Proceedings of the Twentieth International AAAI Conference on Web and Social Media (ICWSM 2026)

Consistency of Responses and Continuations Generated by Large Language Models on Social Media

Wentao Xu[1], **Yuqi Zhu**[1], **Bin Wang**[2], **Wenlu Fan**[1*]

[1]University of Science and Technology of China
[2] Independent Researcher, Beijing, 100000, China

Abstract

Large Language Models (LLMs) show strong text generation capabilities, yet their emotional consistency in social media contexts remains underexplored. This study examines emotional and semantic behaviors of four models: Gemma, Llama 3, Llama 3.3, and Claude in continuation and response tasks using climate change discussions from Twitter and Reddit. Results show that while all models maintain high semantic coherence, they systematically attenuate negative emotions, often shifting them toward neutral or even positive tones. Compared to human-authored content, LLM outputs exhibit reduced emotional intensity and a consistent preference for neutral, rational expression. Despite differences between continuation and response tasks, semantic similarity remains high across models. These findings highlight a general tendency of LLMs toward emotional moderation, offering insights for their deployment in social media and human–AI interaction design.

Introduction

Large Language Models (LLMs) represent one of the most significant yet controversial technological advancements in recent years. These models demonstrate unprecedented and expanding human-like capabilities, particularly in text generation, enabling diverse applications including text summarization (van Schaik and Pugh 2024), translation (Sung et al. 2024), and news writing (Muñoz-Ortiz, Gómez-Rodríguez, and Vilares 2024). Consequently, LLM-based applications have proliferated across domains, from conversational agents (Dam et al. 2024) to educational assistants (Liu, Jiang, and Wei 2025).

Despite their advantages, LLMs raise significant concerns regarding potential negative implications. These include content fabrication, commonly termed "hallucination," which contributes to misinformation propagation (Huang et al. 2023). Furthermore, research indicates that LLM-generated content perpetuates societal biases encountered during training, potentially exacerbating AI fairness issues (Gallegos et al. 2024; Ayoub et al. 2024). Additionally, LLMs can influence human decision-making processes, potentially leading to unintended consequences through emotional manipulation or deception (Park et al. 2024). Given their widespread deployment, careful evaluation of LLMs' text generation capabilities becomes imperative.

LLMs exhibit both task-specificity and context-sensitivity, with performance varying across different applications and contextual settings (Sung et al. 2024; Li, Zhang, and Sun 2023). Consequently, evaluating their text generation capabilities within realistic, socially relevant contexts becomes crucial. Social media platforms, serving as extensive networks for information exchange, provide valuable digital artifacts for such investigations.

In social media contexts, LLM text generation manifests in two primary forms: response tasks (e.g., replies) and continuation tasks (e.g., summarization and dialogue). The generated content influences public perception and engagement on social media platforms. Emotion embedded within text plays a crucial role as it can be rapidly activated and disseminated through extensive social networks, potentially facilitating emotional contagion (Kramer, Guillory, and Hancock 2014). Consequently, emotion serves as a strategic tool for engagement and persuasion in social media environments (Stieglitz and Dang-Xuan 2013; Hamby and Jones 2022).

Previous investigations of emotional effects on social media have employed real-life experiments through content manipulation (Kramer, Guillory, and Hancock 2014). However, it raise ethical concerns regarding manipulation of user content and public discomfort (Boyd 2016). Recent work demonstrates that LLM-based simulations can be a useful tool for modeling social interactions (Gao et al. 2024a): (1) AI agents can serve as safer substitutes for human participants in extreme or sensitive scenarios, and (2) they enable more controlled experimental conditions, facilitating precise examination of relevant variables.

With the widespread adoption of LLMs in generating human-like content, it becomes imperative to understand the consistency of LLM-generated text and its potential societal impact. Moreover, with a growing variety of LLMs being adopted in real-world settings, there remains a lack of systematic, comparative evaluations of how consistent/different these models perform.

Accordingly, this study investigates LLM text generation tasks (response generation and continuation) through systematic analysis of emotional consistency and semantic similarity. By examining these dynamics within climate change communication—a highly polarized and emotionally charged

*Corresponding author, wenlufangfang@gmail.com

domain—this research addresses the following questions:

RQ1: How consistent are the emotions expressed in text generated by LLMs on social media?

RQ2: How does the emotional intensity of text generated by LLMs compare to text on social media?

RQ3: To what extent do LLMs demonstrate semantic similarity between generated text and text on social media?

By answering the research questions, this study has the following findings and contributions:

- We systematically evaluated emotional and semantic consistency between LLM-generated and human-authored social media texts, finding that (1) LLMs' emotional expressions differ significantly from original texts across both continuation and response tasks, and across different model types; (2) LLMs maintain high semantic similarity with original content, indicating strong capabilities in understanding and generating human-like text.
- We found that all models exhibit significantly lower emotional intensity in both tasks, suggesting that LLMs may struggle to convey the full emotional depth of human-authored content.
- We discussed the real-world implications of these emotional differences and semantic proximity — including how LLMs may influence emotional engagement, serve as tools for moderating polarized debates, or, conversely, be misused for emotional manipulation.
- We proposed an ethically grounded simulation framework using LLM agents to explore emotional dynamics around controversial topics on social media, avoiding privacy and consent issues inherent in traditional research (Ferrara and Yang 2015). We also adopted the "LLM-as-judge" approach to assess semantic similarity, reducing reliance on time-consuming and costly human annotations.

Related Works

Evaluation of LLMs generated text

The evaluation of LLM-generated text originates from natural language generation (NLG), defined as the process of computationally producing human-comprehensible text (Sai, Mohankumar, and Khapra 2022). Given the widespread deployment of AI models in text generation, extensive research has explored effective evaluation frameworks for NLG (Sai, Mohankumar, and Khapra 2022). Traditional evaluation metrics, primarily focused on quantifying content overlap between system outputs and references (Gao et al. 2024b), such as BLEU (Papineni et al. 2002) and ROUGE (Lin 2004), have served as standard metrics for automatically assessing output quality in machine translation and summarization tasks. However, these metrics demonstrate limitations when applied to complex, context-dependent tasks, particularly in the current generative AI paradigm (Gao et al. 2024b). Consequently, researchers have developed novel benchmarks for task-specific LLM evaluation (e.g., (Que et al. 2024)), while recent studies have proposed methodologies leveraging LLMs themselves for evaluation purposes (see (Gao et al. 2024b) for a comprehensive review).

The evaluation of LLM-generated text consistency with human behavior represents a fundamental approach to assessing model performance. Alignment with human behavior and response patterns remains a central objective in artificial intelligence development (Russell and Norvig 2016). Consistency is crucial for operational reliability and safety of LLMs, ensuring they can generate contextually appropriate and relatable outputs. Additionally, semantic similarity serves as an established metric for quantifying textual consistency (Chandrasekaran and Mago 2021). Researchers have evaluated LLM output consistency through semantic similarity measures and developed enhancement strategies to improve human alignment (Yang et al. 2024; Raj et al. 2023).

Existing literature predominantly examines distinctive characteristics between LLM- and human-generated text. For instance, (Herbold et al. 2023) conducted comparative analyses of human-written versus ChatGPT-generated essays across dimensions including topical coverage, logical structure, vocabulary usage, and linguistic constructions through human assessment. Beyond manual annotation, (Guo et al. 2023) implemented a mixed-methods approach to analyze LLM/human-generated responses across linguistic dimensions, revealing that LLM outputs demonstrate enhanced logical coherence, comprehensive detail, and reduced bias. (Muñoz-Ortiz, Gómez-Rodríguez, and Vilares 2024) employed quantitative analysis to compare human- and LLM-authored news content across morphological, syntactic, psychometric, and sociolinguistic dimensions.

Text generation on social media context

In the social media environment, LLM text generation offers significant applications, including AI-powered social bots for online discourse participation, discussion summarization tools, and related applications (Li et al. 2024). However, ensuring generated text consistency requires careful consideration of contextual factors and interaction objectives. Social media interactions encompass both response generation (e.g., comment replies) and content continuation (e.g., social bot engagement). While existing research provides empirical evidence comparing human and LLM-generated content, the evaluation of social media-specific tasks, particularly responses and continuations, warrants comprehensive evaluation to understand LLM text generation in dynamic social media contexts.

