of approximately 1.65) compared to Model B's wider variation (standard deviation of approximately 3.31), indicating more reliable empathetic responses.

**Helpfulness** reveals the strongest differentiation between models. Model A consistently outperformed Model B across all three evaluators: 100% vs. 97.14% (Expert 1), 94.28% vs. 97.14% (Expert 2), and 88.57% vs. 88.57% (Expert 3). The ACoT model's structured reasoning approach appears to translate into more practically useful responses for users seeking emotional support.

The **Appraisal Reasoning** scores for Model A (100%, 94.28%, 88.57%) provide crucial insight into the quality of the Chain-of-Thought structure. These high scores across all evaluators validate that the ACoT framework successfully maintains logical coherence and structured thinking while addressing emotional content. **Figure 6.12** is demonstrating the bar chart of the Experts ratings.

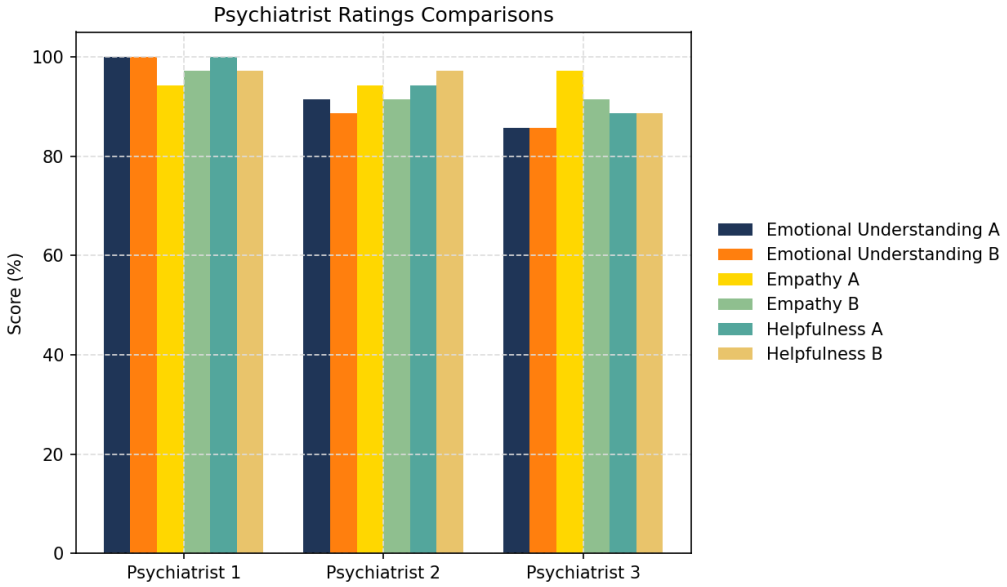


Figure 6.12: Human Evaluation Comparison Between our Qwen1.7B-acot (A) and Baseline Qwen1.7B (B), Highlighting Improvements in Empathetic Response Quality

# Chapter 7

# Discussion and Findings

**Figure 7.1** and **Figure 7.2** showcase the impact of model size on performance improvements for models tuned on the ACoT dataset. Specifically, **figure 7.1** illustrates the change in Empathetic Understanding ($\Delta EU$), while **figure 7.2** presents the change in Emotional Appropriateness ($\Delta EA$) as model parameters increase. Both the results of EU and EA showcase a pattern. Models with 4B parameters have gained maximum growth in both of the assessments. Qwen3 4B achieved almost similar results as its 8B variant in EA assessment. Even models with 1.7B (+bias) have shown significant improvement. Its performance was almost similar to its 4B parameter variant in EA task. Moreover, the EU tasks were much more complicated, but even there we can observe improvement in Qwen 4B .Models with similar parameters such as Gemma3-4B-it-acot also demonstrate advancement in the EU assessment. This suggests that these mid-sized models' (1.7B 4B) Emotional intelligence could be significantly improved by instruction tuning with Appraisal CoT. The reason behind this could be that the model has the maximum impact of our Appraisal CoT dataset in a certain threshold of parameters.

