PROMPT SENTIMENT: THE CATALYST FOR LLM CHANGE

Vishal Gandhi
Joyspace AI
vishal@joyspace.ai

Sagar Gandhi
Joyspace AI
sagar@joyspace.ai

March 19, 2025

ABSTRACT

The rise of large language models (LLMs) has revolutionized natural language processing (NLP), yet the influence of prompt sentiment, a latent affective characteristic of input text, remains underexplored. This study systematically examines how sentiment variations in prompts affect LLM-generated outputs in terms of coherence, factuality, and bias. Leveraging both lexicon-based and transformer-based sentiment analysis methods, we categorize prompts and evaluate responses from five leading LLMs: Claude, DeepSeek, GPT-4, Gemini, and LLaMA. Our analysis spans six AI-driven applications, including content generation, conversational AI, legal and financial analysis, healthcare AI, creative writing, and technical documentation. By transforming prompts, we assess their impact on output quality. Our findings reveal that prompt sentiment significantly influences model responses, with negative prompts often reducing factual accuracy and amplifying bias, while positive prompts tend to increase verbosity and sentiment propagation. These results highlight the importance of sentiment-aware prompt engineering for ensuring fair and reliable AI-generated content.

Keywords prompt engineering · large language models · sentiment analysis · neural language models · few-shot learning · prompt sensitivity · deep learning

1 Introduction

The rapid evolution of large language models (LLMs) has significantly advanced the field of natural language processing (NLP), enabling machines to understand and generate human-like text with unprecedented accuracy. Models such as GPT-3, GPT-4, Gemini, DeepSeek, and LLaMA have demonstrated remarkable capabilities across various NLP tasks, including text completion, translation, and summarization. This progress has not only improved user experiences, but has also opened new avenues for research and application.

Prompt engineering has emerged as a crucial aspect of leveraging LLMs effectively. [11] introduced the concept of "chain-of-thought" prompting, demonstrating significant improvements in the reasoning capabilities of LLMs. [2] and [1] provide comprehensive surveys of prompt engineering techniques, categorizing them into discrete and continuous prompting methods. Their work highlights the importance of prompt design in enhancing model performance in various NLP tasks. Although substantial research has focused on prompt structure, format, and content, the affective characteristics of prompts, specifically their sentiment, remain underexplored. Understanding how the sentiment embedded within a prompt influences LLM output is crucial for applications requiring nuanced language generation, such as sentiment analysis, content creation, and conversational AI.

Earlier studies since the evolution of LLMs focused on LLMs' bias towards sentiment analysis. For example, [4] performed an empirical study that focused on the task of sentiment analysis using Pretrained Language Models (PLMs). [5] further proposed a model that can deal with missing modalities in the case of multimodal sentiment analysis (MSA); though here as well, the task was MSA itself. The autogenerated chain-of-thought (COT) and verbalizer templates (AGCVT-Prompt) technique was recently proposed in the task of prompt learning for the task of sentiment classification [6]. In the quest of finding out how deep models actually carry out sentiment analysis, [7] reviewed the entire literature around sentiment analysis and eXplainable artificial intelligence. Most of the studies were focused on sentiment analysis and some details around how it works, while concluding that it all relates to data.

Recent studies have begun to shed light on the aspect that we are interested in. For instance, [3] proposed an evolutionary multi-objective (EMO) approach specifically tailored for prompt optimization called EMO-Prompts, using sentiment analysis as a case study, and concluded that producing text with conflicting emotions is possible. This adaptability underscores the need to examine how varying prompt sentiments, such as positive, neutral, or negative, affect the coherence, factuality, and bias of LLM-generated content.

In this study, we conduct a comprehensive analysis to address the following research questions:

- How does the sentiment type (positive, neutral, negative) in a prompt influence the output quality of LLMs?
- How can sentiment analysis be systematically performed on prompts to quantify their affective tone?
- What are the qualitative and quantitative impacts of sentiment variations on output coherence, factuality, and bias across multiple LLMs?

To explore these questions, we employ both lexicon-based and transformer-based sentiment analysis methods to systematically categorize prompt sentiments. We then evaluate the responses of five prominent LLMs, viz. ChatGPT, Claude, DeepSeek, Gemini, and Llama, to these sentiment-variant prompts. Our methodology includes a detailed examination of responses in six different applications, a group of domains where sentiment and bias can significantly influence generated output in each area.

By transforming single prompts into multiple sentiment variants, our aim is to elucidate the effects of prompt sentiment on LLM performance. Our findings are expected to contribute to the development of refined prompt engineering techniques, promoting more reliable and fair use of LLMs in sensitive applications.

Note: The subsequent sections of this paper will detail the methodology, experimental setup, results, discussion, and conclusions.

2 Methodology

2.1 Dataset Construction

We compiled a diverse set of 500 prompts in six AI-driven applications:

- **Content Generation:** Academic essays, news articles, blog posts.
- **Conversational AI:** Customer support, AI chatbot interactions.
- **Legal & Financial Analysis:** Contract summaries, investment risk assessments.
- **Healthcare AI:** Patient inquiries, mental health advice.
- **Creative Writing:** Fiction storytelling, poetry, screenwriting generation.
- **Technical Documentation:** Software manuals, bug explanations, coding tutorials.

