Gender Bias in Emotion Recognition by Large Language Models

Maureen Herbert†, Katie Sun†, Angelica Lim and Yasaman Etesam

Simon Fraser University, Burnaby, BC, Canada

Abstract

The rapid advancement of large language models (LLMs) and their growing integration into daily life underscore the importance of evaluating and ensuring their fairness. In this work, we examine fairness within the domain of emotional theory of mind, investigating whether LLMs exhibit gender biases when presented with a description of a person and their environment and asked, "How does this person feel?". Furthermore, we propose and evaluate several debiasing strategies, demonstrating that achieving meaningful reductions in bias requires training based interventions rather than relying solely on inference-time prompt-based approaches such as prompt engineering.

Keywords

Gender Bias in LLMs, Fairness, Emotion Recognition, Emotional Theory of Mind

1. Introduction

As artificial agents increasingly interact with humans, it is essential for them to possess emotional intelligence [1] and be able to perceive and infer human emotions reliably. However, emotion recognition is inherently subjective and our interpretations of others' feelings are shaped by both societal norms and individual perspectives [2]. Plaza-del Arco et al. [3] showed that such biases may also emerge in LLMs when asking LLMs for an emotion label given a situation and a gender. In this paper, we examine these biases using context-rich image descriptions and a multi-label setup, focusing specifically on how gendered perceptions influence the interpretation of emotional expressions. In contrast to Plaza-del Arco et al. [3], we ask the model to identify the other person's emotion rather than describing how the LLM itself would feel in that situation. Gender bias is often shown in subtle ways that reinforce systemic inequalities. A study by Condry and Condry [4] illustrated this effect experimentally: when participants observed identical emotional responses from infants, they tended to describe the behavior as "anger" when the infant was labeled boy, and as "fear" when labeled girl. This finding highlights how observers project gendered stereotypes onto emotional expressions. These biases in humans are reflected in the data that models are trained on, consequently causing the models to learn and reproduce those biases.

Large language models (LLMs), trained on vast datasets of human-generated text, may internalize perceptual biases. Our research investigates how this inheritance manifests in an emotion recognition task and reproduces the gender biases observed in humans. Specifically, we use the NarraCap captions [5] constructed for the EMOTIC [6] dataset as descriptions of context-rich scenarios where people experience different emotions. We input the same caption to LLMs but with different genders and observe whether the models' predictions change accordingly. By analyzing predicted emotion labels across gender modified captions, we examine to what extent different LLMs such as GPT, Mistral [7], and LLaMA [8] contain these subconscious biases that influence the model's perception when analyzing the captions.

In this work, we relied on data augmentation to achieve debiasing. Specifically, we randomly sampled captions from NarraCap [5] with EMOTIC [6] ground truth emotion labels collected from annotators and then expanded them by: 1) swapping the gender and 2) removing the gender from the caption,

Machine Ethics: from formal methods to emergent machine ethics, AAAI-26 Workshop, January 27, 2026, Singapore

†These authors contributed equally.

mrh9@sfu.ca (M. Herbert); ksa138@sfu.ca (K. Sun); angelica@sfu.ca (A. Lim); yetesam@sfu.ca (Y. Etesam)

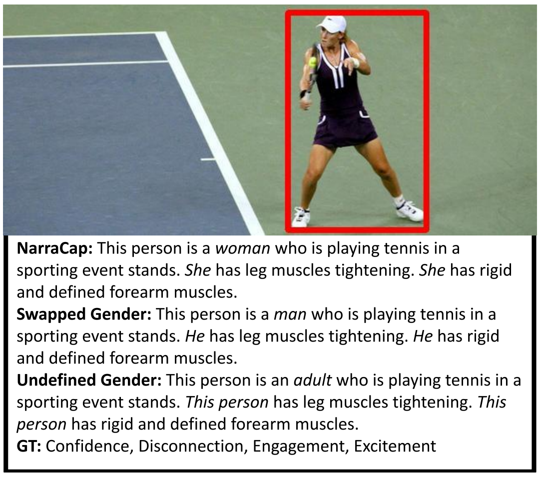

Figure 1: An EMOTIC image [6] with the corresponding NarraCap caption, along with swapped and undefined gender versions. GT represents the ground truth emotion labels chosen by annotators.

while retaining the original ground truth emotion labels for all versions (see Fig. 1). By fine-tuning the model on this augmented dataset, we aimed to desensitize it to gender. To summarize, our contributions are:

- Proposing an investigation of gender biases in LLMs within a multi-label emotion estimation framework.
- Evaluating gender biases in the emotion recognition task across various LLMs, including GPT-4, GPT-5, Mistral, and LLaMA.
- Investigating different debiasing approaches to reduce gender influence in LLMs' emotion predictions, including both inference time prompt-based and training-based methods.

2. Related Work

2.1. Context-based Emotion Recognition

In this paper, we investigate gender bias in LLMs in the context of emotion recognition, which aims to understand a person's apparent emotions. We emphasize that the person's true emotions cannot be reliably inferred, as ground truth is often unavailable; instead, this task focuses on predicting the emotion labels provided by annotators, i.e., "apparent emotion recognition" [2]. Early work largely focused on facial emotion recognition [9, 10, 11, 12], but Barrett et al. (2019) highlighted key challenges, arguing that facial movements alone do not reliably indicate emotion categories across situational contexts in daily life [13]. To incorporate contextual cues, the EMOTIC dataset [6] was introduced as a multi label dataset containing diverse images of people in varying situations experiencing different emotions, alongside a two-branch CNN baseline model to solve the task. Building on this, several approaches have employed attention mechanisms to improve estimation [14, 15], while others have explored multi-branch architectures for distinct context interpretations [16, 17]. More recent methods leverage image captioning to transform visual content into text and extract co-occurrence relationships between words [18, 19]. Recent advances in large language models have introduced their use in contextual emotion recognition, either by first generating captions and feeding them to an LLM [20, 5], or by directly employing multimodal large language models [21, 5, 22].