As the parameters increase, we can observe a modest enhancement in the case of Qwen2 7B instruct acot at the EA category. On the other hand, the Qwen3 8B acot and Qwen2.5 7B instruct acot showcase enhancement in their performance of EU. These results suggest that the Appraisal CoT increases the EA and EU in 7B 8B modestly. Not all tasks of the EU and EU increases; instead, it finds tasks in which it didn't perform well, specifically for that model.

Now, for models Qwen3 0.6B acot, Qwen3 14B acot, and Qwen2.5 14B acot, both EU and EA can be seen to perform neutrally to slightly negatively. This can be explained as models with small parameters such as Qwen3 0.6B would require extensive training to gain Emotional intelligence. Appraisal CoT single handedly will not be able to enhance its capabilities. On the other hand, models with larger parameters (>= 14B), such as Qwen3 14B and Qwen2.5 14B, are already trained on substantial data; thus, they achieve emotional intelligence automatically. In the future, if a more precisely crafted dataset with much more complex data is introduced, the model's performance might also improve.

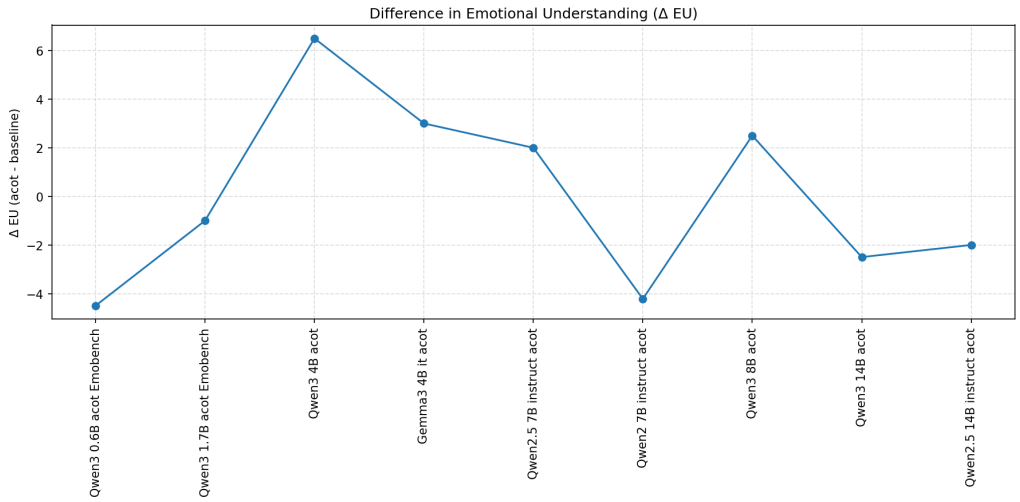


Figure 7.1: Impact of Model Size on Empathetic Understanding ($\Delta EU$) for ACOT-Tuned Models

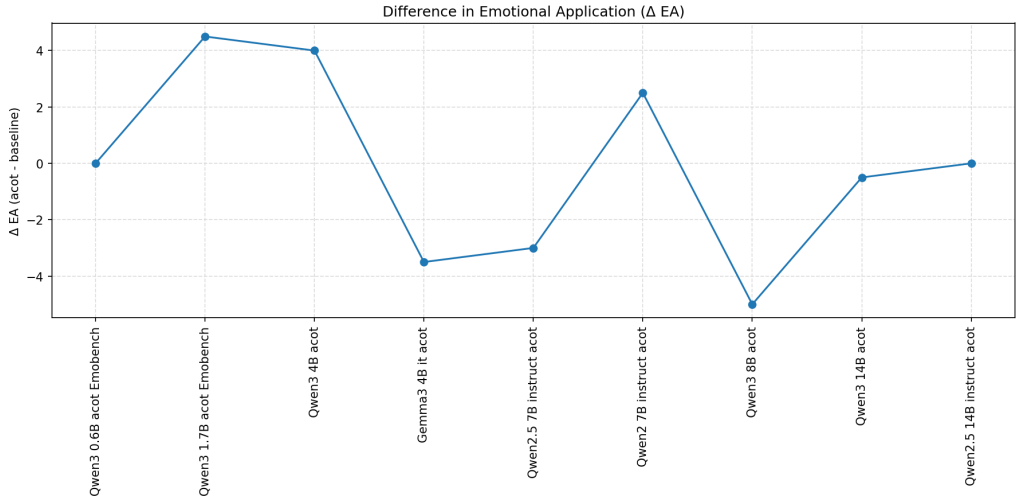


Figure 7.2: Impact of Model Size on Emotional Application ($\Delta EA$) for ACoT-Tuned Models