Although emotion serves as a crucial factor in social media engagement and persuasion, its utilization as an evaluative feature for text generation remains insufficiently explored. Current comparative studies of human and LLM-generated text focus predominantly on static contexts, overlooking emotional dynamics. For instance, comparative analysis of human-written versus LLM-generated news content revealed stronger negative emotional expression in human-authored texts (Muñoz-Ortiz, Gómez-Rodríguez, and Vilares 2024). Similarly, while (Guo et al. 2023) examined response differences through multilingual sentiment classification, this approach presents limitations for comprehensive emotional analysis (e.g., distinguishing between joy and sadness). Given the dynamic nature of social media interactions, evaluation of emotional consistency in information exchange becomes

crucial.

Emotional content permeates social media discourse and functions as a crucial determinant in shaping public opinion (Naskar et al. 2020). Emotion demonstrates high susceptibility to influence and serves as a critical factor in controversial and uncertain social agendas, including epidemics (Lu and Hong 2022), disasters (Chu et al. 2024), and polarizing social issues such as climate change (Brady et al. 2017). In these contexts, emotional responses can exert both beneficial and detrimental effects on public discourse. Climate change discourse, in particular, represents an extensively studied yet remains highly polarized domain, characterized by persistent denialism and skepticism (Treen, Williams, and O'Neill 2020; Whitmarsh 2011). These misconceptions frequently leverage emotional appeals, particularly fear, to influence public perception (Martel, Pennycook, and Rand 2020). Clinical psychology research has established correlations between anger, elevated cortisol responses to stress, and increased vulnerability to misinformation (Sharma, Wade, and Jobson 2023).

Above all, evaluating emotional patterns across response and continuation tasks within climate change discussions on social media provides a crucial framework for comparing LLM and human-generated content in dynamic, real-world scenarios.

Methodology

Experimental Design

Figure 1 illustrates the overall setup of the experimental design. Our core experiment involves using LLMs to interact with human-generated text data, which consists of social media posts collected from real users. In our experiment, we employed three open-source large language models—Gemma[1], Llama 3[2], and Llama 3.3—developed by Google and Meta, respectively, as well as one commercial model, Claude 3.5, developed by Anthropic[3]. Specifically, we utilized the following model variants: Gemma2-27B-Instruct-Q8, Llama3-70B-Instruct, Llama3.3-70B, and Claude-3-5-Haiku-20241022-X. These models were selected due to their strong performance and robustness. We utilized Ollama[4] as a framework to enable the two open source models to run on our local server.

In our study, we primarily used Ollama for LLM inference. For the response task, we directly employed the chat interface without extra prompts, enabling more natural observation of model behavior. The original human post served as the implicit prompt for the model. For the continuation task, we used a concise generation prompt to minimize intervention: *"Assuming you are the author of this text, continue to expand the passage as you understand it."*

Dataset

This study utilized climate change corpora collected from Twitter (now X) and Reddit. We collected data using the Twitter Search API by querying relevant keywords, including "climate change", "climate science", "climate manipulation", "climate Engineering", "climate Hacking", "climate modification", "Global Warming", "carbon footprint", and "The Paris Agreement". For Reddit, we used data maintained by Pushshift from https://the-eye.eu/redarcs/. The Pushshift Reddit dataset consists of two sets of files: submissions and comments (Baumgartner et al. 2020). The same keywords were applied to filter Reddit data, and to compare the differences in emotions, we collected both posts and comments from both platforms.

With the keywords, we obtained 5,768,822 Reddit comments and 76,596,654 tweets from Twitter. We used histograms to understand the basic distribution of data (Figure 2). Twitter and Reddit are two major social media platforms that differ from each other. Twitter is designed to be open, concise, and immediate through short posts and real-time updates, while Reddit is a community-based medium designed for deep, deliberative, and topical discussions(Parsa et al. 2022; Treen et al. 2022a). These differences in platform design not only shape how content is generated and shared, but also attract different user groups with different interests and engagement styles, which provides a valuable resource for understanding different modes of public discussion(Ruan et al. 2022).

We support these distinctions with some quantitative evidence using the average word length, word cloud. The results show that the average length of Reddit text is 656.66 words, while the average length of Twitter text is 246.34 words. This discrepancy reflects fundamental differences in how the two platforms are used. Reddit is structured around topic-based discussions within subreddits, encouraging longer, more detailed posts that often resemble mini-essays or narratives. The word cloud chart (see the appendix7) shows that the expressions of users on the two platforms under the same topic are also different. For example, more people on Reddit are discussing "climate change", while more people on Twitter use "global warming".

To construct datasets representative of the discussion dynamics surrounding climate change over an extended period and to mitigate potential biases introduced by sudden climate change events, we employed a time-stratified sampling approach on the raw data. Specifically, based on the collected corpus of 5,768,822 Reddit comments and 76,596,654 Twitter tweets, we initiated the process by partitioning the data by month. Subsequently, within each monthly segment, we conducted systematic sampling, extracting 200 records from the Twitter data and 100 records from the Reddit data. This method of monthly proportional sampling was implemented to ensure a uniform distribution of the dataset across the temporal dimension, thereby reducing the influence of transient extreme events or fluctuations in topic popularity during specific timeframes. This strategy aims to enable the final dataset to more comprehensively reflect the content and evolving trends of discussions pertaining to climate change across different temporal stages. Ultimately, this process yielded an analytical dataset comprising 12,200 Twitter data points and 10,900 Reddit data points.

[1]https://ai.google.dev/gemma

[2]https://ai.meta.com/llama/license/

[3]https://www.anthropic.com/claude

[4]https://ollama.com/

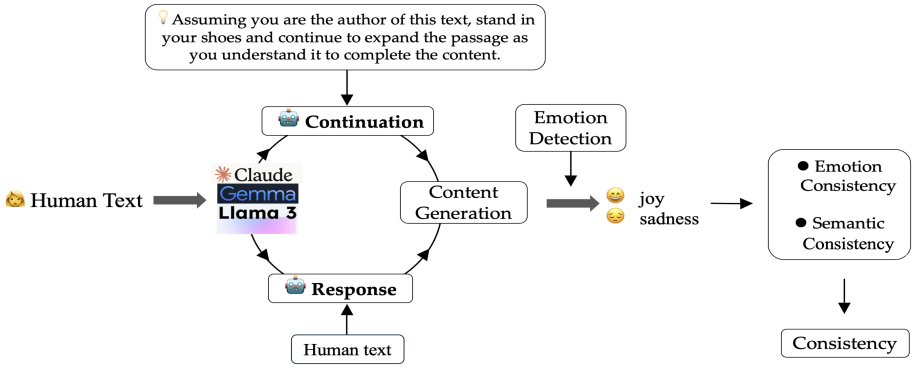
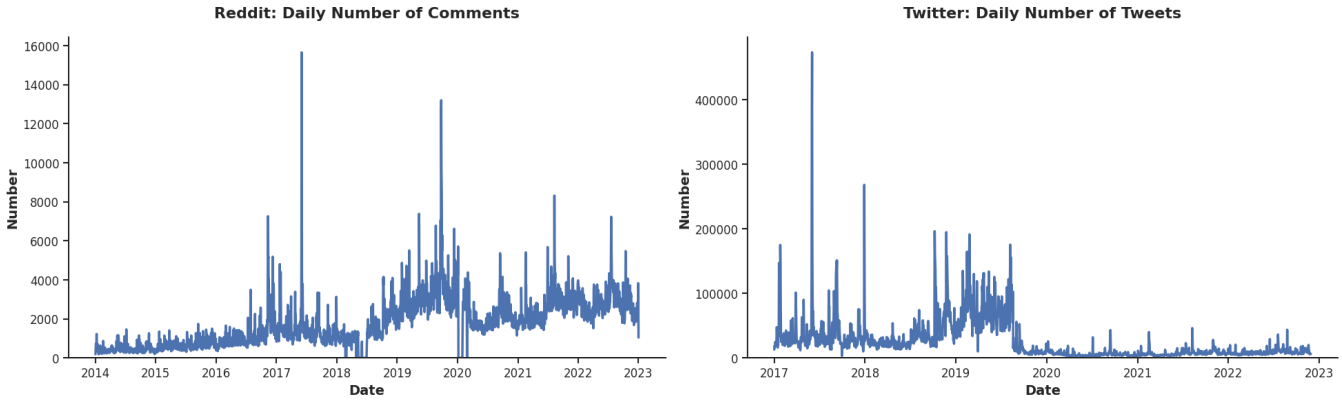


Figure 1: Experimental pipeline of consistency evaluation for LLMs. Our experimental framework starts with inputting human text to four LLMs, which complete two tasks: continuation and response.In the continuation task, models are prompted to extend the text as the original author; in the response task, no explicit prompt is used for natural interaction.After generation, we conduct emotion detection and further analysis.The framework ends with parallel analyses of emotional content and semantic consistency to evaluate how consistent LLM outputs are with the original human input.



a b

Figure 2: Daily data amount of Twitter and Reddit. **a.** Daily comments count of Reddit. **b.** Daily tweets count of Twitter. The x-axis represents the date, and the y-axis represents the frequency.

