sentiment. This neutralization effect aligns with professional expectations in these fields, where objectivity is prioritized over emotional expression.

Notably, healthcare AI responses show increased sentiment amplification, particularly for negative prompts. The amplification may stem from the model's attempt to demonstrate empathy but inadvertently intensifies negative emotional framing.

Our findings suggest that LLMs are more likely to amplify sentiment in subjective domains (e.g., creative writing, journalism) while neutralizing it in objective fields (e.g., legal, finance, technical writing). This domain-dependent behavior has significant implications for the design and deployment of sentiment-aware AI systems.

3.2 Effects on Factual Accuracy

Table 1 presents factual accuracy scores (F) obtained from automated fact-checking (FEVER) and human annotations. Our analysis demonstrates a consistent relationship between prompt sentiment and factual accuracy across all evaluated domains.

Sentiment	Avg. Factual Score (F)	% Change vs. Neutral
Neutral	92.3%	—
Positive	89.7%	-2.8%
Negative	84.5%	-8.4%

Table 1: Factual accuracy scores across different sentiment categories

Prompts with negative sentiment result in the largest factual accuracy decline (∼8.4%). This effect likely stems from the model shifting toward speculative, exaggerated, or alarmist responses when processing negatively framed queries. The correlation between negative sentiment and reduced factuality presents a significant challenge for applications requiring high reliability.

Positive sentiment prompts lead to a smaller (∼2.8%) factuality reduction, often due to overly optimistic framings or subtle embellishments. While this effect is less pronounced than with negative prompts, it nevertheless indicates that any departure from neutral framing compromises factual integrity to some degree.

Across all tasks, models produce the most factually accurate responses when prompted with neutral language, suggesting that emotional content interferes with the model's capacity for precise information retrieval and reasoning. This sentiment bias in factual accuracy poses risks in high-stakes applications such as journalism, finance, and medical AI. Ensuring neutrality in prompts is therefore critical for reliable AI-generated content.

3.3 Effects on Bias

We measured sentiment-induced shifts in bias across LLM outputs, focusing on political framing, gender stereotypes, and racial bias. Our experiments reveal systematic relationships between input sentiment and output bias characteristics.

In news and policy discussions, neutral prompts (e.g., "Discuss economic policies in 2024") produce balanced outputs that present multiple perspectives. However, sentiment-driven prompts (e.g., "Why are 2024 policies failing?") lead to more negatively framed responses, while positive variants (e.g., "How have 2024 policies succeeded?") result in favorable portrayals. This sentiment-driven framing effect occurs despite the model having access to identical background knowledge in all cases.

Sentiment-driven bias amplification is most evident in sensitive topics, where negative prompts increase the likelihood of stereotypical or emotionally charged language. This finding highlights the need for careful prompt engineering in applications addressing socially sensitive issues.

3.4 Sentiment Drift in Long-Form Content

The length and verbosity of LLM-generated responses are influenced by prompt sentiment in ways that reflect human communication patterns. Our analysis of response length across sentiment categories reveals consistent structural effects.

On average, essays and blog-style responses from positive prompts are 8.1% longer than those generated from neutral prompts. This suggests a model preference for elaboration when engaging with affirming or enthusiastic input, mirroring human tendencies toward expansiveness in positive conversational contexts.

In contrast, responses to negative prompts are 17.6% shorter, often displaying signs of disengagement or terseness. This trend is particularly noticeable in long-form content generation, where models may truncate outputs in response to negative sentiment. The greater magnitude of length reduction for negative prompts compared to length increase for positive prompts suggests an asymmetric response to sentiment variation.

4 Discussion

The results presented in Sections 3.1–3.4 provide compelling evidence that prompt sentiment is a critical determinant of large language model (LLM) behavior. In this section, we elaborate on the theoretical, practical, and ethical implications of our findings while also outlining limitations and avenues for future work.

4.1 Theoretical Implications

Our empirical findings bolster theoretical models of affective computing by demonstrating that LLMs not only process semantic content but also internalize and propagate the affective tone of prompts. This behavior is in line with the notion that language models, as statistical approximators of human language, inherently mirror sentiment cues present in the input [8]. The observation that models exhibit domain-dependent sentiment amplification, strong in creative tasks and subdued in technical domains, suggests that underlying training regimes and fine-tuning procedures play a role in modulating affective responses [9, 10].

Furthermore, our results on factual degradation associated with negative sentiment lend support to recent studies that highlight the fragility of LLM outputs under varied prompt conditions [11, 12]. The tendency for negative sentiment to induce speculative or alarmist language underscores the delicate balance between creativity and factuality in neural text generation. These insights reinforce the importance of incorporating sentiment dynamics into existing models of reasoning and language generation, as well as in the design of robust prompt engineering strategies [13].

4.2 Practical Implications

The differential impact of prompt sentiment on output quality has significant practical ramifications across several application domains. For instance, in healthcare AI, the amplification of negative sentiment may inadvertently intensify patient anxiety, suggesting that prompt design in sensitive applications must account for emotional tone to ensure user safety [14]. In journalism and policy analysis, the observed bias shifts could lead to skewed reporting or unbalanced policy discussions if sentiment is not adequately controlled [15].