The most significant finding is that our instruction tuning improves not only surface-level text matching, but also fosters genuine emotional comprehension. The combination of high BLEU (precise language), high BERTScore (semantic comprehension), high ROUGE (complete coverage), and suitably focused Distinct scores paints a persuasive image of models that have learned to be truly compassionate and emotionally intelligent. This multi-metric validation provides significant support for our hypothesis that the Appraisal CoT dataset successfully improves emotional intelligence in language models across various quality parameters.

The ACoT model demonstrates particular strength in empathy consistency and helpfulness, suggesting that structured appraisal reasoning enhances the model's ability to provide reliable, actionable emotional support. While emotional awareness improves in varying degrees, consistent empathy scores suggest that the ACoT framework aids in maintaining a consistent empathic tone across different contexts.

The high appraisal reasoning scores demonstrate that the organized approach does not sacrifice answer quality, but rather provides a dependable framework for emotional AI interactions. This shows that combining cognitive appraisal theory into language model training can result in more consistent and useful answers in emotionally charged situations.

# Chapter 8

# Future Research and Limitations

## 8.1 Limitation

Although CAT-CoT is a significant direction of integrating Cognitive Appraisal Theory (CAT) with large language models (LLMs), there exist several crucial limitations that one should remember:

1. **Limited Conversational Depth:** The dataset includes conversations of a single turn. This simplifies the ability to study emotions in a clear sense; however, it fails to indicate the development process of emotions, their alterations, and complexities in more extended dialogues.

2. **Cultural and Demographic Constraints:** Although CAT is intended to apply to any individual, the understanding and expression of feelings may differ across particular cultures, age groups, or backgrounds. Not all types of users may find the emotional responses effective since the dataset fails to reveal the differences.

3. **Evaluation Scope:** The human evaluation involved only three psychiatrists, despite the model being rated with the help of automated scores and expert opinion. The model should receive feedback from a more diverse group of people to determine how it works effectively.

4. **Narrow Psychological Focus:** CAT is a powerful theory; however, it is not the only means to comprehend emotion. The exclusive reliance on CAT may limit the generality of our framework in handling emotions not well-explained by appraisal-based reasoning. In real life, when a human responds, there can be an unknown number of Psychological algorithms working behind that.

## 8.2 Future Work

Future directions to enhance CAT-CoT to address the existing problems are as follows:

1. **Multi-Turn Emotion Reasoning & Complex scenario:** Future datasets may involve more extended conversations to monitor changes in emotions over time. This would assist the model in conceptualizing emotional development,

re-evaluating emotions, and improving empathy in the process of continued conversation. Furthermore, complex scenarios will enable models to achieve an in-depth understanding.

2. **Integration of Additional Psychological Theories:** CAT can also intersect with other theories of emotion, such as Emotion Regulation Theory, Affective Neuroscience, or the Theory of Constructed Emotion. Combinations of various ideas may make the model less rigid and comprehensive in terms of reasoning based on emotions.

3. **Preference Alignment via Human Feedback:** Large language models may produce emotionally appropriate responses that still fail to align with individual or cultural user preferences. Future efforts could incorporate Reinforcement Learning from Human Feedback (RLHF) or Direct Preference Optimization (DPO), using human-annotated comparisons to align emotional tone, reasoning style, and helpfulness with diverse user expectations.

4. **Cultural and Demographic Appraisal Calibration:** Emotional perspectives on other cultures and social groups should be incorporated into future work. This will enable the model to understand and respond to users across the board better and eliminate the emotional prejudice.

5. **Psychologist-in-the-Loop Fine-Tuning:** Integrating real-time human feedback from clinical experts during tuning or deployment could improve accuracy in high-stakes domains. Such collaboration would help detect hallucinated empathy, insensitive phrasing, or unsafe emotional advice early in the model lifecycle.

