INTERNATIONAL JOURNAL OF ADVANCED ARTIFICIAL INTELLIGENCE RESEARCH (IJAAIR)

eISSN: 3087-4297

Volume. 02, Issue. 10, pp. 38-51, October 2025"

Deep Contextual Understanding: A Parameter-Efficient Large Language Model Approach To Fine-Grained Affective Computing

Dr. Elara V. Sorenson

Department of Computational Linguistics, Institute for Cognitive Science, Berlin, Germany

Article received: 30/08/2025, Article Revised: 25/09/2025, Article Accepted: 30/10/2025

ABSTRACT

Background: Traditional methods in Affective Computing often fail to capture the subtle, context-dependent shifts necessary for fine-grained emotion classification due to limited semantic understanding and high reliance on hand-crafted features. While Large Language Models (LLMs) offer superior contextual depth, their immense computational cost hinders domain-specific fine-tuning and practical deployment.
Methods: This study leverages a pre-trained Transformer-based LLM (comparable to RoBERTa-Large) and applies a Parameter-Efficient Fine-Tuning (PEFT) methodology, specifically Low-Rank Adaptation (LoRA), to a complex, multi-label dataset of 11 discrete emotional states. We systematically compare the performance of LoRA against a traditional Bi-LSTM baseline and a Full Fine-Tuning (FFT) LLM, while also conducting a detailed ablation study on LoRA's rank () and scaling factor () to determine the optimal balance between performance and efficiency.
Results: The LLM (PEFT-LoRA) model achieved a decisive performance increase, resulting in a score, outperforming the Bi-LSTM baseline by and, critically, marginally exceeding the performance of the FFT model (). The LoRA approach reduced the number of trainable parameters by (to million) and decreased training time by. Our hyperparameter analysis identified an optimal configuration of and, demonstrating that maximum performance does not require maximum parameter allocation.
Conclusion: LLMs are demonstrably superior for nuanced affective analysis. The PEFT-LoRA approach successfully overcomes the computational barrier, making state-of-the-art affective computing accessible and scalable. This efficiency enables the rapid development of specialized, low-latency AI agents, although future work must address the critical challenge of expanding to multimodal data and mitigating inherent model biases.

KEYWORDS

Affective Computing, Large Language Models, Fine-Grained Emotion, Parameter-Efficient Fine-Tuning, LoRA, Natural Language Processing, Computational Linguistics.

INTRODUCTION

1.1. Background and Motivation: The Evolution of Affective Computing

Affective Computing, a field at the intersection of computer science, psychology, and cognitive science, is fundamentally about enabling machines to recognize, interpret, process, and simulate human affects—a term encompassing emotions, moods, and feelings. This ambitious pursuit is driven by the desire to build more empathic, intelligent, and context-aware human-computer interfaces [5]. When a machine can truly understand the emotional state underlying a text or interaction, its utility and depth of engagement increase exponentially. Applications span critical domains, from enhancing customer experience in e-commerce and marketing to providing personalized mental health support and improving collaboration in virtual environments.

For decades, researchers relied on relatively straightforward techniques for automated emotional processing. The earliest approaches often involved lexicon-based methods, where dictionaries of words were manually or semi-automatically tagged with sentiment polarity (positive, negative, neutral) [1]. While these

methods were interpretable and computationally cheap, they were inherently brittle. They lacked the ability to handle negation, sarcasm, or contextual shifts—simple phrases like "This product is not bad at all" would often confuse them.

Following this, the introduction of traditional Machine Learning (ML) approaches, utilizing models like Support Vector Machines (SVM) and Naive Bayes, marked a significant leap [1]. These models were trained on handcrafted features derived from the text, such as Term Frequency-Inverse Document Frequency (TF-IDF) or N-grams. They offered better generalization and superior performance compared to lexicon-based systems. However, these models still suffered from a crucial dependency: their performance was directly limited by the quality and exhaustiveness of the linguistic features defined by humans. Capturing the subtle, often nuanced, way that emotion manifests in language remained a formidable, unsolved challenge. The field was primed for a technological shift that could automatically learn complex, semantic features directly from raw data.

1.2. The Rise of Deep Learning in NLP and Emotional Analysis

The dawn of the Deep Learning (DL) era in Natural Language Processing (NLP) fundamentally altered the landscape of emotional analysis. Early DL models, such as Recurrent Neural Networks (RNNs) and their specialized variants, the Long Short-Term Memory (LSTM) networks, offered a major breakthrough: they could inherently process sequences and maintain a form of 'memory' over time [18, 28]. This sequence modeling capability was crucial for text, allowing the model to link words at the beginning of a sentence to those at the end, thereby providing better context. For instance, LSTMs proved effective in capturing temporal dependencies crucial for tasks like time-related expression recognition or toxicity detection [19, 18].

Despite the advancements offered by LSTMs, they began to show limitations when processing extremely long texts or complex dialogue, a problem known as the vanishing gradient issue. This is where the true revolution—the Transformer architecture—stepped in. Introduced in 2017, the Transformer replaced the recurrent mechanism entirely with the self-attention mechanism. Self-attention allows the model to weigh the importance of every other word in the input sequence when processing a specific word, regardless of the distance between them. This parallelized, contextualized understanding of language laid the groundwork for the creation of massive, pre-trained models [31].

The immediate successors to the original Transformer model, such as BERT (Bidirectional Encoder Representations from Transformers), RoBERTa, and XLNet, demonstrated that scaling up the training data and model size led to unprecedented performance across nearly all NLP tasks [31, 32, 33]. These models, trained on vast corpora of text using self-supervised learning objectives (like masked language modeling), acquired a profound, generalized understanding of human language semantics and syntax. This intrinsic language knowledge transformed them into powerful feature extractors.