2.2. Emotional Intelligence and LLMs

Mayer et al. [23] describes emotional intelligence (EI) as the "ability to monitor one's own and others' feelings and emotions, to discriminate among them and to use this information to guide one's thinking and actions." [23]. Language plays a crucial role in emotion perception and reasoning [24, 25, 26, 27, 28], and LLMs have shown emerging capacities in these domains. Early work demonstrated latent capabilities for reasoning in large language models (LLMs) [29], including some sub-tasks on emotion inference [30, 31]. Recent studies have begun systematically assessing LLMs' EI. Psychometric evaluations found above-average EQ scores but notable variation across models [32]. EMOBENCH [33] addressed benchmark limitations by testing emotional understanding and application, revealing substantial human–model gaps. Other work emphasized the need for non-deterministic assessments more aligned with human EI [34]. By contrast, Schlegel et al. [35] reported that frontier models (e.g., GPT-4, Claude, Gemini) outperformed humans on five EI tests and could even generate novel test items. Overall, LLMs display promising yet uneven emotional reasoning abilities, with systematic evaluation of their alignment to human EI still an open challenge.

2.3. Social Bias in LLMs

"Social bias broadly encompasses disparate treatment or outcomes between social groups that arise from historical and structural power asymmetries" [36]. Prior studies reveal varying degrees of such biases in large language models (LLMs) [37, 38]. Social biases have also been studied in specific domains such as auto-generated code [39] recommendation [40, 41], ranking [42], political ideology [43], and gender stereotypes [44, 45, 46]. Plaza-del Arco et al. [3] also show that societal biases and stereotypical patterns appear in emotion attribution across LLMs. Recent work further highlights that while LLMs may appear unbiased under explicit bias benchmarks, they can still harbor implicit biases that remain hidden without more nuanced evaluation [47]. To address these challenges, a variety of bias mitigation techniques have been proposed [36]. Specific strategies include prompt engineering and in-context learning [44, 48], hyperparameter tuning, instruction guiding, and debias tuning [45], as well as model fine-tuning approaches [43].

3. Methodology

Gender bias refers to a preference for or prejudice against one gender over another [49]. In this paper, we investigate gender biases in the emotion recognition task in LLMs. Specifically, we examine whether large language models (LLMs) generate consistent outputs when image captions are modified to reflect different genders.

3.1. Defining and Measuring Gender Bias

In this work, we adopt an equal distribution baseline as our definition of an unbiased model. Specifically, a model is considered gender-unbiased if it predicts each emotion label equally for men and women—that is, maintaining a 50:50 distribution across genders for every emotion. This definition acts as a practical reference point for several reasons:

- It acknowledges that there is no objective "ground truth" distribution of emotions by gender that could serve as an alternative reference.
- It gives us a clear, consistent, quantifiable baseline to measure the bias magnitudes across diverse categories.

We emphasize that this 50:50 baseline is a measurement framework rather than a claim about human emotional expression. While human observers do exhibit gender bias in emotion perception (Condry and Condry [4]), our goal is to quantify and compare biases in LLMs using a consistent, neutral benchmark.

3.2. Dataset

To examine the effect of gender on emotion estimation in LLMs, we follow the "captioning + LLM" methodology proposed by Etesam et al. [5]. This method converts an image of a person into a textual

	GPT4o-mini		GPT5-mini		Mistral-instruct		TinyLLaMA		DeepSeek		LLaMA	
	chi2	p	chi2	p	chi2	p	chi2	p	chi2	p	chi2	p
suffering	3.28	0.07	0.01	0.92	0.58	0.45	0.52	0.47	0.04	0.85	0.28	0.60
pain	0.00	1.00	0.00	1.00	0.31	0.58	0.17	0.68	0.21	0.65	0.00	0.99
sadness	0.69	0.41	0.00	0.95	0.07	0.79	0.01	0.94	0.41	0.52	1.24	0.27
aversion	0.36	0.55	0.12	0.73	1.76	0.19	0.01	0.91	0.00	1.00	0.36	0.55
disapproval	0.00	1.00	0.00	1.00	0.00	1.00	0.48	0.49	0.01	0.94	0.61	0.44
anger	0.78	0.38	0.00	0.97	0.13	0.72	0.28	0.60	0.01	0.93	0.76	0.38
fear	0.05	0.82	0.00	0.99	0.06	0.80	0.06	0.80	0.50	0.48	0.04	0.84
annoyance	0.06	0.81	0.00	1.00	0.12	0.73	0.41	0.52	0.15	0.70	0.13	0.72
fatigue	1.42	0.23	0.05	0.82	1.16	0.28	0.13	0.72	0.00	0.97	1.17	0.28
disquietment	1.04	0.31	0.00	1.00	0.50	0.48	0.00	1.00	0.75	0.39	0.95	0.33
doubt/confusion	**7.49**	**0.01**	0.03	0.87	0.00	1.00	0.00	1.00	0.12	0.73	0.00	0.96
embarrassment	0.00	1.00	0.40	0.53	0.02	0.89	0.41	0.52	0.04	0.85	1.29	0.26
disconnection	0.39	0.53	0.61	0.44	2.12	0.15	1.93	0.16	0.00	0.99	0.10	0.75
affection	1.86	0.17	2.41	0.12	0.03	0.87	0.00	1.00	0.73	0.39	0.05	0.83
confidence	0.25	0.62	0.79	0.37	0.45	0.50	3.56	0.06	0.13	0.72	2.28	0.13
engagement	0.11	0.74	0.01	0.94	0.94	0.33	0.94	0.33	0.35	0.55	0.19	0.67
happiness	0.31	0.58	0.35	0.56	2.29	0.13	0.02	0.87	1.14	0.29	0.02	0.89
peace	0.02	0.90	0.04	0.84	0.96	0.33	0.82	0.36	0.21	0.65	1.90	0.17
pleasure	0.00	1.00	0.16	0.69	**7.28**	**0.01**	0.00	0.99	0.01	0.93	0.11	0.74
esteem	0.50	0.48	0.25	0.62	1.35	0.25	0.35	0.56	1.87	0.17	0.02	0.88
excitement	0.00	0.99	0.01	0.91	0.01	0.94	0.00	0.98	0.046	0.83	0.45	0.50
anticipation	0.01	0.92	0.60	0.44	0.27	0.60	0.01	0.91	0.24	0.62	**8.36**	**0.00**
yearning	0.01	0.91	0.36	0.55	0.00	1.00	0.36	0.55	0.11	0.74	0.07	0.80
sensitivity	0.97	0.33	0.00	0.97	1.17	0.28	0.76	0.38	-	-	**7.09**	**0.01**
surprise	0.06	0.81	0.00	1.00	0.18	0.67	0.43	0.51	0.11	0.73	0.21	0.65
sympathy	0.20	0.65	0.09	0.77	1.66	0.20	0.42	0.52	0.00	0.96	1.04	0.31

