PoETIC: A Re-framing of Context Dependent Emotion Detection

Nirmal Surange, Manish Shrivastava
Language Technologies Research Center
Kohli Center for Intelligent Systems, IIIT Hyderabad, India
nirmal.surange@research.iiit.ac.in, m.shrivastava@iiit.ac.in

Abstract

Emotion classification has been extensively studied, with numerous datasets enabling progress in both textual and multimodal settings. However, most existing text-based resources treat emotion as an utterance-level property, assuming that the emotional content is fully encoded in the sentence itself. This assumption is problematic: in the absence of paralinguistic cues such as prosody, facial expressions, or emojis, textual emotions are often highly context-dependent. Many utterances lack explicit emotion markers, and even when present, such cues may be overridden by broader situational context. Sentence-level emotion annotation, thus, is driven by the annotator's ability to imagine the context in which the given utterance would elicit a given emotion. An utterance may be able to express an emotion completely (Emotion Obvious), or it can express an emotion when imagined in a certain context (Emotion Plausible). Also, for an utterance, certain emotions might be implausible to express given the specific wording of a sentence (Emotion-Implausible). To address these issues, we create a new paradigm for emotion classification by categorizing utterance and emotion pairs into context-dependency classes. We present the PoETIC benchmark dataset, where sentences in the GoEmotions dataset are human-annotated for the three aforementioned classes across seven emotions (Fear, Anger, Sadness, Joy, Disgust, Surprise, and Neutral). We observe that gold-tagged emotions in GoEmotions do not have a clear correlation with human judgment with respect to the ability to express other emotions, given different contexts. Human annotators identify significantly more plausible emotions for a given utterance if asked to imagine a plausible context per utterance-emotion pair. We also present baselines using three popular large language models and two "small" language models in zero-shot and few-shot settings on the benchmark dataset.

Keywords: Emotion Recognition, Benchmark, Language resource

1. Introduction

Emotion classification is a well-established domain within natural language processing (NLP), with significant applications in sentiment analysis, dialogue systems, and affective computing.

Over recent decades, a variety of benchmark datasets have been developed to facilitate the study of emotion detection in textual data. For instance, the ISEAR dataset compiled short statements categorized by fundamental emotions (Scherer and Wallbott, 1994), while EmoBank introduced valence–arousal–dominance (VAD) scores applicable to sentences (Büchel and Hahn, 2017). Furthermore, GoEmotions (Demszky et al., 2020) offered detailed annotations for 27 emotions across 58,000 Reddit comments. More recently, the BRIGHTER dataset expanded this research into multiple languages, featuring human-annotated labels for basic emotions (Muhammad et al., 2025). Concurrently with text-exclusive resources, multimodal datasets such as IEMOCAP (Busso et al., 2008), MELD (Poria et al., 2019), and SEMAINE (McKeown et al., 2011) underscore the critical role of audio-visual cues in emotion recognition. Collectively, these datasets have considerably advanced research in both textual and multimodal emotion classification.

Current textual emotion datasets predominantly employ an utterance-based classification paradigm, where a sentence or short text snippet is directly labeled with one or more emotions, often assuming that the emotional content resides entirely within the utterance itself. This simplification, although practical for supervised learning, neglects the inherently context-dependent nature of emotion interpretation.

Most utterances do not contain strong lexical emotion cues. Even when emotion-laden words are present, their interpretation may be overridden or reshaped by the surrounding context. In the absence of paralinguistic signals such as prosody, emojis, or facial expressions, textual emotion becomes highly context-dependent. The same sentence can plausibly convey different emotions when presented in different contexts. For example, while *"I can't believe this happened"* conveys 'surprise', it could signal joy at a positive surprise, anger at an injustice, or sadness at a sudden, unexpected loss. Without context, the interpretation remains ambiguous.

Therefore, we posit that text-based emotion detection faces the following principal challenges:

- Contextual dependence in text-only environments (in the absence of paralinguistic signals such as prosody or emojis).
- Absence of apparent emotional cues in text can be reinterpreted or overridden by the context.

These observations motivate a shift from

Proceedings of Computational Affective Science (CAS) @ LREC 2026, pages 201–211

utterance-level emotion labeling to modeling contextual dependence. We introduce a task that determines whether an utterance: 1. *"obviously"* conveys the given emotion, 2. *can* convey the given emotion in certain contexts, or 3. *cannot* convey the given emotion in any conceivable context. To allow for context-based emotion interpretation, we formulate the task at the utterance–emotion pair level, and classify each pair into one of the three context-dependence categories (Section 4.1).

We present a new language resource: "**P**lausibility **o**f **E**motion **T**ags for **I**magined **C**ontexts (**PoETIC**)[1], a human-annotated benchmark of 7k utterance-emotion pairs annotated for contextual dependence. The utterances are taken from the test set of the GoEmotions (Demszky et al., 2020) dataset.

Given the strong performance of large language models (LLMs) on emotion classification, it is natural to investigate whether they can achieve human-level performance on this task. Therefore, we also report and discuss baseline results using both proprietary LLMs and self-hosted small language models.

