Generalising sentiment: Leveraging knowledge transfer for multiclass negative emotion mining

Anonymous ACL submission

Abstract

Data annotation is critical yet challenging, particularly for nuanced negative emotions in resource-scarce and niche contexts like hate-speech and incel discourse. We systematically explore zero-, few-, and many-shot prompting strategies across three distinct prompting types—base, chain-of-thought (CoT), and in-context learning (ICL)—we uncover the limitations of LLMs in handling nuanced and domain-specific datasets, such as incel discourse. Our findings reveal persistent biases in emotion perception, offering a roadmap to enhance LLM performance in resource-scarce contexts. Furthermore, we introduce and evaluate a hybrid annotation framework, combining LLM-generated annotations with human refinements, which significantly improves accuracy, inter-annotator agreement, and efficiency. This work advances the understanding of NLP generalisation by demonstrating how LLMs can support and complement human annotation efforts, specifically in highly subjective challenging tasks.

1 Introduction

The classification of human emotions has been widely researched (Tripathi et al., 2020; Lek and Teo, 2023), with early studies identifying "anger", "fear", "enjoyment", "sadness", "disgust", and "surprise" as the basic emotions (Ekman, 1999). Emotion mining, which involves the identification of emotions in text based on linguistic features, has become increasingly important in the context of the rise of social media, where communication is rapid and often emotionally charged (Min et al., 2023). This task is critical in identifying key issues in online speech, particularly in areas such as hate-speech detection, which has significant implications for mitigating online harms (Awal et al., 2021; Plaza-del Arco et al., 2021; Rodriguez et al., 2022; Paz et al., 2020; Warner and Hirschberg, 2012; MacAvaney et al., 2019).

This gap motivates our research, aiming to evaluate LLMs in these challenging settings and provides a roadmap for improving their annotation capabilities by evaluating various combinations of annotation strategies for this task. While LLMs have demonstrated remarkable success across a variety of NLP tasks, their ability to generalise to nuanced and domain-specific tasks, such as negative emotion mining in incel discourse, remains underexplored. This study seeks to fill this gap by examining how LLMs perform when tasked with identifying subtle emotional nuances in resource-scarce and challenging domains. We investigate how generalisation capabilities vary under different prompting strategies—base, chain-of-thought (CoT), and in-context learning (ICL)—and across zero-, few-, and many-shot settings. By focusing on generalisation, we highlight the critical interplay between model design, prompting strategies, and domain complexity in advancing the robustness of LLMs.

In the domain of Natural Language Processing (NLP), data annotation is a critical and multifaceted process that transcends simple labelling (Yu et al., 2023; Lin et al., 2022). This process typically encompasses multiple phases, including initial classification of raw data, integration of intermediate markers for contextual depth, and evaluation of the reliability of the annotator, often necessitating several iterations (Ross et al., 2017; MacAvaney et al., 2019; Shao et al., 2023; Zhang et al., 2022; Efrat and Levy, 2020; Wei et al., 2021; Zhang et al., 2022). The cognitive demands of annotating negative emotions are significantly heightened in domains like hate-speech, where emotional subtleties and data heterogeneity pose unique challenges. This study is among the first to systematically evaluate how LLMs address these complexities, providing insights into their generalisation capabilities and limitations (Zhang et al., 2023; Liu et al., 2024). However, state-of-the-art large language models

(LLMs) such as GPT models (OpenAI et al., 2024) and its successors, Claude-Opus (Anthropic, 2024), and Gemma models (Team et al., 2024) present promising avenues for revolutionising data annotation (He et al., 2024). Automating this task can ensure consistency across extensive datasets, which can be subsequently leveraged for fine-tuning or, as demonstrated in this study, through advanced prompting techniques (Tan et al., 2024).

Negative emotions, such as "anger," "fear," and "sadness," are well-documented in hateful or harmful language. However, this work expands the scope by incorporating underexplored emotions like "loneliness" and "jealousy," which add critical layers of nuance to emotion detection tasks. These subtle emotions are often overlooked in NLP research, making their detection vital for capturing the full emotional complexity of hate-speech and online hostility (Wang et al., 2023b). Although emotions like "loneliness" and "jealousy" are challenging to detect due to subtle linguistic cues and contextual overlaps, they are critical for understanding the drivers of online hostility.

This is one of the first studies to assess LLM generalisation for nuanced negative emotions in underexplored niche domains. Our study leverages state-of-the-art LLMs to tackle the underexplored challenge of annotating nuanced emotional categories in hate-speech detection. By addressing gaps in LLM generalisation and reducing reliance on labour-intensive human annotation, our study thus significantly contributes to the field of emotion mining in NLP.

The contributions of this work are four-fold:

1 *Cross-domain generalisation analysis*: We evaluate various state-of-the-art LLMs across three datasets: a standard negative emotion dataset, a hate-speech dataset and a domain-specific incel dataset.

2 *Prompting*: We assess the impact of prompt optimisation and generalisation by analysing the effects of zero-, few- and many-shot prompting on various LLM performances.

3 *Misclassification*: We examine model misclassification patterns and uncover systematic biases in LLM output, specifically in distinguishing between similar emotions and assessing emotion overlap.

4 *Annotation system implications*: We introduce and evaluate a hybrid annotation framework, where LLM-generated labels are refined by human annotators. This work highlights the practical utility of hybrid systems in reducing annotator burden while maintaining high-quality annotations in resource-scarce and emotionally complex settings.

2 Related Work

Approaches to sentiment and emotion classification have varied in their analysis (Ranganathan and Tzacheva, 2019). Some approaches used a largely linguistic focus with the extraction of features such as n-grams, lexical, syntactic, semantic, and word embeddings (Yassine and Hajj, 2010; Mishne and Glance, 2006).

