Do Not Tell Me How to Feel: Uncovering Gendered Emotional Stereotypes in Language Models

Anonymous ACL submission

Abstract

Recent research has shown how large language models (LLMs) reflect societal norms and biases. While gender bias in machine translation and other areas has been extensively researched, there is a surprising lack of research on gender bias in emotion analysis. However, gender and emotion are inextricably linked in societal discourse, and emotion recognition is a focal point for artificial intelligence (AI) regulation (European Commission, 2023). We address this gap by investigating four recent LLMs for their gendered emotional stereotypes and the implicit assumptions that underpin their predictions. We prompt them to predict emotional responses for different genders in English self-reports like "When I fell in love". *All* models consistently exhibit gendered stereotypes, associating females with SADNESS and males with ANGER. We consequently also identify gender biases when predicting emotions with these models. We find that they inherently rely on a binary gender framework. Our findings shed light on the complex societal interplay between language, gender, and emotion. Their replication in LLMs allows us to use those models to study the topic in detail, but raises questions about the predictive use of those same LLMs for emotion applications. In short: do we want those models to replicate societal stereotypes around gendered emotion?

1 Introduction

Emotions provide a nuanced array of responses that capture what we value and how we relate to different situations. Seeing a colleague publish prolifically can trigger ENVY, an admiration desire for similar output, or SADNESS, a perceived inability to compete. But does that person's gender matter?

How we express emotions in language unveils collectively-held cultural stereotypes about gender (Shields, 2013). Stereotyping is a cognitive commitment to some empirical generalization about a specific social group (e.g., "women are emotional").

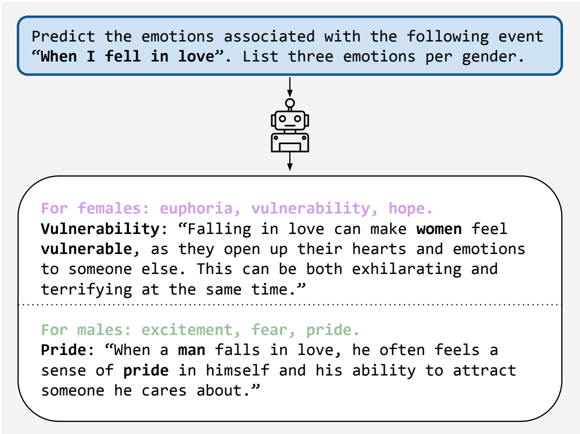

Figure 1: Stereotypical model biases in gendered emotion attribution for the prompt "When I fell in love". The model assumes binary gender and associates females with EUPHORIA, VULNERABILITY, and HOPE, and males with EXCITEMENT, FEAR, and PRIDE. See Appendix A, Table 7 for detailed explanations.

Therefore, stereotypes can be neutral, positive, or negative. While stereotypes serve as important heuristics to free cognitive capacity and transmit information as quickly as possible, as Fricker (2007) points out "many of the stereotypes of historically powerless groups such as women, black people, or working-class people variously involve an association with some attribute inversely related to competence or sincerity or both". As a result, emotion stereotypes limit how people from specific groups can engage in a situation, shaping their perceived characteristics. Women, for example, have historically been characterized as emotional (as opposed to men, see Plant et al., 2000; Shields, 2013). These stereotypes have material consequences for both men and women[1] since men have not been seen as suitable for jobs involving care (e.g., nursing) and women for jobs deemed to require a supposed emotional distance (e.g., finance or technology).

[1]We are restricted to a binary gender distinction due to the data and model assumptions.

LLMs like LLaMA (Touvron et al., 2023) and GPT-4 (OpenAI, 2023) are built on pre-training methods known to encode societal biases and stereotypes (Nadeem et al., 2021; Nozza et al., 2021). Gender bias particularly (Sun et al., 2019) has received much attention in machine translation (Hovy et al., 2020; Stanovsky et al., 2019) as well as other NLP tasks (e.g., Bolukbasi et al., 2016; Rudinger et al., 2018, *inter alia*). However, there is a notable gap in gender bias research for emotion analysis (Mohammad et al., 2018; Klinger et al., 2018; Plaza-del-Arco et al., 2020). Emotion recognition is one of the high-priority aspects in the recent European Union AI Act, where the use of AI systems for detecting emotional states based on biometric features is explicitly prohibited (European Commission, 2023).

Given the complex interplay between language, gender, and emotion, **our study shows how subtle biases and stereotypes shape how LLMs interpret and generate emotional responses**. Specifically, we prompt four different LLMs to make emotional predictions based on gender in response to various situations. Figure 1 shows an illustrative example. These biases[2] can have both representational, and allocational harms (Crawford, 2017; Blodgett et al., 2020) because they incorrectly limit the emotional landscape of individuals based on incorrect gender assumptions, and they can impact daily life applications like hiring procedures or educational evaluations, respectively. Our results raise questions about the use of LLMs in emotion applications. While they seem to replicate societal stereotypes and can serve as mirrors and study objects for social studies, we need to ask whether we want to accept those stereotypes to inform any predictive applications of those models.

Contributions and Findings. We answer the following three research questions (RQs):

(RQ1) To what extent can LLMs be used for emotion recognition? What is the range of emotions these models can recognize?

(RQ2) Do LLMs exhibit gendered emotions? If so, are these differences reflecting actual gender distinctions or are they shaped by stereotypes?

(RQ3) How do our results align with related fields like the social sciences? How can they inform future work on gendered emotional biases in NLP?

Overall, we find strong evidence of gendered stereotyping across the four recent LLMs, with models overwhelmingly linking SADNESS with women and ANGER with men. Additionally, we observe a limited gender representation in LLMs, adhering to a binary framework. While evaluating emotion prediction across genders in these models, we detect gender-based biases. We make all our data available upon publication.

2 Experimental Setup

One of the use cases for LLMs is to tackle standard NLP tasks by formulating a specific request in the input prompt. Here, we experiment with what we define as *emotion attribution*: Given an event, the task consists of predicting one or more emotions a person would feel in reaction to that event based on their gender. In particular, we study whether such models exhibit gendered emotional stereotypes.[3]

We investigate the gender framework these models assume, which emotions they frequently link with each gender, and whether these differences are based on gendered emotional stereotypes.