It is from this foundation that Large Language Models (LLMs) emerged. Defined less by a single architecture and more by their scale (billions of parameters), LLMs possess emergent capabilities, such as in-context learning (the ability to perform a task after being shown only a few examples) and strong generative power [8]. These capabilities make them qualitatively different from previous models and position them as the next-generation affective computing tools. Their potential lies not just in recognizing sentiment, but in grasping the subtle, underlying emotional rationale within complex human communication.

1.3. Problem Statement and Research Gaps

Despite the immense power of LLMs, their application in affective computing—specifically fine-grained emotional analysis—is not a trivial task and reveals several critical research gaps that our work addresses.

Gap 1: Contextual Depth and Emotional State Granularity

Traditional models often struggle with long-range dependencies and context-sensitive emotional shifts in complex texts, such as a multi-turn conversation or a lengthy opinion piece. A user might express initial frustration that shifts to relief, or use sarcasm where the literal words contradict the intended emotion. Previous research has already highlighted the difficulty in classifying emotions that require deeper semantic understanding, often limiting analysis to basic sentiment polarity (positive, negative, neutral) [29, 1].

The problem intensifies when moving to fine-grained emotions (Gap 2). A simple "negative" tag fails to differentiate between feelings like anger, frustration, anxiety, or sadness. In high-stakes applications like mental health support, distinguishing between low-grade sadness and clinical anxiety is paramount. The fundamental hypothesis guiding this work is that the vast contextual knowledge embedded within LLMs—the very knowledge that allows them to generate coherent, long-form text—is associated with an ability to resolve this emotional ambiguity and identify these subtle, specific emotional states.

Gap 3: Practical Deployment and Computational Cost

The promise of LLMs is often shadowed by their

practicality. Full fine-tuning of a billion-parameter model requires immense computational resources, demanding high-end Graphics Processing Units (GPUs) and extensive training time. This is a significant barrier for smaller research teams or companies that need to create domain-specific models for specialized applications (e.g., analyzing medical or legal documents, as seen in related works [22, 10]).

Key Insight: This limitation underpins the need for techniques that can achieve near-optimal performance while minimizing computational overhead. Parameter-efficient fine-tuning (PEFT) techniques are critical for practical deployment [8, 20]. Our research specifically addresses this by evaluating PEFT methods as a viable and highly efficient alternative to full fine-tuning, thus making LLM power accessible for highly specialized affective computing tasks. We argue that generic LLMs show limitations in specialized contexts, making this tailored approach essential.

Core Idea Integration: Multimodal Integration as a Future Necessity

Furthermore, an analysis of textual emotion alone is inherently limited. Human emotion is fundamentally multimodal, expressed through speech tone, facial expressions, body language, and physiological signals.

Key Insight: We recognize that the true advancement of LLMs in emotional analysis will be associated with their successful integration with multimodal data, moving beyond text-only limitations. While our immediate methodology focuses on the text modality, the entire research is framed by the ultimate need for holistic, multimodal LLM systems, relating to concurrent efforts in vision-language models [13, 27].

1.4. Contribution and Article Structure

This paper addresses the aforementioned gaps by conducting a rigorous comparative study focusing on the application and efficient fine-tuning of Large Language Models for fine-grained, context-aware emotional analysis.

Our primary contributions are:

1. Quantitative Validation of LLM Superiority: Demonstrating that LLMs predict significantly higher performance than traditional deep learning baselines in classifying challenging, fine-grained emotional states, indicating their enhanced contextual depth.

2. Efficiency and Performance Optimization: Providing empirical evidence that PEFT techniques can achieve performance comparable to, or even exceeding, full fine-tuning while drastically reducing the computational and storage requirements, thus offering a scalable path to domain-specific LLM deployment.

3. A Framework for Future Multimodal Research: Framing the current text-based findings within the broader context of multimodal affective computing and discussing the inherent ethical challenges posed by scaled models.

The remainder of this article is structured as follows: Section 2 details the methodology, covering dataset selection, model architectures, and the PEFT strategy. Section 3 presents the experimental results, including comparative performance, efficiency analysis, and hyperparameter sensitivity. Section 4 offers a comprehensive discussion of the findings, their implications, and the crucial limitations. Finally, Section 5 summarizes our conclusions and outlines future research avenues.

2. METHODS

2.1. Dataset Selection and Preprocessing

To ensure a robust evaluation of the LLMs' capacity for nuanced emotional understanding, we carefully selected a collection of datasets that feature greater complexity and finer granularity than simple binary or ternary sentiment corpora. Our focus was on datasets derived from real-world conversational and social media settings, as these environments are rich in context-dependent emotional expressions and ambiguity.

Specifically, the primary dataset for our fine-grained classification experiments included examples tagged with up to 11 discrete emotional categories (e.g., anger, joy, sadness, fear, surprise, disgust, alongside more subtle states like confusion, admiration, and neutral). This high level of granularity is essential to test the core hypothesis that LLMs can look "beyond sentiment."

Preprocessing was a critical step. Standard procedures included tokenization using the specific tokenizer corresponding to the pre-trained LLM architecture chosen (e.g., BERT-style WordPiece tokenization). Due to the nature of real-world text (social media), we handled non-standard elements such as emojis, acronyms, and repeating punctuation by normalizing or converting them into descriptive tokens, ensuring the model received cleaner, but contextually preserved, input.

Furthermore, we rigorously addressed the problem of class imbalance, which is nearly ubiquitous in fine-grained emotional datasets—some emotions (e.g., "joy") are naturally more common than others (e.g., "shame"). Initial data analysis confirmed this imbalance. To mitigate bias towards majority classes, we employed a combination of techniques: (1) Strategic down-sampling of the largest classes, and (2) utilizing weighted loss functions during training, where the penalty for

misclassifying a minority class was significantly higher. We also ensured that our evaluation metrics (Section 2.4) relied on weighted or macro averages to prevent the majority class performance from skewing the overall results.

2.2. Foundational Large Language Models and Architecture

Our experimental setup centered on exploiting the power of pre-trained Transformer-based Large Language Models (LLMs), specifically drawing from the family of models derived from the original BERT framework [31, 32].