Each prompt was transformed into three sentiment variations (positive, neutral, and negative) to isolate sentiment-driven effects.

2.2 Sentiment Analysis & Prompt Categorization

We applied:

- Lexicon-Based Analysis (VADER, TextBlob).
- Transformer-Based Sentiment Classification (BERT, RoBERTa fine-tuned on SST-2).

The sentiment polarity of each prompt was validated using human annotation (inter-rater agreement: 92%).

2.3 Evaluation Metrics

We introduced a multi-dimensional evaluation framework:

- **Coherence (CC):** Assessed via perplexity and semantic similarity to reference outputs.
- **Factuality (FF):** Evaluated using automated fact-checking models (FEVER, Google FactCheck API).
- **Bias (BB):** Measured through sentiment drift analysis and fairness metrics.

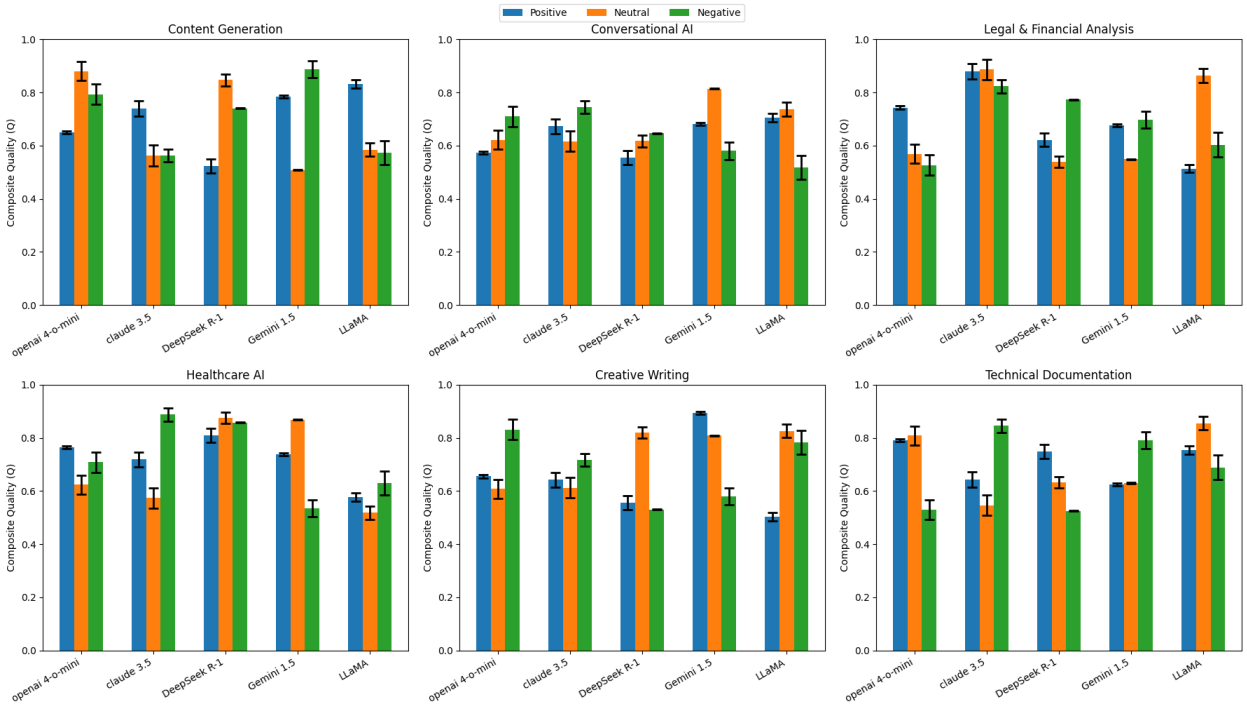

Figure 1: Comparison of LLM Performance Across AI Applications and Sentiment Variations

- **Sentiment Propagation (S_p):** Measures if LLM outputs inherit stronger sentiment than the prompt.

A composite quality score was computed as:

$$Q = \lambda_1 C + \lambda_2 F - \lambda_3 B - \lambda_4 S_p$$

where the λ values were normalizing constants.

In our evaluations, we used the latest versions of leading LLMs, including ChatGPT (v4), Claude (v1.3), DeepSeek (v2.0), Gemini (v1), and LLaMA (v2).

3 Results

We present both quantitative and qualitative analyses of how prompt sentiment influences LLM-generated outputs. Our evaluation spans multiple AI-driven applications, measuring sentiment propagation, factuality, bias, and response structure.

3.1 Impact on Output Sentiment Propagation

Figure 1 illustrates the sentiment propagation ratio (Sp), which quantifies the extent to which LLMs amplify or neutralize input sentiment across different tasks. Our analysis reveals distinct patterns across application domains. Content generation and creative writing tasks exhibit the strongest sentiment propagation, with models not only mirroring the emotional tone of the prompt but also occasionally exaggerating it. This effect is particularly pronounced in storytelling and blog writing applications.

Conversational AI models demonstrate more moderate sentiment propagation characteristics. While these models generally retain the sentiment of the input, they exhibit a tendency to soften extreme emotions, especially for negative prompts.

In contrast, applications focused on legal and financial analysis and technical documentation show significantly lower sentiment propagation. These domain-specific models default to a more formal, neutral tone, regardless of input