2. Related Work

Most of the recent work around emotion classification is on the six basic emotions identified in Ekman (1992): happiness, sadness, fear, anger, disgust, and surprise. These emotions are considered fundamental and universally recognized across cultures, often characterized by distinct physiological and facial expressions.

Several benchmark datasets have addressed emotion classification in a variety of ways and have established emotion classification as a core NLP task. The ISEAR dataset (Scherer and Wallbott, 1994) collected short self-reports of emotional experiences, each labelled with one of seven basic emotions. EmoBank (Büchel and Hahn, 2017) annotated sentences with dimensional valence–arousal–dominance (VAD) scores, providing a graded view of affect. GoEmotions (Demszky et al., 2020) introduced a large-scale dataset of 58k Reddit comments annotated with 27 fine-grained emotions, demonstrating the feasibility of fine-grained classification. More recently, the BRIGHTER dataset (Muhammad et al., 2025) expanded emotion resources across multiple languages, providing high-quality human annotations for the seven basic emotions.

In contrast to text-only resources, multimodal datasets highlight the importance of audio-visual context in emotion recognition. IEMOCAP (Busso et al., 2008) provided dyadic conversations with aligned speech, motion-capture, and text. MELD (Poria et al., 2019) introduced a multimodal corpus of TV dialogues with aligned audio, video, and transcripts for multi-party conversation emotion recognition. Similarly, the SEMAINE database (McKeown et al., 2011) captured emotionally rich human–agent interactions with multimodal annotation.

These datasets have fueled progress in supervised learning for emotion detection but remain utterance-based, providing labels without contextual grounding. However, understanding emotions in the broader context of discourse remains a distant goal. Some resources are available for context-based emotion tagging in dialogues, such as the IEST task at WASSA2018(Klinger et al., 2018) and EmoContext task at SemEval2019(Chatterjee et al., 2019). These datasets highlight how paralinguistic and situational cues disambiguate emotions that text alone cannot resolve. However, they are costly to build and domain-limited, leaving the challenge of context-aware modeling in purely textual settings open. Notably, Yang et al. (2023) attempts to provide a formal definition of "textual context" for emotion identification in LLM prompting, making a case for either re-annotating corpora for emotion identification or using LLMs to synthesize context to help with the task.

Recent advances in LLMs have led to their application to zero-shot and few-shot emotion classification. Demonstrating that, while LLMs can perform well on benchmark datasets, a persistent difficulty is emotion trigger extraction: identifying the implicit contextual assumptions that justify a particular label. Attri et al. (2025) explored the joint task of emotion detection and explanatory span identification in e-commerce reviews to understand what triggers customer emotional responses. They leverage LLMs to unify fine-grained emotion detection and opinion trigger extraction from customer reviews.

In an effort to evaluate LLMs on the Emotion Classification task, Singh et al. (2024) observed that LLMs demonstrate **superior accuracy in predicting emotions**, with GPT-4 consistently outperforming fine-tuned transformer models across diverse datasets. However, they **struggle significantly to identify specific emotion triggers**, showing a limited ability to link emotional labels to the actual events or entities that caused them. Consequently, these models rely more on **topical key phrases and corpus-level cues** than on the deep, appraisal-based understanding characteristic of human emotion processing.

[1] https://github.com/nirmalsurange/PoETIC

Sentence	Emotion	Emotion-Plausibility
I didn't know this was a "reveal" he'd been talking about for a while.	SURPRISE	Emotion-Obvious
I didn't know this was a "reveal" he'd been talking about for a while. **Context:** *Speaker ruined a loved one's big announcement, because of the speaker's misunderstanding*	SADNESS	EMOTION-PLAUSIBLE
I didn't know this was a "reveal" he'd been talking about for a while.	NEUTRAL	Emotion-Implausible
I know. I didn't say I agreed with it did I?	ANGER	Emotion-Obvious
I know. I didn't say I agreed with it did I? **Context:** *a defensive, panicked justification by someone who has been coerced into participating in something illegal, and is now trying to distance themselves from the consequences while fearing retribution from the person in charge.*	FEAR	Emotion-Plausible
It's real and it plays SO OFTEN I CAN'T STAND IT!!	JOY	Emotion-Implausible
It's real and it plays SO OFTEN I CAN'T STAND IT!!	DISGUST	Emotion-Obvious

Table 1: Illustrative examples contrasting context-based emotion plausibility.

3. GoEmotions dataset

One of the most prominent human-annotated datasets available for Emotion Classification is GoEmotions (Demszky et al., 2020). This dataset contains 58k samples, manually annotated for four sentiment categories (Positive, Negative, Neutral, and Ambiguous), which are divided into six coarse Ekman classes (Fear, Anger, Disgust, Sadness, Joy, Surprise) plus Neutral, and are further subdivided into 27 fine-grained emotions. Samples in GoEmotions are sourced from the Reddit[2] platform. The authors of GoEmotions note that while the platform is not representative of the general English-speaking population, the curation and selection steps followed by the authors make the data fairly balanced. They filtered content to reduce profanity and balance emotions, sentiment, length, and source subreddits. The final dataset contains 58009 samples split into 80%, 10%, and 10% for training, development, and test sets, respectively.