To conclude, CAT-CoT demonstrates that Cognitive Appraisal Theory is applicable in the language models to render them more emotion-conscious. However, to truly develop this technology as applicable and secure, future efforts must involve not only the improvement of technical means but also the consideration of psychology, culture, and human experience.

# Chapter 9

# Conclusion

In this thesis, CAT-CoT, an instruction-tuning framework that collectively incorporates Cognitive Appraisal Theory (CAT) into Chain-of-Thought prompting, was proposed. In an analysis of previous studies, we found that large language models (LLMs) have demonstrated impressive conversational skills; however, they often fall short of genuine emotional insight and empathy. Fine-tuning and prompt engineering might temporarily produce sympathetic language, but these models are mostly about pattern matching at the surface and not the cognitive mechanisms through which emotional experience is brought into being. Specifically, they lack an organized appraisal system, which domain researchers identify as the core of human emotion, encompassing judgments of goal relevance, responsibility, and control. Moreover, the lack of data that demonstrates the chains of causal appraisal only increases the distance between shallow empathy and genuine emotional reasoning. Hence, we proposed a three-phase appraisal pipeline: primary appraisal of the relevance of the situation, secondary appraisal of control, and reappraisal to target the model outputs into more psychologically realistic responses. Second, we generated the Appraisal-CoT dataset, which includes clear chains of appraisals alongside responses. We evaluated it using a strict emotional response assessment procedure that combined human evaluation and automated measures, achieving excellent results. Third, instruction-tuned open-source LLMs of different parameter sizes (CAT-CoT) can be provided to demonstrate the influence of the dataset. Our results on the EmoBench showed that a greater improvement is obtained on mid-sized models.

The LLMs improved drastically in both emotional understanding and emotional application. Interestingly, relatively larger models have a modest advantage, whereas the very small and extensive models resulted in minimal or negative payoffs, thereby demonstrating the top-in-the-center ability to be achieved extraordinarily well in our structured reasoning. Additionally, CAT-CoT tuning could improve fluency and semantic relevance while maintaining lexical diversity. Collectively, these results suggest that incorporating appraisal reasoning into LLMs offers a more nuanced representation of feelings without compromising coherence or relevance. Together with automatic measures, human judges rated CAT output as excellent on measures of coherence, naturalness, and empathy, and there was an extraordinarily high agreement on appraisal accuracy. Surprisingly, the model not only used affective language but also diagnosed the causal appraisals that lead the user to the given

emotional response, providing actionable emotional support rather than just sympathetic phrasing. This work has its limitations, even though it represents a step forward. To be specific, our dependence on synthetic single-turn data and the depth of conversations it facilitated gave rise to the necessity of real-life, multi-turn, and diversely culturally rich corpora. However, these limitations offer promising new directions for research rather than weaknesses of the work. Going forward, these systems can provide subtle yet reliable empathy and advice by explicitly modeling the reasons why people feel a certain way. The next step is to integrate additional psychological theories, extend them to multi-turn therapeutic dialogues and develop hybrid symbolic neural architectures that further advance cognitive science and machine learning.

To conclude, this thesis provides a solid foundation for emotionally intelligent AI. It demonstrates that LLMs can internalize an orderly appraisal mechanism, creating coherent, varied, and genuine empathetic responses. CAT-CoT will ultimately bring us one step closer to the golden age, where AI cannot only comprehend human speech but also the full spectrum of human emotions and behaviors.

# Bibliography


[1] W. B. Cannon, "The james-lange theory of emotions: A critical examination and an alternative theory," *The American journal of psychology*, vol. 39, no. 1/4, pp. 106–124, 1927.

[2] W. James, "What is emotion? 1884.," 1948.

[3] S. Schachter and J. Singer, "Cognitive, social, and physiological determinants of emotional state.," *Psychological review*, vol. 69, no. 5, p. 379, 1962.

[4] R. S. Lazarus, *Stress, appraisal, and coping*. Springer, 1984, vol. 464.

[5] C. A. Smith and P. C. Ellsworth, "Patterns of cognitive appraisal in emotion.," *Journal of personality and social psychology*, vol. 48, no. 4, p. 813, 1985.