Table 1
We employed multiple LLMs to perform the emotion recognition task. In this table, we report the Chi-square test values for the association between man and woman variables.

description (NarraCap in Fig. 1) and subsequently performs emotion inference on that text description. We utilize the NarraCap captions [5], which are generated by passing EMOTIC [6] images through CLIP [50] to answer the questions “who”, “what”, “where”, and “how”. EMOTIC contains context-rich images of people experiencing different emotions, with multi-label ground-truth annotations covering 26 categories, which were selected by clustering 400 affect-related words using the ‘visual separability’ criterion [6]. This emphasis on context in EMOTIC makes the generated NarraCap captions rich in contextual information. The answer to the “who” question in NarraCap provides the gender and age of the person, considering only man and woman genders. While it has been shown that these captions can be improved and there is a need for better captions for more accurate emotion estimation [20, 21], it is important to note that the focus of this work is not on the emotion recognition task itself, but on gender biases in emotion recognition. Consequently, in this study, we do not compare the generated emotion labels with ground truth labels, but rather with the generated labels corresponding to the caption with opposite gender.

For our study, we randomly selected 1000 samples from the NarraCap captions corresponding to the EMOTIC validation set and expanded them by 1) swapping the gender (e.g., changing *boy* to *girl*, *man* to *woman*, and *he* to *she*) and 2) neutralizing it (e.g., using *adult* instead of *man* or *woman*, and *this person* instead of *he* or *she*). We want to emphasize that, although the original image distribution may contain an underlying bias toward women or men, this does not affect the proposed approach as we assume a 50–50 distribution. You can see an example in Fig. 1, which shows the original NarraCap caption, the swapped gender version, and the undefined caption, while the ground truth labels remain the same across all three versions. Although the ground truth labels are not used for evaluation, they are required for fine-tuning.

3.3. Evaluating LLMs

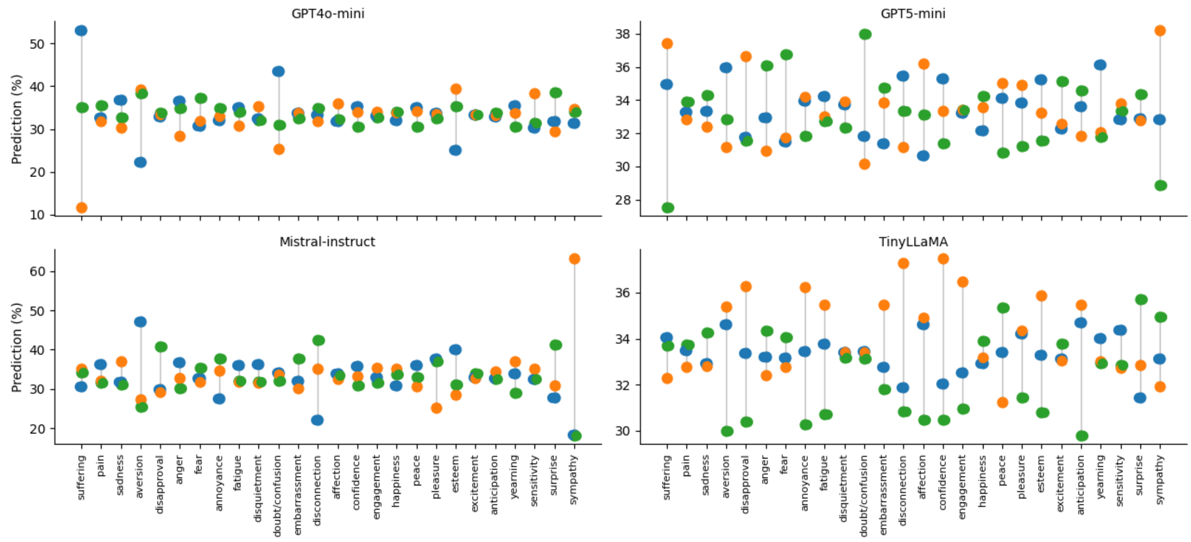

Figure 2: This figure shows the frequency of emotion labels predicted by GPT-4, GPT-5, Mistral, and Tiny LLaMA for captions with man (blue), woman (orange), and undefined (green) genders. To better illustrate the differences across genders, the predictions were normalized based on each emotion label.