We utilize sentences from the test set of this dataset for human annotation of Contextual Dependence across the seven emotion categories (the six Ekman classes plus Neutral) outlined previously. This annotation exercise resulted in a dataset which has an average of 4.55 emotions (Table 3) per utterance as compared to 1.09 for the source GoEmotions Test set.

4. Context Dependency

Contextual dependence of emotion interpretation is an understudied phenomenon that may have a significant bearing on the emotion prediction task. As Table 1 shows, while some sentences can have some apparently obvious emotions, these utterances can be reinterpreted to have different emotions in some imagined contexts.

We aim to study the extent to which the perceived emotion of a sentence depends on its surrounding context and to augment existing emotion recognition resources with explicit annotations indicating context dependence as perceived by a human annotator. While we do not generate contexts relevant to the specific utterance-emotion pairs, we ask the human annotators to evaluate if, under some imagined context, the utterance in question will express the given emotion. This also allows us to consider those utterance-emotion pairs that cannot belong together (**EI**).

Stage	Test
Original sentences	5,427
Emoji removal	-268
Foul language removal	-267
Short/noisy sentence removal	-106
Hashtag / URL removal	-16
Clean sentences	4,770
Selected sentences	1000

Table 2: Dataset Preparation Statistics

Data Preparation. We base our dataset on the Test set of the GoEmotions corpus, which contains 5,427 sentences. To ensure that emotion recognition is based solely on linguistic cues, we first remove sentences containing emojis (268 instances). We then apply a series of filtering steps to eliminate sentences with auxiliary or noisy signals, specifically those containing foul or profane language (267 instances), very short or noisy content (106 instances), and hashtags or URLs (16 instances).

[2]www.reddit.com

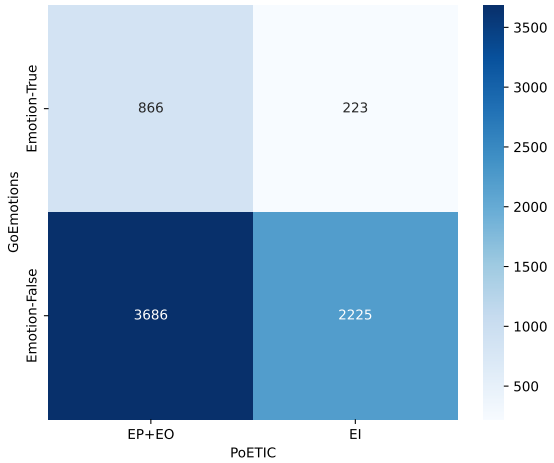


Figure 1: GoEmotions Vs. PoETIC tagging

After applying these filters, we retain 4,770 sentences, from which we randomly select 1,000 to serve as the source for our benchmark. (Table 2).

4.1. Human Annotation

Annotation Protocol Three paid annotators, with native-level proficiency in English, label each utterance-emotion pair for its context dependence. Annotators are provided with an utterance and a candidate emotion label. Their task is to assign exactly one of the following labels to each pair:

- **Emotion-Obvious(EO)**: The sentence clearly and unambiguously expresses the target emotion even in the absence of any additional context
- **Emotion-Plausible(EP)**: The sentence may express the target emotion, but the interpretation is plausible only in the presence of relevant additional contextual information
- **Emotion-Implausible (EI)**: The sentence cannot plausibly express the target emotion in any imaginable context

While it is possible to have a much finer classification ("highly dependent on context", "slightly dependent on context", etc.) to better cover the spectrum of context dependence, we found that increasing the labelset resulted in very poor inter-annotator agreement. This suggests that the cognitive load of making finer distinctions leads to poor performance. The same is true for annotating the entire emotion set of 27 emotions presented in GoEmotions. This is why we utilize the Ekman-emotion mapping provided by GoEmotions to convert the original annotations of 27 emotions into the 6 basic emotions.

Neutral emotion. The label NEUTRAL denotes the absence of all six non-neutral emotions (ANGER, SADNESS, JOY, FEAR, DISGUST, SURPRISE). A sentence cannot be labeled **EO** for both NEUTRAL and any other emotion. If a sentence clearly expresses a non-neutral emotion, it must not be labeled **EO** for NEUTRAL.

4.2. Dataset statistics

The annotation exercise resulted in a final corpus of 7000 annotated utterance-emotion pairs (Table 3). We achieve a Krippendorff's α score of 0.62 for the inter-annotator agreement. This inter-annotator score underlines the cognitive complexity of the task. We take the majority vote across annotators to select our final class label. We find that significant confusion between EI and EP classes, with $20 - 25\%$ of disagreement among annotators for these classes on average.