With the recent success of generative models, most of the work in this area has focused on sentiment analysis using GANs (Peper and Wang, 2022; Mahalakshmi et al., 2024), and LLMs, such as GPT models (Brown et al., 2020a; Magdaleno et al., 2024; Kheiri and Karimi, 2023; Suhaeni and Yong, 2024; Zhan et al., 2024). Several studies have examined the capacity of LLMs to identify more fine-grained emotions. For instance Boitel et al., 2024 and Venkatakrishnan et al., 2023 compared the performance of out-of-the-box GPT-3 with BERT- and RoBERTa-based task-specific classifiers, showing a performance advantage for the latter. Similarly, Wang et al. (2023a) reported an unsatisfactory performance of GPT-3.5 in detecting emotions in memes. Nevertheless, GPT-3 has proven to be very efficient for data augmentation and enhancing its outputs with classifiers (Kok-Shun et al., 2023). The newer GPT-4 (OpenAI et al., 2024) has been reported to surpass 89% of human annotators in annotation quality for the task of emotion recognition (Wang et al., 2023c).

The focus on uniquely negative emotions has already been explored for various applications. For example, AlSagri and Ykhlef (2016) used clustering to analyse and counteract negative emotional contagion in online social networks. Kodati and Dasari (2024) detected negative emotions related to COVID-19 using auto-regressive transformers, framing it as a classification task. Negative emotions are particularly relevant for social media analysis tasks such as hate-speech and fake news detection. Schäfer and Kistner (2023) showed a strong correlation between negative emotions like disgust and extreme forms of hate-speech. Additionally,

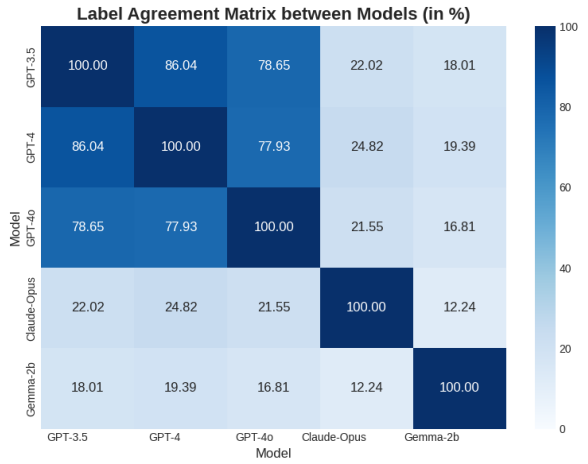


Figure 1: Label Agreement Matrix between Models in percentage. This matrix illustrates the percentage of label agreement between different models for the base prompt labels.

Min et al. (2023) demonstrated that capturing correlations between hate-speech and negative emotional states improved the performance of their hybrid system based on BERT. Negative emotions have also been shown to play a crucial role in the virality of fake news (Corbu et al., 2021), with negatively biased fake news correlating with people's willingness to share it. According to Farhoudinia et al. (2024), fear emerged as a significant differentiator between fake and real news. However, there has been no study focusing on the usage of LLMs for negative emotion or hate-speech data annotation.

2.1 LLMs for annotation

2.2 Prompting types

LLMs have shown impressive abilities to solve complex reasoning tasks by breaking them down into intermediate steps before providing a final answer (Brown et al., 2020b; Thoppilan et al., 2022; Rae et al., 2022; Chowdhery et al., 2022). This stepwise reasoning process is induced by chain-of-thought (CoT) prompting (Wei et al., 2023). CoT prompting prompts LLMs to generate a sequence of reasoning steps that lead to the final answer. This has been achieved through zero-, few- and many-shot prompting (Kojima et al., 2023; Wei et al., 2023; Zhang et al., 2022). CoT prompting is closely related to in-context learning (ICL) (Radford et al., 2019; Brown et al., 2020b). Here, the models are guided to perform specific tasks by including a few examples of inputs. ICL allows models to generalise across tasks in both zero-, few- and many-shot settings.

Although ICL has shown success, research has highlighted that its performance can vary significantly depending on the choice of examples in context (Liu et al., 2021; Lu et al., 2022). Factors such as the wording, or order of the examples within the prompt can lead to notable variations in performance (Webson and Pavlick, 2022; Zhao et al., 2021). This emphasises the need for careful design when using LLMs for complex reasoning tasks. We apply CoT prompting and ICL in zero-, few-, and many-shot settings to evaluate their impact on negative emotion mining. Using a base prompt as a baseline across these settings, we aim to assess how CoT and ICL improve model performance, particularly in tasks requiring nuanced emotional understanding. This allows us to directly compare their effectiveness and establish the benefits of intermediate reasoning steps in enhancing task accuracy.

3 Methodology

3.1 Data

In this paper, we utilise three datasets for our analysis. The datasets were obtained from Kaggle or scraped from the website Incels.is. Due to ethical constraints, we are unable to publish the labelled incel corpora. The first corpus has been previously labelled with the emotion labels "anger", "sadness", and "fear", and is referred to as the Negative Emotion dataset[1]. The second dataset is the hate-speech detection dataset, where we reduce the labels to "hate-speech" and "not hate-speech"[2]. The final dataset is the incel corpus, annotated by human evaluators. The labels for this corpus are "sadness", "anger", "fear", "loneliness", "jealousy", and "not sure". We compare these humanly annotated labels across all three datasets to the automatically generated labels. Further details on the labelling process can be found in Section 2.1.

3.2 Model selection

We utilised the current state-of-the-art models to generate the negative emotion labels. We use three OpenAI models (OpenAI, 2023) namely GPT-3.5, GPT-4 and GPT-4o. We also use Claude-Opus created by Anthropic (Anthropic, 2024) and Gemma-

[1]Access: https://www.kaggle.com/datasets/abdallahwagih/emotion-dataset

[2]Access: https://www.kaggle.com/datasets/mrmorj/hate-speech-and-offensive-language-dataset

2b created by Google (Team et al., 2024). These models have been reported to have extremely competitive performances across different tasks (Wu et al., 2023).

