From Context to Emotion: Leveraging LLMs for Recognizing Implicit Emotions

Hanane Boutouta, Abdelaziz Lakhfif, Ferial Senator, and **Chahrazed Mediani**
Artificial Intelligence laboratory (AI-lab), Department of Computer Science,
Setif 1 University, Ferhat ABBAS, Setif, Algeria
{hanane.boutouta, abdelaziz.lakhfif, ferial.senator, chahrazed.mediani}@univ-setif.dz

Abstract

Implicit Emotion Recognition (IER) is a challenging task in Natural Language Processing (NLP), as it requires identifying emotions that are not directly expressed through explicit emotion words but must be inferred from contextual, situational, or linguistic cues. With the rapid progress of Large Language Models (LLMs), new opportunities have emerged for tackling such complex language understanding tasks. In this work, we investigate the effectiveness of two different architectures of LLMs for IER: masked language models, including BERT and RoBERTa, and causal language models, represented by ChatGPT. We fine-tuned BERT and RoBERTa on benchmark IER datasets, while we evaluated ChatGPT in a zero-shot setting to assess its ability to generalize without task-specific training. Our experiments on the ISEAR and IEST datasets show that fine-tuned masked language models perform strongly on the IER task. At the same time, ChatGPT achieves promising results in zero-shot scenarios, highlighting its potential for emotion recognition tasks with limited or no labeled data.

1 Introduction

Text-based Emotion Recognition (ER) is a fundamental research area in Natural Language Processing (NLP). In recent years, this field has seen important advancements due to increased human-computer interaction, as well as the rapid growth of online social media (Bisogni et al., 2023). ER can be classified into explicit ER (EER) and implicit ER (IER), depending on whether explicit emotional words emerge in the text (Kusal et al., 2021). Different from EER, where emotional words (e.g., happy, angry) occur in the text, in IER, emotions must be inferred from linguistic cues such as contextual descriptions, metaphorical expressions, or situational events without any explicit emotional expression (Klinger et al., 2018). Implicit emotions often require deep semantic understanding to interpret subtle cues, such as sarcasm (Perfect, just what I needed) (Zhu et al., 2025), ambiguous statements (e.g., She looked out the window as the train pulled away, which could imply sadness, longing, or even relief) (Orizu, 2018), or behavioral context (e.g., They all left without me). This makes IER particularly challenging due to subjectivity, cultural variability, and strong dependence on context.

Researchers have proposed several ER approaches, including lexicon-based, machine learning, and deep learning methods. Most of these approaches primarily focus on extracting explicit emotions, whereas recognizing implicit emotions poses a greater challenge, as it demands sophisticated techniques capable of accurately interpreting context and deeply understanding nuanced linguistic patterns.

Recent advancements in Large Language Models (LLMs) have revolutionized NLP by achieving state-of-the-art performance across a wide range of tasks, such as question answering (Goar et al., 2023), machine translation (Hendy et al., 2023), and sentiment analysis (Ding et al., 2022). These models have also demonstrated remarkable capabilities in comprehending, interpreting, and recognizing human emotions (Banimelhem and Amayreh, 2023; Lee et al., 2024). LLMs, trained on large-scale and extensive corpora, have demonstrated a deep understanding of linguistic patterns, contextual dependencies, and even some aspects of world knowledge, enabling them to infer meaning and emotion from text based on the surrounding context in ways that were previously unattainable (Hong et al., 2024; Buscemi and Proverbio, 2024). These capabilities may be particularly useful for tasks like IER, where emotions are not explicitly stated but must be inferred from subtle linguistic cues, situational context, or background knowledge. Unlike traditional methods that rely heavily on explicit emotional keywords or rule-based systems, LLMs

leverage their transformer-based architecture, contextual embeddings, and pretrained knowledge to analyze the interplay of words and sentences, to decode emotional tones and comprehend the complexity of emotions.

In this study, we aim to explore the effectiveness of LLMs, specifically BERT, RoBERTa, and ChatGPT, in the task of IER. We assess the performance of fine-tuned, encoder-based models, including BERT and RoBERTa architectures, to evaluate their suitability and effectiveness for IER. Furthermore, we investigate ChatGPT's capabilities in a zero-shot learning setting to determine its ability to generalize to IER without task-specific fine-tuning. This approach highlights its potential for applications where labeled data is limited or unavailable.

This comparison provides insight into the strengths and weaknesses of these LLMs in capturing implicit emotional cues, contributing to a deeper understanding of their real-world applicability. The main contributions of this paper are summarized as follows:

- We conduct a comparative analysis of two distinct paradigms for IER:
  1. Fine-tuned masked language models (BERT and RoBERTa), and
  2. Zero-shot prompting using a causal language model, specifically ChatGPT.
- We fine-tune BERT and RoBERTa on labeled emotion datasets to evaluate their task-specific performance in recognizing implicit emotions.
- We evaluate the performance of ChatGPT to generalize to the IER task, in a zero-shot setting without the need for task-specific fine-tuning or additional training.
- We provide empirical evidence on the effectiveness, limitations, and generalization capabilities of ChatGPT in contrast to traditional fine-tuned models.