[6] S. Folkman, R. S. Lazarus, R. J. Gruen, and A. DeLongis, "Appraisal, coping, health status, and psychological symptoms.," *Journal of personality and social psychology*, vol. 50, no. 3, p. 571, 1986.

[7] H. G. Wallbott and K. R. Scherer, "Assessing emotion by questionnaire," in *The measurement of emotions*, Elsevier, 1989, pp. 55–82.

[8] P. C. Ellsworth, "Some implications of cognitive appraisal theories of emotion," 1991.

[9] R. S. Lazarus, *Emotion and adaptation*. Oxford University Press, 1991.

[10] R. A. Shweder and M. A. Sullivan, "Cultural psychology: Who needs it?" *Annual review of psychology*, vol. 44, no. 1, pp. 497–523, 1993.

[11] D. Goleman, "Emotional intelligence. why it can matter more than iq.," *Learning*, vol. 24, no. 6, pp. 49–50, 1996.

[12] R. S. Lazarus, "Relational meaning and discrete emotions.," 2001.

[13] I. J. Roseman and C. A. Smith, "Appraisal theory," *Appraisal processes in emotion: Theory, methods, research*, pp. 3–19, 2001.

[14] K. R. Scherer, "Appraisal considered as a process of multilevel sequential checking," *Appraisal processes in emotion: Theory, methods, research*, vol. 92, no. 120, p. 57, 2001.

[15] P. J. Silvia, "Interest and interests: The psychology of constructive capriciousness," *Review of General Psychology*, vol. 5, no. 3, pp. 270–290, 2001.

[16] C. A. Smith and L. D. Kirby, "Toward delivering on the promise of appraisal theory," *Appraisal processes in emotion: Theory, methods, research*, pp. 121–138, 2001.

[17] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[18] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[19] I. Roseman and A. Evdokas, "Appraisals cause experienced emotions: Experimental evidence," *Cognition and emotion*, vol. 18, no. 1, pp. 1–28, 2004.

[20] R. Artstein and M. Poesio, "Inter-coder agreement for computational linguistics," *Computational linguistics*, vol. 34, no. 4, pp. 555–596, 2008.

[21] R. Snow, B. O'connor, D. Jurafsky, and A. Y. Ng, "Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks," in *Proceedings of the 2008 conference on empirical methods in natural language processing*, 2008, pp. 254–263.

[22] S. C. Marsella and J. Gratch, "Ema: A process model of appraisal dynamics," *Cognitive Systems Research*, vol. 10, no. 1, pp. 70–90, 2009.

[23] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan, "A diversity-promoting objective function for neural conversation models," *arXiv preprint arXiv:1510.03055*, 2015.

[24] T. A. O'Neill, "An overview of interrater agreement on likert scales for researchers and practitioners," *Frontiers in psychology*, vol. 8, p. 777, 2017.

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

[26] H. Zhou, M. Huang, T. Zhang, X. Zhu, and B. Liu, "Emotional chatting machine: Emotional conversation generation with internal and external memory," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, 2018.

[27] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *arXiv preprint arXiv:1904.09675*, 2019.

[28] T. Brown, B. Mann, N. Ryder, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[29] Y. Oortwijn, T. Ossenkoppele, and A. Betti, "Interrater disagreement resolution: A systematic procedure to reach consensus in annotation tasks," in *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, 2021, pp. 131–141.

[30] J. Wei, M. Bosma, V. Y. Zhao, *et al.*, "Finetuned language models are zero-shot learners," *arXiv preprint arXiv:2109.01652*, 2021.

[31] A. Lahnala, C. Welch, D. Jurgens, and L. Flek, "A critical reflection and forward perspective on empathy and natural language processing," *arXiv preprint arXiv:2210.16604*, 2022.

[32] L. Ouyang, J. Wu, X. Jiang, *et al.*, "Training language models to follow instructions with human feedback," *Advances in neural information processing systems*, vol. 35, pp. 27 730–27 744, 2022.