Our annotations reveal that sentences can plausibly express substantially more emotions (Table 3) once context dependency is modeled. As expected, context-dependent emotions dominate in the available data, highlighting the importance of contextual reasoning for realistic emotion understanding. When seen in comparison to the source dataset, though, we observe that while most of the sentences get **EO** or **EP** tags for their assigned emotions in GoEmotions, a substantial number of utterance-(tagged)emotion pairs are classified as **EI** by our annotators (Figure 1). While this statistic is interesting, it only points to the subjectivity inherent in the emotion classification task. Table 4 details this further. Here, the 'tagged' row lists the annotation distribution across the three categories for sentences with their GoEmotions assigned tags. Similarly, the 'untagged' row reports the distribution of sentences paired with other emotions (negative).

5. LLMs for Context Dependency

To explore LLMs' ability to identify context dependence of emotions, we performed zero-shot inference on three popular instruction-tuned LLMs. These models are: GPT-5-mini[1], Gemini-2.5-flash[2] and Llama3-70b (GroqLlama)[3] . Appendix-A shows the prompts and hyperparameter settings used for these LLMs. As the models are quite sensitive to prompts, we share the prompts that performed the best for each model. We further evaluate performance against a supervised baseline

[1]GPT-5-mini model by OpenAI. https://platform.openai.com/docs/models

[2]Gemini 2.5–flash model by Google. https://ai.google.dev/models/gemini

[3]Llama 3 70B hosted via Groq. https://console.groq.com/docs/models

	TOTAL	Class Distribution						Avg. Emotions / Sentence	
		EP	(%)	EI	(%)	EO	(%)	GoEmotions	EP+EO
PoETIC	**7,000**	3,474	49.58	2,448	34.94	1,078	15.38	1.09	4.55

Table 3: PoETIC Stats and Class Distribution

	GoEmotions Tag	Context-Dependency Labels		
		Emotion-Plausible	Emotion-Implausible	Emotion-Obvious
PoETIC	Tagged	368	223	498
	Untagged	3,106	2,225	580

Table 4: Relation between original GoEmotions and PoETIC tagging

based on RoBERTa[4], trained for 10 epochs on the PoETIC dataset (80% training, 10% development, and 10% test split), using 5-fold cross-validation.

Table 5 compares the three large language models on PoETIC and the RoBERTa benchmark. As can be seen, GPT-5-mini performs reasonably well, with an accuracy and F1 scores of 0.77. The model(Figure 3a), does fairly well on the Emotion-Plausible (**EP**) class but results in significant misclassifications of the Emotion-Implausible (**EI**) class, showing that the model is not able to disambiguate between these classes. At the same time, Gemini-2.5-flash demonstrates moderate performance, achieving an accuracy of 0.68. While its macro F1 (0.68) is the same as its micro and weighted F1 scores, the confusion matrix (Figure 3b) reveals some confusion between **EP** and **EI** classes.

This shows that GPT-5-mini can reliably distinguish between emotion expressions that require external context and those that are self-contained. The consistency between macro, micro, and weighted F1 indicates robust performance across both frequent and infrequent classes.

Gemini-2.5-flash captures class-level distinctions quite well, but it struggles to consistently identify when emotion interpretation depends on additional context, particularly for **EP** instances.

GroqLlama performs the weakest among the evaluated systems, with an accuracy of 0.59 and a notably lower macro F1 score of 0.53. The model exhibits confusion between **EP** and **EI** categories (Figure 3c), showing limited sensitivity to subtle contextual cues that disambiguate emotion dependence. This aligns with the hypothesis that smaller or less instruction-tuned models may rely more heavily on surface-level lexical cues rather than deeper contextual reasoning.

We also consider a majority voting (MajorityVote in Table 5) scheme to consolidate the annotations by all the models. As we can see, the combined oracle of the models does not result in a significant improvement in performance. We find that a large number of utterance-emotion pairs were misclassified by all the models (Table 6).

Overall, these results highlight that context-dependence classification is highly sensitive to a model's contextual reasoning capabilities. While state-of-the-art models approach human-level reliability on this task, weaker models exhibit systematic confusions that disproportionately affect context-dependent emotion categories. This performance gap highlights the importance of evaluating emotion understanding not only in isolation but also through the lens of contextual grounding.

6. Local LLMs for Context Dependency

While models with $> 20b$ parameters are considered 'experts' for almost all NLP tasks, most real-world use cases often require comparable performance from significantly smaller models. We explore this by choosing two popular local LLMs (Qwen3-8b[3] and Llama-3-8b[4]) and test them on PoETIC. Table 7 reports the performance of Qwen3-8B and Llama-3-8B under zero-shot and few-shot settings.

In the zero-shot regime, both models achieve comparable accuracy (Qwen3-8B: 0.5; Llama-3-8B: 0.47), with Llama-3-8B exhibiting a slightly higher macro F1 (0.22 vs. 0.18), indicating marginally better class balance despite lower overall accuracy. However, both models show low macro F1 scores, reflecting difficulty in handling minority classes (**EO**).