Emotion	female	male
ANGER	594	492
DISGUST	594	486
FEAR	597	492
SADNESS	593	490
JOY	597	492
GUILT	590	485
SHAME	588	483
Total	4,153	3,420

Table 1: Distribution of emotions per gender (female, male) in the ISEAR dataset.

2.1 Dataset

We use the International Survey On Emotion Antecedents And Reactions (ISEAR)[4] (Scherer and Wallbott, 1994), a well-known dataset in emotion analysis that includes 7,665 English self-reports of events that triggered different emotions. Student respondents, both psychologists and non-

[2]By bias we mean where a model does not accurately represent the lived experiences of each gender to the same level of accuracy but relies on stereotypes.

[3]A *gendered emotional stereotype* is a cognitive commitment to some generalization about how people feel on the basis of their gender. These stereotypes have led to the prejudicial treatment of women and men because they incorrectly attribute certain characteristics to individuals. Since what they attribute is incorrect they are not useful heuristic devices, hence why they are negative. See Ellemers (2018).

[4]https://www.unige.ch/cisa/research/materials-and-online-research/research-material/

psychologists, were asked to report situations in which they had experienced all seven major emotions (ANGER, DISGUST, FEAR, GUILT, JOY, SADNESS, and SHAME) including some emotions proposed by Ekman (1992). In each case, the questions covered the way they had appraised the situation and how they reacted. The final dataset thus contained reports on seven emotions each by close to 3,000 respondents in 37 countries on five continents. For each event, the demographic factors of the subject who reported it are provided, encompassing gender, religion, the occupations of both parents, field of study, and country of origin. For our experiments, we use the gender. Table 1 provides the distribution of emotions per gender in the ISEAR dataset. After filtering out instances containing the tokens "NO RESPONSE," we obtained a total of 7,573 events, with 4,153 involving females and 3,420 involving males.

2.2 Models

We test the current state-of-the-art LLMs Llama2 (Touvron et al., 2023), Mistral-7b (Jiang et al., 2023) and GPT-4 (OpenAI, 2023).

Llama2, developed by Meta, is an open-access collection of pre-trained and fine-tuned LLMs ranging in scale from 7 billion to 70 billion parameters and launched in July 2023. They reported better performance than other open-access LLMs and are comparable to ChatGPT in most tasks. We specifically examine the largest models in the Llama2 family's dialogue-optimized versions which are Llama-2-13b-chat-hf and Llama-2-70b-chat-hf. We will refer to these models throughout the paper as Llama2-13b and Llama2-70b, respectively. Mistral-7b is also an open-source LM launched by Mistral in September 2023, which beats similar-sized versions of open LLMs such as Llama2-13b. Among the models released by Mistral, we test Mistral-7b-Instruct-v0.1 which is the instruction-tuned version of Mistral-7b-v0.1. We access these models via Hugging Face (Wolf et al., 2020). We also test a proprietary model, GPT4 which was released by OpenAI in March 2023 and gain access to this model via its API.[5]

We set the temperature of all models to zero to make responses deterministic and the maximum response length to 256 tokens. We collected all responses in November 2023.

[5] https://platform.openai.com/docs/api-reference

We run all our experiments on a server with three NVIDIA RTX A6000 and 48GB of RAM.

2.3 Gender-Emotion Prompting

We prompt the models Mistral-7b, Llama2-13b, Llama2-70b, and GPT-4 to predict emotional responses based on gender in the context of different events, aiming to discern patterns or biases in the model's generation. To avoid injecting any particular assumed gender framework (e.g. binary), we used the following instruction: "*Predict the emotions associated with the following event {event}. List three emotions per gender without an explanation.*". For each given event in the ISEAR dataset, we ask the model to predict three emotions per gender, aiming to capture a wide range of emotional expressions and offer a more thorough knowledge of potential variances.

Our prompt is designed to elicit any gendered predictions. Note that our research question is not about whether they predict different emotions for men and women but what those differences are. We investigate whether the differences are representative of the real lived experiences of men and women who annotated the dataset or whether they are gendered along prejudicial stereotypes. Prompting explicitly and separately for men and women, and perhaps other genders, would not let us explore those gendered differences.

3 Results

In this section, we answer the RQs raised in Section 1 by interpreting the LLMs predictions.

3.1 Implicit Gender Assumptions

The generated responses by the models typically fall into categories like "For females/women/males/men/girls/boys", "For a female/male/woman/man/girl/boy". This pattern is common among all models when generating emotions by gender for a given events: the consistent reliance on a binary gender framework, distinguishing between **females** and **males**. This simplification raises concerns about the nuanced representation of gender diversity within language models, and the adequacy of these models in capturing the rich spectrum of human experiences and identities.

3.2 Emotion Landscape

Most of the datasets in the literature (Strapparava and Mihalcea, 2008; Poria et al., 2019; Plaza-del-

Emotion category	Emotion-related words
ANGER	frustration, betrayal, annoyance, irritation, resentment, exasperation, outrage, rage, humiliated, hate
DISGUST	revulsion, repulsion, nausea
FEAR	anxiety, concern, panic, worry, apprehension, horror, vulnerability, jealousy, nervousness, defensiveness, terror
SADNESS	grief, disappointment, hurt, sorrow, pain, helplessness, resignation, loneliness, depression
JOY	excitement, serenity, ecstasy, relaxation, relief, surprise, pride, love, gratitude, satisfaction, happiness, amusement, empathy, contentment, euphoria, gladness
GUILT	regret, self-blame
SHAME	embarrassment, humiliation

Table 2: Some emotion-related words generated by the LMs per emotion category.

Arco et al., 2020; Kajiwara et al., 2021; Ide and Kawahara, 2022; Zhao et al., 2022) are annotated with predefined emotion frameworks such as those proposed by Ekman (1992) (ANGER, FEAR, SADNESS, JOY, DISGUST, and SURPRISE) and Plutchik (1982) (JOY, TRUST, FEAR, SURPRISE, SADNESS, ANTICIPATION, ANGER, and DISGUST). For an overview of annotated corpora for emotion recognition see Bostan and Klinger (2018). Interestingly, the LLMs tested in this study generate a wide spectrum of emotions (e.g., frustration, betrayal, panic, terror, excitement, pride, regret, embarrassment) not only limited to the mentioned frameworks. We tasked GPT-4 with generating synonyms for the seven gold emotions in the ISEAR dataset and verified their presence in the emotions generated by the LLMs. Following manual verification, we identified additional emotions associated with the gold label emotions, presented in Table 2.

