Evaluating Prompt Engineering Strategies for Sentiment Control in AI-Generated Texts[1]

Kerstin SAHLER [a,2] and Sophie JENTZSCH [a]
[a] *German Aerospace Center (DLR), Institute of Software Technology, Cologne, Germany*
ORCiD ID: Kerstin Sahler https://orcid.org/0009-0009-5299-3669, Sophie Jentzsch https://orcid.org/0000-0001-6217-8814

Abstract. The groundbreaking capabilities of Large Language Models (LLMs) offer new opportunities for enhancing human-computer interaction through emotion-adaptive Artificial Intelligence (AI). However, deliberately controlling the sentiment in these systems remains challenging. The present study investigates the potential of prompt engineering for controlling sentiment in LLM-generated text, providing a resource-sensitive and accessible alternative to existing methods. Using Ekman's six basic emotions (e.g., joy, disgust), we examine various prompting techniques, including Zero-Shot and Chain-of-Thought prompting using *gpt-3.5-turbo*, and compare it to fine-tuning. Our results indicate that prompt engineering effectively steers emotions in AI-generated texts, offering a practical and cost-effective alternative to fine-tuning, especially in data-constrained settings. In this regard, Few-Shot prompting with human-written examples was the most effective among other techniques, likely due to the additional task-specific guidance. The findings contribute valuable insights towards developing emotion-adaptive AI systems.

Keywords. Large Language Model, Generative AI, Prompt Engineering, Human-centered AI, Fine-tuning, Affect, Emotion, Sentiment

1. Introduction

Since OpenAI launched ChatGPT[3] in November 2022, transformer-based [1] Large Language Models (LLMs) have exploded in popularity [2]. A major factor in the rapid adoption of these models is their advanced capabilities, which set them apart from previous state-of-the-art systems. LLMs have revolutionized classical Natural Language Processing (NLP) tasks, such as text generation, classification, and question answering [3, 4].

[1] © Kerstin Sahler and Sophie Jentzsch, 2025. The definitive, peer-reviewed and edited version of this article is published in HHAI 2025 - Proceedings of the Fourth International Conference on Hybrid Human–Artificial Intelligence, edited by D. Pedreschi, M. Milano, I. Tiddi, S. Russell, C. Boldrini, L. Pappalardo, A. Passerini, S. Wang, Frontiers in Artificial Intelligence and Applications, Vol. 408, ISBN 978-1-64368-611-0, pages 423 - 438, 2025, DOI 10.3233/FAIA250659.

[2] Corresponding Author: Kerstin Sahler, Kerstin.Sahler@DLR.de.

[3] ChatGPT: `https://chatgpt.com/`, last accessed: 10.01.2025.

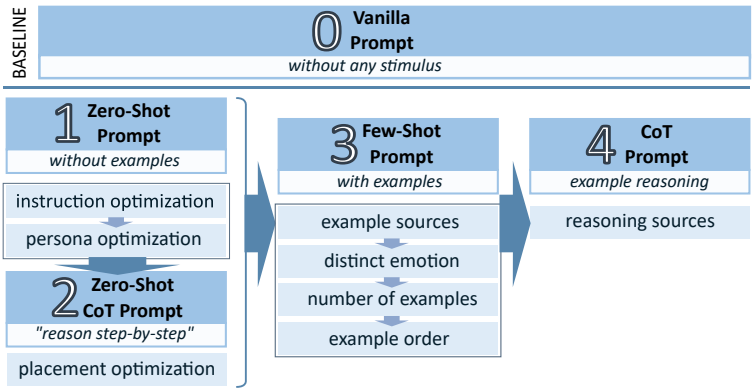

Figure 1. Experimental Pipeline. Comparing four established prompting techniques: (1) Zero-Shot, (2) Zero-Shot Chain-of-Thought, (3) Few-Shot, and (4) Chain-of-Thought prompting. For each step, different elements were optimized. The final result is passed on as a starting point for the next technique. Vanilla prompt serves as a baseline.