Introducing few-shot prompting leads to some performance gain for Qwen3-8B, improving accuracy to 0.52 and macro F1 to 0.36. There are balanced predictions across classes, as evident by the reduced off-diagonal mass in the confusion matrix (Figure 4). Inexplicably, Llama-3-8B exhibits a severe performance degradation in the few-shot

[4]RoBERTa-base model https://huggingface.co/FacebookAI/roberta-base

[3]https://huggingface.co/Qwen/Qwen3-8B

[4]https://huggingface.co/meta-llama/Meta-Llama-3-8B

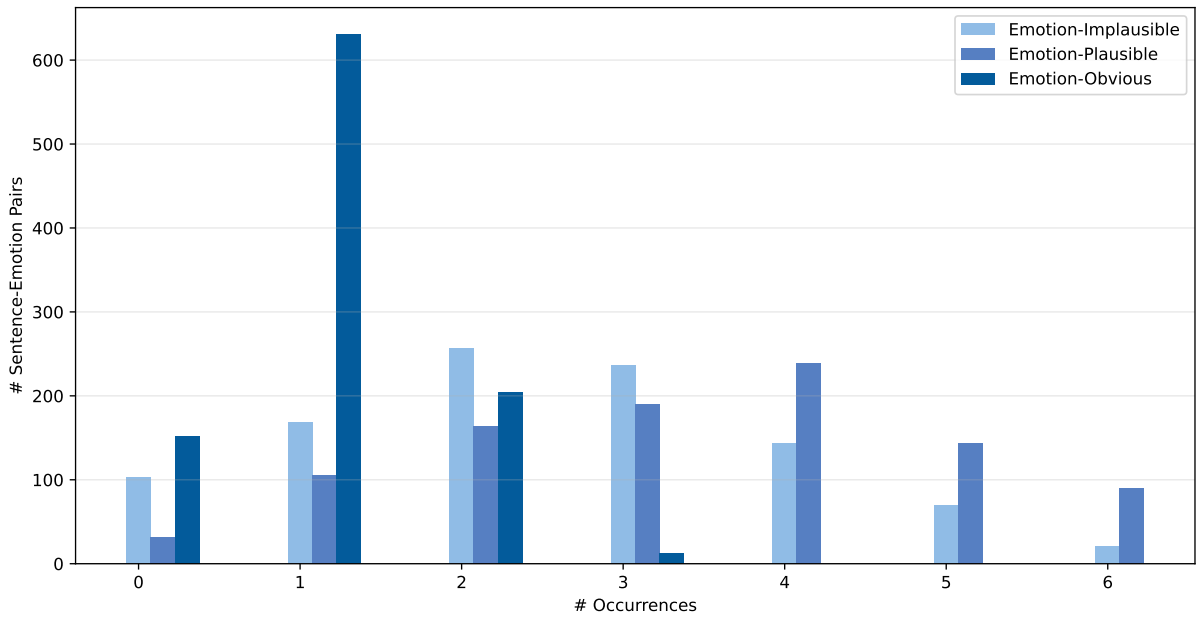


Figure 2: Context-Dependency Distribution

Model	Acc.	F1$_{micro}$	F1$_{macro}$	F1$_{weighted}$	Prec.	Rec.
GPT-5-mini	0.77	0.77	0.77	0.77	0.78	0.77
Gemini-2.5-flash	0.68	0.68	0.68	0.68	0.69	0.68
GroqLlama	0.59	0.59	0.53	0.54	0.64	0.59
MajorityVote	0.79	0.79	0.79	0.79	0.81	0.79
RoBERTa	0.52	0.52	0.27	0.40	0.34	0.52

Table 5: LLMs' performance comparison against PoETIC for context-dependence classification

Sentence	Emotion	MajorityVote	Human
Where were they? I was trying to track them down and couldn't remember where I saw them.	NEUTRAL	**EO**	**EI**
I know you're joking, but there are people here either stupid or desperate enough to believe and perpetuate such idiocy.	SURPRISE	**EO**	**EI**
Would you like a friend? I sure could use one.	JOY	**EI**	**EP**
Not sure if guys like that are better or worse than regular Incels.	SURPRISE	**EI**	**EP**

Table 6: Illustrative examples contrasting model-predicted classes with human annotations.

Model	Setting	Acc.	F1$_w$	F1$_m$
RoBERTa	Supervised	0.52	0.40	0.27
Qwen3-8B	Zero-shot	0.5	0.35	0.18
Llama-3-8B	Zero-shot	0.47	0.35	0.22
Qwen3-8B	Few-shot	**0.52**	**0.51**	**0.36**
Llama-3-8B	Few-shot	0.35	0.19	0.18

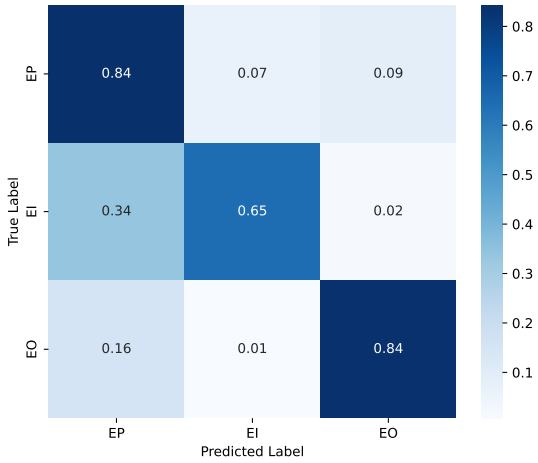
Table 7: Model performances under zero-shot and few-shot settings.

setting (accuracy 0.35, macro F1 0.18), characterized by near-collapse to a single class prediction.