Emotion Labeling

In this study, we developed a methodology to analyze emotions in cross-platform social media data using a deep neural network-based model. We employed the*j-hartmann/emotion-english-distilroberta-base*[5] model from Hugging Face to examine the emotional content of both original texts and content generated by large language models. This model, built upon the RoBERTa-base architecture, is a fine-tuned checkpoint of "DistilRoBERTa-base" the datasets contain emotion labels for texts from Twitter, Reddit, student self-reports, and utterances from TV dialogues.

The *emotion-english-distilroberta-base* model identifies seven distinct emotion categories:joy, surprise, neutral, anger, disgust, fear and sadness. The model outputs probability scores for each category, which serve as quantitative measures of emotional content for subsequent analysis. While previous studies used sentiment analysis (negative, positive, and neutral) for evaluating differences between human and LLMs-generated text (Guo et al. 2023), our emotion-based approach provides a more granular and nuanced understanding of the underlying emotional states in posts.

Semantic Consistency

We design a evaluation framework that uses a rubric-guided expert-like LLM to score consistency and faithfulness. Here we hired the model:gemini-2.0-flash-thinking-exp-1219 as our expert LLM. This allows us to directly assess whether a model-generated response is logically aligned with the user's comment, and please see the appendix for specific prompt design1. After obtaining the evaluation results of the LLM-as-judge, the results are reviewed by a human evaluator. Human evaluator judges whether the evaluation results of the LLM are reasonable and accurate. In the end, the consistency rate between the two is maintained at 75%.

[5]https://huggingface.co/j-hartmann/emotion-english-distilroberta-base

Automatic evaluations of instruction following abilities in LLMs has recently received significant attention (Zheng et al. 2023). Given the significant time and cognitive effort required for human evaluation of large-scale generated content, we adopted a stratified random sampling approach to ensure representativeness across different model-platform-task combinations. Our full dataset consists of different sub-corpora, each corresponding to a specific combination of model (e.g., Gemma, Llama 3, Llama 3.3, Claude), platform (Twitter, Reddit), and task type (response or continuation), with approximately 10,000 instances per sub-corpus. To balance coverage and feasibility, we randomly sampled 30 instances from each sub-corpus, resulting in a total of 480 samples for evaluation.

Manual expert evaluation is costly. This hybrid evaluation strategy combines manual judgment with the assistance of a LLM. Although the sample size is small, it covers diversity across platforms, tasks, and models. We acknowledge in the limitations section that this small sample size may affect the generalizability of our findings.

Results

Emotion Dynamics of the Original Text in Downstream Tasks

In this study, we examined the emotional transitions between human-generated text and LLM outputs in downstream tasks. We categorize 7 emotions as positive emotionally oriented and negatively oriented as well as neutral emotions as defined by itself, as follows (Robinson 2008): Positive emotions: anticipation, joy, love, optimism, surprise, trust (Vaillant 2008); Negative emotions: anger, disgust, fear, pessimism, sadness

Analysis of Figure 3a reveals that in Gemma's continuation tasks, 34% of the texts initially labeled as 'anger' maintained their emotional, and 38% of the texts originally categorized as 'fear' also preserved the 'fear' emotion in the continuation task. For other emotion categories, including 'disgust,' 'sadness,' 'neutral,' 'joy,' and 'surprise,' the proportions of texts that retained the same emotional label were 22%, 13%, 40%, 10%, and 1%, respectively, demonstrating a capability for emotion preservation.

It is noteworthy that for all original emotion labels, a considerable proportion of the texts shifted to a neutral label in the continuation task. Concurrently, a significant portion of texts also transitioned to the 'anger' label during the continuation. These findings suggest that the Gemma model possesses a certain ability to recognize and perpetuate the original emotions. In subsequent tasks, it exhibits a systematic tendency to convert various emotional expressions into a neutral sentiment. The results in Figure 3a also, to some extent, indicate the model's sensitivity to negative emotions in the continuation task, particularly anger.

In analyzing the emotional shifts within Gemma's response tasks, we observed significant changes in emotional conversion. For texts with originally positive emotions such as 'joy' and 'surprise,' as well as those with negative emotions like 'anger,' 'disgust,' 'fear,' and 'sadness,' over 50% of their emotion labels were converted to neutral in the response task. Our analysis reveals a systematic bias in Gemma towards rational, neutral sentiment in these response tasks. Notably, a portion of the original emotions still transformed into 'anger' in the responses; for instance, 'disgust' showed a 21% conversion rate to 'anger,' while 'surprise' had a 14% conversion rate. This suggests Gemma's sustained sensitivity to anger-related content across both response and continuation tasks.

Figure 3c and Figure 3d illustrate the performance of Llama 3, showing that it has the same ability to recognize and preserve emotions in continuation and response tasks, and is able to maintain the original emotional valence more consistently. Similarly, we see a neutral emotion bias in the Llama3 model.

In the response and continuation tasks of the commercial model Claude, we observed that more original text emotions were more neutral in downstream tasks. At the same time, in the same model family, we observed that the Llama 3.3 model was also more inclined to generate neutral content in response and continuation than Llama3.

Beyond Reddit, our analysis also included Twitter data, which revealed distinct discourse patterns surrounding climate change topic. Figure 4 illustrates the emotion shift between the original Twitter content and the replies generated by the LLM. The emotion performance patterns of the Gemma, Llama and Claude models on Twitter texts in continuation and response tasks yielded two key insights: Firstly, Gemma demonstrated heightened sensitivity to content related to anger; secondly, for models exhibited a systematic bias towards neutral sentiment when operating in interactive conversational contexts. third, both Claude and Llama3.3 models tend to generate content with neutral emotions. In addition, it is worth noting that in the continuation task of Llama3.3, its ability to continue emotions is better than Llama3.

Resources of LLMs' Generated Content Emotions

We analyzed the emotional sources of LLM-generated content by examining the relationship between input and output emotions. As shown in Figure 5a, the results show that Gemma did not simply copy or continue the original emotion of the input text, but perform significant emotion modulation. A core finding is that the model has a general tendency to moderate negative initial emotion. Specifically, when the original text carries negative emotion (as shown by the blue bar in the figure 5, representing anger, disgust, fear, or sadness), a considerable proportion of the generated content emotion will turn into neutral (source shown by the green bar) or even positive emotion (source shown by the red bar, representing joy or surprise). For example, in the continuation and response tasks of multiple models (such as Gemma, Llama, Claude), as well as on the Reddit and Twitter datasets, it can be observed that the generated text with "neutral" or "joy" emotion has a significant negative emotion contribution in its original emotion input. In addition, the model also shows a general trend towards neutral emotion, that is, the generated neutral emotion text comes from a wide range of original positive, negative, and neutral inputs. Even if the original input is positive, the generated content has a certain probability of being adjusted to neutral. These emotion conversion patterns