Our study further indicates that even subtle shifts in prompt sentiment can alter factual accuracy, which is critical for applications in legal and financial analysis where precision is paramount. This reinforces calls for the development of sentiment-aware prompt filtering and post-processing mechanisms that can mitigate unintentional bias and preserve output reliability [16, 17].

4.3 Ethical and Social Considerations

Beyond technical performance, the ethical implications of sentiment-induced bias are profound. As LLMs become increasingly integrated into decision-making systems, the risk of amplifying societal biases through emotionally charged prompts cannot be overlooked. Our findings highlight the need for a robust ethical framework that guides the deployment of sentiment-sensitive AI systems, ensuring that they align with shared human values and avoid propagating harmful stereotypes [12, 14].

This study also contributes to ongoing discussions about AI fairness by demonstrating that even seemingly benign variations in input language can lead to systematic shifts in bias. Such insights advocate for a holistic approach to AI evaluation, where performance metrics are supplemented with rigorous bias and fairness assessments [15, 16].

4.4 Limitations and Future Work

While our analysis offers significant insights, several limitations warrant discussion. First, our dataset of 500 prompts, although diverse, represents a fraction of the potential prompt space. Future work should consider larger, more varied datasets, including cross-lingual and multicultural contexts, to fully understand the global implications of sentiment dynamics in LLM outputs [10, 13].

Second, the automated fact-checking methods used in this study, while state-of-the-art, are not infallible. The integration of more robust, human-in-the-loop evaluations could further validate our findings on factual accuracy and bias.

Furthermore, exploring the interaction between prompt sentiment and other prompt engineering techniques, such as chain-of-thought prompting [11], may yield deeper insights into the interplay between reasoning and sentiment.

Finally, our work primarily focuses on static sentiment classifications. Future research should explore the adaptation of dynamic sentiments, where LLMs adjust their responses in real time based on evolving emotional contexts. Such adaptive mechanisms could pave the way for more resilient and context-aware AI systems, capable of mitigating adverse effects while capitalizing on the benefits of sentiment-aware interactions [9, 17].

5 Conclusion

This study reveals that prompt sentiment significantly influences LLM output quality, factual accuracy, and bias propagation. Our experiments provide empirical evidence that LLMs amplify prompt sentiment, with the most pronounced effects observed in subjective domains (e.g., journalism and creative writing) and a neutralizing effect in more objective fields (e.g., finance and legal analysis). Notably, negative sentiment prompts are associated with a substantial decrease in factual accuracy (approximately -8.4%), introducing risks for applications in news reporting, financial analysis, and healthcare.

The implications of our work extend beyond immediate technical concerns and touch on broader societal issues. As AI-generated content becomes ubiquitous, ensuring that outputs are both factually robust and ethically sound is paramount. Our findings advocate for an interdisciplinary approach that integrates insights from natural language processing, ethics, and human-computer interaction to design AI systems that are not only high-performing but also socially responsible [8, 10, 14].

Our key contributions include:

- Empirical proof that LLMs amplify prompt sentiment, with stronger effects in subjective domains compared to objective fields.
- Quantitative evidence demonstrating that negative prompt sentiment leads to a significant reduction in factual accuracy.
- Theoretical and practical recommendations for developing sentiment-aware prompt engineering techniques to achieve fair, reliable, and context-appropriate AI-generated content.

Looking ahead, we propose future research on cross-linguistic sentiment effects in LLMs, strategies to mitigate sentiment-induced biases, and real-time sentiment-aware model adaptations to ensure more robust factual responses. By addressing these challenges, future work can pave the way for AI systems that balance creative expression with precision and ethical integrity.

In summary, this conclusion underscores the necessity for continued exploration into the interplay between sentiment and AI output, setting the stage for the development of more nuanced, responsible, and effective large language models.

References

[1] Gu, J., Han, Z., Chen, S., Beirami, A., He, B., Zhang, G., Liao, R., Qin, Y., Tresp, V., and Torr, P. A systematic survey of prompt engineering on vision-language foundation models. *arXiv preprint arXiv:2307.12980*, 2023.

[2] Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., and Chadha, A. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*, 2024.

[3] Baumann, J. and Kramer, O. Evolutionary multi-objective optimization of large language model prompts for balancing sentiments. In *International Conference on the Applications of Evolutionary Computation (part of EvoStar)*, pages 212–224, 2024.

[4] Mao, R., Liu, Q., He, K., Li, W., and Cambria, E. The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection. In *IEEE Transactions on Affective Computing*, 14(3):1743–1753, 2022.

[5] Yu, Y., Zhang, D., and Li, S. Unified multi-modal pre-training for few-shot sentiment analysis with prompt-based learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 189–198, 2022.

[6] Gu, X., Chen, X., Lu, P., Li, Z., Du, Y., and Li, X. AGCVT-prompt for sentiment classification: Automatically generating chain of thought and verbalizer in prompt learning. In *Engineering Applications of Artificial Intelligence*, 132:107907, 2024.