The remainder of this paper is structured as follows: Section 2 presents a concise review of related work on IER and recent developments in LLMs. Section 3 introduces the datasets used in our experiments. Section 4 provides some details about the experiments' setup, including the models employed and the different methodological approaches used. Section 5 presents and discusses the results of our experiments. Finally, Section 6 concludes the paper and outlines potential directions for future research.

2 Related Work

IER has emerged as a complex and less explored task within the field of NLP. Unlike EER, which relies on identifying overt emotion words, IER requires understanding contextual and semantic cues to infer emotional states. This task has been addressed using various approaches, including rule-based, classical machine learning, deep learning, and, more recently, transformer-based approaches (Alswaidan and Menai, 2020).

Early efforts relied on knowledge-based and rule-based approaches. For instance, EmotiNet linked events to emotions through commonsense knowledge, while cognitive-theory-inspired rules attempted to capture implicit affective states (Balahur et al., 2011, 2012; Udochukwu and He, 2015). Classical algorithms, such as Support Vector Machines (SVM) and Naive Bayes (NB), were combined with lexical features, syntactic patterns, and semantic resources to infer implicit emotions (Balahur et al., 2012; Riahi and Safari, 2016; Khoshnam and Baraani-Dastjerdi, 2022). These methods, although somewhat effective, have had difficulty with generalization due to the complexity of implicit emotional expressions and the absence of explicit emotional words.

The emergence of deep learning (DL) has introduced new avenues for recognizing implicit emotions in textual data, leveraging neural architectures such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to capture complex linguistic and contextual patterns. Models such as LSTM and BiLSTM incorporated with attention mechanisms demonstrate improved performance by capturing temporal dependencies and contextual information (Rozental et al., 2018; Balazs et al., 2018; Chronopoulou et al., 2018; Rathnayaka et al., 2018; Zhou and Wu, 2018; Witon et al., 2018; Pecar et al., 2018; Fei et al., 2019). Recently, transformer-based models like BERT (Devlin et al., 2018) further enhanced the IER task by leveraging pre-trained embeddings and self-attention mechanisms, making them well suited for understanding implicit cues (Khoshnam et al., 2022; Qian et al., 2023; Boutouta et al., 2025).

Transformer-based LLMs, such as ChatGPT, have significantly expanded the possibilities of the NLP field, demonstrating remarkable performance across a wide range of tasks. These tasks include text understanding and generation (Mitrović et al., 2023; Gao et al., 2024), machine translation (Peng

et al., 2023), sentiment analysis (Buscemi and Proverbio, 2024), and semantic role labeling (Senator et al., 2025). Their strong generalization capabilities and the ability to capture contextual nuances enable more accurate emotion identification without requiring additional training (Kadiyala, 2024; Banimelhem and Amayreh, 2023; Lee et al., 2024; Hong et al., 2024; Liu et al., 2024). A recent study by Hong et al. (2024) introduced a method that addresses the complex and ambiguous nature of human emotions by using LLMs for ER. The approach considers multiple emotion labels and the intricate nature of emotional expressions. Another work proposed EmoLLMs (Liu et al., 2024), a series of open-source instruction-following LLMs fine-tuned for comprehensive affective analysis. These models are trained on a diverse dataset covering various classification and regression tasks related to emotions, enhancing their applicability in ER tasks. In another study (Wake et al., 2023), the performance of ChatGPT in the area of emotion detection was assessed on a variety of datasets, including IEMOCAP and DailyDialog. ChatGPT was able to classify text with emotional labels in both zero-shot and fine-tuning settings.

Despite these significant advances in ER, existing studies have predominantly focused on EER, with limited attention given to the IER task. To the best of our knowledge, none of the existing works have comprehensively addressed the unique challenges of IER, nor have they fully examined the potential of LLMs, such as ChatGPT, within this context.

3 Datasets

Two datasets were used: the WASSA-2018 Implicit Emotions Shared Task (IEST) dataset (Klinger et al., 2018) and the International Survey on Emotion Antecedents and Reactions (ISEAR) dataset (Scherer, 2005). Both datasets are widely used for ER, but differ significantly in terms of domain, format, and emotion expression. Table 5 in the Appendix A presents a brief comparison between the IEST and ISEAR datasets, while the label distributions are shown in Fig. 1 and Fig. 2.

3.1 IEST

The IEST dataset[1], introduced by Klinger et al. (2018), was developed for the WASSA-2018 Implicit Emotions Shared Task. It is a large automatically labeled dataset of 191,731 English tweets, split into 153,600 for training, 9,600 for validation, and 28,800 for testing. Each tweet is annotated with one of Ekman's six basic emotions: anger, disgust, fear, joy, sadness, or surprise. Given computational constraints, only the testing set of the IEST dataset was used in this study.

To simulate implicit emotion scenarios, each tweet in the dataset has had its explicit emotion word masked and replaced with a placeholder token [#TARGETWORD#]. This design forces models to rely only on contextual cues to infer the underlying emotion, making it particularly suited for research on emotion understanding in indirect and implicit expressions. Some examples from the dataset are provided in Table 6 in the Appendix A.