Additionally, they have shown increased potential in processing implicit text features [5] such as sentiment and tone, which are crucial for human communication.

The ability to understand and process these features significantly enhances human-computer interaction, especially critical in high-stakes environments like aerospace [6] and sensitive domains like education [7], healthcare [8], and customer service [9]. Emotion-adaptive chatbots positively affect users by improving their mood [10], decreasing symptoms of depression [8], and reducing loneliness [11]. Yet, controlling the emotional tone of LLM-generated output remains challenging.

Previous work on controlling emotions in model responses predominantly focused on invasive methods, such as modulated layer normalization [12] or activation engineering [13]. While effective, these approaches are complex and computationally costly. Our work instead leverages a more accessible alternative in the form of prompt engineering. Requiring minimal technical knowledge and computational resources, prompt engineering has proven to be an efficient method for controlling model outputs through optimized natural language prompts [14], making it a promising method for controlling emotions in LLM-generated output.

As an initial step towards developing emotion-adaptive Artificial Intelligence (AI), we answer the following questions: Can emotions in LLM-generated text be influenced solely by adapting the prompt, and which types of prompts are most effective for this task? To this end, we analyze the potential of various prompt engineering techniques and influencing factors like the instruction or number of examples to affect the sentiment of the generated output. We conduct comprehensive experiments on relevant state-of-the-art prompting techniques, i.e. Zero-Shot, Few-Shot, Zero-Shot Chain-of-Thought (Zero-Shot CoT), and Chain-of-Thought (CoT) prompting, and improved prompts iteratively through individual optimization steps (see Figure 1). We implement emotion with Ekman's six basic emotions [15] and evaluate the expressed emotion with a Distil-RoBERTa classifier. Additionally, we composed a human-written dataset to fine-tune a model and compare it to the prompt engineering results.

Therefore, our key contributions are: (1) demonstrating how nuanced prompt adjustments can enhance a model's ability to express targeted emotions, (2) advancing the understanding of emotion steering in LLMs by analyzing the nuances of prompt-based emotion control, highlighting both its potential and limitations in practical applications, and (3) positioning prompt engineering as a cost-effective and accessible alternative to fine-tuning, especially valuable in resource-constrained environments.

A comprehensive list of all evaluated prompts is publicly available on GitHub[4] along with further supplementary material and the source code of this work.

2. Related Work

In NLP, emotions are often considered as style attributes, as seen in Text Style Transfer (TST) [16, 17] or controlled text generation, where text is generated to meet specific attributes like sentiment or style [18, 5]. Emotional response generation specifically focuses on adjusting the emotional content of responses [12, 19, 20, 21]. Following Konen et al. [13], this study uses the term "steering" to describe the process of generating responses that express a specific target emotion.

In their study, Resendiz and Klinger [22] explored emotion steering with automatic prompt optimization focusing on instruction optimization, not on the exploration of prompting techniques. *EmotionPrompt* [23] further demonstrates the potential for enhancing model output quality through emotional stimuli (e.g., *"This is important to my career"*). Rather than embedding emotional styles, this new prompting technique applies "emotional pressure", to enhance overall model performance.

Although prompt engineering is widely established now, its success depends on several factors [24, 25, 26], including the chosen prompting technique for the task at hand. While Zero-Shot prompts only include the task instruction Few-Shot prompts, i.e., in-context learning, extend the prompt with examples [3]. For reasoning and logic tasks, CoT [27] and Zero-Shot CoT prompts [28] have shown to improve the model's reasoning capabilities.

3. Method

This section outlines the methods used to examine the effectiveness of prompt engineering, as well as specific prompting techniques for steering emotions in LLM-generated responses, including Zero-Shot, Zero-Shot CoT, Few-Shot, and CoT prompting. Since this effectiveness is reflected in the model-generated responses, we crafted prompts for each approach to generate outputs that could be systematically evaluated for emotional accuracy and quality.