On our curated dataset, we compare different LLMs by examining the distributions of predicted emotion labels for captions with woman, man, and undefined subjects. A model without gender bias should treat these captions similarly, producing comparable distributions across genders. We assessed the following models on these three sets: Mistral, TinyLLaMA, GPT-4o-mini, GPT-5-mini, DeepSeek, and LLaMA. This evaluation is performed in a zero-shot setting, without any task-specific training. The same prompt is used for all models:

```
From this list of emotions: {EMOTIC 26 labels} pick the most likely
emotions this person feels simultaneously.
Return ONLY comma-separated emotions. No explanations.
Caption: {NarraCap caption}
Emotions:
```

3.4. Debiasing LLMs

We then used the open-source Mistral-instruct-7B [7] model and applied different techniques to reduce gender bias. These techniques include inference time prompt based approaches, such as prompt engineering, in-context learning, and chain-of-thought reasoning, as well as fine-tuning methods.

3.4.1. Prompt Engineering

To encourage the model to generate emotion labels without considering the person's gender, we add *"Disregard any gender bias you have."* to the prompt.

3.4.2. In-context Learning

We also experimented with enhancing the prompt by adding examples to our original prompt. To assist with debiasing, we include two similar captions in which the only difference is the person's gender, while the expected emotion labels remain unchanged. This indicates that changing the gender should not influence the predicted labels:

```
<original prompt> +
Example:
Caption:
The woman wiped her eyes and smiled softly as she looked at the photo.
Emotions: Sadness, Happiness, Peace, Yearning, Sensitivity, Engagement

Example:
Caption: The man wiped his eyes and smiled softly as he looked at the
photo.
Emotions: Sadness, Happiness, Peace, Yearning, Sensitivity, Engagement
```

3.4.3. Chain-of-Thought (CoT)

Etesam et al. [21] showed that the chain-of-thought technique can help large vision-language models infer people's emotions more accurately. Here, we explore whether this technique can also help LLMs be less emotionally biased. To this end, we use this prompt:

```
From this list of emotions: {EMOTIC 26 labels} pick the most likely
emotions this person feels simultaneously.
Explain the reasoning behind your choice(s) and then give the emotion
label(s).
Example:
Caption: "The woman wiped her eyes and smiled softly as she looked at
the photo."
Reasoning and emotion labels: She feels the pain of missing someone:
Sadness. She wishes she could be with the person or relive the moment:
Yearning. She smiles softly, recalling a joyful memory: Happiness, Peace.
The photo evokes deep emotional response: Sensitivity. She is fully
absorbed in the memory: Engagement.
Caption: {NarraCap caption}
Reasoning and emotion labels:
```

3.4.4. Fine-tuning using Emotion Labels (FT1)

We also explored fine-tuning the model using quantization [51] and LoRA [52]. For this task, we randomly selected 100 samples from the NarraCap captions in the validation set that were not included in our 1,000 test samples. As shown by [5], only 100 samples are sufficient to fine-tune these models. Following the same procedure used to create our test data, we expanded these 100 samples into 300 caption–emotion label pairs by changing or removing gender from the captions while keeping the same set of emotion labels. We also randomly shuffled the emotion labels to eliminate any effects of order. We then fine-tuned Mistral on this dataset, enabling the model to learn that similar captions with different genders should yield the same set of emotion labels. The prompt we use during fine tuning is similar to our original prompt.

3.4.5. Fine-tuning using Gender and Emotion Label (FT2)

This is a two-step fine-tuning process. In the first step, given the following prompt:

```
Question: {NarraCap Caption} This person is feeling {ground truth
labels}. Is this person a male or female?
Answer:
```