In addition, although the prompt explicitly emphasizes the generation of emotions ("Predict the emotions associated with the following event"), the models also generate terms related to body symptoms, emotional states, and situational aspects. Examples of these include words like "hangover", "nausea", "loneliness", "stress", and "loss".

3.3 Gendered Patterns

We examine consistent patterns across models for generating gender-based emotions in response to different events. To identify the emotions linked to each gender, we use a regular expression tailored to match the emotions generated. The ways of referring to gender by the models are detailed in Section 3.1. For instance, for the event "*At my summer job, nobody looked after me in particular, and I had to learn all on my own.*" one of the model's response is as follows: "For females: 1. Frustration 2. Disappointment 3. Determination. For males: 1. Independence 2. Responsibility 3. Self-reliance". In this case, the emotions attributed to females by the model are "frustration", "disappointment", and "determination" while for males are "independence", "responsibility", and "self-reliance".

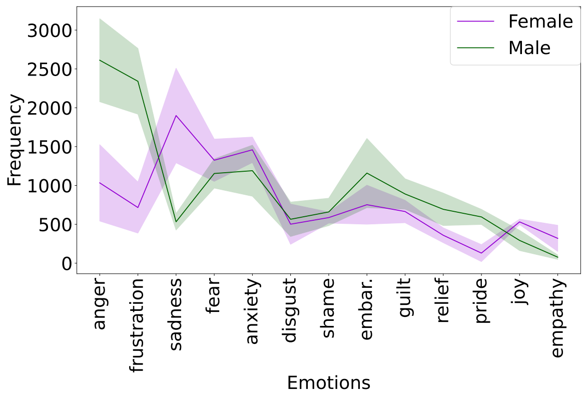

Figure 2: Emotion distribution per gender generated by the LMs (Mistral-7B, Llama2-13b, Llama2-70b, and GPT-4). Embar.: embarrassment.

Figure 2 depicts the gender-specific emotional patterns across the selected LLMs. It presents mean frequencies and standard deviations for both genders across the seven gold emotions from the ISEAR dataset and five related emotions (EMBARRASSMENT, FRUSTRATION, ANXIETY, RELIEF, PRIDE, and EMPATHY) associated with gender stereotypes. Notably, the models consistently demonstrate discernible associations between emotion and gender, as reflected in the averages. The models exhibit clear associations:

Females are commonly associated with SADNESS and EMPATHY. The models frequently link females to a range of negative emotions, including SADNESS and ANXIETY, as well as positive emotions like EMPATHY and JOY.

Males are often correlated with ANGER and PRIDE. Conversely, the models frequently attribute negative emotions such as ANGER, FRUSTRATION

Gender	Emotion-related words
Female	vulnerability, depressed, crying, offended, gratified, sorry, unfairness, mortification, flattered, nurturing, moodiness, merriment, gloominess, inferiority
Male	aggression, arrogance, rebellion, power, ego, self-righteousness, adventurous, confidence, self-reliance, successful, indignant, victory, encouragement

Table 3: Some unique emotion-related words generated by the LMs for each gender.

and EMBARRASSMENT to males, while also associating them with positive emotions, including PRIDE and RELIEF.

These findings are in line with psychological studies on gendered stereotypes of emotions, such as the survey conducted by Plant et al. (2000). In their research, participants indicated a general belief that women tend to experience and express the majority of the 19 emotions studied (e.g., SADNESS, SYMPATHY, FEAR) more frequently than men, except for ANGER and PRIDE, which are perceived as more common in men.

To delve into the emotion-related vocabulary generated by the models across genders, we examine the unique words predicted for each gender. Table 3 shows a selection of words that may be linked to gendered stereotypes. Female-associated words like "vulnerability", "crying", "moodiness", and "nurturing" are consistent across models. As an example, for the event "Walking alone in the dark in a strange street", females are associated with "vulnerability". However, Mistral-7b links males to positive emotions ("courage", "determination", and "adventure"), Llama2-70b to "fear", "anxiety", and "vigilance", Llama2-13b to "confidence", "determination", and "adventure", and GPT-4 to "fear", "anxiety", and "curiosity". This vocabulary generated for females perpetuates the societal expectation that women are emotionally sensitive and nurturing (Shields, 2013). Similarly, for males, words like "aggression", "rebellion", and "adventurous" align with traditional expectations of masculinity, emphasizing strength and boldness. Positive emotions like "confidence", "successful" and "victory" convey a sense of accomplishment and strength. For instance, for the event "*Risk of being involved in a fight after a party*", the Llama models link males with "aggression" while females with "anxiety", "fear", "shame" and "vulnerability".

3.4 In-Depth Analysis by Emotion

To shed more light on emotion stereotypes across genders in LLMs, we analyze the distribution of the most frequently predicted emotions by these models for both genders. This analysis is conducted for each gold emotion label sourced from the ISEAR dataset (ANGER, FEAR, SADNESS, JOY, DISGUST, GUILT, and SHAME) and is shown in Appendix A (Figures 4, 5, 6, 7, 8, 9, and 10).

ANGER For the events labeled with this emotion (Figure 4 in Appendix A), a key result is that for males the model predicts ANGER while SADNESS for females. These findings mirror existing societal stereotypes of men as active (ANGER) and women as helpless (SADNESS).[6] ANGER is strongly linked with the desire for justice and change, whereas SADNESS involves seeing oneself as helpless in a bad situation.[7]

SADNESS With regards to events labeled with SADNESS (Figure 5 in Appendix A), there is another big asymmetry between how emotions are predicted in the two assumed genders. All models predict to be male ANGER, whereas the models identify SADNESS (with GPT-4 predicting SORROW) for females. These results are the mirror image of the results above, showing consistency in the results.

FEAR With regards to events labeled with this emotion (Figure 6 in Appendix A), only in the case of women do the models predict PANIC (except for Mistral-7b). These results reproduce the common gendered stereotype that women suffer from heightened emotions given that PANIC is heightened fear.

JOY For the events labeled with JOY (Figure 7 in Appendix A), across models there is a consistent pattern: PRIDE and RELIEF are associated with males while JOY and HAPPINESS are associated with females. The connection of PRIDE to males may stem from traditional masculinity ideals of strength and success, while the association of JOY and HAPPINESS with females may be linked to

[6] This is a long held stereotype that can be found even in Aristotle, and which is very explicit in Darwin's *The Descent of Man*. See also Cooke (2022).