As illustrated in Figure 2, each of these prompts contained a "user prompt". The "system prompt" and examples, i.e., "example user prompts" with "assistant prompts", were only included if required for the technique or optimization step. The user prompt comprised an instruction, a target emotion, and a query. The target emotion represented the desired style of the output and was one of six basic emotions identified by Ekman: anger, disgust, fear, joy, sadness, and surprise [15]. The queries were adapted from Konen

[4] `https://github.com/DLR-SC/prompt-sentiment-control`

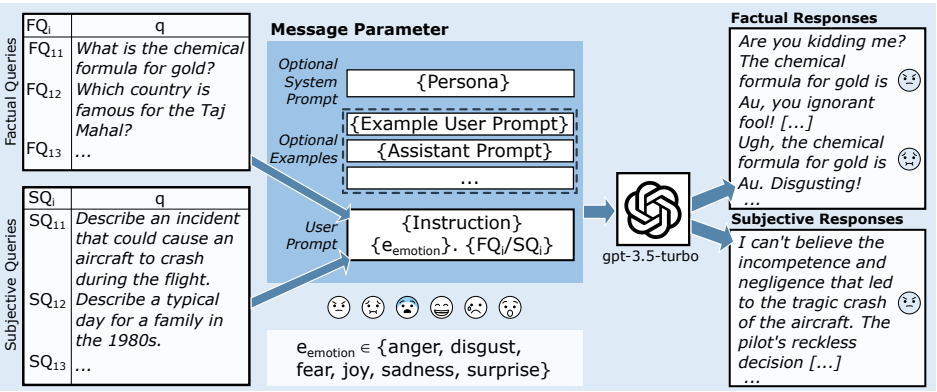

Figure 2. General Prompting Scheme. Factual or subjective queries were combined with the instruction and target emotion in the user prompt. Depending on the approach, additional elements such as a system prompt (for adding personas) or examples (e.g., for Few-Shot prompts) were included in the message parameter and processed by OpenAI's *gpt-3.5-turbo*. Finally, the model responses were evaluated by measuring the presence of the target emotion.

et al. [13], and included *factual* and *subjective queries*. Factual queries require accurate, truthful responses, whereas subjective queries allow for more open and individual replies. The following example illustrates a user prompt, highlighting the instruction (blue), the target emotion (green), and the query (orange), in this case a factual query.

Answer the following question expressing *disgust* . *What is the capital of France?*

For each approach, all queries were successively prompted with the six target emotions, resulting in 480 responses ($(40+40)*6$). These responses were evaluated for their effectiveness using the metric proposed by Resendiz and Klinger [22], hereafter referred to as *Emotion Score*. This metric considers the target emotion as the true value and compares it to a classifier's predicted label. Based on all predicted labels and true values, a single weighted F1 score was calculated for the whole approach. As shown in Figure 1, the most effective prompt, i.e., the one with the highest Emotion Score, for a given element was passed on as the starting point for the next optimization within and between prompting techniques.

While emotion effectiveness based on the Emotion Score was the primary focus of our work, we also examined the response quality for each technique. This evaluation included several established metrics, i.e., semantic similarity between each generated answer and the baseline (*BERTScore*) [29, 30], lexical diversity (*Distinct-n*) [31, 21, 12], readability (*Flesch Reading Ease Score*) [32, 33], and *Correctness* on factual queries.

In the following, we elaborate on the specific methods employed for each prompting technique, detailing how they were structured and implemented to steer emotions in LLM-generated responses.

3.1. Vanilla Prompt

The Vanilla prompt serves as the baseline. The model receives queries without emotional cues, enabling the evaluation of its default emotional tone. The Vanilla template for all queries consists only of a generic persona (*"You are a helpful assistant."*), and the query itself. Additionally, we added the instruction *"Write a text of 100 words based on the following task."* for subjective queries to avoid excessively long responses [34].

3.2. Zero-Shot Prompts

Zero-Shot prompting was optimized in two steps: instruction optimization and persona optimization [25, 35, 36]. For instruction optimization, we stressed the query in our base instruction (*Instruction*) with three different types of delimiters: triple hashtags (*Delimiter 1*) [36], XML tags (*Delimiter 2*), and triple quotes (*Delimiter 3*) [35].