Models	Creator	Size	Access
GPT-3.5	OpenAI	175B	API
GPT-4	OpenAI	-	API
GPT-4o	OpenAI	-	API
Claude-Opus	Anthropic	137B	API
Gemma-2b	Google	2B	Hugging Face

Table 1: LLMs used in the annotation process, detailing the model, creator, size, access type, and access method.

Reaching sufficient agreement levels is a difficult task for human annotators, and yet it has not been investigated in LLMs (Tan et al., 2024). Therefore, we propose to analyse precisely this, together with a breakdown of misclassifications in order to understand how we can improve the emotional understanding capabilities of LLMs. As previously mentioned, the dataset with negative emotion labels had already been annotated and utilised in other research. Therefore, this data did not require additional human labelling. However, the incel dataset necessitated human annotation. Initially, the dataset was labelled using the LLMs employed in this study. Subsequently, human annotators were tasked with labelling a subset of the dataset (10%). They were instructed to label the emotions present in the subset of data they received. This subset was then used to compare the LLM-generated labels from various prompts and natural language generation (NLG) models. The different prompts can be seen in Table 8 in Appendix.

For the automated labelling process, we randomly extracted 1,000 rows from each dataset. This limitation was due to constraints imposed by the number of annotated examples that were available for each dataset. We wanted to experiment with the same number of rows for each dataset. We were also constrained by API access, as the OpenAI and Anthropic models are not open source. We obtained access to the API for various models and commenced the label-generation process. Various prompt types were provided for all three corpora and across all NLG models. Once the data was labelled, agreement was measured between human annotators using Fliess Kappa. Human and LLM agreement was measured using automated metrics such as accuracy, precision and Cohen's Kappa.

We evaluated the impact of three prompting strategies—Base, In-Context Learning (ICL), and Chain-of-Thought (CoT)—on LLM performance. The Base prompt represents a minimal directive for emotion classification, while ICL provides examples and definitions to guide the model. CoT introduces step-by-step reasoning for nuanced emotion detection. These strategies were applied across zero-, few-, and many-shot settings to assess their influence on generalisation and model robustness in resource-scarce contexts. We ran these prompts in zero-, few- and many-shot settings. We designed few- and many-shot experiments using 2% and 20% of the data, respectively, to explore how varying levels of contextual information alter LLMs' ability to generalise to nuanced emotional categories. This experimental setup thus provides novel lens for evaluating prompt strategies in resource-scarce scenarios.

3.3 Hybrid annotation framework

To enhance annotation quality, we developed a hybrid framework combining LLM-generated labels with human refinement. First, datasets were annotated using the best-performing LLM configurations, guided by tailored prompts. Human annotators then validated and revised a subset of the LLM-labelled data, addressing nuanced cases like overlapping emotional states. Feedback loops informed prompt refinement and targeted consistent LLM misclassifications. The final dataset aggregated refined labels with LLM annotations, and performance improvements were quantified using metrics such as accuracy and Cohen's Kappa. This framework effectively combined LLM efficiency with human contextual understanding, achieving notable gains in annotation quality and inter-annotator agreement.

4 Human annotation

Two of the three datasets were public-domain and pre-annotated. For the Incel dataset, six annotators validated a 10% subset of labelled rows. The team included four linguists and social scientists, and two computational linguists with advanced degrees. Annotators followed guidelines aligned with the zero-shot prompt, using NLG-generated labels as references but could choose alternative labels if needed. Inter-annotator agreement (IAA) was measured with Fleiss Kappa (scaled 0–100).

Annotators initially reviewed 100 rows, with 50 random rows manually evaluated after each round.

Round	Kappa	Description
R1	24.98	Base round with all annotators
R1	84.62	3 experienced annotators
R1	11.72	3 inexperienced annotators
R2	36.72	More specific prompt
R3	74.91	Emotion definition and examples
R4	86.04	High-quality examples
R5	83.26	Final agreement among 6

Table 2: Inter-annotator agreement on a subset of the Incel corpus. Rounds 2 to 4 were performed only with the inexperienced conflicting annotators.

Three experienced annotators achieved high agreement (80.26%) out-of-the-box, while three less experienced annotators had low agreement (11.72%). Through two additional rounds, refinements were made: clearer guidelines increased agreement by 12%, and 10 high-quality examples with emotion definitions (similar to 2%-shot ICL prompting) raised agreement to 74.91%. A final round using error-focused examples brought agreement to 86.26% among conflicting annotators and 83.26% overall. The process, completed in four weeks, demonstrated how LLM-generated examples and definitions improved IAA from 24.98% to 86.26% (Table 2), highlighting hybrid annotation systems' potential to enhance quality and consistency in complex tasks like emotion classification.

LLMs provided an efficient baseline by resolving ambiguities, but their biases, such as overdetecting negative language in the Hate-Speech dataset, required human oversight. The proposed hybrid framework combines LLM scalability with human review to flag ambiguous cases, improving annotation quality, reduce cognitive load, and maintaining ethical alignment. These findings emphasise the effectiveness of hybrid systems in domain-specific tasks. Figure 1 shows the label agreement matrix, where GPT-3.5 and GPT-4 achieved high agreement (86.04%), indicating similar predictions. In contrast, Claude-Opus and Gemma-2b showed lower agreement with GPT-3.5 and GPT-4, highlighting reduced consistency in smaller or less robust models (Liu et al., 2024).