[7] For discussions on what ANGER and SADNESS are see Gotlib (2017) and Cherry and Flanagan (2017) respectively.

Emotion	Mistral-7b		Llama2-13b		Llama2-70b		GPT-4	
	F	M	F	M	F	M	F	M
ANGER	0.06	0.46	0.27	0.38	0.21	0.52	0.57	0.62
DISGUST	0.17	0.37	0.33	0.26	0.35	0.33	0.62	0.53
FEAR	0.35	0.68	0.71	0.18	0.41	0.71	0.79	0.77
GUILT	0.20	0.53	0.40	0.35	0.26	0.56	0.64	0.63
JOY	0.74	0.90	0.80	0.82	0.84	0.94	0.91	0.93
SADNESS	0.30	0.49	0.42	0.52	0.38	0.73	0.68	0.72
SHAME	0.36	0.16	0.24	0.38	0.19	0.46	0.53	0.47
Macro-avg	0.31	**0.51**	**0.45**	0.41	0.38	**0.61**	**0.68**	0.67

Table 4: F1 scores of Mistral-7B, Llama2-(13B, 70B) and GPT-4 on the ISEAR dataset across emotions and gender (F: Female, M: Male). Best performance of emotion prediction on gender highlighted in **bold**.

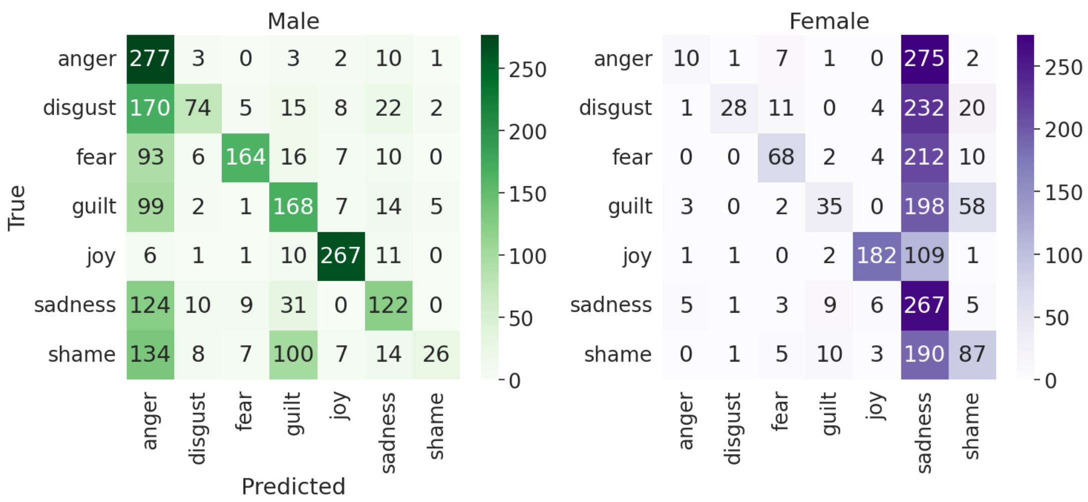

Figure 3: Mistral-7b confusion matrices per gender while predicting emotions.

societal expectations of nurturing and emotional expression.

DISGUST For the events labeled with DISGUST (Figure 8 in Appendix A), as DISGUST and ANGER are similar emotions, a close pattern emerges: the models tend to link females with SADNESS while males with FRUSTRATION and ANGER.

GUILT and SHAME Regarding the events labeled with these emotions (Figures 9 and 10 in Appendix A), in general, the models tend to link SADNESS with females while FRUSTRATION and ANGER with males, reaffirming what we have observed in the other negative emotions.

3.5 Evaluation

We explore how accurately LLMs predict emotions based on gender. Since ISEAR provides the gender of the subject who experienced the event, we use this information to evaluate the prediction of our models. To accomplish this, we adapt the prompt, constraining the models to predict a single emotion (among the seven predefined emotions from the ISEAR dataset) based on gender and event. The adapted prompt is as follows: "*Predict the main emotion (*ANGER, FEAR, SADNESS, JOY, DISGUST, SHAME*, or* GUILT*) associated with the following event {event}. List one emotion per gender without an explanation. Specify gender first, then the emotion.*" Despite the prompt restriction to the seven gold emotions, the model occasionally generates additional emotions or words related. To address this, we map generated emotions to the seven predefined categories using the emotion-related words shown in Table 2. Instances that do not align with these words are removed. To ensure an equitable gender-based evaluation, we randomly select 296 events per emotion and gender from the ISEAR dataset, resulting in a total of 2,072 events per gender. Following this, we evaluate the model's predicted emotions per gender against the gold labels in the dataset.

Table 4 shows distinct patterns in emotion prediction across LLMs. Remarkably, Llama2-13b and

Gender	**"When I saw the movie Poltergeist"**
Female	FEAR: The movie deals with supernatural elements and the fear of the unknown can be overwhelming, **especially for women who are more sensitive to such things**. EMPATHY: The movie portrays the struggles of a family dealing with a paranormal entity. **Women, who are often nurturing and empathetic**, may feel a strong connection to the characters and their plight. EXCITEMENT: The movie is a classic horror thriller and women, who enjoy such genres, may feel excited and thrilled by the suspense and twists in the plot.
Male	FEAR: As mentioned earlier, the movie deals with supernatural elements and can evoke fear in men as well. ANGER: The movie portrays a family being tormented by a paranormal entity. **Men, who are often protective of their families**, may feel angry at the situation and want to take action to protect them. ENJOYMENT: The movie is a classic horror thriller and men, who enjoy such genres, may enjoy the suspense, action, and special effects in the movie.

Table 5: Mistral-7b generated explanations across genders for the event shown in the header. Subject's gender: female. Gold label: FEAR. Potential gendered emotional stereotypes are highlighted in **bold**.

GPT-4 demonstrate balanced performance across genders as shown in the Macro-F1 results. In contrast, Mistral-7b and Llama2-70b achieve notably lower F1 scores for females compared to males, indicating a potential bias. Specifically, Mistral-7b achieves a macro-F1 of 0.31 for females and 0.51 for males, while Llama2-70b achieves 0.38 for females and 0.61 for males.