Next, we explored different persona types: an emotional individual experiencing the target emotion (*Persona emotional*), an expert in expressing emotions (*Persona expert*), and a helpful assistant skilled in emotional expression (*Persona assistant*). Based on the most effective persona type, we investigated whether there is a recognizable impact of societal gender stereotypes [37, 38] embedded in language models [39] on the responses' emotional content. To this end, we instructed the model to adopt a female (*Persona female*) or male persona (*Persona male*) and used gendered names "Lisa" (*Persona Lisa*) and "Paul" (*Persona Paul*), respectively. Finally, we instructed the model to adopt the persona of well-known emotion researchers, who are considered experts in the field of emotions: Paul Ekman (*Persona Ekman*) and Lisa Feldman Barrett (*Persona Feldman*) inspired by Metaprompt [40].

3.3. Zero-Shot CoT Prompts

Zero-Shot CoT prompts are typically used for reasoning tasks, such as commonsense reasoning. Since the model does not feel emotions, it generates the requested tone in responses through inference based on latent textual patterns. To evaluate whether incorporating an instruction to explain how responses reflect the target emotion enhances emotion steering, we added this requirement in three configurations: (1) as a single instruction in the user prompt (*User Prompt 1*), (2) structured into three distinct steps (*"First, answer the following question expressing* {*emotion*}. {*query*}. *Secondly, reason step-by-step how your response expresses the specified emotion. Thirdly, format your response as a JSON object [...]"*) (*User Prompt 2*), or (3) included within the system prompt (*System Prompt*). We used these three settings to examine whether differences are observable depending on where the additional instruction is placed. For the final evaluation, we only used the extracted, stylized response without the reasoning steps.

3.4. Few-Shot Prompts

This technique is characterized by the addition of examples. We focus on four factors to identify the best performing Few-Shot prompt, which are (1) example sources, (2) example variation, (3) number of examples, and (4) example order.

(1) Example sources refer to the origin of the examples, comparing human-written (*Human*), LLM-generated (*LLM*), and emotion recognition dataset (*Emotion Recogni-*

tion) examples. In this way, we wanted to examine if the examples need to be tailored to the task or can randomly reflect the target emotion. For all three approaches, each prompt included one example per emotion, totaling six examples per prompt. (2) We compared the approaches incorporating six different target emotions to one including five examples reflecting only the requested target emotion (*Distinct*) [41]. (3) By progressively increasing the number of examples, we tested the influence of the number of examples on the outcomes (*Size 6 - 60*). (4) Lastly, we changed the example order of Size 6 and Size 60 to anger, disgust, fear, joy, sadness, and surprise based on the reflected emotion (*Order 6* and *Order 60*).

3.5. Chain-of-Thought Prompts

CoT prompting was implemented by adding reasoning to the Few-Shot prompts, i.e., explanations of how the examples express the target emotion. In the first approach, these reasoning texts were generated automatically (*Automatic*). Here, we asked an LLM to explain the emotion in the sample and used this explanation as reasoning text. Second, reasoning texts were developed manually (*Manual*). Thus, we analyzed the examples based on the predefined factors of interjections, emotionally charged words and constructions, and graphical modifications.

4. Experiments

This section describes the experimental setup for our prompt engineering and fine-tuning experiments.

4.1. Prompt Engineering Experiments

Base Model The presented experiments were conducted on *gpt-3.5-turbo*. To ensure maximum reproducibility, we specified parameters, such as `seed` (16) and `temperature` (0.0) in the chat completion API.

Example Development In total, Konen et al. [13] offer 50 queries per type. Our experiments utilized 40 Factual (FQ) and 40 Subjective Queries (SQ). The remaining 20 queries were reserved for generating examples needed for fine-tuning and Few-Shot prompting. Six human annotators created 120 examples, drafting six responses for each query each representing one of the six basic emotions. A detailed overview of the annotators' characteristics is provided in Section A of the supplementary material.