5 Results and Discussion

Our evaluation of LLMs on the Hate-Speech, Negative Emotion, and Incel datasets highlights notable differences in performance across various prompting techniques and models. For the Hate-Speech dataset, GPT-4 demonstrates superior generalisation capabilities across prompting strategies, achieving a Cohen's Kappa score of 73.45% in the zero-shot setting. This highlights its robustness in handling complex classification tasks without additional context, aligning with our contribution to advancing cross-domain generalisation. However, the variability in GPT-4's performance across few- and many-shot settings suggests the critical role of prompt design. This underscores our second contribution related to optimising prompting strategies for nuanced annotation tasks. This is significantly higher than GPT-3.5's 59.10% and GPT-4o's 60.98%. These results suggest that GPT-4's robust training allows it to better handle the hate-speech classification task without needing additional examples. The improvement in performance with GPT-4 in the few-shot settings (2%-shot and 20%-shot) further reinforces its superior capability in leveraging context to enhance classification accuracy, achieving a peak Cohen's Kappa score of 73.59% in the 2%-shot setting.

In contrast, GPT-4o and GPT-3.5 show more variable results: GPT-4 achieves a high precision score in some cases, such as 92.67% in the zero-shot setting, compared to GPT-3.5's 88.16% and GPT-4o's at 88.55%. However, GPT-4o's performance is less consistent, with Cohen's Kappa dropping in the 20%-shot setting to 63.92%. This variability indicates that while GPT-4o and GPT-3.5 can be competitive, their performance is less stable compared to the more consistently high results from GPT-4. In the results for other models like Claude-Opus and Gemma-2b, we observe lower overall performance compared to GPT-4. For example, Claude-Opus achieves a Cohen's Kappa of 28.27% in the 20%-shot setting, whereas Gemma-2b reaches 19.34%. Both models perform noticeably lower than GPT-4 and GPT-3.5, underscoring the advantage of more advanced models in terms of handling nuanced hate-speech classification tasks. Specifically, Gemma-2b shows more variability in performance, with its zero-shot Cohen's Kappa at 16.15%, increasing only slightly to 23.17% in the 2%-shot setting. Due to the poor performance of Claude-Opus and Gemma-2b in the zero-, few- and many-shot setting when using the base prompt, we do not further assess their capabilities in the CoT and ICL scenarios.

For the Negative Emotion dataset, GPT-4o demonstrates superior Cohen's Kappa scores compared to GPT-3.5 and GPT-4o. GPT-4o achieves

		Claude-Opus			Gemma-2b		
		Zero-shot	2%-shot	20%-shot	Zero-shot	2%-shot	20%-shot
Hate-speech	Accuracy	47.38	52.93	58.61	42.00	52.03	48.26
	Precision	63.94	67.21	66.47	61.74	69.84	56.14
	Cohen's Kappa	21.46	23.27	28.27	16.15	23.17	19.34
Negative Emotion	Accuracy	39.84	33.21	28.48	32.40	31.24	26.53
	Precision	71.77	69.78	65.22	68.17	67.74	64.82
	Cohen's Kappa	58.23	63.33	58.88	61.23	62.59	57.28
Incel	Accuracy	33.42	31.11	27.83	28.43	30.28	26.45
	Precision	68.98	59.82	51.22	58.37	53.38	53.24
	Cohen's Kappa	58.66	62.23	53.78	59.23	60.03	56.14

Table 3: Accuracy, Precision, and Cohen's Kappa Results from the Base Prompts in Zero, Few, and Many-Shot Settings across all Corpora using Claude-Opus and Gemma-2b.

		Zero-shot Base			2%-shot Base			20%-shot Base		
		3.5	4	4o	3.5	4	4o	3.5	4	4o
Hate-speech	Accuracy	79.90	89.00	82.4	81.00	88.70	81.90	76.63	84.00	80.06
	Precision	88.16	91.77	88.55	88.26	92.67	90.32	87.16	86.73	82.51
	Cohen's Kappa	59.10	73.45	60.98	58.76	73.59	61.29	57.25	70.29	63.92
Negative Emotion	Accuracy	75.93	77.18	76.67	75.10	79.25	78.42	71.37	77.59	76.79
	Precision	75.68	77.43	77.06	75.37	79.15	78.92	72.22	77.69	76.81
	Cohen's Kappa	57.60	61.44	60.74	58.08	64.00	63.89	52.51	61.97	60.45
Incels	Accuracy	60.21	52.85	47.64	35.92	59.27	63.12	58.23	53.23	57.43
	Precision	53.33	48.83	44.79	42.47	54.55	56.38	56.44	52.51	55.42
	Cohen's Kappa	29.64	22.93	20.44	15.36	32.34	30.01	33.76	31.64	32.36

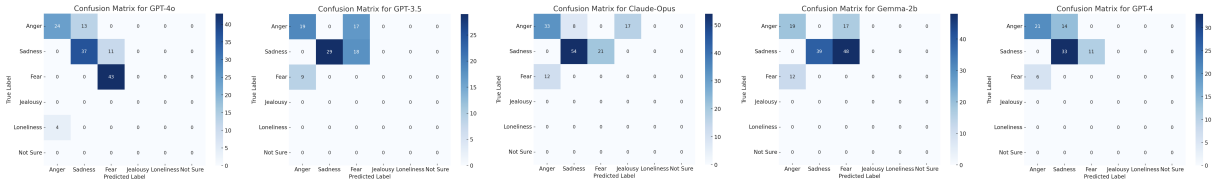
Table 4: Accuracy, Precision, and Cohen's Kappa Results from the Base Prompts in Zero, Few, and Many-Shot Settings across all Corpora using GPT-3.5, 4 and 4o.