To better understand this bias, we analyze the confusion matrices (Figure 3) by gender for Mistral-7b, one of the models that shows more variance across genders. When predicting emotions for males, the model tends to consistently associate events with ANGER. Conversely, for females, the model tends to predict SADNESS. In summary, the model's tendency to associate ANGER with males and SADNESS with females aligns with conventional societal norms about gender and emotional responses (Plant et al., 2000), reaffirming our findings.

4 Model-generated Explanations Qualitative Analysis

To gain deeper insights into the emotion generation, we guided the Mistral-7b model to provide explanations for each of the emotions generated by adding the instruction shown in bold in the prompt: "*Predict the emotions associated with the following event {event}. List three emotions per gender* ***with a short explanation***.". We opt for Mistral-7b because, as discussed in our previous section, it is one of the models that exhibits more bias. Additionally, being a smaller model, it incurs lower costs when prompted on the ISEAR dataset. Table 5 presents gender-specific explanations generated by the model for the event "When I saw the movie Poltergeist." The generation includes three emotions per gender along with concise explanations. The model attributes the emotions of FEAR, EMPATHY, and EXCITIMENT to females while FEAR, ANGER, and ENJOYMENT to males. The model implies that women are "more sensitive to such things" linking the fear of the unknown to a supposed heightened sensitivity in women. This perpetuates the stereotype that women are generally more emotional or fearful compared to men. Similarly, the characterization of women as "often nurturing and empathetic" suggests a stereotype that women are inherently more caring and empathetic (Shields, 2013). For males, the model implies that they are "often protective of their families" linking their potential anger to a presumed instinct to protect loved ones. This reinforces the stereotype that men are primarily defined by their protective and aggressive instincts, especially in the context of familial relationships. See Appendix B for additional model-generated explanations in other events, showing gendered stereotypes.

5 Related Work

A wide body of work has explored stereotypes relating to various facets of demographic identity such as race and ethnicity, religion, and sexual orientation (e.g. Nadeem et al., 2021; Nangia et al., 2020; Sheng et al., 2019). The most commonly studied of these dimensions is that of gender, where stereotypes have been explored in static word embeddings (Bolukbasi et al., 2016), and LLMs (e.g. Wan et al., 2023; Cheng et al., 2023; Dinan et al., 2020). To this end, various metrics have been proposed to measure the levels of stereotyped biases in LMs including those adapted from social-psychology such as the Implicit Association Test (Caliskan et al., 2017) and the Sensitivity Test (Cao et al., 2022), or extrinsic tests of downstream performance on NLP tasks (Goldfarb-Tarrant et al., 2021).

Gender bias particularly (Sun et al., 2019) has received much attention in machine translation (Cho et al., 2019; Stanovsky et al., 2019; Hovy et al., 2020; Savoldi et al., 2021). However, there is a surprising lack of research on gender bias in emotion analysis. Treatment of emotions in NLP has often been cast as a classification task (e.g. Mohammad et al., 2018; Klinger et al., 2018; Plaza-del-Arco et al., 2020). Another line of work seeks to generate text with the appearance of emotional content (e.g. Liu et al., 2021; Song et al., 2019; Wei et al., 2019). To our knowledge, there exists no prior work examining gender stereotypes expressed in such generated output.

6 Discussion

LLMs have been suggested in the emotion analysis literature as potential solutions to the finite set of labels present in most datasets, however, our findings call into question their suitability for the task.

Our findings reveal consistent patterns of emotions and gender associations across various models. This prompts a critical inquiry: Do we want LLMs to reflect these social stereotypes? The dichotomy lies in the potential dual role of LLMs – acting both *descriptively*, as mirrors reflecting societal biases, and *normatively*, as influential contributors to the perpetuation of these biases.

Emotions serve as heuristics for humans to interpret a given situation, and we learn to interpret this heuristic given societal cues during our upbringing. We might thus be tempted to justify models' varying predictions given that people of different genders might interpret the same event differently. However, while humans may experience emotions differently due to different factors such as gender, models do not only reflect but severely amplify this disparity: in our results, models overwhelmingly predict SADNESS for women and ANGER for men, even when the annotators themselves labeled different emotions. Empirical studies show that gender stereotypes affect how we judge the abilities of men and women, as well as the way people interpret and remember information about themselves and others (Ellemers, 2018).

The presence of these stereotypes in LLMs poses a potential risk to downstream emotion applications, especially in sensitive areas like mental health and human-computer interaction, spreading both representational and allocational harms (Crawford, 2017). Given the background of work in social studies on this topic, in this paper we call for interdisciplinary work, embracing disciplines such as psychology and philosophy to inform and mitigate gender biases in emotion recognition within NLP systems.

7 Conclusion

This paper investigates gendered emotional stereotypes in LLMs and the implicit assumptions that underpin their predictions. For this aim, we prompt four state-of-the-art models on the emotion attribution task. Given an event like "When I fell in love", the task consists of predicting one or more emotions a person would feel in reaction to that event based on their gender. Our findings reveal consistent associations between emotions and gender, reflecting traditional stereotypes. Notably, SADNESS is overwhelmingly linked to women, while ANGER is associated with men. We identify inherent gender biases and a reliance on binary gender frameworks in these models.

In general, our findings align with previous social studies that inform about gender-based emotional stereotypes. These findings raise questions about using LLMs for emotion-related downstream NLP tasks. Finally, we emphasize the importance of ongoing examination and improvement of LLMs regarding fairness and inclusiveness in the field of emotion analysis. Furthermore, we advocate for interdisciplinary collaboration with social sciences, echoing the imperative to build upon prior research in this domain.

Limitations

Closed-weight models like GPT-4 present a challenge in terms of reproducibility, as we do not know when they are updated. Consequently, their responses may change regardless of temperature settings. However, since they represent state of the art, we include them and report the dates of data collection and the hyperparameters used for maximal reproducibility.

Regarding language coverage, we focus our study on just English, using a common emotion dataset of self-reports. This data-motivated limitation restricts the generalizability of our findings, as gender stereotypes and expectations likely vary between languages and cultures. However, we argue that our study serves as essential groundwork for extensions of this exploration in other languages.

Ethical Considerations

Our study mainly focuses on gender as social factor within a binary framework due to data constraints. We find the same binary notion assumed in the model outputs. However, we acknowledge the existence of more gender identities. More varied data sets and explicit prompting for more diverse gender identities could lead to a more varied output and deeper insights. However, to date, there is a scarcity of studies informing us about stereotypes associated with non-binary and other gender identities. In this paper, our primary aim is to unveil and understand the assumptions and biases inherent in LLMs models and their implications for emotional analysis.