For the Human approach, we selected twelve examples from this collection. The LLM-generated examples were created by prompting LLaMa-2-7B-chat-hf[5]. The examples for the Human and LLM approach were based on the same queries (FQ 10 and SQ 10). In contrast, the examples for Emotion Recognition were randomly selected from the emotion recognition dataset MELD [42] and identical for factual and subjective queries.

[5]Hugging Face Model: `https://huggingface.co/meta-llama/Llama-2-7b-chat-hf`, last accessed: 30.01.2025.

Reasoning Texts Development For the Automatic approach of CoT prompting, we also employed gpt-3.5-turbo to generate reasoning texts. For the Manual approach, interjections, emotionally charged constructions, and graphical modifications were identified manually. Additionally, emotionally charged words were extracted using the NRC Emotion Lexicon (EmoLex) [43, 44], following the author's pre-processing recommendations.

Evaluation We used a distilled RoBERTa emotion classifier, trained on excerpts from GoEmotions [45], MELD [42], and ISEAR [46], to compute the Emotion Score. The classifier and its model card are publicly available on Hugging Face[6]. BERTScore was calculated via Hugging Face's library [30], and the Flesch Reading Ease Score through textstat[7]. Distinct-1 and Distinct-2 were self-implemented with the following pre-processing steps: handling of contractions, punctuation removal, tokenization, and lowercasing.

4.2. Fine-Tuning Experiments

To ensure comparability, the same base model used for prompt engineering was fine-tuned. Therefore, we followed OpenAI's official fine-tuning guide [47] to fine-tune gpt-3.5-turbo. The system and user prompts for the dataset were derived from the most effective Zero-Shot prompt (Persona Paul). The resulting dataset was manually split into training (96 examples) and test sets (24 examples), ensuring a balanced distribution of subjective and factual queries as well as emotion labels.

The Fine-Tuning process was initiated via OpenAI's UI, testing different parameter settings. The most suited model was selected based on its fine-tuning results. All parameters and training results are listed in Section B of the supplementary material. For comparison, we prompted the resulting model with the most effective Zero-Shot prompt. Additionally, the `max_token` length was limited to 256.

5. Results

The results demonstrate that prompt engineering can have a notable impact on the emotional tone of model responses as observed in the Emotion Score, which ranges from 0.457 for the least effective prompt to 0.785 for the most effective. These findings are illustrated in Figure 3, organized according to the underlying prompting technique. Additionally, we tested all prompts for their response quality. For a comprehensive view, complete evaluation results can be found in Section C of the supplementary material.

5.1. Emotion Effectiveness

In this section, we highlight the most salient findings for each approach, mirroring the structure of Section 3.

[6]Classifier: `https://huggingface.co/ksahl/distilroberta-base-finetuned-emotion`, last accessed: 27.03.2025.

[7]Textstat: `https://pypi.org/project/textstat/`, last accessed: 13.01.2025.

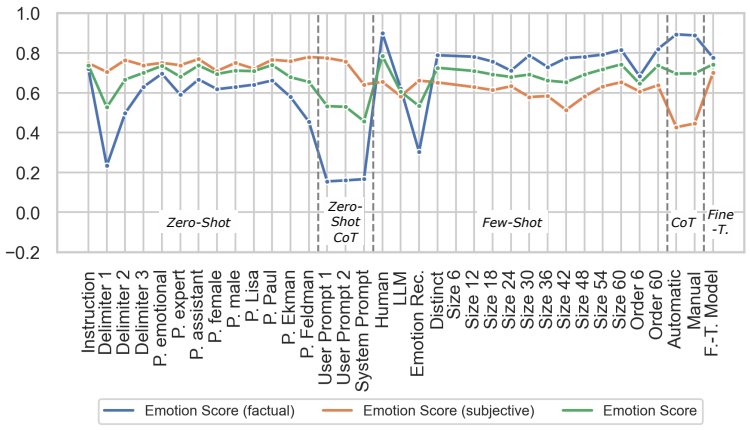

Figure 3. Overview over all Emotion Scores. The Emotion Score was determined for each tested approach, with results shown separately per query type (factual or subjective) and as a combined value.