	Zero-shot		Prompt-eng		In-context		CoT		FT1		FT2	
	chi2	p	chi2	p	chi2	p	chi2	p	chi2	p	chi2	p
suffering	0.58	0.45	0.28	0.60	0.01	0.92	0.00	1.00	0.01	0.92	0.01	0.92
pain	0.31	0.58	0.26	0.61	0.00	1.00	0.06	0.80	0.01	0.92	0.02	0.90
sadness	0.07	0.79	0.67	0.41	2.41	0.12	0.02	0.88	0.04	0.84	0.00	0.98
aversion	1.76	0.19	0.00	1.00	**4.29**	**0.04**	1.93	0.16	0.00	0.98	0.00	0.96
disapproval	0.00	1.00	0.03	0.86	0.28	0.60	0.02	0.90	0.00	0.99	0.03	0.86
anger	0.13	0.72	0.01	0.92	0.00	1.00	0.59	0.44	-	-	-	-
fear	0.06	0.80	0.15	0.69	0.07	0.80	0.35	0.55	0.00	1.00	0.16	0.69
annoyance	0.12	0.73	0.18	0.67	0.12	0.73	0.10	0.75	0.01	0.94	0.02	0.89
fatigue	1.16	0.28	2.07	0.15	**4.88**	**0.03**	0.46	0.50	0.11	0.74	0.00	0.96
disquietment	0.50	0.48	0.07	0.79	1.59	0.21	0.55	0.46	0.00	1.00	0.02	0.88
doubt/confusion	0.00	1.00	0.01	0.94	0.19	0.66	1.57	0.21	0.35	0.56	0.32	0.57
embarrassment	0.02	0.89	0.61	0.43	0.16	0.68	1.01	0.31	0.15	0.70	0.98	0.32
disconnection	2.12	0.15	0.02	0.88	2.00	0.16	1.12	0.29	0.01	0.92	0.10	0.75
affection	0.03	0.87	0.25	0.62	0.00	1.00	1.47	0.23	0.11	0.74	0.03	0.86
confidence	0.45	0.50	0.97	0.33	2.99	0.08	0.03	0.85	0.00	0.98	0.00	0.96
engagement	0.94	0.33	0.36	0.55	0.25	0.61	0.14	0.71	0.00	0.96	0.00	0.99
happiness	2.29	0.13	2.15	0.14	**6.47**	**0.01**	**5.77**	**0.02**	0.01	0.91	0.00	0.97
peace	0.96	0.33	0.61	0.43	0.00	1.00	0.26	0.61	0.17	0.68	0.18	0.67
pleasure	**7.28**	**0.01**	**3.88**	**0.05**	2.92	0.09	0.44	0.51	0.00	0.96	0.04	0.85
esteem	1.35	0.25	0.84	0.36	**6.21**	**0.01**	0.42	0.52	0.07	0.79	0.01	0.91
excitement	0.01	0.94	0.00	0.98	1.33	0.25	0.13	0.72	0.00	0.94	0.02	0.90
anticipation	0.27	0.60	0.09	0.77	1.33	0.25	1.32	0.25	0.00	0.96	0.00	0.99
yearning	0.00	1.00	0.31	0.58	0.00	1.00	0.39	0.53	0.00	0.94	0.00	1.00
sensitivity	1.17	0.28	3.64	0.06	**5.33**	**0.02**	3.19	0.07	0.00	0.97	0.10	0.75
surprise	0.18	0.67	0.00	1.00	0.00	1.00	0.02	0.88	0.01	0.91	0.00	1.00
sympathy	1.66	0.20	1.69	0.19	2.10	0.15	1.21	0.27	0.00	0.94	0.00	0.99

Table 2
We applied different debiasing techniques, using Mistral Instruct-7B as the base model for all methods. In this table, we report the results of a Chi-square test to determine whether there is a statistically significant association between man and woman variables. Please note that the fine-tuned methods did not predict the label “anger” for any of the captions.

We expect the model to give similar probabilities to a man and woman. To this end, we fine-tune the model by minimizing the KL divergence between the logarithmic softmax of generated logits and a target distribution where all tokens have zero probability except for an equal high probability (0.5) assigned to “man” and “woman”. After this, we sequentially fine-tune the model using the emotion labels (see previous section “Fine-tuning using Emotion Labels”). In this method, we generate two sets of LoRA weights, one for each step. For evaluation, we load the base model and apply both sets of LoRA weights.

4. Experiments

We passed the captions to different LLMs and collected the frequency of each predicted emotion label. It is important to note that some models generated emotion labels outside the 26 predefined categories (e.g., exhaustion). We excluded those labels and only considered predictions that matched the 26 EMOTIC emotion labels. The experiments were conducted using a single NVIDIA GeForce RTX 3090 GPU.

4.1. Evaluation Metric

For the evaluation metric, we either report the number of times an emotion is predicted for different genders, normalized for each emotion (Fig. 2), or use the Chi-square (χ^2) test. The Chi-square test is a statistical method used to determine whether there is a significant association between categorical variables (in this case, man and woman). It compares the observed frequencies in each category to the expected frequencies if there were no relationship between the variables. The χ^2 value (Chi-square statistic) measures how much the observed data deviate from the expected values-the larger it is, the greater the difference. The p-value indicates the probability that such a difference could occur by chance; a small p-value suggests that the observed association is statistically significant, meaning it is unlikely to have occurred randomly. Some models failed to predict certain labels entirely, for example, DeepSeek did not predict "sensitivity" for either man or woman. In such cases, it is not possible to compute (χ^2).

4.2. LLMs

For LLMs, we configured `do_sample=False` and `max_new_tokens=64`. When employing chain-of-thought prompting, however, we set `max_new_tokens=256`.

4.2.1. Mistral

We used Mistral-7B-Instruct-v0.3[1], an LLM which is an instruction tuned version of Mistral-7B-v0.3.

4.2.2. TinyLLaMA

We used TinyLlama-1.1B[2], a pretrained LLaMA model with 1.1 billion parameters trained on 3 trillion tokens. The model was trained over a 90-day period using 16 NVIDIA A100-40GB GPUs.

4.2.3. LLaMA

We used llama-3.3-70b-versatile[3], a model based on Meta's LLaMA 3.3 and fine-tuned for helpfulness and safety.

4.2.4. GPT-4

We used GPT-4o mini[4] which is a smaller, faster, and cheaper version of GPT-4o that handles text and image inputs. This is the only multi modal model included in our experiments.

4.2.5. GPT-5

We utilized GPT-5 mini[5] which is a smaller, faster, and cheaper version of GPT-5.

4.2.6. Deepseek

We utilized deepseek-chat[6] which is a Chat completion model.

4.3. Fine-tuning

For both of our fine-tuning approaches, we used QLoRA and loaded the model with `load_in_4bit=True`, using `Mistral-7B-Instruct-v0.3` as the base model. The LoRA parameters were set as follows: $r = 16$, lora_alpha = 8, and (target_modules = `["q_proj", "k_proj", "v_proj", "o_proj"]`.