[33] J. Wei, X. Wang, D. Schuurmans, *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[34] K. L. Aw, S. Montariol, B. AlKhamissi, M. Schrimpf, and A. Bosselut, "Instruction-tuning aligns llms to the human brain," *arXiv preprint arXiv:2312.00575*, 2023.

[35] A. Bao, Y. Zeng, and E. Lu, "Mitigating emotional risks in human-social robot interactions through virtual interactive environment indication," *Humanities and Social Sciences Communications*, vol. 10, no. 1, pp. 1–9, 2023.

[36] H. Cai, X. Shen, Q. Xu, *et al.*, "Improving empathetic dialogue generation by dynamically infusing commonsense knowledge," in *Findings of the Association for Computational Linguistics: ACL 2023*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds., Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 7858–7873. DOI: 10.18653/v1/2023.findings-acl.498. [Online]. Available: https://aclanthology.org/2023.findings-acl.498/.

[37] F. Fu, L. Zhang, Q. Wang, and Z. Mao, "E-core: Emotion correlation enhanced empathetic dialogue generation," *arXiv preprint arXiv:2311.15016*, 2023.

[38] J.-t. Huang, M. H. Lam, E. J. Li, *et al.*, "Emotionally numb or empathetic? evaluating how llms feel using emotionbench," *arXiv preprint arXiv:2308.03656*, 2023.

[39] Y. Qian, W.-N. Zhang, and T. Liu, "Harnessing the power of large language models for empathetic response generation: Empirical investigations and improvements," *arXiv preprint arXiv:2310.05140*, 2023.

[40] G. Yeo and K. Jaidka, "The peace-reviews dataset: Modeling cognitive appraisals in emotion text analysis," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 2822–2840.

[41] J. P. Beno, "Electra and gpt-4o: Cost-effective partners for sentiment analysis," *arXiv preprint arXiv:2501.00062*, 2024.

[42] X. Cao, T. Meng, and X. Bi, "Iecck: Integrating emotion cause and common knowledge for empathetic dialogue generation," *International Journal of Machine Learning and Cybernetics*, pp. 1–13, 2024.

[43] Y. Chen, H. Wang, S. Yan, *et al.*, "Emotionqueen: A benchmark for evaluating empathy of large language models," *arXiv preprint arXiv:2409.13359*, 2024.

[44] J.-t. Huang, M. H. Lam, E. J. Li, *et al.*, "Apathetic or empathetic? evaluating llms' emotional alignments with humans," *Advances in Neural Information Processing Systems*, vol. 37, pp. 97 053–97 087, 2024.

[45] Y. Kang and J. Kim, "Chatmof: An artificial intelligence system for predicting and generating metal-organic frameworks using large language models," *Nature communications*, vol. 15, no. 1, p. 4705, 2024.

[46] J.-C. Klie, R. E. d. Castilho, and I. Gurevych, "Analyzing dataset annotation quality management in the wild," *Computational Linguistics*, vol. 50, no. 3, pp. 817–866, 2024.

[47] W. Li, T. Sun, K. Qian, and W. Wang, "Optimizing psychological counseling with instruction-tuned large language models," *arXiv preprint arXiv:2406.13617*, 2024.

[48] Z. Li, G. Chen, R. Shao, Y. Xie, D. Jiang, and L. Nie, "Enhancing emotional generation capability of large language models via emotional chain-of-thought," *arXiv preprint arXiv:2401.06836*, 2024.

[49] H. Liang, L. Sun, J. Wei, *et al.*, "Synth-empathy: Towards high-quality synthetic empathy data," *arXiv preprint arXiv:2407.21669*, 2024.

[50] Z. Ma, Y. Mei, Y. Long, Z. Su, and K. Z. Gajos, "Evaluating the experience of lgbtq+ people using large language model based chatbots for mental health support," in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1–15.

[51] P. Raile, "The usefulness of chatgpt for psychotherapists and patients," *Humanities and Social Sciences Communications*, vol. 11, no. 1, pp. 1–8, 2024.

[52] S. Sabour, S. Liu, Z. Zhang, *et al.*, "Emobench: Evaluating the emotional intelligence of large language models," *arXiv preprint arXiv:2402.12071*, 2024.