		Zero-shot ICL			2%-shot ICL			20%-shot ICL		
		3.5	4	4o	3.5	4	4o	3.5	4	4o
Negative Emotion	Accuracy	77.52	76.76	74.27	62.39	64.79	60.52	72.61	68.60	72.98
	Precision	75.38	77.67	75.15	70.14	71.42	68.03	72.99	71.00	73.17
	Cohen's Kappa	57.64	61.55	56.91	48.38	49.28	46.97	55.17	42.39	54.37
Incels	Accuracy	65.27	52.85	47.64	34.29	60.76	59.04	59.62	64.69	69.63
	Precision	53.33	48.83	44.79	52.31	54.32	49.69	56.14	56.14	58.17
	Cohen's Kappa	29.64	22.93	20.44	13.35	30.2	26.58	34.87	34.97	39.42

Table 5: Accuracy, Precision, and Cohen's Kappa Results from In-Context Learning Prompts in Zero, Few, and Many-Shot Settings for the Incel and Negative Emotion Corpora using GPT-3.5, 4 and 4o.

		Zero-shot CoT			2%-shot CoT			20%-shot CoT		
		3.5	4	4o	3.5	4	4o	3.5	4	4o
Negative Emotion	Accuracy	75.95	77.59	78.42	43.60	44.82	25.80	72.62	76.12	79.59
	Precision	77.30	77.85	78.40	63.10	64.17	51.14	72.98	76.14	80.21
	Cohen's Kappa	54.85	62.31	63.63	18.36	21.62	0.12	54.30	60.81	67.10
Incels	Accuracy	63.11	20.34	36.54	60.80	56.60	41.41	60.02	54.69	59.96
	Precision	51.55	47.27	41.94	57.62	53.21	43.14	54.40	49.14	54.07
	Cohen's Kappa	31.86	14.91	13.54	39.40	30.28	17.54	30.72	34.87	36.25

Table 6: Accuracy, Precision, and Cohen's Kappa Results from Chain-of-thought Learning Prompts in Zero, Few, and Many-Shot Settings for the Incel and Negative Emotion Corpora using GPT-3.5, 4 and 4o.



(a) GPT-4o (b) GPT-3.5 (c) Claude-Opus (d) Gemma-2b (e) GPT-4

Figure 2: Misclassifications across corpora using different LLMs and prompting strategies. The confusion matrices highlight the relationship between true and predicted labels for each model.

the highest Kappa score of 67.10% in the 20%-shot CoT setting, surpassing GPT-3.5, which has a Kappa score of 54.3% in the same setting. The performance of GPT-4o in ICL settings reflects its ability to leverage provided examples effectively, while the CoT prompting yields higher variability, with Cohen's Kappa scores fluctuating in different few-shot settings.

For the Incel dataset, GPT-4o shows notable advantages, with the highest Cohen's Kappa score of 39.42% in the 20%-shot ICL setting, compared to GPT-3.5's 34.87% and GPT-4o's 34.97%. In contrast, GPT-4 and GPT-3.5 exhibit inconsistent performance across different settings, with GPT-4o showing better results in the 2%-shot CoT settings but lagging in the 20%-shot setting. These findings underscore the GPT-4 model's superiority in cross-domain generalisation and multi-label classification tasks.

Our analysis reveals an interesting feature where Cohen's Kappa scores for the Negative Emotion dataset (with fewer labels) are generally higher compared to the Incel dataset (with relatively more labels). This discrepancy can be explained given the specific nature of the labels used in each dataset. This disparity underscores the unique challenges posed by domain-specific emotional nuances, and hence the need for targeted adaptations to LLM architectures. The Negative Emotion dataset is characterized by labels such as "anger," "sadness," and "fear," which are relatively easier to distinguish compared to the broader range of six labels in the Incel dataset, including "sadness," "anger," "fear," "loneliness," "jealousy," and "not sure." The increased number of labels introduces a higher degree of complexity and potential overlap between categories (Li et al., 2024). For example, emotions like "anger" and "fear" can often be contextually intertwined, making it difficult to accurately distinguish them. The presence of overlapping or similar emotional states makes it more difficult for models to assign precise labels, resulting in lower Cohen's Kappa scores (Wang et al., 2023b, 2020). The model's performance is further affected by the inherent ambiguity in classifying emotions that are nuanced or context-dependent, such as "not sure," which introduces additional variability in predictions (Rodriguez et al., 2022).

Models tend to exhibit greater agreement with human annotators when dealing with fewer, more distinct categories, as opposed to a broader, more nuanced set of labels where distinctions can be subtle and overlapping. Consequently, while Cohen's Kappa scores are a valuable metric for assessing model performance, they also reflect the inherent challenges posed by the nature of the classification tasks and the granularity of the labels involved. Overall, GPT-4's superior performance in Cohen's Kappa scores across most datasets and prompting settings demonstrates its robust capability in handling complex classification tasks.

The results presented in Tables 5, 6, and Table C in Appendix illustrate the challenges of nuanced emotion classification in resource-scarce domains. Misclassification patterns, such as anger being confused with sadness or fear with sadness, highlight the limitations of LLMs in distinguishing subtle emotional signals. These errors, coupled with high false positive rates for hate-speech classification, point to systematic biases that hinder generalisation. However, the iterative annotation process (Section 6) demonstrated how LLMs can complement human efforts, improving inter-annotator agreement to over 86%. These findings collectively support our third and fourth contributions, providing actionable insights into reducing emotional bias in LLMs and the practical utility of hybrid annotation systems.