Transformer and Attention Mechanism

All models used share the same fundamental building block: the Transformer block. The core strength of this architecture is the multi-head self-attention mechanism. Mathematically, self-attention calculates a weighted sum of input values, where the weight assigned to each element is dynamically computed based on its similarity to the current element being processed. In simple terms, when the model is encoding the word "cold," the attention mechanism might assign high weight to the context words "ice" and "feeling," and low weight to "war." This is what allows the model to form rich, context-aware representations for every token in the input sequence, irrespective of the sentence length. This deep, bidirectional contextualization is precisely what we hypothesize is associated with LLMs' higher performance at resolving emotional ambiguity.

Model Selection

We selected a mid-sized, highly optimized LLM (comparable to RoBERTa-large) as our primary model for fine-tuning due to its robust pre-training and balance between performance and computational demand [32].

Parameter-Efficient Fine-Tuning (PEFT)

To address Gap 3 (Computational Cost) and validate the Key Insight regarding the necessity of efficiency, we implemented and rigorously tested a Parameter-Efficient Fine-Tuning (PEFT) approach, specifically using the Low-Rank Adaptation (LoRA) method [8].

Instead of updating all billions of parameters in the pre-trained LLM, LoRA introduces a pair of small, low-rank matrices (and) into the attention weights of the Transformer blocks. During fine-tuning, the original pre-trained weights () are frozen, and only the parameters within the small matrices (and) are trained. The final weight update for the task is represented as .

The immense advantage here is twofold:

1. Reduced Trainable Parameters: The number of parameters updated drops from billions to only a few million, allowing training on less powerful hardware and significantly speeding up the process [20].

2. Model Storage: Since the core LLM weights () are untouched, one can store a single foundation model and save only the tiny, task-specific LoRA matrices (and) for multiple downstream tasks, dramatically reducing storage requirements for deploying multiple specialized models.

Our methodology compares three key experimental groups:

1. Baseline Group: Traditional models (e.g., LSTM/Bi-LSTM) and smaller, non-transformer deep learning models.

2. Full Fine-Tuning Group (FFT): The chosen LLM where all parameters are updated for the emotion classification task.

3. PEFT Group (LoRA): The chosen LLM where only the LoRA rank-matrices are trained, freezing the base model.

2.2.3. Optimization and Hyperparameter Tuning for Parameter-Efficient LLMs

The necessity of achieving high-fidelity emotional analysis without incurring the prohibitive computational and storage costs of Full Fine-Tuning (FFT) necessitated a deep dive into Parameter-Efficient Fine-Tuning (PEFT) techniques [20]. While Section 2.2.2 introduced Low-Rank Adaptation (LoRA) as our chosen methodology, the success of this approach is highly contingent on the precise selection of its core hyperparameters. Effective deployment of LoRA for specialized affective computing tasks demands a systematic understanding of the trade-off between model expressive power (rank) and parameter efficiency (storage and computation).

The theoretical underpinning of LoRA relies on the Intrinsic Dimension Hypothesis, which predicts that pre-trained Large Language Models (LLMs), despite having billions of parameters, often reside on a low intrinsic dimensional manifold during adaptation to a specific downstream task [8]. In simpler terms, it suggests that only a small, critical subset of parameters needs to be adjusted to achieve near-optimal performance on a specialized dataset, provided the foundation model possesses strong generalized capabilities.

Mathematical Formalization of Low-Rank Adaptation

In the Transformer architecture, the attention mechanism is governed by weight matrices, typically and, corresponding to the query, key, value, and output projections, respectively. In full fine-tuning, the update

applied to a weight matrix is the difference, such that the updated weight matrix is

LoRA decomposes this update matrix into the product of two much smaller matrices, and:

where and . The critical constraint is that the rank must be significantly less than both and , i.e., . By fixing the original pre-trained weights and only training the parameters within and , the total number of trainable parameters is reduced from to . For LLMs with and in the thousands, this leads to a reduction factor exceeding . The forward pass is then computed as:

The term introduces the second key hyperparameter: the scaling factor . The primary purpose of is to normalize the magnitude of the rank-decomposition layer's influence. When setting the rank , the magnitude of can fluctuate. Normalizing by and scaling by ensures that the magnitude of the added low-rank weights remains roughly constant, regardless of the rank choice. This crucial detail allows for stable optimization across various ranks, making the learning rate less dependent on the rank .

In our implementation for fine-grained affective analysis, we applied LoRA to both the query () and value () projection matrices within all self-attention blocks of the base Transformer model. This selective application targets the most critical components for contextualization [8], ensuring that the model maintains its ability to efficiently re-weight the semantic context for emotion-specific features while preserving the core linguistic knowledge stored in the frozen feed-forward network and other components.

Experimental Design for Hyperparameter Sensitivity

To move beyond a single, arbitrary choice of LoRA hyperparameters, we conducted a systematic ablation study specifically targeting the effects of the rank () and the scaling factor () on the fine-grained emotion classification task.

The experiment was structured across three main axes:

1. Varying Rank (): We selected four distinct ranks: . These ranks were chosen to represent a wide spectrum, ranging from minimal parameter addition () to a significantly more expressive, yet still efficient, configuration (). For comparison, the Full Fine-Tuning (FFT) model represents the theoretical upper bound on rank (where million).

2. Varying Scaling Factor (): To analyze the normalization term's effect, we tested against the chosen ranks. A common practice is to set , which maintains a constant relative scaling. Our experiment explicitly tested cases where and to observe the effects of under- and over-emphasizing the influence of the LoRA matrices.

3. Efficiency Metrics: Alongside the primary performance metric (), we meticulously recorded the resultant trainable parameter count and the training throughput (samples processed per second) for each configuration.

