Emotional-Tone Fairness Across Language Levels in Large Language Models

DUO CHEN, YIHAO ZHANG, YUOH LEE, Case Western Reserve University, USA

This project investigates and mitigates emotional-tone bias in instruction-tuned large language models (LLMs) across different language proficiency levels. We examine whether models respond with systematically different levels of sentiment, empathy, or politeness when presented with semantically equivalent inputs written in standard versus non-standard or learner English. Using paired data from corpus dataset, we will analyze responses generated by different LLMs like LLaMA-3-Instruct, Gemma-Instruct, and Mistral-Instruct. Sentiment and emotion will be quantified using VADER, NRC Emotion Lexicon, and Perspective API metrics. To address potential disparities, we will explore some possible mitigation strategies such as prompt-based style normalization and re-ranking responses by sentiment parity. The study aims to reveal whether linguistic variation leads to unequal emotional treatment in AI communication and propose practical interventions to promote fairness and inclusivity in multilingual user interactions.

Additional Key Words and Phrases: Responsible AI, Fairness, Emotional Bias, Language Proficiency, Large Language Model

ACM Reference Format:
Duo Chen, Yihao Zhang, Yuoh Lee. 2025. Emotional-Tone Fairness Across Language Levels in Large Language Models. In *Proceedings of Project Proposal (CSDS 447).* ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Motivation

Large Language Models (LLMs) are increasingly used in interactive systems such as educational tutors, customer support agents, and mental health chatbots. These systems are expected to communicate not only accurately but also with empathy, politeness, and positive emotional tone. However, many users worldwide communicate in non-standard, dialectal, or second-language English.

If LLMs produce more supportive or respectful responses to grammatically fluent users than to those writing with less formal or non-native structures, this could create a new form of linguistic discrimination. Such disparities may discourage non-native speakers, reinforce feelings of exclusion, and reduce trust in AI systems.

This issue is particularly critical because instruction-tuned LLMs—optimized for human interaction—are increasingly deployed in real-world applications. Ensuring that these models treat users with equal emotional consideration, regardless of writing style or proficiency, is essential for building fair and inclusive AI communication. By studying and mitigating emotional-tone bias across language levels, this project contributes to the broader goals of responsible AI development and equitable human–AI interaction.

2 Related Work

Research on bias and fairness in large language models (LLMs) has grown rapidly in recent years. Comprehensive surveys such as Gallegos et al. [2] highlight that bias can manifest across social, demographic, and linguistic dimensions, influencing both model behavior and generated content. Several studies have focused on linguistic variation and its impact on fairness. Lin et al. [5] introduced a dialect fairness benchmark to examine reasoning robustness across dialectal forms,

Author's Contact Information: Duo Chen, Yihao Zhang, Yuoh Lee, dxc830@case.edu,yxz3114@case.edu,yxl3225@case.edu, Case Western Reserve University, Cleveland, OH, USA.

CSDS 447, 2025, Case Western Reserve University
2025. https://doi.org/XXXXXXX.XXXXXXX

while Hofmann et al. [4] and Okpala et al. [9] demonstrated that dialectal or informal English often leads to degraded or more negative responses from LLMs. Similarly, Radaideh and Abbas [10] quantified sentiment-based bias across demographic groups, providing a foundation for sentiment-driven bias evaluation.

Beyond structural or performance bias, emotional behavior in LLMs has become a new area of interest. Bardol and Smirnova [1] explored how emotional framing of prompts alters model affective responses, and Ma and Chen [7] used emotion lexicons to measure implicit sentiment and personality bias in LLM outputs. In parallel, Mozicov et al. [8] studied emotional decision-making patterns in language models, while Gehman et al. [3] examined toxicity and politeness disparities using the Perspective API, establishing a precedent for politeness-based fairness metrics.

Recent works have also proposed fairness auditing and mitigation approaches. Tian et al. [11] applied prompt tuning to evaluate and reduce sentiment bias, and Liu et al. [6] investigated sentiment neutrality in LLM-generated responses as a fairness indicator.

However, most existing research focuses on demographic or dialectal bias rather than differences across language proficiency levels. Our work extends this line of research by analyzing *emotional-tone fairness*—how instruction-tuned LLMs express varying sentiment, emotion, and politeness toward fluent and non-fluent English users—and by exploring lightweight mitigation strategies to reduce such disparities.

3 Problem Definition

3.1 Research Gap

Prior studies on fairness in large language models (LLMs) have primarily examined social or reasoning disparities, such as gender bias, racial prejudice, or performance degradation under dialectal variation. While these studies highlight important fairness concerns, they largely focus on factual accuracy or toxicity rather than emotional or affective differences in responses. Emotional-tone fairness—whether an LLM provides equally empathetic, polite, or supportive responses across different language users—remains underexplored. Specifically, no prior research has systematically investigated whether instruction-tuned LLMs express different emotional tones when interacting with users who write in non-standard or learner English compared to fluent or formal English. This unexplored area represents a critical gap in responsible AI, particularly for multilingual and cross-cultural communication systems.