[1]https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3
[2]https://huggingface.co/TinyLlama/TinyLlama-1.1B-Chat-v1.0
[3]https://console.groq.com/docs/model/llama-3.3-70b-versatile
[4]https://platform.openai.com/docs/models/gpt-4o-mini
[5]https://platform.openai.com/docs/models/gpt-5-mini
[6]https://docs.deepseekapi.io/deepseek-api/chat/

5. Results

	Model	Female	Male	Undefined
LLMs	GPT-4o mini	4737	4734	4817
	GPT-5 mini	4610	4553	4575
	DeepSeek	3820	3759	3861
	TinyLLaMA	12356	12531	11961
	LLaMA	5869	5843	5851
	Mistral Instruct	3008	2920	2978
De-biased	Prompt eng	3192	3120	3178
	In-context	5204	5106	5191
	CoT	8711	8488	8830
	FT-labels	20890	20826	20772
	FT-logits&labels	20085	20056	20046

Table 3
Number of predicted labels for each model and gender.

We evaluated the frequency of emotion labels predicted by multiple LLMs for captions with man, woman, and undefined genders. Fig. 2 visualizes these distributions for 4 of these models (GPT-4o mini, GPT-5 mini, Mistral-instruct, TinyLLaMA), normalized per emotion label to highlight relative differences across gender variants. We observe deviations in the predicted emotion distributions across man, woman, and undefined captions. It is important to note that the total number of predictions per model and for each gender varies across different models (see Table 3). Specifically, among the various LLMs, TinyLLaMA tends to predict more labels, and fine-tuned models also produce a higher number of predictions across all settings. Furthermore, except for TinyLLaMA and the fine-tuned models, the other models tend to predict fewer labels for captions referring to men compared to those referring to women or undefined genders.

We applied Chi-square tests to examine whether there was a statistically significant association between predicted emotion labels and gender (man vs. woman) across different models (Table 1). Among these models, GPT5-mini, TinyLLaMA and DeepSeek show no significant gender bias ($p \geq 0.05$), whereas GPT4o-mini, Mistral instruct, and LLaMA exhibit bias for at least one emotion. We observe that different models exhibit varying degrees of bias toward specific emotions, which may be attributed to differences in the data on which they were trained. Table 2 reports the χ^2 statistics and corresponding p-values for prompting and fine-tuning methods using Mistral Instruct-7B as the base model. Certain methods exhibited significant deterioration; for example, the in-context learning technique introduces significant biases across different emotion labels. Notably, *aversion*, *fatigue*, *happiness*, *esteem*, and *sensitivity* were significant under in-context prompting ($p \leq 0.05$); however, the significance of *pleasure* was reduced compared to original Mistral Instruct. The other prompt-based mitigation techniques, while not as detrimental as in-context learning, still did not improve the bias, which is in line with the findings of Kuan and Lee [53]. However, fine-tuned models (FT1 and FT2) eliminated detectable bias, with all emotions yielding non-significant associations. It is noteworthy that the fine-tuned models did not predict the emotion *anger*, which is reflected by missing values in Table 2. These results suggest that, while prompting methods do not mitigate gender-related patterns in emotion estimation, fine-tuning reduces potential gender bias across the evaluated emotional categories to some extent.

6. Conclusion

In this paper, we evaluated gender biases in LLMs on the apparent emotion recognition task. We observed that most of the models exhibit significant gender biases for at least one emotion label. We also proposed different techniques to mitigate these biases, including inference-time prompting and fine-tuning. Our results show that while inference-time prompting did not improve the biases, fine-tuning techniques can be an effective way to mitigate them.

Limitations

Although our results demonstrate measurable gender-related emotional biases in LLMs, it is important to consider several limitations when interpreting them.

Our study uses text captions from static visual scenes, which only partially capture real-world emotional complexity. In natural interactions, emotion perception depends on tone, body language, context, and interpersonal dynamics, so the biases we observe may not fully generalize to multi modal or conversational settings.

Different LLMs tend to output varying numbers of emotion labels per caption. Since EMOTIC allows multiple concurrent emotions, this variability can influence the normalized frequency distributions and χ^2 statistics, potentially masking bias-related patterns.

We restrict our analysis to the 26 EMOTIC emotion categories and a limited set of models (GPT-4, GPT-5, Mistral, LLaMA, TinyLLaMA, DeepSeek). Expanding the study to include other models such as Claude or Gemini, or alternative emotional taxonomies (e.g., Plutchik's wheel) could reveal different trends.

The NarraCap dataset and our augmentation procedure include only binary gender categories (man/woman), so non-binary and gender-diverse identities are not represented. This reflects dataset limitations rather than theoretical intent, and future work should aim to cover all gender diversities.

It is worth considering that emotional expression may vary across genders. In such a scenario, an LLM's gender-based probabilistic estimates of emotion could be justified. However, we do not have data to either confirm or deny this.

Overall, these limitations suggest that the reported biases and mitigation outcomes should be interpreted as indicative rather than conclusive. Future research should explore multi modal evaluations, more diverse gender representations, larger training corpora, and more comprehensive emotional frameworks to better understand and mitigate bias in emotional theory of mind within LLMs.

7. Ethical Impact Statement

This research explores gender biases in LLMs for the apparent emotion recognition task. In this section, we discuss the ethical implications of this work.

Potential Negative Societal Impact. LLM-powered software can unintentionally spread and reinforce stereotypes, including gender biases, when such biases exist in the models. As these tools are put to more widespread use, they may influence public perception and behavior in subtle but significant ways. While the LLMs examined in this study improve accessibility, they also allow for widespread deployment without adequate safeguards. This could unintentionally amplify latent biases on a large scale, potentially shaping users' perceptions, decisions, and interactions even through applications that appear neutral or unbiased. The authors do not support applying this research in ways that perpetuate harmful stereotypes or deepen gender biases.