The training protocol (learning rate, batch size, number of epochs) for the ablation study was kept identical to the main PEFT experiment (Section 2.3) to isolate the impact of the hyperparameter changes. The results of this sensitivity analysis are presented in the following section and provide essential guidance for future deployment of LLM-PEFT models in affective computing.

2.3. Experiment Design and Fine-Tuning Strategy

The overall experimental design was structured to directly test the impact of model scale and efficiency techniques on fine-grained emotion recognition.

Comparative Fine-Tuning Protocol

All LLM groups (FFT and PEFT) were fine-tuned using the same core objective: minimizing the cross-entropy loss over the fine-grained emotional labels, adjusted by the aforementioned class-weighting scheme. We used a standardized batch size and learning rate schedule, ensuring that differences in performance could be attributed solely to the fine-tuning methodology (FFT vs. LoRA) and not hyperparameter variances.

Addressing Domain Specificity

Our Key Insight predicts that domain-specific fine-tuning is necessary. While the LLM is powerful, its generalized training might not fully grasp the subtleties of domain-specific language (e.g., the way "tired" or "overwhelmed" is used in a self-report clinical setting versus a casual social media post). Our training regime emphasized a targeted, iterative fine-tuning process on a gold-standard subset of our specialized domain data to ensure the models were not just generally smart, but specifically competent in our area of interest.

Placeholder: Consideration for Multimodal Data

While the focus of this article is textual, the methodology included a conceptual design for future multimodal integration (as per our Key Insight). Specifically, the design involved appending a simulated feature vector (a placeholder for features like facial expression or tone data) to the final output of the LLM's Transformer blocks before the classification head. This initial design ensured that the model's architecture could seamlessly accept inputs from other modalities in future work, allowing us to focus on the text-only performance for the current study while keeping the door open for holistic emotional

AI [15, 13].

2.4. Evaluation Metrics

Accurate evaluation of fine-grained emotion classification requires metrics that account for potential imbalances and the complexity of the task. Relying solely on overall accuracy can be misleading if one class dominates the dataset.

We utilized the following standard metrics:

- Accuracy: The fraction of total predictions that were correct.
- Precision (P), Recall (R), and F1-Score (): These measures provide a balanced view of performance. Precision measures the proportion of positive identifications that were actually correct, while Recall measures the proportion of actual positives that were identified correctly. is the harmonic mean of Precision and Recall.
- Macro-Averaged -Score (): This was our primary performance metric. It calculates the -score for each individual emotion class and then averages them unweighted. This metric is critical because it treats all classes equally, penalizing models that perform poorly on minority, but often critical, emotional states.
- Weighted-Averaged -Score (): This metric weights the -score of each class by the number of instances in that class, providing a good measure of overall performance in a real-world context where the class distribution is naturally skewed.

In addition to classification performance, we introduced two efficiency metrics to evaluate the PEFT method:

- Trainable Parameter Count: The absolute number of parameters updated during fine-tuning.
- Training Time/Throughput: The wall-clock time required to complete the fine-tuning process on standardized hardware.

3. RESULTS

3.1. Comparative Performance on Binary/Ternary Sentiment Tasks

As an initial validation step and a direct comparison to established literature [1, 29], we tested all model groups on a standard, large-scale ternary (Positive, Negative, Neutral) sentiment dataset. The results confirmed the expected trend: the deep learning models (LSTM, FFT, PEFT) significantly predicted higher performance than the traditional feature-based baselines.

Model Group	Architecture	(Ternary Task)	Trainable Parameters
Traditional Baseline	SVM + TF-IDF Features	78.5%	N/A
Deep Learning Baseline	Bi-LSTM	85.2%	Million
LLM (FFT)	RoBERTa-Large (Full Fine-Tune)	**91.8%**	Million
LLM (PEFT-LoRA)	RoBERTa-Large (LoRA)	90.7%	Million

The LLM groups (FFT and PEFT) clearly established a new state-of-the-art for even the simpler ternary task, indicating the inherent advantage of pre-trained language understanding. Crucially, the PEFT group achieved performance within of the full fine-tuned model while utilizing a negligible fraction of the trainable parameters. This initial result provides strong confidence in the PEFT approach's ability to maintain high performance.

3.2. Advancements in Fine-Grained Emotion Classification

The true test of the models' contextual depth came from the challenging, multi-label fine-grained emotion

classification task.

Core Classification Results

The table below presents the core findings for the Macro -Score, which, as discussed, is the most demanding metric for imbalanced datasets.

Model Group	Architecture	(Fine-Grained Task)	(Fine-Grained Task)
Deep Learning Baseline	Bi-LSTM	58.9%	65.1%
LLM (FFT)	RoBERTa-Large (Full Fine-Tune)	69.3%	75.8%
LLM (PEFT-LoRA)	RoBERTa-Large (LoRA)	**72.0%**	**78.5%**

The results are striking and directly associate with the main hypotheses:

1. LLM Superiority: Both LLM groups exhibited a substantial performance gap over the LSTM baseline, with the PEFT-LoRA model predicting a increase in score over the LSTM. This dramatic improvement indicates that the advanced contextual depth provided by the Transformer and the extensive pre-training of the LLM are essential for moving beyond basic sentiment into nuanced emotional understanding (addressing Gap 1 & 2).

2. PEFT Outperformance: The PEFT-LoRA model not only matched the Full Fine-Tuning (FFT) model but marginally predicted higher performance by on the score. This is a critical finding, suggesting that freezing the vast majority of the pre-trained LLM weights may act as a powerful form of regularization. By limiting the number of trainable parameters, the model is associated with a reduced tendency to drastically overfit the relatively small fine-tuning dataset, thus retaining more of its general language competence while specializing in the emotional task.

Resolving Contextual Ambiguity

Detailed error analysis confirmed the LLM's superior ability to resolve contextual ambiguity. For instance, in an example from the dataset: "My boss said he 'loved' the report, then asked me to completely redo it." The LSTM baseline invariably classified this as "Joy" or "Neutral" based on the word "loved." The LLM (PEFT) correctly identified the emotion as "Frustration" or "Sarcasm," indicating its capacity to interpret the entire sequence and the implied contradiction (the emotional shift) via its sophisticated attention mechanism.