3.2 ISEAR

The ISEAR dataset[2], introduced by Scherer (2005), is a manually labeled dataset collected as part of a psychological study aimed at exploring emotional experiences across cultures. The data were gathered from over 3,000 participants in 26 countries, all of whom had university-level education and were fluent in English. Each participant was asked to describe situations in which they had personally experienced one of seven emotions: joy, fear, anger, sadness, disgust, shame, and guilt. In total, the dataset contains approximately 7,666 instances, making it one of the most widely cited benchmarks for ER in psychology and affective computing tasks. Examples from the dataset are provided in Table 7 in Appendix A.

3.3 Data pre-processing

To align the ISEAR dataset with the IER task, we applied an additional filtering step: we ensured that none of the selected instances contained explicit emotion words. This pre-processing step allows us to reuse the ISEAR as a proxy dataset for IER, focusing only on instances where emotions must be inferred from the described context rather than directly stated. Furthermore, both datasets were subjected to standard pre-processing steps, including the removal of HTML tags, URLs, emojis, and extra spaces, as well as the correction of inconsistent punctuation.

[1] https://implicitemotions.wassa2018.com/data/

[2] https://github.com/sinmaniphel/py_isear_dataset

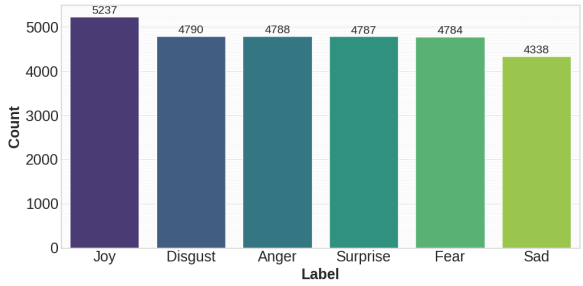

Figure 1: Distribution of emotion labels in the IEST dataset.

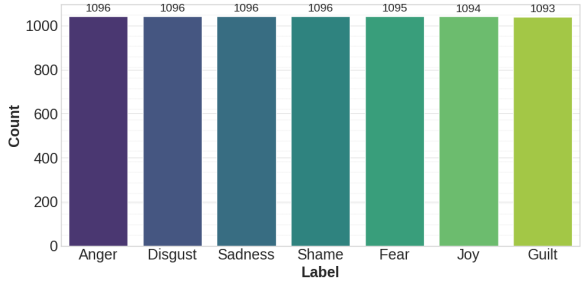

Figure 2: Distribution of emotion labels in the ISEAR dataset.

4 Experimental Methodology

We investigate two prominent approaches for text classification in the context of IER: (1) fine-tuning masked language models, and (2) prompt-based interaction with causal language LLMs. For the first approach, we employ BERT and RoBERTa, both pre-trained transformer encoders that are fine-tuned on task-specific data. These models have been widely recognized for their ability to capture contextual semantics and perform well across a range of NLP tasks. In this setup, the models are initialized with pre-trained weights and then fine-tuned using supervised learning on labeled emotion data. In contrast, the second approach utilizes ChatGPT, a large, decoder-based LLM, accessed via zero-shot prompting. Rather than fine-tuning the model, we interact with ChatGPT using carefully crafted prompts that define the task and specify the desired output format. This method evaluates ChatGPT's ability to generalize to the IER task without the need for additional training or fine-tuning.

By comparing these two paradigms, we aim to assess the trade-offs in performance, flexibility, and data efficiency when applied to implicit emotion classification.

4.1 Models

4.1.1 BERT

A state-of-the-art NLP model introduced by Google in 2018 (Devlin et al., 2018) revolutionized the field by leveraging a bidirectional transformer architecture, which allows it to capture context from both the left and right of a word simultaneously. BERT is pre-trained on large text corpora using two key objectives: Masked Language Modeling (MLM), where it predicts randomly masked words within a sentence, and Next Sentence Prediction (NSP), where it learns to determine whether one sentence logically follows another. These pre-training tasks enable BERT to develop a deep understanding of both semantic meaning and syntactic structure in natural language.

4.1.2 RoBERTa

Developed by Facebook AI in 2019 upon the foundational BERT architecture (Liu, 2019). RoBERTa improves and optimizes BERT's pre-training process by removing the NSP objective, training on significantly larger datasets, and employing dynamic masking during pre-training. These enhancements lead to improved performance on a wide range of natural language understanding tasks.

4.1.3 ChatGPT

An advanced LLM developed by OpenAI in November 2022 (OpenAI), based on the Generative Pre-trained Transformer (GPT) architecture, a causal variant of the transformer neural network that has become the industry standard for a wide range of NLP tasks (Gillioz et al., 2020). Unlike masked language models, GPT models are trained in an autoregressive manner to predict the next token in a sequence, enabling strong generative and contextual reasoning abilities. ChatGPT was trained on a vast and diverse corpus, including academic texts, literary works, and large-scale web content, which equips it with broad linguistic and world knowledge. One of its key features is its ability to generate coherent, contextually relevant, and human-like responses to user input. Through interactive prompt-based querying, ChatGPT can adapt flexibly to new tasks without the need for additional fine-tuning.