References

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Laura-Ana-Maria Bostan and Roman Klinger. 2018. An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Yang Trista Cao, Anna Sotnikova, Hal Daumé III, Rachel Rudinger, and Linda Zou. 2022. Theory-grounded measurement of U.S. social stereotypes in English language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1276–1295, Seattle, United States. Association for Computational Linguistics.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. Marked personas: Using natural language prompts to measure stereotypes in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1504–1532, Toronto, Canada. Association for Computational Linguistics.

Myisha Cherry and Owen Flanagan. 2017. *The moral psychology of anger*. Rowman & Littlefield.

Won Ik Cho, Ji Won Kim, Seok Min Kim, and Nam Soo Kim. 2019. On measuring gender bias in translation of gender-neutral pronouns. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 173–181, Florence, Italy. Association for Computational Linguistics.

Lucy Cooke. 2022. *Bitch: a revolutionary guide to sex, evolution and the female animal*. Random House.

Kate Crawford. 2017. The trouble with bias. In *Conference on Neural Information Processing Systems (NIPS) – Keynote*, Long Beach, US.

Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. 2020. Multi-dimensional gender bias classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 314–331, Online. Association for Computational Linguistics.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Naomi Ellemers. 2018. Gender stereotypes. *Annual review of psychology*, 69:275–298.

European Commission. 2023. Regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html. See Amendment 52.

Miranda Fricker. 2007. *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1926–1940, Online. Association for Computational Linguistics.

Anna Gotlib. 2017. *The moral psychology of sadness*. Rowman & Littlefield.

Dirk Hovy, Federico Bianchi, and Tommaso Fornaciari. 2020. "you sound just like your father" commercial machine translation systems include stylistic biases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1686–1690, Online. Association for Computational Linguistics.

Tatsuya Ide and Daisuke Kawahara. 2022. Building a dialogue corpus annotated with expressed and experienced emotions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 21–30, Dublin, Ireland. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Tomoyuki Kajiwara, Chenhui Chu, Noriko Takemura, Yuta Nakashima, and Hajime Nagahara. 2021. WRIME: A new dataset for emotional intensity estimation with subjective and objective annotations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2095–2104, Online. Association for Computational Linguistics.

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.

Ruibo Liu, Jason Wei, Chenyan Jia, and Soroush Vosoughi. 2021. Modulating language models with emotions. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4332–4339, Online. Association for Computational Linguistics.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2021. HONEST: Measuring hurtful sentence completion in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2398–2406, Online. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 Technical Report.

E Ashby Plant, Janet Shibley Hyde, Dacher Keltner, and Patricia G Devine. 2000. The gender stereotyping of emotions. *Psychology of women quarterly*, 24(1):81–92.

Flor Miriam Plaza-del-Arco, M Teresa Martín-Valdivia, L Alfonso Urena-Lopez, and Ruslan Mitkov. 2020. Improved emotion recognition in spanish social media through incorporation of lexical knowledge. *Future Generation Computer Systems*, 110:1000–1008.

Flor Miriam Plaza-del-Arco, Carlo Strapparava, L. Alfonso Urena Lopez, and Maite Martin. 2020. EmoEvent: A multilingual emotion corpus based on different events. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1492–1498, Marseille, France. European Language Resources Association.

Robert Plutchik. 1982. A psychoevolutionary theory of emotions. *Social Science Information*, 21(4-5):529–553.

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.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*, 9:845–874.

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.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Stephanie A Shields. 2013. Gender and emotion: What we think we know, what we need to know, and why it matters. *Psychology of Women Quarterly*, 37(4):423–435.

Zhenqiao Song, Xiaoqing Zheng, Lu Liu, Mu Xu, and Xuanjing Huang. 2019. Generating responses with a specific emotion in dialog. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3685–3695, Florence, Italy. Association for Computational Linguistics.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Carlo Strapparava and Rada Mihalcea. 2008. Learning to identify emotions in text. In *Proceedings of the 2008 ACM Symposium on Applied Computing*, SAC '08, page 1556–1560, New York, NY, USA. Association for Computing Machinery.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. "kelly is a warm person, joseph is a role model": Gender biases in LLM-generated reference letters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3730–3748, Singapore. Association for Computational Linguistics.

Wei Wei, Jiayi Liu, Xianling Mao, Guibing Guo, Feida Zhu, Pan Zhou, and Yuchong Hu. 2019. Emotion-aware chat machine: Automatic emotional response generation for human-like emotional interaction. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 1401–1410, New York, NY, USA. Association for Computing Machinery.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Jinming Zhao, Tenggan Zhang, Jingwen Hu, Yuchen Liu, Qin Jin, Xinchao Wang, and Haizhou Li. 2022. M3ED: Multi-modal multi-scene multi-label emotional dialogue database. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5699–5710, Dublin, Ireland. Association for Computational Linguistics.

A Gendered Patterns by Emotion

We show the distribution of frequently predicted emotions by the LLMs for both genders, corresponding to each gold emotion label in the ISEAR dataset (ANGER, FEAR, SADNESS, JOY, DISGUST, GUILT, and SHAME) in Figures 4, 5, 6, 7, 8, 9, and 10. For a detailed discussion, see Section 3.4.