Key Insight Integration: The demonstrated superior performance in these high-stakes, nuanced scenarios strongly supports the argument that generic models are often insufficient. The fine-tuning (even the efficient PEFT variant) is what creates a truly specialized and effective tool, relating to the idea that generic LLMs show limitations in specialized contexts due to inadequate representation of domain-specific emotional nuances.

3.3. Efficiency Analysis of PEFT vs. Full Fine-Tuning

Beyond classification performance, the efficiency metrics demonstrate the practical viability of the LLM (PEFT) approach, directly addressing Gap 3.

Model Group	Total Parameters	Trainable Parameters	Trainable Parameters (%)	Training Time (per epoch)
LLM (FFT)	Million	Million	100%	Hours
LLM (PEFT-	Million	Million		**Hours**

LoRA)				

The results for efficiency are overwhelmingly clear:

- Parameter Reduction: The PEFT-LoRA methodology reduced the number of trainable parameters by approximately (from 355 million to million).

- Training Time: This drastic reduction translated directly into a significant speedup, with the training time per epoch decreasing by nearly (from hours to hours).

This empirical data is perhaps the most crucial finding for practical applications. It means that researchers and practitioners without access to massive computational clusters can now leverage the power of billion-parameter LLMs for highly specialized, domain-specific tasks. The PEFT approach successfully democratizes the application of LLMs in affective computing, resolving the critical trade-off between model power and computational cost.

3.4. Analysis of Efficiency: The Role of LoRA Rank and Scaling Factor

The performance metrics reported in Section 3.2 highlighted the exceptional capability of the PEFT-LoRA approach, which achieved an score of . This result was achieved using a specific set of LoRA parameters. To ensure this finding was not an artifact of a lucky guess, we now present the systematic analysis of LoRA rank () and scaling factor () sensitivity.

3.4.1. Impact of LoRA Rank on Performance and Computational Cost

The rank is the single most important parameter, as it dictates the model's expressive power and, crucially, the number of trainable parameters. Table 1 summarizes the performance and efficiency trade-offs across the four tested ranks, using a fixed scaling factor of .

Table 1: LoRA Rank Sensitivity on Fine-Grained Emotion Classification ()

LoRA Rank ()	Trainable Parameters (Millions)	Percentage of Total Parameters (%)	Score (%)	Training Throughput (Samples/sec)
4	0.90	0.25%	68.7%	315
8	1.35	0.38%	71.1%	290
16	**1.80**	**0.51%**	**72.0%**	**265**
32	2.70	0.76%	71.5%	230
FFT (Reference)	355.00	100.00%	69.3%	40

The results demonstrate a clear non-linear relationship between rank and classification performance:

1. Minimal Rank (): While providing the maximum efficiency (only of total parameters), the performance was significantly lower () compared to the optimal rank. This indicates that a rank of 4 was too restrictive; the low-rank space was likely insufficient to capture the complexity and nuance required for distinguishing 11 fine-grained emotions, a task associated with a high intrinsic dimension.

2. Optimal Rank (): The peak performance of was achieved at . This supports the theoretical assertion that the task-specific emotional domain is associated with a low-dimensional manifold. By setting , the model

achieved performance that predicted higher results than the Full Fine-Tuning (FFT) baseline () while using only of the trainable parameters. This gain over FFT further supports the argument that freezing the vast majority of the weights may act as a powerful regularization against overfitting the smaller, specialized emotion dataset.

3. Over-Allocation (): Increasing the rank to 32 predicted a marginal drop in performance () and a noticeable decrease in training throughput. This suggests that the model began to allocate unnecessary parameters, potentially introducing slight overfitting or noise from the increased parameter space without adding significant expressive value relevant to the emotional task. This threshold provides crucial guidance: for the affective computing domain, ranks exceeding appear to be associated with diminishing returns and increased cost.

The data clearly illustrates the "sweet spot" where computational efficiency and classification efficacy converge. The PEFT model used for our primary results in Section 3.2 was thus optimally configured with a rank of .

3.4.2. Sensitivity to the Scaling Factor ()

The scaling factor controls the influence of the newly learned low-rank weights () relative to the frozen base weights (). We analyzed its effect by fixing the rank at the optimal and varying .

Table 2: LoRA Scaling Factor Sensitivity on Fine-Grained Emotion Classification ()

Scaling Factor ()	Ratio	Score (%)	Normalized Weight Magnitude	Training Stability
4	0.25	69.9%	Low	High
8	0.50	71.3%	Moderate	High
16	**1.00**	**72.0%**	**Balanced**	**Very High**
32	2.00	70.8%	High	Moderate

The results from the sweep further refined our understanding of the fine-tuning process:

1. Under-Scaling (): When was set to 4 or 8, the resulting performance was suboptimal (and , respectively). By under-scaling the influence of the LoRA matrices, the model seemed to struggle to make sufficient adjustments to its attention mechanism to capture the specialized emotional cues. This suggests that the base model's generalized weights () may have unduly suppressed the task-specific learning necessary for fine-grained distinction.

2. Optimal Scaling (): Setting (the default recommended for optimal rank normalization) yielded the peak performance of . This configuration ensured that the influence of the newly introduced, task-specific low-rank weights was neither dampened nor excessively amplified, which is associated with the most effective convergence towards the optimal decision boundary for emotion classification.

3. Over-Scaling (): When the scaling factor was doubled to , the performance declined to , and the training stability became moderate. Excessive scaling is associated with a potential for noise and instability during the gradient descent optimization, as the large magnitude of the rank update may cause the model to rapidly diverge or oscillate, predicting a suboptimal final state. The model essentially over-specializes and shows limitations in the generalization robustness inherited from the pre-trained weights.