4.2 Evaluation Approaches

4.2.1 Fine-tuning encoder-based models

Fine-tuning involves adapting the pre-trained language models to a specific task by training them

on a smaller, task-specific dataset. This process requires significantly less data compared to training a model from scratch, thanks to the rich linguistic knowledge already encoded in the pre-trained model parameters. Regarding encoder-based models, we explored a range of hyperparameters configurations to optimize performance. Specifically, we experimented with different learning rates (1e-5, 2e-5, and 3e-5), batch sizes (16 and 32), training durations (ranging from 3 to 6 epochs), and maximum sequence lengths (64, 128, and 512). Additionally, we compared different model variants, including base and large versions, to assess their suitability for the IER task.

Each configuration was evaluated using a 10% development split of the training data, and the optimal setup was selected based on the macro F1-score. The chosen hyperparameters were validated across various random seeds to ensure robustness. Table 1 summarizes both the tested and optimal hyperparameter configurations.

We used the pre-trained "bert-base-uncased" and "roberta-base" models from the Huggingface Transformers library. The models consist of 12 transformer layers, a hidden size of 768, and 12 attention heads. For both models, we appended a dense layer with a softmax activation function for classification. The models were trained for 4 epochs using a batch size of 32 and a maximum sequence length of 128. Training was performed using the Adam optimizer and categorical cross-entropy loss. We evaluated these models on held-out test sets comprising 10% of the IEST and ISEAR datasets.

4.2.2 Prompt Design for Zero-Shot IER

For ChatGPT, we evaluated its zero-shot performance on test sets consisting of 600 and 700 instances from the IEST and ISEAR datasets, respectively (100 instances per emotion). As a proprietary model, ChatGPT was accessed via its chatbot interface using the GPT-4 Turbo version. To eliminate potential influence from prior context, each input was submitted in a separate chat session, ensuring full isolation between predictions.

Carefully designed zero-shot prompts are essential for enabling LLMs to generalize effectively across diverse domains (Team et al., 2023). We prompted ChatGPT with a text sample, a predefined list of emotion labels, task-specific instructions, and a set of output constraints. The prompts were iteratively designed and refined to align with the task's unique demands, namely, detecting emotional states without the presence of explicit emotion words. Early versions of the prompt included basic task instructions. However, we observed improved performance when the implicit nature of the task was explicitly stated, when emotion label choices were clearly specified, and when the model's role was defined. For example, we experimented with formulations such as "the emotion is implied rather than stated." Additionally, we consistently framed the model as an "expert in implicit emotion recognition" at the beginning of each interaction to guide its behavior.

After multiple iterations, the final prompt adopted was:

> ***Role***: *You are an expert in implicit emotion recognition.*
> ***Prompt***: *The following sentence contains an emotion that is expressed implicitly. Based on context alone, identify the most likely emotion. Choose only one from: [Emotion List].*
> *Respond with the emotion without any explanation.*
> ***Text***: *[example text]*

This final format was selected after testing several prompt versions on a development subset of the IEST and ISEAR datasets, evaluating performance manually and through agreement with gold-standard labels. We observed that prompting clarity, emotion list formatting, and explicit task framing significantly affected model responses.

4.3 Evaluation Metrics

Classification problem's performance is evaluated using a set of metrics. In our case, we use the accuracy and the macro average precision, recall, and F1-score. Each metric is defined in accordance with the following equations: (1), (2), (3), and (4), respectively. Where TP, TN, FP, and FN represent the number of True Positives, True Negatives, False Positives, and False Negatives, respectively.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1 - score = 2 \cdot \frac{Precision * Recall}{Precision + Recall} \quad (4)$$

Hyperparameter	Tested Values	Optimal Value
Learning Rate	1e-5, 2e-5, 3e-5	1e-5
Loss Function	Categorical Cross-Entropy	Categorical Cross-Entropy
Optimizer	Adam	Adam
Batch Size	16, 32	32
Epochs	3, 4, 5, 6	4
Max Length	64, 128, 512	128

Table 1: Hyperparameter settings

5 Results and Discussion

Table 2 provides a concise overview of the performance of encoder-based LLMs (BERT and RoBERTa) and the decoder-based LLM (ChatGPT), using different approaches (fine-tuning and zero-shot prompting) across the IEST and ISEAR datasets.