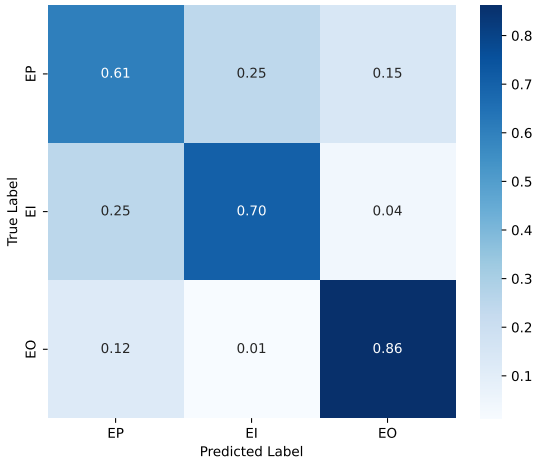
Overall, Qwen3-8B demonstrates robust adaptation with few-shot learning, achieving the best performance across all metrics, while Llama-3-8B shows inconsistent behavior, highlighting model-specific differences in prompt conditioning and generalization.

7. Discussions

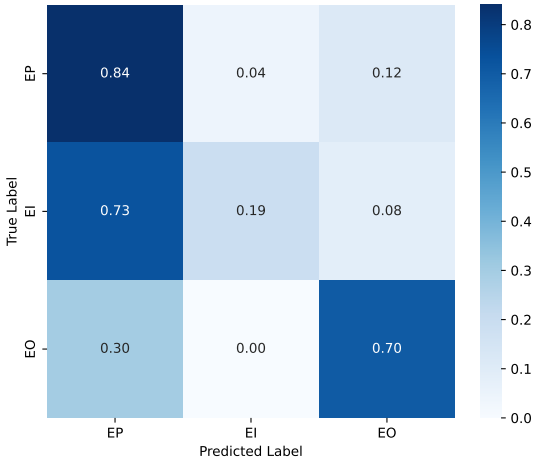
Prior work, such as Lecourt et al. (2025), reports that large language models (LLMs) achieve relatively modest performance on the GoEmotions dataset. This suggests that directly assigning an utterance to a single emotion label—chosen from a large set—poses a substantial challenge, partic-



(a) GPT-5-mini



(b) Gemini-2.5-flash



(c) GroqLlama3-70B

Figure 3: Normalized confusion matrices for different models on PoETIC.

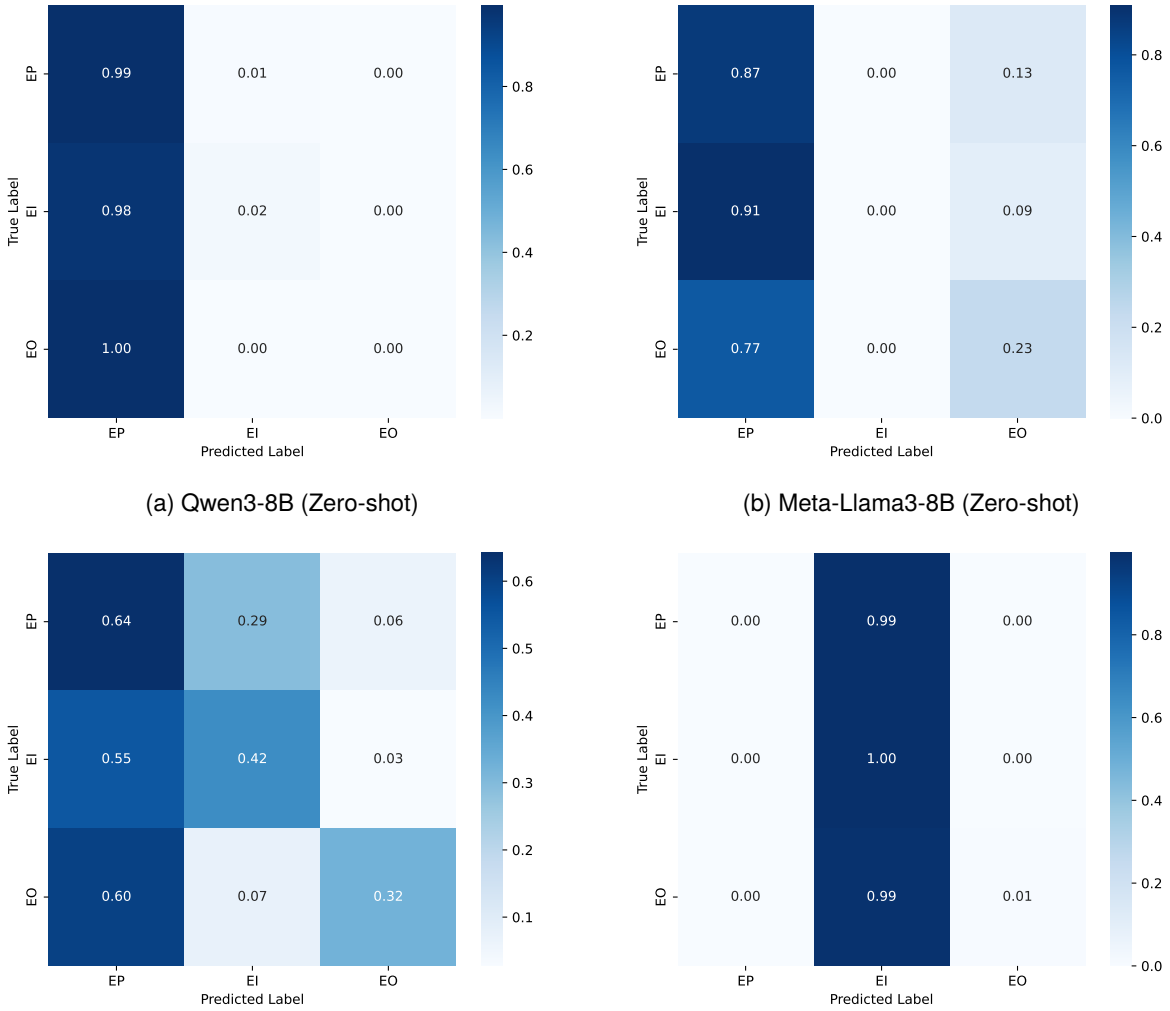
ularly when multiple emotions may be contextually plausible. We argue that the utterance-based emotion tagging formulation conflates emotion recognition with contextual ambiguity, making the task inherently difficult for both models and annotators.