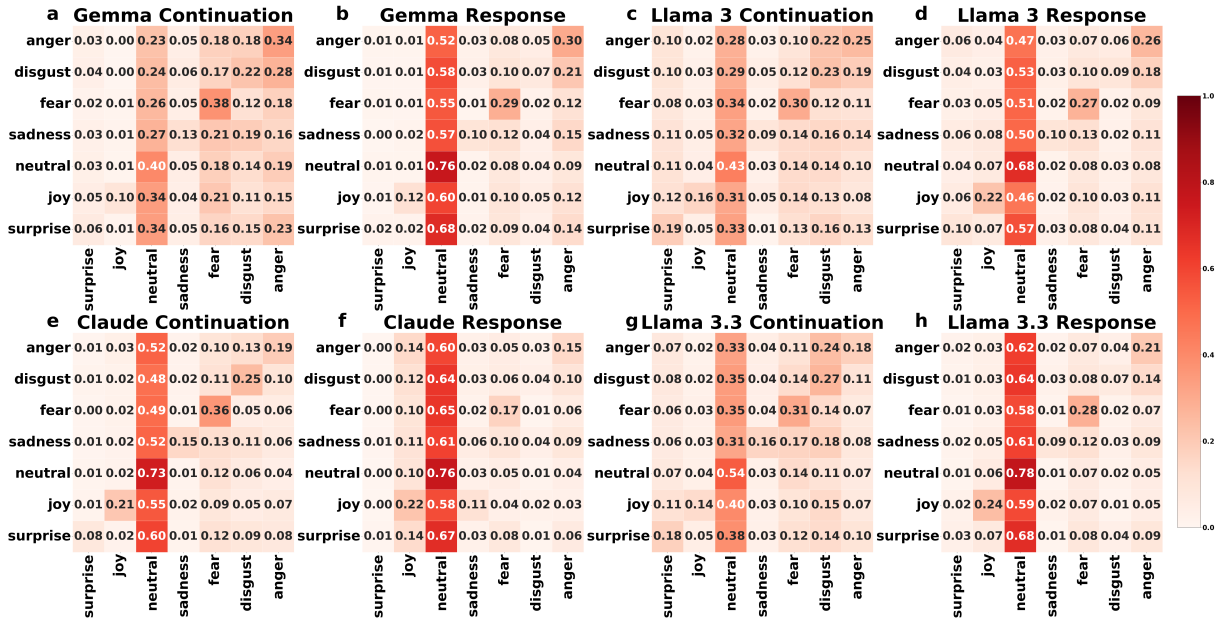


Figure 3: Emotional Transition Analysis of LLM Response and Continuation Tasks in Reddit Comments. Panels **a**–**h** show emotional transitions in content generated by Gemma, Llama, and Claude in continuation and response tasks. The y-axis represents human source emotions, and the x-axis shows emotions in LLM outputs. Cell values indicate the proportion of emotional transitions between original and generated text. For instance, in Figure **3a**, the value 0.34 in the anger-to-anger cell means 34% of originally angry inputs retained anger in Gemma's continuation outputs. Darker cell shading corresponds to higher emotional transition frequencies.

show high consistency between different LLMs, task types (text continuation and response), and data sources (Reddit and Twitter), suggesting that this may be an inherent characteristic or common training result of current mainstream LLMs in emotion processing.

Comparative Analysis of Emotional Intensity between LLMs and Human Text

We analyzed the differences in emotional intensity between LLM-generated and human-authored content, focusing on whether LLMs exhibit higher or lower emotional intensity. Emotional content was quantified using a probabilistic model assigning normalized scores (0 to 1) to each emotional category, interpreted as intensity scores (Miyazaki et al. 2024). These scores were grouped into five intensity levels. Statistical analysis included ANOVA to compare group differences and Tukey's post-hoc test for significant pairwise variations(Elnaggar, Mohamed, and Gehan 2024).

Analysis of variance (ANOVA) results presented in Table 1 indicate statistically significant differences ($P < 0.01$) across the nine groups in emotional intensity values for anger, disgust, fear, sadness, joy, surprise and neutral on Twitter. Similar significant variations were also observed in the Reddit dataset for the seven emotions mentioned above. Tukey's post-hoc analysis identified several significant differences across three comparison categories: within-model, between-model, and model-to-human comparisons.

Each emotion score represents its degree of emotion, and the higher the score, the stronger the emotion. As shown in

Platform	Emotions	F statistic	P value
Reddit	anger	114.2381	<0.001
	joy	80.899	<0.001
	disgust	29.8564	<0.001
	surprise	24.9967	<0.001
	fear	17.3482	<0.001
	sadness	5.5149	<0.001
	neutral	325.2502	<0.001
Twitter	anger	202.1090	<0.001
	joy	34.4011	<0.001
	disgust	22.4241	<0.001
	surprise	51.5744	<0.001
	fear	75.0905	<0.001
	sadness	9.2834	<0.001
	neutral	552.6669	<0.001

Table 1: ANOVA test of different model emotions

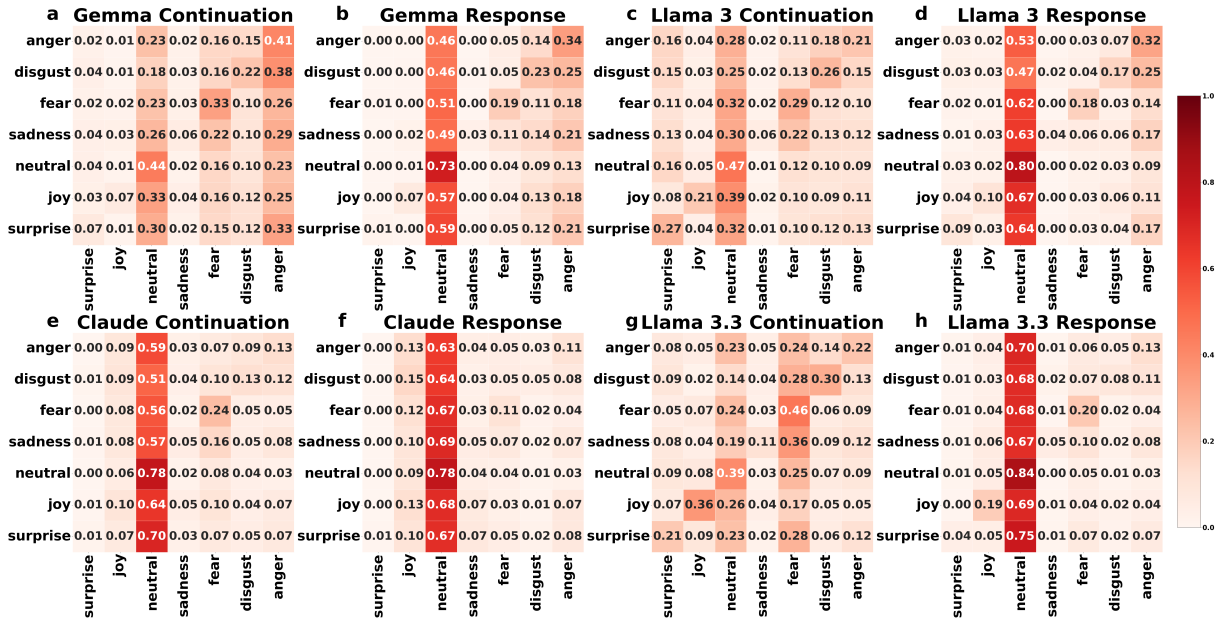


Figure 4: Emotional Transition Analysis of LLM Response and Continuation Tasks in Twitter Comments. Panels **a**–**h** illustrate emotional transitions in content generated by Gemma, Llama, and Claude during Twitter continuation and response tasks. The y-axis shows original human emotions, and the x-axis shows emotions in LLM outputs. Each cell value represents the emotional transition proportion, with darker red shades indicating higher frequencies.

the table2, intra-model comparisons indicate that emotional scores for certain emotions—such as anger and fear—are generally higher in continuation tasks than in response tasks. This suggests that LLMs tend to adopt a more neutral emotional tone when operating as conversational agents. Furthermore, when compared to the original human-authored texts,shown in the table3, most LLM-generated outputs exhibit lower emotional intensity, particularly in dimensions such as anger, fear, disgust, and joy. This may indicate that LLMs express emotions in a more restrained and subdued manner relative to humans.

Cross-model comparisons also reveal that emotional expression varies significantly depending on task type and data source. For instance, within the Llama model family, Llama 3 demonstrates higher emotional scores than Llama 3.3, suggesting a greater tendency toward heightened emotional expression.

These findings suggest two key insights: first, LLMs demonstrate systematic suppression of certain negative emotions, particularly in continuation tasks; second, the response task appears to operate under distinct generative mechanisms, resulting in differential emotional expression patterns. Furthermore, the consistent reduction in optimism across all LLM-generated texts relative to human-authored content indicates a systematic constraint in LLMs' capability to fully capture and convey positive emotional states.