5.1 Adaptation and Generalization

ChatGPT achieves the highest accuracy (77.14%) and F1-score (77.00%) on the ISEAR dataset, outperforming fine-tuned BERT and RoBERTa on the same dataset, which achieve an accuracy of 70.36% and 70.84%, respectively. However, on the IEST dataset, fine-tuned RoBERTa performs best (66.88% accuracy, 66.67% F1-score), while ChatGPT's performance drops significantly (54.17% accuracy, 54.93% F1-score). These results highlight a fundamental distinction between generalization and adaptation in IER. ChatGPT, as a causal language model, leverages broad pre-training to generalize well on datasets like ISEAR, where contextual cues align with its prior knowledge. In contrast, fine-tuned BERT and RoBERTa models, as masked language models, demonstrate superior adaptation to domain-specific constraints in IEST, where emotional keywords are masked, and cues are subtle. The masked modeling architecture, coupled with task-specific fine-tuning, equips these models with the ability to capture fine-grained contextual dependencies tailored to the dataset's structure, whereas ChatGPT's causal generation approach, optimized for predicting the next token, may be less effective in such constrained contexts. This performance gap underscores how model architecture and training paradigms interact with dataset characteristics to shape success in IER.

5.2 Performance Variation Across Datasets

As we show in Fig. 3, all models consistently performed better on the ISEAR dataset than on the IEST dataset. A possible reason for this finding is the contrast between the two datasets. While the ISEAR dataset was originally developed for general ER, we adapted it for the IER task by excluding any instances containing explicit emotion words (as noted in Section 3.3). This ensured that emotional states had to be inferred from contextual and situational cues rather than directly stated. Nevertheless, ISEAR dataset remains more clear, formal, consisting of well-structured, self-reported emotional experiences. These descriptions tend to be complete, coherent, and grammatically consistent. In contrast, the IEST dataset is derived from social media (tweets), which are often informal, fragmented, noisy, and contextually ambiguous. In addition, tweets may include slang, sarcasm, or cultural references that are not easily interpreted without broader context. This shift in genre presents additional challenges for IER, as models must not only infer unstated emotions but also navigate less structured and noisier linguistic input. We include representative examples from both datasets and a comparative table in Appendix A to illustrate these variations.

5.3 Emotional Implicitness

When considering the implicit emotional expression in each dataset, the IEST dataset represents masked emotion as a proxy for implicit emotion, where explicit emotion words were originally present in the sentence but have been deliberately removed. This deliberate omission weakens contextual support, forcing models to infer emotions from incomplete or ambiguous linguistic cues. In contrast, ISEAR contains naturally implicit emotions embedded within coherent, narrative-style descriptions of personal experiences. These richer and more structured contexts provide clearer situational signals, which both fine-tuned models and ChatGPT exploited to infer emotions more effectively. This distinction highlights how the availability and

Approach	Model	ISEAR		IEST	
		Acc (%)	F1 (%)	Acc (%)	F1 (%)
Fine-tuned	BERT	70.36	69.66	62.02	61.52
	RoBERTa	70.84	70.34	**66.88**	**66.67**
Zero shot	ChatGPT	**77.14**	**77.00**	54.17	54.93

Table 2: Performance comparison of different models on ISEAR and IEST datasets.

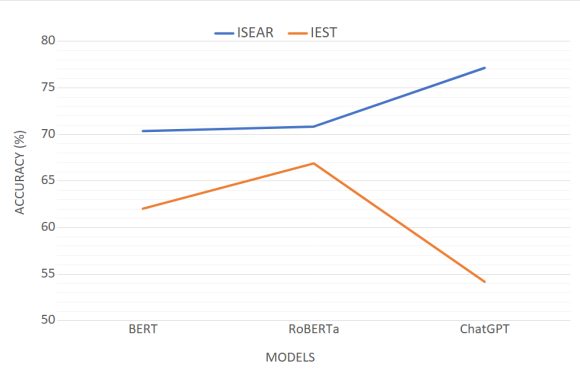

Figure 3: Performance comparison of different models on ISEAR and IEST datasets.

quality of contextual information directly shape the difficulty of implicit emotion recognition, with IEST posing a greater challenge due to its sparse and less informative cues.

5.4 Performance on Individual Emotions

Table 3 presents the performance of the three models for each emotion on the ISEAR dataset. As indicated, ChatGPT demonstrates superior performance for the majority of the emotions. Specifically, it achieves an F1-score of 94% for the emotion 'joy' and 84% for 'fear,' significantly outperforming the fine-tuned BERT and RoBERTa models. This demonstrates its strong ability to recognize emotions with clear contextual cues. However, all models show relatively poor performance on shame compared to other emotions, with F1-scores of 47%, 51%, and 59% for BERT, RoBERTa, and ChatGPT, respectively. The lower scores for shame reflect the challenge of detecting emotions that are highly implicit, underscoring the critical role of contextual clarity in IER.

The results on the IEST dataset are presented in Table 4, revealing a different trend compared to the ISEAR dataset. In this case, RoBERTa achieves the best performance across most emotions, with F1-scores of 77%, 78%, and 58% for the emotions 'joy', 'fear', and 'anger', respectively. These results significantly outperform those of the BERT and ChatGPT models.