In contrast, we reformulate the problem in terms of contextual dependence. Rather than requiring the model to select the correct emotion in isolation, our task asks whether there exists any plausible context in which a given utterance could convey a specified emotion. This re-framing reduces the burden of exhaustive disambiguation and instead focuses on contextual plausibility. While this remains a non-trivial cognitive task, it is arguably more aligned with the generative and inferential strengths of LLMs.

Although we adopt discrete categories for contextual dependence in this work, the notion itself naturally admits a graded interpretation. Contextual dependence can be viewed as a continuum, ranging from utterances that require minimal contextual elaboration to express an emotion to those that demand highly imaginative or contrived contexts. Modeling contextual dependence as a rating or ordinal prediction task is, therefore, a promising direction for the future.

8. Limitations

The dataset is derived from English Reddit data, inheriting platform-specific linguistic norms and demographic biases from GoEmotions. As a result, the observed patterns of contextual dependence may not generalize to other domains, languages, or interaction settings. Contextual dependence is modeled using three discrete categories. While this formulation improves annotation reliability, contextual dependence is inherently continuous, and many sentence–emotion pairs lie near category boundaries, as reflected in moderate inter-annotator agreement. More fine-grained or scalar annotations could better capture these nuances.



(a) Qwen3-8B (Zero-shot)

(b) Meta-Llama3-8B (Zero-shot)

(c) Qwen3-8B (Few-shot)

(d) Meta-Llama3-8B (Few-shot)

Figure 4: Normalized confusion matrices for zero-shot and few-shot settings on PoETIC.

Also, the annotations rely on imagined rather than explicit contexts. Judgments, therefore, depend on annotators' world knowledge and cultural assumptions, introducing subjectivity that may vary across annotator populations. Due to resource constraints, model evaluation is restricted to zero-shot and few-shot prompting. The strong performance of proprietary LLMs, while encouraging, raises questions about robustness and interpretability that suggest further investigation through cross-domain and adversarial evaluations. Overall, these limitations point toward future work on broader datasets, graded annotations, explicit context modeling, and more rigorous evaluation settings.

9. Use of AI Agents

AI Agents, primarily ChatGPT, have been used for grammar correction and paraphrasing in order to improve the language quality. The paraphrased outputs were thoroughly checked to ensure correctness. Copilot was used for code completion occasionally.

Anuj Attri, Arnav Attri, Suman Banerjee, Amey Patil, Muthusamy Chelliah, Nikesh Garera, and Pushpak Bhattacharyya. 2025. LLMs as architects and critics for multi-source opinion summarization. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 69–101, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Sven Büchel and Udo Hahn. 2017. Emobank: Studying the impact of annotation perspective and representation format on dimensional emo-

tion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 578–585.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. 2019. Semeval-2019 task 3: Emocontext contextual emotion detection in text. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 39–48.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Paul Ekman. 1992. Are there basic emotions? *Psychological Review*, 99(3):550–553.

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.

Florian Lecourt, Madalina Croitoru, and Konstantin Todorov. 2025. 'only chatgpt gets me': An empirical analysis of gpt versus other large language models for emotion detection in text. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 2603–2611.

Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. 2011. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE transactions on affective computing*, 3(1):5–17.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine de Kock, Nirmal Surange, Daniela Teodorescu, Ibrahim Said Ahmad, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino D. M. A. Ali, Ilseyar Alimova, Vladimir Araujo, Nikolay Babakov, Naomi Baes, Ana-Maria Bucur, Andiswa Bukula, Guanqun Cao, Rodrigo Tufiño, Rendi Chevi, Chiamaka Ijeoma Chukwuneke, Alexandra Ciobotaru, Daryna Dementieva, Murja Sani Gadanya, Robert Geislinger, Bela Gipp, Oumaima Hourrane, Oana Ignat, Falalu Ibrahim Lawan, Rooweither Mabuya, Rahmad Mahendra, Vukosi Marivate, Alexander Panchenko, Andrew Piper, Charles Henrique Porto Ferreira, Vitaly Protasov, Samuel Rutunda, Manish Shrivastava, Aura Cristina Udrea, Lilian Diana Awuor Wanzare, Sophie Wu, Florian Valentin Wunderlich, Hanif Muhammad Zhafran, Tianhui Zhang, Yi Zhou, and Saif M. Mohammad. 2025. BRIGHTER: BRIdging the gap in human-annotated textual emotion recognition datasets for 28 languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8895–8916, Vienna, Austria. Association for Computational Linguistics.

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

Klaus R Scherer and Harald G Wallbott. 1994. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of personality and social psychology*, 66(2):310.

Smriti Singh, Cornelia Caragea, and Junyi Jessy Li. 2024. Language models (mostly) do not consider emotion triggers when predicting emotion. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 603–614, Mexico City, Mexico. Association for Computational Linguistics.

Daniel Yang, Aditya Kommineni, Mohammad Alshehri, Nilamadhab Mohanty, Vedant Modi, Jonathan Gratch, and Shrikanth Narayanan. 2023. Context unlocks emotions: Text-based emotion classification dataset auditing with large language models. In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE.

A. Prompts

GPT-5-mini.

```
You are an expert NLP evaluator.
You must classify how dependent
the text is on surrounding context
to express the given emotion.

Allowed labels (SHORT FORM):
A = Emotion-Obvious
B = Emotion-Plausible
C = Emotion-Implausible

Label meanings:
A: The emotion is clearly expressed
without needing additional context.
B: The emotion is ambiguous or
barely present, and typically needs
outside context.
C: Even with added context, this
text cannot plausibly express the
target emotion.

Special rule for NEUTRAL:
- Since "NEUTRAL" means complete
absence of the 6 target emotions
  {ANGER, SADNESS, JOY, FEAR,
  DISGUST, SURPRISE},
  a sentence cannot be "A"
  (Emotion-Obvious) for both
  NEUTRAL and another emotion.

STRICT OUTPUT REQUIREMENTS:
- You must output ONLY a JSON
object of the form:
  {"label":"A"} or {"label":"B"}
  or {"label":"C"}
- Do NOT output any text before or
after the JSON.
- Do NOT explain your choice.
- Do NOT add any additional keys.
- If unsure, pick the best label; do
NOT invent new labels.

-----------------
Text: {text}
Emotion: {emotion}
```

Gemini 2.5–flash.

```
You are an expert NLP evaluator. You
must classify how dependent the text
is on surrounding context to express
the given emotion.

Allowed labels:
Emotion-Obvious
Emotion-Plausible
Emotion-Implausible

Label meanings:
Emotion-Obvious: The emotion is
clearly expressed without needing
additional context.
Emotion-Plausible: The emotion is
ambiguous or barely present, and
typically needs outside context.
Emotion-Implausible: Even with added
context, this text cannot plausibly
express the target emotion.

Special rule for NEUTRAL:
- Since "NEUTRAL" means complete
absence of the 6 target emotions
  {ANGER, SADNESS, JOY, FEAR,
  DISGUST, SURPRISE},
  a sentence cannot be
  "Emotion-Obvious" for both
  NEUTRAL and another emotion.

STRICT OUTPUT REQUIREMENTS:
- You must output EXACTLY ONE label
- Do NOT output any text before or
after the label.
- Do NOT explain your choice.
- If unsure, pick the best label;
do NOT invent new labels.

-----------------
Text: "{sentence}"
Emotion: "{emotion}"
```

Hyperparameters	Gemini2.5Flash	GPT-5-mini	Llama3-70B
temperature	0.0	1.0	0.1
top_p	1.0	1.0	1.0
top_k	40	-	-
max_tokens	8192	1800	3

Table 8: Model hyperparameter setting

Groq (Llama3-70B).

```
Classify the dependency of the
following sentence on its prior
context to express the given emotion.

Allowed classes with definitions:
EO: The emotion is clearly expressed
without needing additional context.
EP: The emotion is ambiguous or barely
present, and typically needs outside
context.
EI: Even with added context, this
text cannot plausibly express the
target emotion.

Reply ONLY with the class label.
Do NOT include any explanation or
additional text.

Example output: EO

-----------------
Sentence: {sentence}
Emotion: {emotion}
Class-label?
```