Evaluating Semantic Consistency of LLM-Generated Content in Social Media Contexts

In LLM-human interactions, we analyzed models' ability to maintain topical coherence and contextual relevance using scores labeled by LLM-as-judge (Zheng et al. 2023) to measure semantic and logic alignment with the original content, providing a framework for evaluating semantic and logic fidelity across contexts.

Figure 6 plotted the frequency of each score (ranging from 1 to 5) separately for each task. This allows us to examine whether certain tasks tend to produce more consistent model outputs. The results suggest that while most tasks concentrate around higher scores, there are subtle differences in how each subtask is rated by the evaluator. We find that most models are capable of generating high-quality content in both response and continuation tasks, indicating that current LLMs possess strong capabilities in understanding and interacte with human language. However, subtle differences remain across models. For instance, the Llama 3.3 model demonstrates consistently strong performance in content generation across both types of datasets. Comparized with Llama 3' result, this may also indicate to some extent the impact of different model sizes on the quality of generated content. Additionally, model performance varies across platforms. For example, the Claude model shows higher fidelity and consistency in its outputs on the Reddit dataset, which may be attributed to the more complex linguistic environment of Twitter.

Discussion

This study utilized the Twitter and Reddit datasets on climate change to systematically evaluate three aspects of LLM-generated text versus original social media posts: (1) emotional consistency in both continuation and response tasks (and differences in different models), (2) emotional intensity across those tasks, and (3) logic similarity and context-

Emotion	Within Group[1]				Between Groups[2]								Platform
	W1	W2	W3	W4	B1	B2	B3	B4	B5	B6	B7	B8	
anger	$>^{***}$	$<^{***}$	$>^{***}$	-	$>^{**}$	-	$>^{***}$	-	$<^{***}$	$<^{*}$	$<^{***}$	$>^{***}$	
fear	$<^{***}$	-	$<^{**}$	$<^{***}$	$<^{***}$	$>^{***}$	-	$>^{*}$	-	-	-	-	
sadness	$>^{**}$	$<^{***}$	$<^{**}$	-	$>^{***}$	-	$>^{***}$	-	$<^{**}$	$>^{***}$	$>^{***}$	-	
joy	-	-	$>^{***}$	$>^{***}$	$>^{***}$	-	$>^{***}$	-	$>^{*}$	$<^{**}$	-	$>^{***}$	Reddit
surprise	-	$>^{**}$	-	-	$>^{***}$	$<^{**}$	-	-	$>^{***}$	$<^{*}$	-	-	
disgust	$<^{*}$	$<^{***}$	$<^{**}$	$<^{***}$	-	-	-	$>^{***}$	-	-	-	-	
neutral	$>^{***}$	$>^{***}$	$>^{**}$	$>^{***}$	-	$<^{***}$	$<^{***}$	$>^{***}$	$<^{***}$	$>^{*}$	$<^{***}$	$>^{***}$	
anger	-	-	$>^{***}$	$>^{***}$	$>^{***}$	-	$<^{***}$	-	$>^{*}$	$>^{**}$	-	$>^{***}$	
fear	$>^{***}$	$>^{***}$	$>^{**}$	$>^{***}$	-	$>^{***}$	$>^{***}$	$>^{***}$	$<^{***}$	$<^{***}$	$>^{***}$	$>^{***}$	
sadness	$<^{***}$	-	$<^{*}$	$<^{***}$	$<^{***}$	$<^{***}$	-	$>^{*}$	-	-	-	-	
joy	-	$>^{***}$	-	-	$>^{***}$	$>^{**}$	-	-	$>^{***}$	$>^{*}$	-	-	Twitter
surprise	$>^{*}$	$<^{***}$	$<^{**}$	-	$>^{***}$	-	$<^{***}$	-	$<^{**}$	$<^{***}$	$<^{***}$	-	
disgust	$>^{***}$	$<^{***}$	$>^{***}$	-	$>^{**}$	-	$<^{***}$	-	$<^{***}$	$>^{*}$	$>^{***}$	$>^{***}$	
neutral	$<^{*}$	$<^{***}$	$<^{**}$	$<^{***}$	-	-	-	$>^{***}$	-	-	-	-	

Table 2: Tukey's post-hoc test of LLM-generated content emotion values

[1] Within-group comparisons: W1 = Gemma continuation vs response; W2 = Llama continuation vs response; W3 = Claude continuation vs response; W4 = Llama3.3 continuation vs response.

[2] Between-group comparisons:B1 = Gemma continuation vs Llama continuation; B2 = Gemma continuation vs Claude continuation; B3 = Llama continuation vs Claude continuation; B4 = Llama continuation vs Llama3.3 continuation; B5 = Gemma response vs Llama response; B6 = Gemma response vs Claude response;B7 = Llama response vs Claude response; B8 = Llama response vs Llama3.3 response.

[3] The stars indicate significance levels: *** for $p < 0.001$, ** for $p < 0.01$, and * for $p < 0.05$.

[4] Symbols $<$ and $>$ indicate that the value in the previous column is lower or higher.

[5] "-" indicates no statistically significant difference.

awareness between model outputs and original social media texts. By combining these analyses, we aim to uncover the potential effects on emotional dynamics that LLMs have when implemented in online discussion.

In the continuation task, Gemma consistently transformed most emotions toward anger, showing a bias toward negative expression and intensifying anger. However, it maintained optimism and surprise, suggesting some ability to preserve emotional valence. On the other hand, Llama better preserved original emotions like anger, anticipation, fear, optimism, and sadness, with fewer emotional shifts, reflecting stronger emotional continuity (RQ1).

By combining the comparative analyses of emotional consistency and intensity, this study demonstrates that in both response and continuation tasks, LLMs systematically shift emotional content toward neutrality and lower its intensity across most emotional categories compared to original texts. Particularly, generated outputs display a higher prevalence and intensity of "neutral" manner, especially in response tasks rather than continuation tasks. This pattern shows that LLMs do have effect on the change of emotion when engaged in online discussion, and they're turning the discussion in a emotionally-moderate way.

As aforementioned, emotion plays a critical role in navigating social media discussion, especially in controversial topics like climate change. While the climate change discussion online is quite polarized (Treen et al. 2022b), such a tendency to neutrality could help defuse heated exchanges and foster calmer, less contentious dialogue. Hence, even though LLMs cannot keep the emotional consistency as human, its preferences toward moderation could benefit the highly polarized and emotionally debatable topics online, and lead the discussion more neutrally.

However, beyond this "neutral" tendency, what should be aware is that the negative emotions like "anger" and "fear" also showed increase in LLM-generated continuations and responses in most of the cases, which indicates that these models can amplify such emotional cues. Anger and fear are among the most salient emotional frames in climate discourse (Nabi, Gustafson, and Jensen 2018; Davidson and Kecinski 2022), and are significantly associated with climate activism, engagement, and has great effect on people's attitudes and behaviors (Stanley et al. 2021; Gregersen, Andersen, and Tvinnereim 2023). An uptick in fear and anger in LLM outputs may therefore shape users' perceptions of climate change. Considering the echo chamber effect on social media, LLMs' propensity of fear and anger could increase people's perceived threat by climate change, which in turn leads to higher awareness and action intentions towards climate crisis.