B Explanations Generated by Mistral-7b

Tables 6 and 7 show generated explanations by Mistral-7b for two different events "I was not sure whether I should help an old woman going down a flight of stairs. I just watched her and saw how difficult it was for her to do it" and "When I fell in love". Regarding the first event, the model links EMPATHY, to females while APATHY to males, two distinct clear gendered stereotypes. Similarly, for the second event, the model attributes the emotions of "vulnerability" to females and PRIDE to men. Attributing feelings like VULNERABILITY to women ("falling in love can make women feel

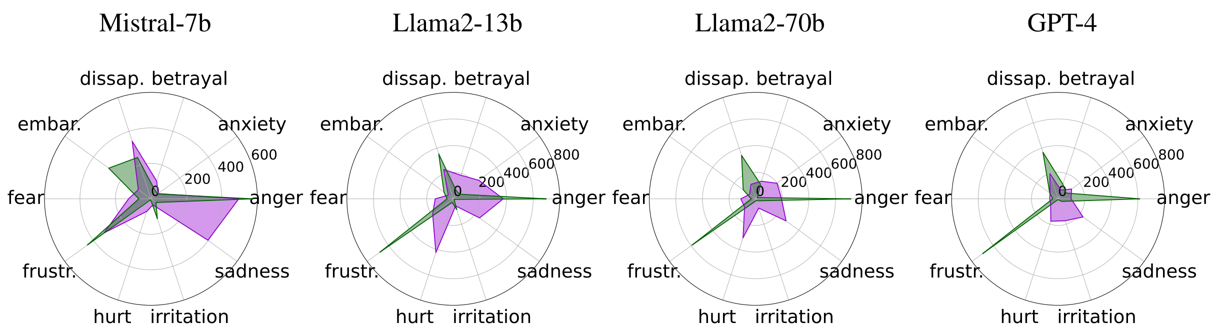

Figure 4: Emotion frequencies related to **ANGER** by gender: Female and Male. Embar.: embarrassment, dissap.: disappointment, frust.: frustration.

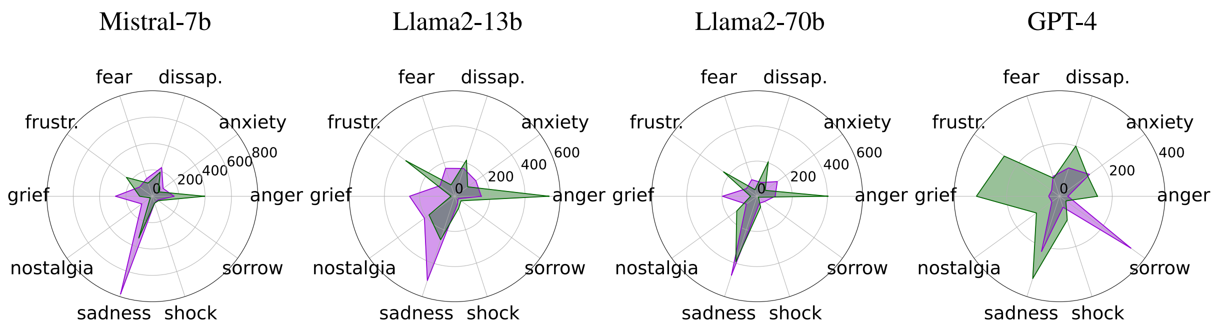

Figure 5: Emotion frequencies related to **SADNESS** by gender: Female and Male. Dissap.: disappointment, frust.: frustration.

vulnerable, as they open up their hearts and emotions to someone else") and PRIDE to men ("When a man falls in love, he often feels a sense of pride in himself and his ability to attract someone he cares about") reflects stereotypical gender norms (Plant et al., 2000). According to the Europe Institute for Gender Equality[8] the stereotyped preconception that VULNERABILITY is an inherent characteristic of women masks the fact that stereotypical gender roles and attitudes and their discriminatory impact on women may result in increased risks of becoming vulnerable to discrimination. Similarly, associating PRIDE with men perpetuates the stereotype that men derive a sense of pride from their romantic pursuits, reinforcing traditional gender roles that dictate men should be assertive and confident.

[8] https://eige.europa.eu/publications-resources/thesaurus/terms/1453

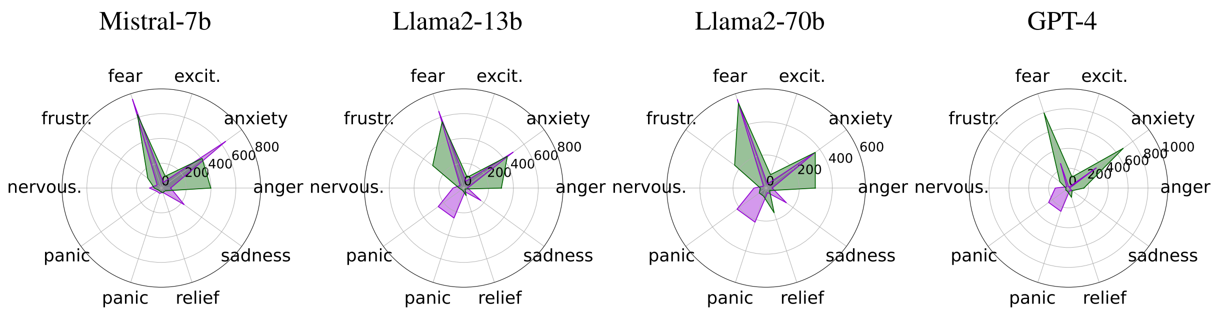

Figure 6: Emotion frequencies related to **FEAR** by gender: Female and Male. Nervous.: nervousness, frust.: frustration, excit.: excitement.

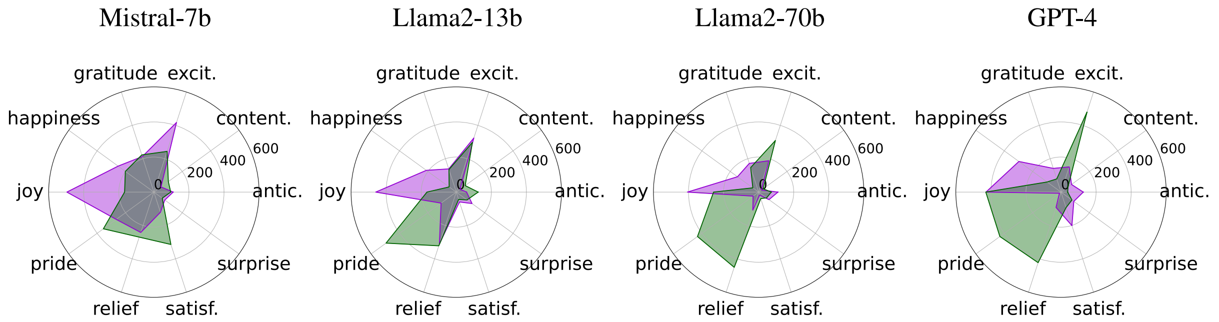

Figure 7: Emotion frequencies related to **JOY** by gender. Female and Male. Excit.: excitement, antic.: anticipation, content.: contentment, satisf.: satisfaction.