Limits of Generalizability. The gender bias identified in this study is specific to the pre-trained models we evaluated, which include both open-source models and proprietary models accessed via API. Each model was trained on its own dataset and likely reflects societal biases present in that data. Our study focuses on detecting these biases in specific contexts and may not capture the full range of cultural variations in emotional expression, particularly across non-Western cultures and marginalized communities.

Other Issues This work relies on pre-trained LLMs, including Mistral-7B-Instruct-v0.3, TinyLlama-1.1B, llama-3.3-70b-versatile, GPT-4o mini, and GPT-5 mini. Additionally, we conducted lightweight fine-tuning on the pre-trained Mistral-7B-Instruct-v0.3 model. We recognize that the large-scale computational resources used for pre-training LLMs contribute to a tangible environmental footprint. It is important to acknowledge this impact responsibly.

8. Declaration on Generative AI

We used ChatGPT-5.2 free version (https://chatgpt.com) for formatting assistance (e.g., swapping table columns) and for checking the grammar and flow of our written text.

Acknowledgments

This research was supported by the Canada CIFAR AI Chairs Program and the Rajan Scholar Research Fund.

References

[1] A. Bera, T. Randhavane, D. Manocha, The emotionally intelligent robot: Improving socially-aware human prediction in crowded environments, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019, pp. 0–0.

[2] Y. Etesam, Vision language models for environmental and emotional awareness (2025).

[3] F. M. Plaza-del Arco, A. C. Curry, A. Curry, G. Abercrombie, D. Hovy, Angry men, sad women: Large language models reflect gendered stereotypes in emotion attribution, arXiv preprint arXiv:2403.03121 (2024).

[4] J. Condry, S. Condry, Sex differences: A study of the eye of the beholder, Child development (1976) 812–819.

[5] Y. Etesam, Ö. N. Yalçın, C. Zhang, A. Lim, Contextual emotion recognition using large vision language models, in: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2024, pp. 4769–4776.

[6] R. Kosti, J. M. Alvarez, A. Recasens, A. Lapedriza, Context based emotion recognition using emotic dataset, IEEE transactions on pattern analysis and machine intelligence 42 (2019) 2755–2766.

[7] A. Q. Jiang et al., Mistral 7b, arXiv (2023).

[8] H. Touvron, T. Lavril, G. Izacard, et al., Llama: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971 (2023).

[9] M. Pantic, L. J. Rothkrantz, Expert system for automatic analysis of facial expressions, Image and Vision Computing 18 (2000) 881–905.

[10] B. C. Ko, A brief review of facial emotion recognition based on visual information, sensors 18 (2018) 401.

[11] W. Mellouk, W. Handouzi, Facial emotion recognition using deep learning: review and insights, Procedia Computer Science 175 (2020) 689–694.

[12] W.-S. Chu, F. De la Torre, J. F. Cohn, Selective transfer machine for personalized facial expression analysis, IEEE transactions on pattern analysis and machine intelligence 39 (2016) 529–545.

[13] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, S. D. Pollak, Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements, Psychological Science in the Public Interest 20 (2019) 1–68.

[14] N. Le, K. Nguyen, A. Nguyen, B. Le, Global-local attention for emotion recognition, Neural Computing and Applications 34 (2022) 21625–21639.

[15] A. Mittel, S. Tripathi, Peri: Part aware emotion recognition in the wild, in: Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI, Springer, 2023, pp. 76–92.

[16] T. Mittal, P. Guhan, U. Bhattacharya, R. Chandra, A. Bera, D. Manocha, Emoticon: Context-aware multimodal emotion recognition using frege's principle, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 14234–14243.

[17] D. Yang, S. Huang, S. Wang, Y. Liu, P. Zhai, L. Su, M. Li, L. Zhang, Emotion recognition for multiple context awareness, in: European conference on computer vision, Springer, 2022, pp. 144–162.