We also noticed significant variation when examining performance based on individual emotion labels. For example, in the zero-shot experiments on the ISEAR dataset, the recognition performance (F1-score) for 'joy' was around 94%, while it was below 60% for 'shame'. Similarly, on the IEST dataset, the F1-score for 'fear' was around 66%, while it was below 46% for 'anger'. In the fine-tuning approach, we observed that IER performance varied significantly across datasets, even for similar emotions. For instance, in the ISEAR dataset, the recognition performance (F1-score) for 'joy' was around 86% and 92% for BERT and RoBERTa models, respectively, while in the IEST dataset, the F1-scores for the same emotion (joy) were only around 70% and 77% for the same models, respectively. Notably, this tendency is also observed in the zero-shot condition with ChatGPT. For example, in the ISEAR dataset, the F1-score for 'anger' was around 73%, while in the IEST dataset, it was only around 45% with ChatGPT. This contrast in performance demonstrated that IER is a challenging task, with performance varying significantly depending on emotions, datasets, and models.

6 Conclusions and future works

In this study, we examined the effectiveness of two different architectures of LLMs for recognizing implicit emotions: masked language models, including BERT and RoBERTa, via a series of fine-tuning experiments, and causal language models, represented by ChatGPT, using a zero-shot prompting approach. The models were tested on two datasets: IEST and ISEAR. Both datasets are widely used for ER, but they differ significantly in terms of domain, format, and emotion expression. Our findings indicate that BERT-based fine-tuned models, particularly RoBERTa, excel at capturing implicit emotional cues. In contrast, zero-shot ChatGPT delivers promising results for certain emotion categories but struggles with more com-

Model	Joy			Fear			Anger			Sadness			Disgust			Shame			Guilt		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
BERT	0.82	0.91	0.86	**0.78**	0.78	0.78	0.63	0.57	0.60	0.70	0.83	0.76	0.76	0.76	0.76	0.65	0.37	0.47	0.58	**0.72**	0.64
RoBERTa	0.93	0.91	0.92	0.65	0.79	0.71	0.61	0.63	0.71	0.72	0.83	0.78	0.71	**0.80**	0.75	0.63	0.43	0.51	**0.69**	0.57	**0.78**
ChatGPT	**0.96**	**0.92**	**0.94**	**0.78**	**0.90**	**0.84**	**0.72**	**0.73**	**0.73**	**0.74**	**0.92**	**0.82**	**0.91**	0.70	**0.79**	**0.67**	**0.53**	**0.59**	0.65	**0.72**	0.68

Table 3: Performance comparison of different models per emotion on ISEAR dataset.

Model	Joy			Fear			Anger			Sad			Disgust			Surprise		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
BERT	0.66	**0.75**	0.70	0.64	**0.80**	0.71	**0.57**	0.47	0.51	0.60	0.53	0.57	**0.67**	0.53	0.57	0.57	**0.64**	**0.60**
RoBERTa	**0.79**	**0.75**	**0.77**	**0.78**	0.78	**0.78**	0.55	**0.62**	**0.58**	**0.63**	**0.66**	**0.65**	0.65	**0.59**	**0.62**	0.61	0.59	**0.60**
ChatGPT	0.65	0.61	0.63	0.77	0.57	0.66	0.36	**0.62**	0.45	0.48	0.53	0.50	0.60	0.52	0.56	**0.66**	0.40	0.50

Table 4: Performance comparison of different models on emotion classification on IEST dataset.

plex and context-dependent cases, where its performance declines noticeably. These results highlight the strengths of fine-tuned, medium-sized language models in handling IER tasks, while also underscoring the potential of zero-shot LLMs for emotions that are simpler or positively valenced. However, progress in IER remains constrained by the scarcity of high-quality datasets. Emotions are often conveyed indirectly, and building datasets that capture this nuance without relying on explicit markers is inherently challenging. Despite its limitations, the IEST dataset serves as a practical proxy by simulating implicitness through masked emotion words, offering a controlled evaluation setting.

Future research would benefit from the development of more diverse and realistic datasets for IER, as current resources are limited and often fail to capture the nuanced and context-dependent nature of implicit expressions. In addition to zero-shot prompting, we will explore alternative strategies such as few-shot learning and fine-tuning with LLMs, aiming to combine the adaptability of prompt-based approaches with the task-specific precision of supervised learning. Finally, addressing the persistent challenge of detecting socially complex emotions, such as *shame*, *guilt*, remains an important direction for future investigation, as these emotions often rely on subtle discourse cues and cultural context.

Limitations

Despite the promising results presented in this study, some limitations should be considered. First, the evaluation was restricted to two benchmark datasets, ISEAR and IEST, which, although widely used in the field of ER, may not comprehensively reflect the variability of implicit emotional expressions encountered in real-world scenarios, particularly in social media or multilingual contexts. This raises concerns regarding the generalization of the findings. Second, while LLMs exhibit an ability to capture certain contextual and cultural cues, their comprehension remains limited in the presence of more nuanced expressions such as sarcasm, idioms, or domain-specific references, which are common in implicit emotional content. Lastly, the lack of interpretability remains a critical challenge, particularly with generative models like ChatGPT, which operate as black boxes. This opacity hinders the ability to understand or explain model decisions, posing a barrier to trust and transparency in practical applications.

References

Nourah Alswaidan and Mohamed El Bachir Menai. 2020. A survey of state-of-the-art approaches for emotion recognition in text. *Knowledge and Information Systems*, 62(8):2937–2987.