3.2 Main Contributions

This project aims to address this research gap through both analysis and mitigation. The main contributions are as follows:

- We define and formalize the concept of **emotional-tone fairness across language levels**, focusing on equality of sentiment, empathy, and politeness in model responses.
- We construct a controlled evaluation framework using paired sentences from the **JFLEG** dataset, ensuring semantic equivalence while varying grammar and fluency.
- We measure emotional disparities across multiple instruction-tuned LLMs (**LLaMA-3-Instruct, Gemma-Instruct, and Mistral-Instruct**) using sentiment and emotion analysis tools.
- We propose and test some possible mitigation strategies, including prompt-based style normalization and re-ranking based on sentiment parity, to reduce observed disparities.
- We provide empirical findings and recommendations for developing more inclusive and emotionally fair conversational AI systems.

4 Method Sketch

We propose a framework to evaluate emotional-tone bias in instruction-tuned large language models (LLMs) across different language proficiency levels. The approach compares two groups—**fluent** (standard English) and **non-fluent** (informal, learner, or dialectal English)—that express the same meaning. By feeding these paired inputs into multiple LLMs and analyzing their responses using sentiment, emotion, and politeness metrics, we quantify emotional disparities and identify potential bias in model behavior. Finally, we explore some possible simple mitigation strategies, such as prompt-based style normalization and sentiment-aware re-ranking, to reduce potential emotional bias.

4.1 Evaluation metric

In order to evaluate emotional bias in Large Language Models (LLMs), the first essential step is to quantify their emotional output. To achieve this, we separate emotional tone into three primary dimensions: **Sentiment Polarity**, **Emotion Category**, and **Politeness/Empathy**. Each dimension captures a different aspect of how an LLM expresses affective responses toward users.

(1) Sentiment Polarity. This dimension measures the overall valence of a response—whether it conveys a positive, negative, or neutral attitude. We apply the *VADER* sentiment analyzer to compute polarity scores ranging from -1 (very negative) to $+1$ (very positive). For each model, the average sentiment difference between responses to fluent and non-fluent inputs (ΔSentiment) serves as an indicator of bias.

(2) Emotion Category. Beyond general sentiment, models may express different discrete emotions such as joy, sadness, or anger. To capture this, we use the *NRC Emotion Lexicon*, which maps words to eight basic emotion categories (joy, trust, fear, surprise, sadness, disgust, anger, anticipation). We compute the proportion of each emotion present in model outputs and compare these distributions across language levels.

(3) Politeness and Empathy. Since fairness in emotional tone also involves respectful and supportive communication, we assess politeness and empathy using the *Perspective API*. This tool provides continuous scores (0–1) for attributes such as toxicity, politeness, and respect. We analyze whether the average politeness score differs systematically between groups, using the difference ΔPoliteness as another bias indicator.

Together, these three dimensions provide a comprehensive framework for quantifying emotional-tone fairness in LLMs. Significant differences across language levels in any of these metrics would indicate potential emotional bias in model behavior.

4.2 Dataset

To evaluate emotional-tone fairness across language levels, we consider several established corpora representing different linguistic variations. The **JFLEG** dataset provides learner–corrected English pairs for testing fluency-based bias. The **GYAFC** corpus contrasts formal and informal English, suitable for examining stylistic tone bias. For dialectal variation, the **TwitterAAE** dataset includes African-American and Standard American English samples. Depending on scope and feasibility, one or more of these datasets will be used to capture fluency, formality, and dialect dimensions in assessing emotional fairness.

References

[1] Anja Bardol and Tatiana Smirnova. 2025. ChatGPT Reads Your Tone and Responds Accordingly—Emotional Framing Induces Bias in LLM Outputs. *arXiv preprint arXiv:2507.21083* (2025).

[2] Joaquin Gallegos, Ruixi Liu, and Diyi Yang. 2023. Bias and Fairness in Large Language Models: A Survey. *arXiv preprint arXiv:2309.00770* (2023).

[3] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. *Findings of the Association for Computational Linguistics: EMNLP* (2020).
[4] Jan Hofmann, Jonas Wulff, and Iryna Gurevych. 2024. Covert Prejudice in Large Language Models. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2024).
[5] Yuxin Lin, Yi Chen, and Zhiyong Wu. 2024. One Language, Many Gaps: Evaluating Dialect Fairness and Robustness of Large Language Models in Reasoning Tasks. *arXiv preprint arXiv:2410.11005* (2024).
[6] Wenqi Liu, Min Zhang, and Xinyu Huang. 2025. Exploring Fairness and Explainability in LLM-Generated Responses. *Journal of Learning Analytics* (2025).
[7] Zhiqiang Ma and Hao Chen. 2024. Leveraging Implicit Sentiments: Enhancing Reliability and Validity in Psychological Trait Evaluation of Large Language Models. *Frontiers in Artificial Intelligence* (2024).
[8] Dan Mozicov, Weiqi Zhou, and Xinlei Wang. 2024. EAI: Emotional Decision-Making of LLMs in Strategic Settings. In *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS)*.
[9] Chuka Okpala, Samuel Johnson, and Zeerak Waseem. 2025. Large Language Model Annotation Bias in Hate Speech Detection. *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)* (2025).
[10] Bilal Radaideh and Mohamed Abbas. 2025. Fairness and Social Bias Quantification in Large Language Models for Sentiment Analysis. *SSRN Electronic Journal* (2025).
[11] Zhen Tian, Ming Guo, and Xiangyu Li. 2023. Soft-Prompt Tuning for Large Language Models to Evaluate Bias. *arXiv preprint arXiv:2306.04735* (2023).