[18] J. Chen, T. Yang, Z. Huang, K. Wang, M. Liu, C. Lyu, Incorporating structured emotion commonsense knowledge and interpersonal relation into context-aware emotion recognition, Applied Intelligence 53 (2023) 4201–4217.
[19] W. de Lima Costa, E. Talavera, L. S. Figueiredo, V. Teichrieb, High-level context representation for emotion recognition in images, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2023, pp. 326–334.
[20] V. Yang, A. Srivastava, Y. Etesam, C. Zhang, A. Lim, Contextual emotion estimation from image captions, in: 2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII), IEEE, 2023, pp. 1–8.
[21] Y. Etesam, Ö. N. Yalçin, C. Zhang, A. Lim, Emotional theory of mind: Bridging fast visual processing with slow linguistic reasoning, in: 2024 12th International Conference on Affective Computing and Intelligent Interaction (ACII), IEEE, 2024, pp. 10–18.
[22] J. Zhang et al., Vision-language models for vision tasks: A survey, PAMI (2024).
[23] J. D. Mayer, P. Salovey, D. R. Caruso, L. Cherkasskiy, Emotional intelligence., Cambridge University Press, 2011.
[24] K. A. Lindquist, J. K. MacCormack, H. Shablack, The role of language in emotion: Predictions from psychological constructionism, Frontiers in psychology 6 (2015) 444.
[25] M. D. Lieberman, N. I. Eisenberger, M. J. Crockett, S. M. Tom, J. H. Pfeifer, B. M. Way, Putting feelings into words, Psychological science 18 (2007) 421–428.
[26] K. A. Lindquist, L. F. Barrett, E. Bliss-Moreau, J. A. Russell, Language and the perception of emotion., Emotion 6 (2006) 125.
[27] K. A. Lindquist, M. Gendron, What's in a word? language constructs emotion perception, Emotion Review 5 (2013) 66–71.
[28] M. Gendron, K. A. Lindquist, L. Barsalou, L. F. Barrett, Emotion words shape emotion percepts., Emotion 12 (2012) 314.
[29] J. Huang, K. C.-C. Chang, Towards reasoning in large language models: A survey, in: 61st Annual Meeting of the Association for Computational Linguistics, ACL 2023, Association for Computational Linguistics (ACL), 2023, pp. 1049–1065.
[30] R. Mao, Q. Liu, K. He, W. Li, E. Cambria, The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection, IEEE Transactions on Affective Computing (2022).
[31] M. Sap, R. LeBras, D. Fried, Y. Choi, Neural theory-of-mind? on the limits of social intelligence in large lms, arXiv preprint arXiv:2210.13312 (2022).
[32] X. Wang, X. Li, Z. Yin, Y. Wu, J. Liu, Emotional intelligence of large language models, Journal of Pacific Rim Psychology 17 (2023) 18344909231213958.
[33] S. Sabour, S. Liu, Z. Zhang, J. M. Liu, J. Zhou, A. S. Sunaryo, J. Li, T. Lee, R. Mihalcea, M. Huang, Emobench: Evaluating the emotional intelligence of large language models, arXiv preprint arXiv:2402.12071 (2024).
[34] D. Dalal, G. Negi, D. Picca, Llms and emotional intelligence: Evaluating emotional understanding through psychometric tools, in: 33rd ACM Conference on User Modeling, Adaptation and Personalization, 2025, pp. 323–328.
[35] K. Schlegel, N. R. Sommer, M. Mortillaro, Large language models are proficient in solving and creating emotional intelligence tests, Communications Psychology 3 (2025) 80.
[36] I. O. Gallegos, R. A. Rossi, J. Barrow, M. M. Tanjim, S. Kim, F. Dernoncourt, T. Yu, R. Zhang, N. K. Ahmed, Bias and fairness in large language models: A survey, Computational Linguistics 50 (2024) 1097–1179.
[37] Y. Bai, J. Zhao, J. Shi, T. Wei, X. Wu, L. He, Fairmonitor: A four-stage automatic framework for detecting stereotypes and biases in large language models, arXiv preprint arXiv:2308.10397 (2023).
[38] J. Zhao, M. Fang, S. Pan, W. Yin, M. Pechenizkiy, Gptbias: A comprehensive framework for evaluating bias in large language models, arXiv preprint arXiv:2312.06315 (2023).
[39] L. Ling, F. Rabbi, S. Wang, J. Yang, Bias unveiled: Investigating social bias in llm-generated code, in: AAAI Conference on Artificial Intelligence, volume 39, 2025, pp. 27491–27499.

[40] A. Tommasel, Fairness matters: A look at llm-generated group recommendations, in: 18th ACM Conference on Recommender Systems, 2024, pp. 993–998.
[41] H. Li, D. Shen, C. Wang, Y. Liu, J. Gu, Can llms enhance fairness in recommendation systems? a data augmentation approach, in: 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2025, pp. 570–580.
[42] Y. Wang, X. Wu, H.-T. Wu, Z. Tao, Y. Fang, Do large language models rank fairly? an empirical study on the fairness of llms as rankers, arXiv preprint arXiv:2404.03192 (2024).
[43] L. Lin, L. Wang, J. Guo, K.-F. Wong, Investigating bias in llm-based bias detection: Disparities between llms and human perception, arXiv preprint arXiv:2403.14896 (2024).
[44] S. Dwivedi, S. Ghosh, S. Dwivedi, Breaking the bias: Gender fairness in llms using prompt engineering and in-context learning, Rupkatha Journal on Interdisciplinary Studies in Humanities 15 (2023).
[45] X. Dong, Y. Wang, P. S. Yu, J. Caverlee, Disclosure and mitigation of gender bias in llms, arXiv preprint arXiv:2402.11190 (2024).
[46] A. Sorokovikova, P. Chizhov, I. Eremenko, I. P. Yamshchikov, Surface fairness, deep bias: A comparative study of bias in language models, arXiv preprint arXiv:2506.10491 (2025).
[47] X. Bai, A. Wang, I. Sucholutsky, T. L. Griffiths, Explicitly unbiased large language models still form biased associations, National Academy of Sciences 122 (2025) e2416228122.
[48] G. Chhikara, A. Sharma, K. Ghosh, A. Chakraborty, Few-shot fairness: Unveiling llm's potential for fairness-aware classification, arXiv preprint arXiv:2402.18502 (2024).
[49] T. Sun, A. Gaut, S. Tang, Y. Huang, M. ElSherief, J. Zhao, D. Mirza, E. Belding, K.-W. Chang, W. Y. Wang, Mitigating gender bias in natural language processing: Literature review, arXiv preprint arXiv:1906.08976 (2019).
[50] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: 38th International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.
[51] T. Dettmers et al., Qlora: Efficient finetuning of quantized llms, NeurIPS 36 (2024).
[52] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al., Lora: Low-rank adaptation of large language models., ICLR 1 (2022) 3.
[53] C.-Y. Kuan, H.-y. Lee, Gender bias in instruction-guided speech synthesis models, arXiv preprint arXiv:2502.05649 (2025).