In terms of semantic consistency, the results of LLM-as-

Platform	Emotion	C1	C2	C3	C4	C5	C6	C7	C8
Reddit	anger	$<^{***}$	$<^{***}$	$<^{***}$	$<^{***}$	$<^{***}$	$<^{***}$	$<^{***}$	$<^{***}$
	fear	$<^{***}$	-	-	-	-	-	$<^{***}$	-
	sadness	-	-	-	-	$>^{*}$	$>^{***}$	-	-
	joy	-	-	$<^{***}$	-	-	$<^{**}$	$<^{**}$	$<^{***}$
	surprise	-	-	$<^{***}$	$<^{***}$	-	-	$<^{***}$	-
	disgust	$>^{***}$	$>^{***}$	$>^{***}$	$>^{***}$	$>^{**}$	$>^{***}$	-	$>^{***}$
	neutral	$>^{***}$	$<^{***}$	$>^{***}$	$<^{**}$	$<^{***}$	$<^{***}$	-	$<^{***}$
Twitter	anger	-	-	$<^{***}$	-	-	$<^{**}$	$<^{**}$	$<^{***}$
	fear	$>^{***}$	$<^{***}$	$>^{***}$	$<^{**}$	$<^{***}$	$<^{***}$	-	$<^{***}$
	sadness	$<^{***}$	-	-	-	-	-	$<^{***}$	-
	joy	-	-	$<^{***}$	$<^{***}$	-	-	$<^{***}$	-
	surprise	-	-	-	-	$>^{*}$	$>^{***}$	-	-
	disgust	$<^{***}$	$<^{***}$	$<^{***}$	$<^{***}$	$<^{***}$	$<^{***}$	$<^{***}$	$<^{***}$
	neutral	$>^{***}$	$>^{***}$	$>^{***}$	$>^{***}$	$>^{**}$	$>^{***}$	-	$>^{***}$

Table 3: Tukey's post-hoc test of emotion values between LLM-generated content and human text

Column definitions:

C1: Gemma continuation vs original; C2: Gemma response vs original; C3: Llama continuation vs original; C4: Llama response vs original; C5: Claude continuation vs original; C6: Claude response vs original; C7: Llama3.3 continuation vs original; C8: Llama3.3 response vs original

Notes:

- The stars indicate significance levels: *** for $p < 0.001$, ** for $p < 0.01$, and * for $p < 0.05$.
- Symbols $<$ and $>$ indicate that the value in the previous item is lower or higher.
- "–" indicates no statistically significant difference.

judge (Figure 6) show that LLMs generated text could generally keep the consistency to social media texts. This comparative findings proves a critical evidence from the perspective of emotion that LLMs could generally "understand" the context of a social media post and produce continuations or replies that align both thematically and logically with the source content. However, the consistency varies due to different types of LLMs, while most of the scores are high, some specific model like Claude's performances are dispersed, featuring a nontrivial tail into other scores. It indicates that some models may be more sensitive to prompt phrasing or exhibit less robust world-knowledge integration. The findings from semantic consistency imply that LLMs nowadays can generally preserve semantic and logic coherence in social-media dialogue, which makes the LLMs-generated content hard to distinguish on social media.

Taken together, the emotional and semantic consistency examined in this study could be seen as a technological improvement of context-awareness in LLMs' text generation capabilities. However, it could also raise practical concerns to general social media users for the risks of LLMs being manipulated for some malicious or intentional purposes, such as misinformation.

Effectively managing public emotions during controversial discussion online is crucial for governance of public opinion. Recent advances in Artificial Intelligence Generated Content (AIGC), driven by Generative AI (GAI) technology, have garnered attention beyond computer science (Cao et al. 2023). Given the increasing integration of LLMs into daily life, their emotional characteristics significantly influence opinion leadership, as emotional content shapes public perception and discourse framing.

Future implications suggest that while LLMs generate semantically coherent content, there is potential to improve alignment with nuanced human contexts. Research should focus on refining LLMs' understanding of implicit meaning and contextual subtleties to enhance user experience and broaden application domains.

Limitation and Future Work

This study improves our understanding of emotional dynamics in human-AI interaction, yet it also has limitations that warrant future research.

First, our experiments only used Reddit and Twitter data. Other platforms such as YouTube, Instagram, and TikTok have distinct user behaviors and content structures (Hilde A. M. Voorveld and Bronner 2018).

Second, the underlying reasons for emotional inconsistency in LLMs need further investigation. We found that emotions differ across models and tasks: neutral emotion was dominant but varied between continuation and response tasks and across models; the overall proportion of anger increased in all cases; and different models showed distinct emotional patterns. However, since the training data and in-

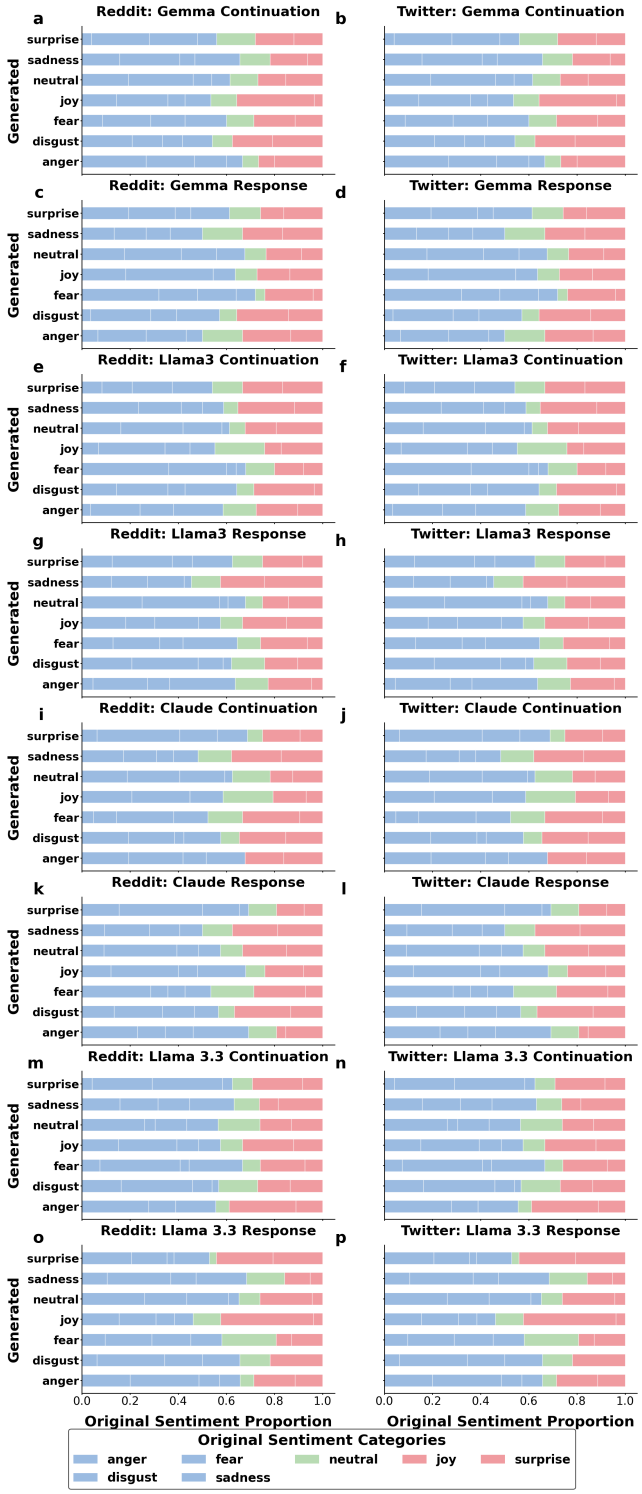


Figure 5: Emotional source analysis of LLM-generated content across platforms. Panels a–n show emotional transitions in Gemma, Llama, and Claude models' continuation/response tasks on Reddit and Twitter data. Red bars denote positive original emotions, blue bars negative emotions, and green bars neutral emotions. The y-axis lists emotional categories in original and generated content.

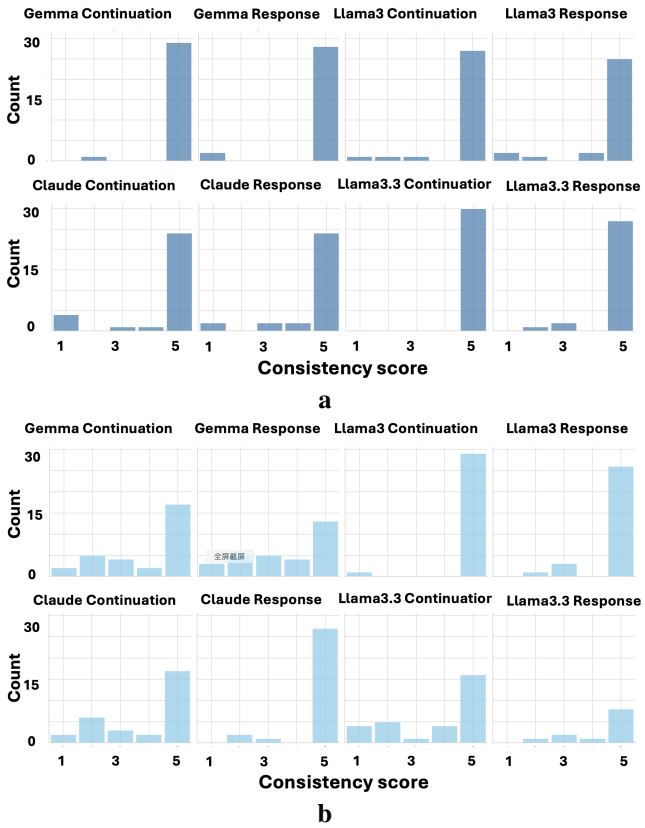


Figure 6: Stacked histograms in panels **a** (Reddit) and **b** (Twitter) show the distribution of faithfulness and consistency scores for continuations and responses generated by Gemma, Llama, and Claude.The x-axis represents rating categories: 5 = Excellent, 4 = Good, 3 = Fair, 2 = Poor, 1 = Very Poor.

ternal weights of these models are not public, we cannot determine whether these differences arise from pretraining data biases, fine-tuning strategies, or contextual factors. Future work could conduct interpretability analyses, controlled comparisons, and replications on open datasets to explore the mechanisms behind these emotional changes.