Alexandra Balahur, Jesús M Hermida, and Andrés Montoyo. 2011. Detecting implicit expressions of sentiment in text based on commonsense knowledge. In *Proceedings of the 2nd workshop on computational approaches to subjectivity and sentiment analysis (WASSA 2.011)*, pages 53–60.

Alexandra Balahur, Jesús M Hermida, and Andrés Montoyo. 2012. Detecting implicit expressions of emotion in text: A comparative analysis. *Decision support systems*, 53(4):742–753.

Jorge Balazs, Edison Marrese-Taylor, and Yutaka Matsuo. 2018. IIIDYT at IEST 2018: Implicit emotion classification with deep contextualized word representations. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 50–56, Brussels, Belgium. Association for Computational Linguistics.

Omar Banimelhem and Wlla Amayreh. 2023. The performance of chatgpt in emotion classification. In *2023 14th International Conference on Information and Communication Systems (ICICS)*, pages 1–4. IEEE.

Carmen Bisogni, Lucia Cimmino, Maria De Marsico, Fei Hao, and Fabio Narducci. 2023. Emotion recognition at a distance: The robustness of machine learning based on hand-crafted facial features vs deep learning models. *Image and Vision Computing*, 136:104724.

Hanane Boutouta, Abdelaziz Lakhfif, Ferial Senator, and Chahrazed Mediani. 2025. A transformer-based hybrid model for implicit emotion recognition in arabic text. *Engineering, Technology & Applied Science Research*, 15(3):23834–23839.

Alessio Buscemi and Daniele Proverbio. 2024. Chatgpt vs gemini vs llama on multilingual sentiment analysis. *arXiv preprint arXiv:2402.01715*.

Alexandra Chronopoulou, Aikaterini Margatina, Christos Baziotis, and Alexandros Potamianos. 2018. NTUA-SLP at IEST 2018: Ensemble of neural transfer methods for implicit emotion classification. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 57–64, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Shafiq Joty, Boyang Li, and Lidong Bing. 2022. Is gpt-3 a good data annotator? *arXiv preprint arXiv:2212.10450*.

Hao Fei, Yafeng Ren, and Donghong Ji. 2019. Implicit objective network for emotion detection. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 647–659. Springer.

Ge Gao, Jongin Kim, Sejin Paik, Ekaterina Novozhilova, Yi Liu, Sarah T Bonna, Margrit Betke, and Derry Tanti Wijaya. 2024. Enhancing emotion prediction in news headlines: Insights from chatgpt and seq2seq models for free-text generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5944–5955.

Anthony Gillioz, Jacky Casas, Elena Mugellini, and Omar Abou Khaled. 2020. Overview of the transformer-based models for nlp tasks. In *2020 15th Conference on computer science and information systems (FedCSIS)*, pages 179–183. IEEE.

Vishal Goar, Nagendra Singh Yadav, and Pallavi Singh Yadav. 2023. Conversational ai for natural language processing: An review of chatgpt. *International Journal on Recent and Innovation Trends in Computing and Communication*, 11:109–117.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Xin Hong, Yuan Gong, Vidhyasaharan Sethu, and Ting Dang. 2024. Aer-llm: Ambiguity-aware emotion recognition leveraging large language models. *arXiv preprint arXiv:2409.18339*.

Ram Mohan Rao Kadiyala. 2024. Cross-lingual emotion detection through large language models. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 464–469.

Fereshteh Khoshnam and Ahmad Baraani-Dastjerdi. 2022. A dual framework for implicit and explicit emotion recognition: An ensemble of language models and computational linguistics. *Expert Systems with Applications*, 198:116686.

Fereshteh Khoshnam, Ahmad Baraani-Dastjerdi, and MJ Liaghatdar. 2022. Cefer: A four facets framework based on context and emotion embedded features for implicit and explicit emotion recognition. *arXiv preprint arXiv:2209.13999*.

Roman Klinger, Orphée De Clercq, Saif Mohammad, and Alexandra Balahur. 2018. IEST: WASSA-2018 implicit emotions shared task. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 31–42, Brussels, Belgium. Association for Computational Linguistics.

Sheetal Kusal, Shruti Patil, Ketan Kotecha, Rajanikanth Aluvalu, and Vijayakumar Varadarajan. 2021. Ai based emotion detection for textual big data: Techniques and contribution. *Big Data and Cognitive Computing*, 5(3):43.

Sanghyub John Lee, Hyunseo Tony Lee, and Kiseong Lee. 2024. Enhancing emotion detection through chatgpt-augmented text transformation in social media text. In *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*, pages 872–879. IEEE.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Zhiwei Liu, Kailai Yang, Qianqian Xie, Tianlin Zhang, and Sophia Ananiadou. 2024. Emollms: A series of emotional large language models and annotation tools for comprehensive affective analysis. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5487–5496.

Sandra Mitrović, Davide Andreoletti, and Omran Ayoub. 2023. Chatgpt or human? detect and explain. explaining decisions of machine learning model for detecting short chatgpt-generated text. *arXiv preprint arXiv:2301.13852*.