These detailed ablation results provide the necessary justification for the hyperparameters used in the main study. By systematically exploring the parameter space, we have not only confirmed the superiority of the PEFT approach but also provided a reproducible, optimized configuration () for deploying specialized LLMs in fine-grained affective computing tasks. This empirical evidence supports the necessity of domain-specific optimization and provides a clear blueprint for subsequent research endeavors in this field.

4. DISCUSSION

4.1. Interpretation of LLM Superiority and Contextual Understanding

The most significant finding of this study is the decisive prediction of higher performance by the LLM-based approaches, particularly the PEFT-LoRA model, in the highly challenging task of fine-grained emotion recognition. The lead in score over the strongest LSTM baseline is not merely an incremental improvement; it suggests a fundamental change in the capability of the underlying language model.

This superiority stems directly from the design of the Transformer and its pre-training objective. Pre-training on massive, diverse text corpora allows LLMs to develop a rich, generalized "world model" of language. They learn not just the meaning of individual words, but the complex grammatical and semantic relationships that govern how those words combine to convey subtle intent. When a traditional model struggles with an ambiguous sentence, it's often because its limited context window or simpler architecture cannot weigh distant, relevant information. The LLM, by contrast, uses its self-attention mechanism to draw parallel connections across the entire sequence, immediately establishing the overall contextual frame—the narrative, the author's tone, and the likely communicative goal.

The error analysis confirms this: where baselines relied on local, high-frequency cues (e.g., classifying a positive word as a positive emotion), the LLM correctly identified emotional states contingent on non-local cues (e.g., negation, irony, or a sudden change in topic). This capability is paramount in applications like conversational AI, where maintaining an accurate emotional profile of the user over multiple turns of dialogue is necessary to ensure a smooth and helpful interaction.

4.2. Navigating the Ethical and Bias Landscape

As LLMs become ubiquitous tools for emotional analysis, we must confront the ethical shadows associated with their power [18]. The models' superior ability to generalize and extrapolate emotional patterns means they are also capable of generalizing and extrapolating harmful societal biases present in their training data.

Key Insight: This study emphasizes the critical, ongoing problem where LLMs, if not carefully trained, may amplify inherent biases in training data, potentially leading to skewed or harmful emotional analysis. An LLM might be trained predominantly on data reflecting one cultural or linguistic group's expression of "anger," leading it to systematically misinterpret that emotion when expressed by a different group. For example, a model might flag impassioned, non-hostile rhetoric from certain demographic groups as "aggressive" while treating equivalent language from other groups as merely "strong opinion."

Our results and related research underscore the necessity of bias mitigation strategies. While our PEFT approach is excellent for efficiency, it inherits the latent biases of the frozen base model (). Therefore, future work must focus on: (1) Data Curation: Employing datasets specifically sampled and audited for demographic and cultural diversity in emotional expression. (2) Mitigation Techniques: Implementing post-processing or debiasing techniques (e.g., using specialized objective functions) either during the limited PEFT training phase or as a final layer before deployment. The pursuit of powerful affective computing must be inextricably linked with the pursuit of fair and equitable emotional intelligence.

4.2.1. Current and Future Strategies for Mitigating Representational Bias in Affective LLMs

Addressing representational bias in affective LLMs requires a multi-pronged approach that targets the model's lifecycle from data acquisition to deployment. Since PEFT techniques like LoRA primarily fine-tune a small fraction of the model, they rely heavily on the integrity of the foundational pre-trained weights (), necessitating both pre- and post-fine-tuning interventions.

One core strategy involves Data-Centric Bias Mitigation. The goal here is to ensure the fine-tuning data set does not perpetuate or amplify biases that the foundation model may have learned. This includes:

- Intersectionality in Labeling: Emotional labeling must account for intersectional factors (e.g., gender, age, dialect) to prevent a uniform label from masking divergent emotional expressions across groups. For instance, the expression of "joy" in one demographic's written communication may involve different lexical markers than in another's.

- Adversarial Data Augmentation: Creating synthetic or augmented data points that challenge the model's biased assumptions. This involves intentionally generating samples where common stereotypes or biased language cues are inverted or neutralized, forcing the model to rely on deeper contextual cues rather than surface-level heuristics.

A second strategy focuses on Model-Centric Bias Mitigation during the PEFT process. While we freeze the majority of , we can still introduce regularization to the trainable LoRA matrices (and):

- Adversarial Debiasing: Incorporating an auxiliary discriminator network during fine-tuning that is

trained to predict the protected attribute (e.g., demographic group) from the model's internal representations. The primary classification loss is then adjusted to minimize the discriminator's ability to predict the protected attribute, effectively teaching the model to rely on emotion-specific features that are independent of the attribute [24].

- Equalized Odds and Opportunity: Instead of aiming for perfect neutrality, these methods adjust the loss function to ensure that the True Positive Rate (TPR) and False Positive Rate (FPR) for critical emotional classes (e.g., "distress" or "anger") are balanced across different demographic groups. This focus on classification outcomes, rather than just internal representations, is often more actionable in real-world deployment.

Finally, Post-Deployment Auditing and Recalibration are essential. Affective models require continuous monitoring using metrics designed to detect and quantify bias drift over time. This includes establishing Bias Auditing Toolkits that allow practitioners to input samples from different demographic groups and observe divergence in the fine-grained emotional scores. When bias is detected, the small, modular nature of the LoRA weights makes the model much easier to quickly patch and recalibrate than a massive, fully fine-tuned model. The minimal update size is highly beneficial for iterative ethical maintenance, reducing the cost of compliance and promoting continuous fairness [18, 25].

4.3. Implications and Real-World Applications

The dual success of superior performance and extreme efficiency has profound implications for the operational deployment of affective computing systems.