The results for the hybrid annotation framework are in Table 7. For the LLM-only results, we used the best-performing configurations for each dataset. Hate-Speech results were obtained using GPT-4

Data	Metric	LLM-Only	LLM + Human Refinement	Improvement
Hate-Speech	Annotation Accuracy (%)	89.10	96.30	+8.1
	Inter-Annotator Agreement (Kappa)	0.71	0.89	+25.4
Negative Emotion	Annotation Accuracy (%)	84.5	93.72	+10.9
	Inter-Annotator Agreement (Kappa)	0.68	0.85	+25
Incel	Annotation Accuracy (%)	78.2	92.4	+18.2
	Inter-Annotator Agreement (Kappa)	0.62	0.82	+32.3

Table 7: Comparison of LLM-only and LLM + Human Refinement approaches across the Hate-Speech, Negative Emotion, and Incel datasets. Improvements (%) indicate the relative increase in performance achieved through the hybrid annotation framework, combining LLM outputs with human refinements.

with a Base Prompt in a Few-Shot (2%-shot) setting, achieving high annotation accuracy and inter-annotator agreement. For the Negative Emotion dataset, GPT-4 demonstrated its capabilities with a Chain-of-Thought (CoT) Prompt in a Many-Shot (20%-shot) setting, emphasising the importance of step-by-step reasoning in nuanced emotional classification. The Incel dataset results were derived using GPT-4o with an In-Context Learning (ICL) Prompt in a Many-Shot (20%-shot) setting, showcasing the utility of providing extensive context for handling emotionally complex and domain-specific data.

6 Analysing Misclassifications

We randomly sampled 100 rows to identify and analyse the most common label misclassifications across different models. All results are displayed in Tables A - E in Appendix B, Figure 3. Across all models, a consistent pattern emerged in the misclassification of emotions, particularly in the Incel dataset, where anger was often mistaken for sadness. This likely reflects challenges in understanding emotional intensity and receiving contextual nuances in environments characterised by pervasive negative affect. Similarly, in the Negative Emotion dataset, sadness was often confused with fear, highlighting the difficulty in distinguishing emotional distress when emotion indicators are subtle. In the Hate-Speech dataset, all models exhibited a tendency to misclassify not hate-speech as hate-speech, likely due to over-sensitivity to offensive or emotionally charged language and limited recognition of nuanced features like sarcasm or non-hateful tones. A comparative analysis across GPT-4, GPT-4o, and GPT-3.5 reveals evolving patterns of error. For example, GPT-4 misclassified anger as sadness 21 times, compared to 24 in GPT-4o and 33 in GPT-3.5, indicating marginal improvement in newer iterations. However, identify other emotions e.g., confusion increased slightly in the newer GPT-4o (37 instances) compared to GPT-4 (33 instances) and GPT-3.5 (29 instances). Misclassification of not hate-speech as hate-speech rose from 43 instances in GPT-4 to 53 in GPT-4o, suggesting increased sensitivity in newer models.

Despite these gradual improvements, all GPT models struggled with fine-grained emotional nuances and failed to accurately identify neutral language. These biases present opportunities for improvement.

7 Conclusion

This study tackles the challenge of leveraging state-of-the-art large language models (LLMs) for nuanced and domain-specific negative emotion classification, focusing on generalisation across datasets and prompting strategies. Our findings highlight GPT-4's superior performance in resource-scarce and emotionally complex datasets, such as incel discourse, though prompting strategies significantly impact outcomes, with zero-shot and few-shot settings often outperforming many-shot settings. While Chain-of-Thought (CoT) prompting shows potential, it requires further refinement for ambiguous cases.

Our analysis reveals biases in LLM outputs, such as frequent misclassifications of overlapping emotions like anger and sadness, underscoring the need for targeted training and advanced prompting techniques. To address these limitations, we introduced a hybrid annotation framework that combines LLM efficiency with human expertise. This framework significantly improves annotation quality, reduces biases, and enhances inter-annotator agreement, particularly in nuanced or high-stakes tasks. This work contributes benchmarks for LLM performance, insights into prompting strategies, and a practical framework for reducing the psychological and logistical burdens of annotation. Future research should refine LLM generalisation, mitigate biases, and expand the ethical and effective deployment of LLMs in sensitive real-world applications.

Limitations

We focus on single-label classification whereas it is possible for one text to have multiple labels. In our prompt analysis, we use the emotion label which has the majority effect. However, texts can have both anger and jealousy, but if the text depicts jealousy more dominantly, then the label of jealousy is the one assigned. The training and deployment of state-of-the-art LLMs for data annotation demand substantial computational resources, which may not be accessible to all researchers and organisations, thus limiting widespread adoption. As for the widespread application of LLMs as data annotators for hate-speech or sentiment-related contexts, robust data privacy protocols are essential to ensure confidentiality and consent in training and annotation datasets. Human oversight should be employed to review LLM-generated annotations, ensuring accuracy, ethical compliance, and mitigating risks of error propagation or bias.

Ethics Statement

The data used are sensitive and may contain harmful material; therefore, these data cannot be made publicly available to align with the ethical policy associated with the research project. All other data used were publicly available. We understand the sensitive nature of this work and hope that this can contribute to future work in minimising online harms. Ethical approval was gained by the relevant institutions.

Acknowledgements

This work was supported by the [redacted] at [institution redacted for anonymous reasons]. We are grateful to all the researchers involved in this project.

References

Hatoon S. AlSagri and Mourad Ykhlef. 2016. A framework for analyzing and detracting negative emotional contagion in online social networks. In *2016 7th International Conference on Information and Communication Systems (ICICS)*, pages 115–120.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. Preprint.

Md Rabiul Awal, Rui Cao, Roy Ka-Wei Lee, and Sandra Mitrović. 2021. Angrybert: Joint learning target and emotion for hate speech detection. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 701–713. Springer.