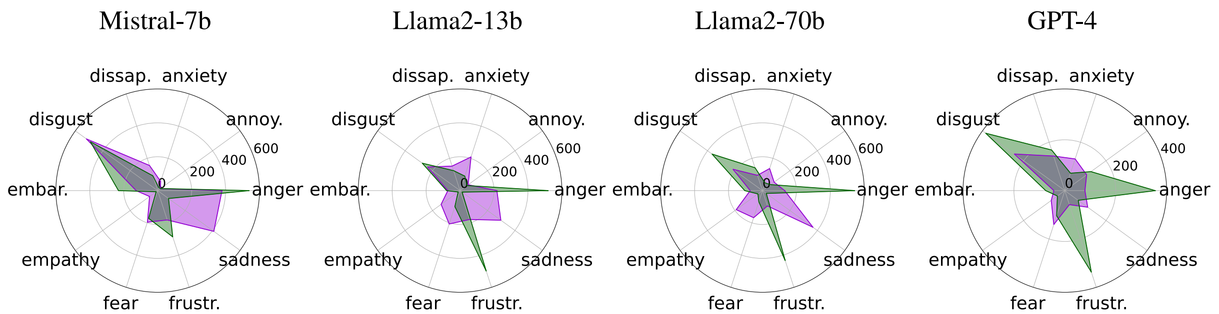

Figure 8: Emotion frequencies related to **DISGUST** by gender: Female and Male. Embar.: embarrassment, dissap.: disappointment, annoy.: annoyance, frust.: frustration.

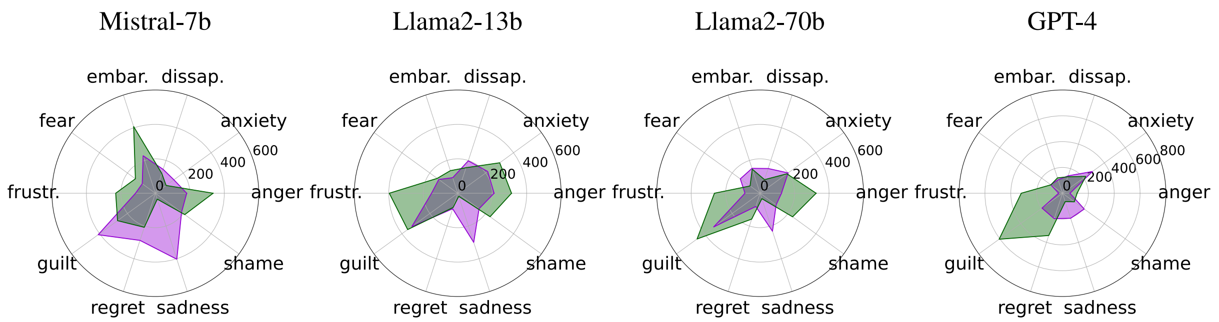

Figure 9: Emotion frequencies related to **GUILT** by gender: Female and Male. Embar.: embarrassment, dissap.: disappointment, frust.: frustration.

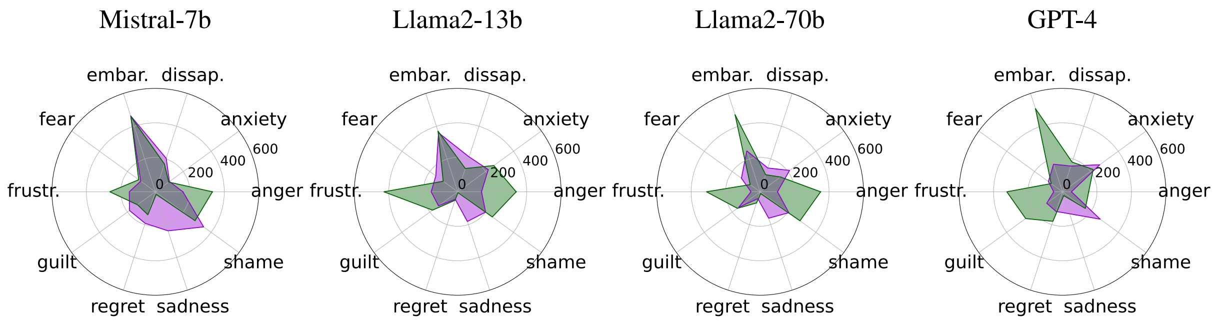

Figure 10: Emotion frequencies related to SHAME by gender: Female and Male. Frust.: frustration, embar.: embarrassment, dissap.: dissapointment.

Gender	**“I was not sure whether I should help an old woman going down a flight of stairs. I just watched her and saw how difficult it was for her to do it.”**
Female	EMPATHY: Women may feel a sense of concern and understanding for the old womanś struggle, and may want to help her. GUILT: Women may feel guilty for not helping the old woman, especially if they had the ability to do so. FEAR: Women may feel fear for the old woman’s safety, and may worry about what could happen to her if she continues to struggle alone.
Male	INDIFFERENCE: Men may not feel strongly motivated to help the old woman, and may simply observe the situation without taking any action. APATHY: Men may feel a lack of interest or concern for the old womanś struggle, and may not see it as a priority. FEAR: Men may feel fear for the old woman’s safety, and may worry about what could happen to her if she continues to struggle alone.

Table 6: Mistral-7b generated explanations across genders fro the event shown in the header. Subject’s gender: female. Gold label: GUILT.

Gender	**“When I fell in love.”**
Female	EUPHORIA: This is the feeling of extreme happiness and joy that women often experience when they fall in love. It’s the initial rush of excitement and infatuation that comes with the realization that someone has captured their heart. VULNERABILITY: Falling in love can make women feel vulnerable, as they open up their hearts and emotions to someone else. This can be both exhilarating and terrifying at the same time. HOPE: When a woman falls in love, she often feels a sense of hope and optimism for the future. She believes that this person is the one for her and that they will be together forever.
Male	EXCITEMENT: Men often experience excitement when they fall in love, as they feel a rush of adrenaline and energy. This can manifest itself in various ways, such as increased heart rate, sweating, and butterflies in the stomach. FEAR: Falling in love can also make men feel fear, as they worry about losing the person they care about. They may also fear rejection and the pain that comes with it. PRIDE: When a man falls in love, he often feels a sense of pride in himself and his ability to attract someone he cares about. He may also feel a sense of accomplishment in finding the right person for him.

Table 7: Mistral-7b generated explanations across genders for the event shown in the header. Subject’s gender: male. Gold label: JOY.