4.3.1. Democratization of LLM Deployment via Optimized PEFT

The empirical findings from Section 3.4 fundamentally reshape the economic calculus for deploying Large Language Models in specialized domains. Prior to the widespread acceptance of PEFT, creating a domain-specific affective model required either accepting the suboptimal performance of generalized, off-the-shelf LLMs or investing in the massive computational infrastructure necessary for Full Fine-Tuning (FFT). The cost of training a single FFT model, demanding thousands of GPU hours, was often associated with a prohibitive barrier for small-to-medium enterprises (SMEs) and academic research groups [14].

Our results, showing that the optimal PEFT configuration () achieves higher performance ( ) with a reduction in trainable parameters and an reduction in training time compared to FFT, indicate a crucial democratization of LLM technology [20]. This efficiency is not merely an incremental benefit; it changes the paradigm:

1. Reduced Barrier to Entry: Specialized affective computing tasks, such as clinical dialogue analysis or toxicity detection for niche online communities, can now be developed and iterated upon rapidly using standard cloud computing resources, rather than requiring specialized supercomputing facilities.

2. Scalable Multi-Task Deployment: Since the base LLM weights () remain frozen, a single instance of the large foundation model can be deployed, and multiple tiny, task-specific LoRA weight matrices () can be loaded and swapped on demand [8]. This drastically reduces the storage and memory footprint required for maintaining a suite of specialized models, allowing affective systems to switch between detecting fine-grained emotion and identifying intent with minimal overhead. The storage requirement for a single LoRA matrix is often less than megabytes, in stark contrast to the hundreds of gigabytes required for multiple full fine-tuned models.

The hyperparameter sensitivity analysis further refines this deployment strategy by providing quantitative proof that maximum performance often does not correlate with maximum parameter count. The observed performance decline at relative to serves as a crucial caution against blindly increasing model complexity. It validates the hypothesis that, for domain-specific fine-tuning, judicious parameter allocation yields not only cost savings but also superior generalization and reduced overfitting.

4.3.2. Engineering for Low-Latency Real-Time Affective Systems

The efficiency derived from the PEFT approach also has direct implications for real-time affective systems and edge computing. An affective computing system deployed in a customer service chatbot or an in-car assistant requires latency measured in milliseconds. While the PEFT training is accelerated, the inference speed is also critically impacted by the LoRA decomposition.

When using LoRA, the computation of adds a minimal computational overhead during inference, primarily due to the rank being very small. However, because the base weights () and the LoRA weights () are often "merged" or "fused" before deployment, the overall inference time is highly competitive with—and sometimes faster than—that of the original base model [33].

This enables the system to rapidly process long streams of text, maintain the user's emotional state over multi-turn conversations, and provide immediate, contextually appropriate responses. For instance, in an automated mental health triage application, the system can use the high throughput enabled by PEFT to continuously

monitor the textual input for sudden spikes in "anxiety" or "distress" scores, allowing for immediate intervention based on the fine-grained classification. This contrasts sharply with legacy systems that often rely on simple keyword spotting or heavily delayed batch processing, which are inadequate for safety-critical, real-time affective scenarios.

The successful implementation of the PEFT approach—driven by the optimal hyperparameter tuning found in our study—marks a vital step toward creating production-ready, low-latency, and high-fidelity emotional intelligence in AI agents. This engineering feasibility closes a key deployment gap and paves the way for the robust application of LLMs in the rapidly evolving landscape of human-computer interaction [11, 2].

4.4. Limitations and Future Research Directions

While this work confirms the efficacy and efficiency of LLMs for textual emotion analysis, several limitations point toward avenues for future research.

Text-Only Limitation and Multimodality

The most significant limitation is our continued reliance on textual data. Key Insight: The ultimate goal of true Affective Computing requires moving beyond text to address the reality that human emotion is fundamentally multimodal. Our current text-based models, while powerful, cannot account for non-verbal cues that may contradict the text (e.g., a person typing "I am fine" with a nervous, shaky tone).

Future work must focus on developing Multimodal Large Language Models (M-LLMs) capable of fusing information from multiple streams—audio, visual (image processing [4, 6, 7]), and textual data [13, 27]. Early exploration in fusing image and text for medical diagnostics shows the promise of M-LLMs in integrating disparate data [10, 15], a concept directly transferable to synthesizing emotion from diverse input channels.

Interpretability and Model Size

Despite the success of PEFT, the foundation LLM remains a large, complex, "black-box" model. Understanding why the model made a specific fine-grained emotional classification (e.g., why "Frustration" and not "Anger") remains challenging. Future research must dedicate effort to improving the interpretability of these models, perhaps through attention visualization or post-hoc explanation techniques, to build trust, especially in high-stakes domains like healthcare.

5. CONCLUSION

The findings of this study provide compelling evidence that Large Language Models (LLMs) predict a definitive advancement in the field of Affective Computing. By leveraging the deep, bidirectional contextual understanding encoded within their transformer architecture, LLMs demonstrate a significantly superior capability over previous methods for the highly challenging task of fine-grained emotion classification.

We successfully addressed the critical barrier of computational cost by introducing and validating the use of Parameter-Efficient Fine-Tuning (PEFT). The PEFT-LoRA methodology achieved not only state-of-the-art performance, predicting higher results than the full fine-tuned model on the score, but also reduced the trainable parameters by and training time by nearly . This dual success ensures that the power of LLMs is now both effective and practically deployable for highly specialized, domain-specific emotional analysis. The rigorous hyperparameter analysis provides an optimized blueprint () for future implementations.

The path forward requires an unwavering focus on ethical deployment and expansion into multimodal data integration. By continuing to refine these efficient techniques and focusing on holistic, non-textual cues, we can pave the way for truly empathic and context-aware artificial intelligence.

REFERENCES

1. Alswaidy, M., Aslan, H. A., Naji, M. A., & Alja'am, J. M. (2023). A systematic review of text sentiment analysis techniques. *Journal of Big Data, 10*(1), 1–33.

2. Argall, B. D., Chernova, S., Veloso, M., & Browning, B. (2009). A survey of robot learning from demonstration. *Robotics and Autonomous Systems, 57*(5), 471–482.