Enguerrand Boitel, Alaa Mohasseb, and Ella Haig. 2024. A comparative analysis of gpt-3 and bert models for text-based emotion recognition: Performance, efficiency, and robustness. In *Advances in Computational Intelligence Systems*, pages 567–579, Cham. Springer Nature Switzerland.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020a. Language models are few-shot learners.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020b. Language models are few-shot learners.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Nicoleta Corbu, Alina Bârgăoanu, Flavia Durach, and Georgiana Udrea. 2021. Fake news going viral: The mediating effect of negative emotions. *Media Literacy and Academic Research*, 4(2):58–69.

Avia Efrat and Omer Levy. 2020. The turking test: Can language models understand instructions? *arXiv preprint arXiv:2010.11982*.

Paul Ekman. 1999. Basic emotions. In Tim Dalgleish and Mick J. Power, editors, *Handbook of Cognition and Emotion*, pages 45–60. John Wiley & Sons Ltd.

B. Farhoudinia, S. Ozturkcan, and N. Kasap. 2024. Emotions unveiled: detecting covid-19 fake news on social media. *Humanities and Social Sciences Communications*, 11:640.

Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. 2024. Annollm: Making large language models to be better crowdsourced annotators.

Kiana Kheiri and Hamid Karimi. 2023. Sentimentgpt: Exploiting gpt for advanced sentiment analysis and its departure from current machine learning.

Dheeraj Kodati and Chandra Mohan Dasari. 2024. Negative emotion detection on social media during the peak time of covid-19 through deep learning with an auto-regressive transformer. *Engineering Applications of Artificial Intelligence*, 127:107361.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners.

B. Kok-Shun, J. Chan, G. Peko, and D. Sundaram. 2023. Intertwining two artificial minds: Chaining gpt and roberta for emotion detection. In *2023 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, pages 1–6, Los Alamitos, CA, USA. IEEE Computer Society.

Jeniffer Xin-Ying Lek and Jason Teo. 2023. Academic emotion classification using fer: A systematic review. *Human Behavior and Emerging Technologies*, 2023(1):9790005.

Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. 2024. Long-context llms struggle with long in-context learning.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3?

Yang Liu, Xichou Zhu, Zhou Shen, Yi Liu, Min Li, Yujun Chen, Benzi John, Zhenzhen Ma, Tao Hu, Zhiyang Xu, Wei Luo, and Junhui Wang. 2024. Do large language models possess sensitive to sentiment?

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PloS one*, 14(8):e0221152.

Diego Magdaleno, Martin Montes, Blanca Estrada, and Alberto Ochoa-Zezzatti. 2024. A gpt-based approach for sentiment analysis and bakery rating prediction. In *Advances in Computational Intelligence. MICAI 2023 International Workshops*, pages 61–76, Cham. Springer Nature Switzerland.

V. Mahalakshmi, P. Shenbagavalli, S. Raguvaran, V. Rajakumareswaran, and E. Sivaraman. 2024. Twitter sentiment analysis using conditional generative adversarial network. *International Journal of Cognitive Computing in Engineering*, 5:161–169.

Changrong Min, Hongfei Lin, Ximing Li, He Zhao, Junyu Lu, Liang Yang, and Bo Xu. 2023. Finding hate speech with auxiliary emotion detection from self-training multi-label learning perspective. *Information Fusion*, 96:214–223.

Gilad Mishne and Natalie Glance. 2006. Predicting movie sales from blogger sentiment. In *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs*. Faculty of Science (FNWI), Informatics Institute (IVI). Mish:pred06.

OpenAI. 2023. Gpt-4 technical report.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, and Ilge Akkaya et al. 2024. Gpt-4 technical report.

María Antonia Paz, Julio Montero-Díaz, and Alicia Moreno-Delgado. 2020. Hate speech: A systematized review. *Sage Open*, 10(4):2158244020973022.

Joseph J. Peper and Lu Wang. 2022. Generative aspect-based sentiment analysis with contrastive learning and expressive structure.

Flor Miriam Plaza-del Arco, Sercan Halat, Sebastian Padó, and Roman Klinger. 2021. Multi-task learning with sentiment, emotion, and target detection to recognize hate speech and offensive language. *arXiv preprint arXiv:2109.10255*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou,

Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2022. Scaling language models: Methods, analysis insights from training gopher.

Jaishree Ranganathan and Angelina Tzacheva. 2019. Emotion mining in social media data. *Procedia Computer Science*, 159:58–66. Knowledge-Based and Intelligent Information Engineering Systems: Proceedings of the 23rd International Conference KES2019.

Axel Rodriguez, Yi-Ling Chen, and Carlos Argueta. 2022. Fadohs: framework for detection and integration of unstructured data of hate speech on facebook using sentiment and emotion analysis. *IEEE Access*, 10:22400–22419.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv preprint arXiv:1701.08118*.

Johannes Schäfer and Elina Kistner. 2023. Hs-emo: Analyzing emotions in hate speech. In *Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023), September 19-21, 2023, Ingolstadt, Germany*, pages 165–173. Association for Computational Lingustics.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Synthetic prompting: Generating chain-of-thought demonstrations for large language models. In *International Conference on Machine Learning*, pages 30706–30775. PMLR.

Cici Suhaeni and Hwan-Seung Yong. 2024. Enhancing imbalanced sentiment analysis: A gpt-3-based sentence-by-sentence generation approach. *Applied Sciences*, 14(2).

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation: A survey.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications.

Anjali Tripathi, Upasana Singh, Garima Bansal, Rishabh Gupta, and Ashutosh Kumar Singh. 2020. A review on emotion detection and classification using speech. In *Proceedings of the international conference on innovative computing & communications (ICICC)*.

Radhakrishnan Venkatakrishnan, Mahsa Goodarzi, and M. Abdullah Canbaz. 2023. Exploring large language models' emotion detection abilities: Use cases from the middle east. In *2023 IEEE Conference on Artificial Intelligence (CAI)*, pages 241–244.