[53] O. Sotolar, "Empo: Theory-driven dataset construction for empathetic response generation through preference optimization," *arXiv e-prints*, arXiv–2406, 2024.

[54] X. Wang, B. Gao, Y. Dai, *et al.*, "Cognition chain for explainable psychological stress detection on social media," *arXiv preprint arXiv:2412.14009*, 2024.

[55] J. Yang and D. Jurgens, "Modeling empathetic alignment in conversation," *arXiv preprint arXiv:2405.00948*, 2024.

[56] G. Yeo and K. Jaidka, "Analyzing language for insights into cognitive and social processes," *Yeo, G. & Jaidka, K.(in press). Analyzing language for insights into cognitive and social processes. In T. Reimer, L. van Swol., & A. Florack (Eds.), The Routledge Handbook of Communication and Social Cognition. Routledge/Taylor and Francis*, 2024.

[57] H. Zhan, A. Zheng, Y. K. Lee, J. Suh, J. J. Li, and D. C. Ong, "Large language models are capable of offering cognitive reappraisal, if guided," *arXiv preprint arXiv:2404.01288*, 2024.

[58] T. Zhang, F. Ladhak, E. Durmus, P. Liang, K. McKeown, and T. B. Hashimoto, "Benchmarking large language models for news summarization," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 39–57, 2024.

[59] Y. Zhang, X. Yang, X. Xu, *et al.*, "Affective computing in the era of large language models: A survey from the nlp perspective," *arXiv preprint arXiv:2408.04638*, 2024.

[60] H. Zhao, L. Li, S. Chen, *et al.*, "Esc-eval: Evaluating emotion support conversations in large language models," *arXiv preprint arXiv:2406.14952*, 2024.

[61] T. D. Belay, A. H. Ahmed, A. Grissom II, *et al.*, "Culemo: Cultural lenses on emotion–benchmarking llms for cross-cultural emotion understanding," *arXiv preprint arXiv:2503.10688*, 2025.

[62] L. F. Föyen, E. Zapel, M. Lekander, E. Hedman-Lagerlöf, and E. Lindsäter, "Artificial intelligence vs. human expert: Licensed mental health clinicians' blinded evaluation of ai-generated and expert psychological advice on quality, empathy, and perceived authorship," *Internet Interventions*, vol. 41, p. 100 841, 2025.

[63] H. Hu, Y. Zhou, L. You, *et al.*, "Emobench-m: Benchmarking emotional intelligence for multimodal large language models," *arXiv preprint arXiv:2502.04424*, 2025.

[64] K. Khalid Saifullah, F. Azmain, and H. Hye, "Sentiment analysis in software engineering: Evaluating generative pre-trained transformers," *arXiv e-prints*, arXiv–2505, 2025.

[65] Y. Li, J. Yao, J. B. S. Bunyi, A. C. Frank, A. Hwang, and R. Liu, "Counselbench: A large-scale expert evaluation and adversarial benchmark of large language models in mental health counseling," *arXiv preprint arXiv:2506.08584*, 2025.

[66] R. M. Murphy, D. A. Dongelmans, N. F. de Keizer, *et al.*, "Creation of a gold standard dutch corpus of clinical notes for adverse drug event detection: The dutch ade corpus," *Language Resources and Evaluation*, pp. 1–17, 2025.

[67] K. Schlegel, N. R. Sommer, and M. Mortillaro, "Large language models are proficient in solving and creating emotional intelligence tests," *Communications Psychology*, vol. 3, no. 1, pp. 1–14, 2025.

[68] L. Wang, J. Li, C. Yang, *et al.*, "Sibyl: Empowering empathetic dialogue generation in large language models via sensible and visionary commonsense inference," in *Proceedings of the 31st International Conference on Computational Linguistics*, 2025, pp. 123–140.

[69] A. Yang, A. Li, B. Yang, *et al.*, "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025.

[70] G. C. Yeo and K. Jaidka, "Beyond context to cognitive appraisal: Emotion reasoning as a theory of mind benchmark for large language models," *arXiv preprint arXiv:2506.00334*, 2025.

[71] S. Hong, J. Sun, and H. Chen, "A third-person appraisal agent: Learning to reason about emotions in conversational contexts,"