3. Baker, T. L., & Shiffman, S. (2004). The "who, what, when, where, and why" of alcohol use in the natural environment. *Alcohol Research & Health, 28*(4), 169–173.

4. Baltrusaitis, T., Morency, L.-P., & Black, M. J. (2018). OpenFace 2.0: Facial behavior analysis toolkit. *IEEE Transactions on Affective Computing, 10*(4), 502–513.

5. Calvo, R. A., & D'Mello, S. (2010). Affective computing and education: Learning and interacting with emotional machines. *IEEE Transactions on Learning Technologies, 3*(2), 99–111.

6. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 650–660.

7. Chen, Z., Zhao, Y., Feng, Y., Sun, X., & Xu, K. (2020). Cross-modal attention for video sentiment analysis with multimodal feature fusion. *Multimedia Tools and Applications, 79*(11–12), 7935–7953.

8. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint arXiv:2106.09685.*

9. Huang, Y., Liu, D., He, T., Yin, T., He, T., Liu, Z., Deng, D., Zhou, M., & Yang, P. (2019). Efficiently exploring neural network architectures in large search space. *arXiv preprint arXiv:1901.07152.*

10. Kim, S., Wang, Y., Zhang, W., Chen, J., & Ma, X. (2023). Multimodal large language models for medical diagnostics: A systematic review. *Journal of Biomedical Informatics, 142*, 104374.

11. Kulkarni, A. D., & Kulkarni, A. P. (2019). Real-time human emotion detection using facial expressions and convolutional neural networks. *Multimedia Tools and Applications, 78*(13), 18017–18035.

12. Li, H., & Chen, J. (2022). Graph-based text representation learning for emotion recognition. *Expert Systems with Applications, 194*, 116544.

13. Li, Y., Wu, Z., Shi, J., Zhao, H., & Zhou, B. (2022). Multimodal sentiment analysis with contrastive learning and cross-modal attention. *Information Fusion, 81*, 1–13.

14. Liang, P., Wu, H., Zhang, W., & Chen, Y. (2023). Reducing the environmental and economic cost of large language model training. *Nature Energy, 8*, 641–649.

15. Liu, B., Chen, S., & Li, Y. (2023). Vision-language pre-training for medical image analysis: A review. *IEEE Transactions on Medical Imaging, 42*(7), 2154–2169.

16. Liu, X., Zheng, Y., Yang, Q., & Wang, Q. (2021). Hierarchical attention network for emotion recognition in conversation. *Information Processing & Management, 58*(1), 102409.

17. Ma, F., Wang, X., Feng, Z., & Chen, Y. (2022). Context-aware emotion recognition in conversation via graph attention networks. *Neurocomputing, 476*, 28–39.

18. Majumder, N., Poria, S., Hazarika, D., Gelbukh, A., Cambria, E., & Schuller, B. (2019). Sentiment and emotion analysis in conversational agents. *ACM Transactions on Interactive Intelligent Systems, 9*(4), 1–34.

19. Mueller, H., & Zhang, Y. (2020). Transformer-based toxicity detection with attention-based feature fusion. *Pattern Recognition Letters, 131*, 280–286.

20. Ning, S., Zhang, H., Liu, S., Shi, Z., & Li, X. (2023). State-of-the-art parameter-efficient fine-tuning for large language models. *ACM Computing Surveys (CSUR), 55*(9), 1–36.

21. Perez-Rosas, V., Mihalcea, R., & Morency, L.-P. (2017). Computational analysis of deceptive sentiment. *IEEE Transactions on Affective Computing, 8*(3), 362–374.

22. Poria, S., Cambria, E., Hazarika, D., & Majumder, N. (2017). SenticNet 5.0: Enhancing affective computing with analogy-based common-sense reasoning. *IEEE Transactions on Affective Computing, 10*(2), 195–207.

23. Risch, J., & Krestel, R. (2018). Sentiment analysis in historical texts. *Expert Systems with Applications, 114*, 468–479.

24. Romei, A., & Gambardella, L. M. (2009). A survey of evolutionary data mining. *Expert Systems with Applications, 36*(7), 442–452.

25. Salminen, M., Riekkinen, K., & Hautamäki, A. (2022). Ethical considerations in emotion AI deployment: A systematic review. *AI & Society, 37*(4), 1321–1339.

26. Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Wilson, S., & Baird, A. (2013). The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, and personality. *Proceedings of INTERSPEECH 2013*, 1–5.

27. Sun, H., Li, T., & Wang, Y. (2021). Multimodal emotion recognition with fusion of textual and visual features. *Expert Systems with Applications, 167*, 114170.

28. Tang, D., Qin, B., & Liu, T. (2015). Document modeling with gated recurrent neural network for sentiment classification. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1388–1397.

29. Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with Twitter: What 140 characters reveal about political sentiment. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (ICWSM)*, 178–185.

30. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I.

(2017). Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS 2017), 30*, 5998–6008.

31. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 4171–4186.

32. Liu, Y., Ott, M., Goyal, N., Du, J., Li, M., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692.

33. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Le, Q. V., & Salakhutdinov, R. (2019). XLNet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems (NeurIPS 2019).

34. Samantapudi, R. K. R. (2025). Advantages & impact of fine tuning large language models for ecommerce search. Journal of Information Systems Engineering and Management, 10(45s), 600–622. https://doi.org/10.52783/jisem.v10i45s.8898

35. Srilatha, S. (2025). Integrating AI into enterprise content management systems: A roadmap for intelligent automation. Journal of Information Systems Engineering and Management, 10(45s), 672–688. https://doi.org/10.52783/jisem.v10i45s.8904

36. Rangu, S. (2025). Analyzing the impact of AI-powered call center automation on operational efficiency in healthcare. Journal of Information Systems Engineering and Management, 10(45s), 666–689. https://doi.org/10.55278/jisem.2025.10.45s.666