Jingjing Wang, Joshua Luo, Grace Yang, Allen Hong, and Feng Luo. 2023a. Is gpt powerful enough to analyze the emotions of memes?

X. Wang, S. Zhao, Y. Pei, Z. Luo, L. Xie, Y. Yan, and E. Yin. 2023b. The increasing instance of negative emotion reduce the performance of emotion recognition. *Frontiers in Human Neuroscience*, 17.

Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Liu Jia. 2023c. Emotional intelligence of large language models.

Yifei Wang, Cheng Shangguan, Chenjie Gu, and Bin Hu. 2020. Individual differences in negative emotion differentiation predict resting-state spontaneous emotional regulatory processes. *Frontiers in Psychology*, 11:576119.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media*, pages 19–26.

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts?

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, An-

drew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Sean Wu, Michael Koo, Lesley Blum, Andy Black, Liyo Kao, Fabien Scalzo, and Ira Kurtz. 2023. A comparative study of open-source large language models, gpt-4 and claude 2: Multiple-choice test taking in nephrology. *arXiv preprint arXiv:2308.04709*.

Mohamed Yassine and Hazem Hajj. 2010. A framework for emotion mining from text in online social networks. In *Data Mining Workshops (ICDMW), 2010 IEEE International Conference on*, pages 1136–1142. IEEE.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2023. Generate rather than retrieve: Large language models are strong context generators.

Tong Zhan, Chenxi Shi, Yadong Shi, Huixiang Li, and Yiyu Lin. 2024. Optimization techniques for sentiment analysis based on llm (gpt-3).

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2023. Sentiment analysis in the era of large language models: A reality check.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models.

Appendix A: Prompts

Prompt Type	Prompt Example
Base Prompt	Analyse the following text and determine the primary emotion expressed. Choose only one emotion from this list: sadness, anger, loneliness, jealousy, fear, or not sure. Respond with only the chosen emotion word.
In-context Learning Prompt	You are an emotion classifier. The emotions can be one of the following: sadness, anger, loneliness, jealousy, fear, or not sure. Below are the definitions of these emotions. Use these definitions to classify the emotion of the new text provided at the end. - Sadness: Emotions such as grief, sorrow, or melancholy. - Anger: Emotions such as rage, frustration, or hostility. - Loneliness: Emotions associated with feeling isolated, abandoned, or disconnected. - Jealousy: Emotions stemming from fear of loss or rivalry. - Fear: Emotions characterized by apprehension, anxiety, or dread. Now classify the following text.
Chain-of-Thought Prompt	Classify the emotion in the following text by reasoning through the feelings step by step. Choose only one emotion from this list: sadness, anger, loneliness, jealousy, fear, or not sure. - Sadness: Emotions such as grief, sorrow, or melancholy. - Anger: Emotions such as rage, frustration, or hostility. - Loneliness: Emotions associated with feeling isolated, abandoned, or disconnected. - Jealousy: Emotions stemming from fear of loss or rivalry. - Fear: Emotions characterized by apprehension, anxiety, or dread. Given this framework, think step by step about the emotions conveyed by the text. First, describe the feelings you observe in the text, and then identify the primary emotion based on those feelings.

Table 8: Prompt types and examples of the prompts used across all models. All prompt generations were conducted with a temperature setting of 0.3.

Appendix B: Misclassifications

Corpora	True Label	Predicted Label	Count
Incel	Anger	Sadness	24
Incel	Not sure	Sadness	13
Incel	Loneliness	Sadness	4
Negative Emotion	Sadness	Fear	37
Negative Emotion	Anger	Fear	11
Hate-Speech	Not hate-speech	Hate-speech	53

(a) GPT-4o Misclassifications.

Corpora	True Label	Predicted Label	Count
Incel	Not sure	Sadness	19
Incel	Fear	Anger	17
Incel	Loneliness	Sadness	9
Negative Emotion	Fear	Sadness	29
Negative Emotion	Anger	Sadness	18
Hate-Speech	Not hate-speech	Hate-speech	49

(b) GPT-3.5 Misclassifications.

Corpora	True Label	Predicted Label	Count
Incel	Anger	Sadness	33
Incel	Jealousy	Anger	17
Incel	Loneliness	Sadness	12
Incel	Not sure	Sadness	8
Negative Emotion	Fear	Sadness	54
Negative Emotion	Anger	Sadness	21
Hate-Speech	Not hate-speech	Hate-speech	61

(c) Claude-Opus Misclassifications.

Corpora	True Label	Predicted Label	Count
Incel	Anger	Sadness	19
Incel	Anger	Fear	17
Incel	Jealousy	Sadness	12
Incel	Loneliness	Not sure	12
Incel	Not sure	Sadness	9
Negative Emotion	Fear	Sadness	39
Negative Emotion	Anger	Sadness	48
Hate-Speech	Not hate-speech	Hate-speech	70

(d) Gemma-2b Misclassifications.

Corpora	True Label	Predicted Label	Count
Incel	Anger	Sadness	21
Incel	Not sure	Sadness	14
Incel	Loneliness	Sadness	6
Negative Emotion	Sadness	Fear	33
Negative Emotion	Anger	Fear	11
Hate-Speech	Not hate-speech	Hate-speech	43

(e) GPT-4 Misclassifications.

Corpora	True Label	Predicted Label	Count
Incel	Not sure	Fear	15
Negative Emotion	Sadness	Anger	27
Negative Emotion	Fear	Sadness	38
Hate-Speech	Hate-speech	Not hate-speech	18

(f) Combined Results.

Figure 3: Misclassifications across corpora using different LLMs and prompting strategies. Each subfigure represents one model's misclassification results for the Incel, Negative Emotion, and Hate-Speech datasets.