OpenAI. Chatgpt. Accessed: 2025-02-23.

Udochukwu Orizu. 2018. *Implicit emotion detection in text*. Ph.D. thesis, Aston University.

Samuel Pecar, Michal Farkas, Marián Šimko, Peter Lacko, and Maria Bielikova. 2018. Nl-fiit at iest-2018: Emotion recognition utilizing neural networks and multi-level preprocessing. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 217–223.

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards making the most of chatgpt for machine translation. *arXiv preprint arXiv:2303.13780*.

Yanjun Qian, Jin Wang, Dawei Li, and Xuejie Zhang. 2023. Interactive capsule network for implicit sentiment analysis. *Applied Intelligence*, 53(3):3109–3123.

Prabod Rathnayaka, Supun Abeysinghe, Chamod Samarajeewa, Isura Manchanayake, and Malaka Walpola. 2018. Sentylic at iest 2018: Gated recurrent neural network and capsule network-based approach for implicit emotion detection. *arXiv preprint arXiv:1809.01452*.

Nooshin Riahi and Pegah Safari. 2016. Implicit emotion detection from text with information fusion. *Journal of Advances in Computer Research*, 7(2):85–99.

Alon Rozental, Daniel Fleischer, and Zohar Kelrich. 2018. Amobee at iest 2018: Transfer learning from language models. *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*.

Klaus R Scherer. 2005. What are emotions? and how can they be measured? *Social science information*, 44(4):695–729.

Ferial Senator, Abdelaziz Lakhfif, Imene Zenbout, Hanane Boutouta, and Chahrazed Mediani. 2025. Leveraging chatgpt for enhancing arabic nlp: Application for semantic role labeling and cross-lingual annotation projection. *IEEE Access*, 13:3707–3725.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Orizu Udochukwu and Yulan He. 2015. A rule-based approach to implicit emotion detection in text. In *Natural Language Processing and Information Systems: 20th International Conference on Applications of Natural Language to Information Systems, NLDB 2015, Passau, Germany, June 17-19, 2015, Proceedings 20*, pages 197–203. Springer.

Naoki Wake, Atsushi Kanehira, Kazuhiro Sasabuchi, Jun Takamatsu, and Katsushi Ikeuchi. 2023. Bias in emotion recognition with chatgpt. *arXiv preprint arXiv:2310.11753*.

Wojciech Witon, Pierre Colombo, Ashutosh Modi, and Mubbasir Kapadia. 2018. Disney at iest 2018: Predicting emotions using an ensemble. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 248–253.

Qimin Zhou and Hao Wu. 2018. Nlp at iest 2018: Bilstm-attention and lstm-attention via soft voting in emotion classification. In *Proceedings of the 9th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 189–194.

Li'an Zhu, Junjie Peng, and Huiran Zhang. 2025. Text-based sarcasm detection with emoji contradictory clues assisting. In *Advanced Intelligent Computing Technology and Applications*, pages 248–260, Singapore. Springer Nature Singapore.

A Appendix

Aspect	IEST	ISEAR
Type of Emotion	Implicit (emotion word masked)	Explicit (emotion word present)
Annotation	Automatically labeled	Manually labeled
Emotion Labels	Ekman's six basic emotions: anger, disgust, fear, joy, sadness, surprise	anger, disgust, fear, joy, sadness, shame, guilt
Genre	social media (Twitter)	Survey responses
Style	Informal, noisy, fragmented sentence	formal, structured, complete sentences
Text Length	Short (tweets, < 280 characters)	Medium (1–3 sentences per instance)
Size	191,731 instances	7,666 instances
Purpose	IER	ER
Contextual Clues	Sparse; relies on social and situational context	Rich descriptions of emotional experiences

Table 5: Comparison between the IEST and ISEAR datasets.

Emotion	Tweet
Anger	I get impatient and [#TARGETWORD#] when I'm hungry.
Disgust	So many people looked at me just [#TARGETWORD#] when I said that mustaches are hot.
Fear	So [#TARGETWORD#] that I'm not good enough
Joy	you're gonna be [#TARGETWORD#] when you realize you deserve to be.
Sadness	Very [#TARGETWORD#] when he goes on these tirades
Surprise	They just jealous, they get [#TARGETWORD#] when she pull up.

Table 6: Example tweets from the IEST dataset, with the emotion word masked as [#TARGETWORD#].

Emotion	Example
Joy	An encounter with a man whom I love, after a very long separation.
Fear	After mischieviously ringing on the chemist's trade-entrance doorbell and getting caught by him.
Anger	At my Summer job, nobody looked after me in particular and I had to learn all on my own.
Sadness	After I had lived with my boyfriend in a foreign country for half a year, I saw that it was impossible for me to stay with him (for economic reasons). We separated although I loved him.
Disgust	A mother who shouts at her child for nothing.
Shame	During carnaval I danced for a few minutes normally I don't dance because I am rigid in my moving around during a dance, I stopped very soon.
Guilt	I speak harshly to my parents though they only mean my own good.

Table 7: Example from the ISEAR dataset for each emotion label.