Third, our study may face potential data contamination by using public Twitter and Reddit data, which could overlap with LLMs' pre-training corpora. This may lead model outputs to reflect memorized patterns rather than genuine reasoning, limiting result generalizability.While we cannot fully quantify this issue, we encourage future work to mitigate contamination by using data after model cutoff dates or clearly outside training sets.Additionally, social media data may contain automated posts and bot interactions that bias results.Future research should adopt stricter data validation to improve dataset quality and reduce confounding factors.

Fourth, since human expert evaluation is expensive, we adopt a hybrid evaluation strategy that combines human judgment with LLM assistance based on a small but diverse sample across platforms, tasks, and models. While this approach achieves high-quality and interpretable scores, we acknowledge that the limited sample size may not capture the full

distribution of consensus behaviors across the datasets. Future work could extend this approach, such as scalable pure LLM evaluation and automatic divergence detection for selective human review.

Conclusion

In this study, we evaluated how large language models process emotional and semantic content in social media. We compared four models—Gemma, Llama3, Claude, and Llama3.3—in continuation and response tasks, and identified distinct patterns in their emotional expression and contextual understanding.

Across both tasks, LLMs tended to shift emotions toward neutrality and reduce emotional intensity, especially in response tasks, indicating a preference for moderate language. LLM-as-judge results showed that generated content generally maintained thematic and logical consistency with the original text. While most models performed stably, variations in models such as Claude reflected sensitivity to prompt wording and robustness differences. The neutrality bias of LLMs may help reduce polarization and promote rational dialogue, while the amplification of anger and fear brings both potential benefits (e.g., social mobilization) and risks (e.g., manipulation and echo chambers).

These findings clarify key emotional dynamics of LLMs and provide implications for designing and applying LLMs in emotion-sensitive scenarios.

Code and Data Availability

All code and materials are available on GitHub: https://github.com/Lena-Van/LLMs-emotion-and-semanticy

Acknowledgments

We are grateful to Dr. Chenyang Wang of USTC for fruitful discussions.

References

Ayoub, N. F.; Balakrishnan, K.; Ayoub, M. S.; Barrett, T. F.; David, A. P.; and Gray, S. T. 2024. Inherent Bias in Large Language Models: A Random Sampling Analysis. *Mayo Clinic Proceedings: Digital Health*, 2(2): 186–191.

Baumgartner, J.; Zannettou, S.; Keegan, B.; Squire, M.; and Blackburn, J. 2020. The Pushshift Reddit Dataset. *CoRR*, abs/2001.08435.

Boyd, D. 2016. Untangling research and practice: What Facebook's "emotional contagion" study teaches us. *Research Ethics*, 12(1): 4–13.

Brady, W. J.; Wills, J. A.; Jost, J. T.; Tucker, J. A.; and Van Bavel, J. J. 2017. Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, 114(28): 7313–7318.

Cao, Y.; Li, S.; Liu, Y.; Yan, Z.; Dai, Y.; Yu, P. S.; and Sun, L. 2023. A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT. arXiv:2303.04226.

Chandrasekaran, D.; and Mago, V. 2021. Evolution of semantic similarity—a survey. *ACM Computing Surveys (CSUR)*, 54(2): 1–37.

Chu, M.; Song, W.; Zhao, Z.; Chen, T.; and Chiang, Y.-c. 2024. Emotional contagion on social media and the simulation of intervention strategies after a disaster event: a modeling study. *Humanities and Social Sciences Communications*, 11(1): 1–15.

Dam, S. K.; Hong, C. S.; Qiao, Y.; and Zhang, C. 2024. A complete survey on llm-based ai chatbots. *arXiv preprint arXiv:2406.16937*.

Davidson, D. J.; and Kecinski, M. 2022. Emotional pathways to climate change responses. *Wiley Interdisciplinary Reviews: Climate Change*, 13(2): e751.

Elnaggar, M.; Mohamed, K.; and Gehan, S. 2024. Effectiveness of Gamified Cooperation and Competition Strategies in a Blended Learning Environment for Developing EFL Business Writing Skills for TVET Learners. *European Scientific Journal*, 30: 107–128.

Ferrara, E.; and Yang, Z. 2015. Measuring emotional contagion in social media. *PloS one*, 10(11): e0142390.

Gallegos, I. O.; Rossi, R. A.; Barrow, J.; Tanjim, M. M.; Kim, S.; Dernoncourt, F.; Yu, T.; Zhang, R.; and Ahmed, N. K. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 1–79.

Gao, C.; Lan, X.; Li, N.; Yuan, Y.; Ding, J.; Zhou, Z.; Xu, F.; and Li, Y. 2024a. Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications*, 11(1): 1–24.

Gao, M.; Hu, X.; Ruan, J.; Pu, X.; and Wan, X. 2024b. Llm-based nlg evaluation: Current status and challenges. *arXiv preprint arXiv:2402.01383*.

Gregersen, T.; Andersen, G.; and Tvinnereim, E. 2023. The strength and content of climate anger. *Global Environmental Change*, 82: 102738.

Guo, B.; Zhang, X.; Wang, Z.; Jiang, M.; Nie, J.; Ding, Y.; Yue, J.; and Wu, Y. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*.

Hamby, A.; and Jones, N. 2022. The effect of affect: An appraisal theory perspective on emotional engagement in narrative persuasion. *Journal of Advertising*, 51(1): 116–131.

Herbold, S.; Hautli-Janisz, A.; Heuer, U.; Kikteva, Z.; and Trautsch, A. 2023. A large-scale comparison of human-written versus ChatGPT-generated essays. *Scientific reports*, 13(1): 18617.

Hilde A. M. Voorveld, D. G. M., Guda van Noort; and Bronner, F. 2018. Engagement with Social Media and Social Media Advertising: The Differentiating Role of Platform Type. *Journal of Advertising*, 47(1): 38–54.

Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*.

Kramer, A. D.; Guillory, J. E.; and Hancock, J. T. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24): 8788–8790.

Li, J.; Tang, T.; Zhao, W. X.; Nie, J.-Y.; and Wen, J.-R. 2024. Pre-trained language models for text generation: A survey. *ACM Computing Surveys*, 56(9): 1–39.

Li, Y.; Zhang, Y.; and Sun, L. 2023. Metaagents: Simulating interactions of human behaviors for llm-based task-oriented coordination via collaborative generative agents. *arXiv preprint arXiv:2310.06500*.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Liu, J.; Jiang, B.; and Wei, Y. 2025. LLMs as Promising Personalized Teaching Assistants: How Do They Ease Teaching Work? *ECNU Review of Education*, 20965311241305138.

Lu, D.; and Hong, D. 2022. Emotional contagion: Research on the influencing factors of social media users' negative emotional communication during the COVID-19 pandemic. *Frontiers in psychology*, 13: 931835.

Martel, C.; Pennycook, G.; and Rand, D. G. 2020. Reliance on emotion promotes belief in fake news. *Cognitive research: principles and implications*, 5: 1–20.

Miyazaki, K.; Uchiba, T.; Kwak, H.; An, J.; and Sasahara, K. 2024. The impact of toxic trolling comments on anti-vaccine YouTube videos. *Scientific Reports*, 14(1): 5088.

Muñoz-Ortiz, A.; Gómez-Rodríguez, C.; and Vilares, D. 2024. Contrasting linguistic patterns in human and llm-generated news text. *Artificial Intelligence Review*, 57(10): 265.

Nabi, R. L.; Gustafson, A.; and Jensen, R. 2018. Framing climate change: Exploring the role of emotion in generating advocacy behavior. *Science Communication*, 40(4): 442–468.

Naskar, D.; Singh, S. R.; Kumar, D.; Nandi, S.; and Rivaherrera, E. O. d. l. 2020. Emotion dynamics of public opinions on twitter. *ACM Transactions on Information Systems (TOIS)*, 38(2): 1–24.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Park, P. S.; Goldstein, S.; O'Gara, A.; Chen, M.; and Hendrycks, D. 2024. AI deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(5).

Parsa, M. S.; Shi, H.; Xu, Y.; Yim, A.; Yin, Y.; and Golab, L. 2022. Analyzing Climate Change Discussions on Reddit. In *2022 International Conference on Computational Science and Computational Intelligence (CSCI)*, 826–832.

Que, H.; Duan, F.; He, L.; Mou, Y.; Zhou, W.; Liu, J.; Rong, W.; Wang, Z. M.; Yang, J.; Zhang, G.; et al. 2024. Hellobench: Evaluating long text generation capabilities of large language models. *arXiv preprint arXiv:2409.16191*.

Raj, H.; Gupta, V.; Rosati, D.; and Majumdar, S. 2023. Semantic consistency for assuring reliability of large language models. *arXiv preprint arXiv:2308.09138*.

Robinson, D. L. 2008. Brain function, emotional experience and personality. *Netherlands Journal of Psychology*, 64: 152–168.

Ruan, T.; Kong, Q.; McBride, S. K.; Sethjiwala, A.; and Lv, Q. 2022. Cross-platform analysis of public responses to the 2019 Ridgecrest earthquake sequence on Twitter and Reddit. *Scientific reports*, 12(1): 1634.

Russell, S. J.; and Norvig, P. 2016. *Artificial intelligence: a modern approach*. Pearson.

Sai, A. B.; Mohankumar, A. K.; and Khapra, M. M. 2022. A survey of evaluation metrics used for NLG systems. *ACM Computing Surveys (CSUR)*, 55(2): 1–39.

Sharma, P. R.; Wade, K. A.; and Jobson, L. 2023. A systematic review of the relationship between emotion and susceptibility to misinformation. *Memory*, 31(1): 1–21.

Stanley, S. K.; Hogg, T. L.; Leviston, Z.; and Walker, I. 2021. From anger to action: Differential impacts of eco-anxiety, eco-depression, and eco-anger on climate action and wellbeing. *The Journal of Climate Change and Health*, 1: 100003.

Stieglitz, S.; and Dang-Xuan, L. 2013. Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior. *Journal of management information systems*, 29(4): 217–248.

Sung, M.; Lee, S.; Kim, J.; and Kim, S. 2024. Context-aware LLM translation system using conversation summarization and dialogue history. *arXiv preprint arXiv:2410.16775*.

Treen, K.; Williams, H.; O'Neill, S.; and and, T. G. C. 2022a. Discussion of Climate Change on Reddit: Polarized Discourse or Deliberative Debate? *Environmental Communication*, 16(5): 680–698.

Treen, K.; Williams, H.; O'Neill, S.; and Coan, T. G. 2022b. Discussion of climate change on Reddit: Polarized discourse or deliberative debate? *Environmental Communication*, 16(5): 680–698.

Treen, K. M. d.; Williams, H. T.; and O'Neill, S. J. 2020. Online misinformation about climate change. *Wiley Interdisciplinary Reviews: Climate Change*, 11(5): e665.

Vaillant, G. E. 2008. Positive emotions, spirituality and the practice of psychiatry. *Mens sana monographs*, 6(1): 48.

van Schaik, T. A.; and Pugh, B. 2024. A field guide to automatic evaluation of llm-generated summaries. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2832–2836.

Whitmarsh, L. 2011. Scepticism and uncertainty about climate change: Dimensions, determinants and change over time. *Global environmental change*, 21(2): 690–700.

Yang, J.; Chen, D.; Sun, Y.; Li, R.; Feng, Z.; and Peng, W. 2024. Enhancing semantic consistency of large language models through model editing: An interpretability-oriented approach. In *Findings of the Association for Computational Linguistics ACL 2024*, 3343–3353.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36: 46595–46623.

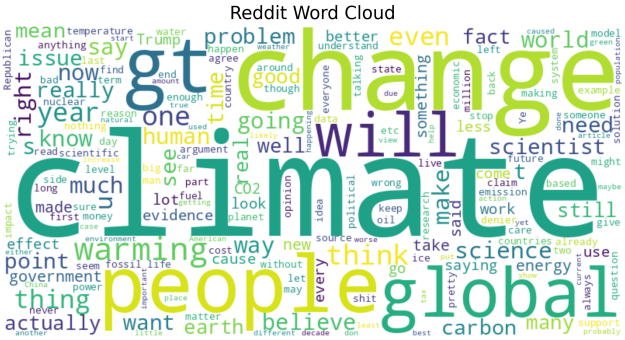
Ethics Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, our research question advance science without violating social contracts.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes, the abstract and introduction reflect the paper's contributions and scope
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, the aim of our study is about the emotional and semantic consistency of large language models, which we measure using sentiment scores and cosine similarity in our methodology.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, our research considered the population-specific distribution.
   (e) Did you describe the limitations of your work? Yes, we talked our limitations in the last part.
   (f) Did you discuss any potential negative societal impacts of your work? No.
   (g) Did you discuss any potential misuse of your work? No
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? Yes
   (b) Have you provided justifications for all theoretical results? Yes, all theoretical results have related justification.
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Yes
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Yes
   (e) Did you address potential biases or limitations in your theoretical framework? Yes
   (f) Have you related your theoretical results to the existing literature in social science? Yes
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? NA
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes, it's reflected in our code
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? NA
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes
   (b) Did you mention the license of the assets? Yes, our Twitter data was collected before X closed the data sharing channel and is open source data.Our Reddit data, on the other hand, is collected at https://the-eye.eu/redarcs/, and according to https://redditinc.com/blog/2023apiupdates, Reddit allows scientific researchers to access their data
   (c) Did you include any new assets in the supplemental material or as a URL? NA
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? NA
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? No
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? No
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? No
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? NA
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA

(d) Did you discuss how data is stored, shared, and deidentified? NA

Appendices



a

b

Figure 7: Word cloud about Reddit (panel **a**) and Twitter (panel **b**) platforms.

Listing 1: Prompt Function

```
def build_prompt(premise, hypothesis):
    return f"""You are an expert evaluator scoring the faithfulness and consistency of an LLM's response to a user's comment on a scale of 1 to 5.
**User Comment (Premise):**
"{premise}"
**LLM Response (Hypothesis):**
"{hypothesis}"

**Evaluation Task:**

Assign a score from 1 to 5 based on how faithful and consistent the LLM Content is to the User Comment. Faithfulness here means being relevant, factually accurate (or appropriately handling factual disagreements/user misinformation), and generating constructively to the user's input, even if disagreeing with their stance. Use the rubric below. Provide the score and a brief explanation justifying your score.

**Scoring Rubric:**

* **5 (Excellent):** Fully relevant, factually consistent. If user expresses skepticism or potential misinformation, the response addresses it constructively and accurately (e.g., by suggesting reliable sources) without being preachy or dismissive. Excellent conversational handling.
* **4 (Good):** Mostly relevant and factually consistent. Appropriately handles user stance/skepticism (like the example of suggesting sources), even if it implies disagreement. Minor awkwardness in phrasing allowed. Overall, a reasonable and faithful response.
* **3 (Fair):** Addresses the topic but might have minor factual inconsistencies OR handles user stance/skepticism in a less ideal way (e.g., slightly dismissive, ignores a key part of the comment). The connection is clear but flawed.
* **2 (Poor):** Partially relevant or contains significant factual contradictions/unsupported claims OR severely mishandles user stance (e.g., aggressive disagreement, irrelevant lecture). Weak connection to the user's comment.
* **1 (Very Poor):** Irrelevant, directly contradicts clear facts without basis, hallucinates, or is conversationally inappropriate in its handling of the user's comment.

**Output:**
Score (1-5): [Your Score]
Explanation: [Your brief justification referencing the rubric]"""
```