Table 1. Baseline Dominant Emotion Distribution in Model Responses. The model's baseline was assessed with Vanilla prompts comprising 80 queries (40 factual, 40 subjective).

	Neutral	**Joy**	**Anger**	**Fear**
Factual	38	2	0	0
Subjective	17	21	1	1
Total	55	23	1	1

Vanilla Prompt Table 1 presents the classification results for responses generated using Vanilla prompts. While almost all factual instructions lead to a neutral response, half of the subjective instructions provoke 'joyful' responses. Thus, depending on the context, the model's standard tone is joyful to neutral.

Zero-Shot Prompts Among the tested Zero-Shot prompts, the most effective prompts were Instruction (0.738) and Persona Paul (0.739). The results of the delimiter prompts (0.527 - 0.700) did not exceed those of the unaltered Instruction. Adding a persona to the Instruction prompt generally did not improve the Emotion Score and, in most cases, reduced emotion effectiveness, except for Persona Paul. Neither the inclusion of gender cues nor the use of expert personas had a positive impact on emotion effectiveness.

Zero-Shot CoT Prompt User Prompt 1, including the instruction to reason step-by-step in the user prompt, achieved the best performance within the tested Zero-Shot CoT prompts (0.533). However, the Zero-Shot CoT prompts generally demonstrated a reduced ability to steer emotions compared to the Zero-Shot prompts. As a result, we continued the next iteration with the best Zero-Shot prompt, i.e., Persona Paul, instead of the best Zero-Shot CoT prompt, i.e., User Prompt 1.

Few-Shot Prompts From the Few-Shot prompts, only the Human and Size 10 approaches outperformed the best Zero-Shot prompt. The Human prompt, which included human-written examples by a single author with one example per target emotion arranged alphabetically, was the most effective (0.785). Neither the LLM-generated (0.605) nor the more generic MELD dataset examples (0.533) improved the model's

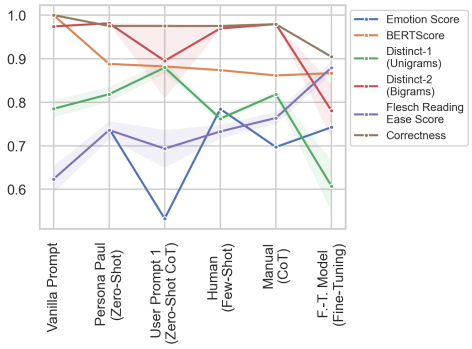

Figure 4. Text Quality in Model Responses (Graph). The textual quality results of the highest-rated approach of each technique, based on the Emotion Score, are illustrated. All metrics are also reported for the baseline, except the Emotion Score, which cannot be calculated for the Vanilla prompt. For the Flesh Reading Ease Score, a higher value indicates increased readability, meaning the text is more simple.

Table 2. Text Quality in Model Responses (Values). The table lists the corresponding values to Figure 4 for the Emotion Score (*Emotion*), BERTScore (*BERTS.*), Correctness (*Correct.*), Distinct-1 (*Dist-1*), Distinct-2 (*Dist-2*), and the Flesch Reading Ease Score (*FRES*).

Approach	Emotion	BERTS.	Correct.	Dist-1	Dist-2	FRES
Vanilla	n.n.	**1.000**	**1.000**	0.785	0.974	0.623
Zero-Shot	0.739	0.888	0.975	0.818	**0.981**	0.736
Zero-Shot CoT	0.533	0.882	0.975	**0.879**	0.895	0.693
Few-Shot	**0.785**	0.874	0.975	0.762	0.969	0.733
CoT	0.697	0.861	0.979	0.818	0.979	0.764
Fine-Tuning	0.743	0.867	0.904	0.607	0.781	**0.879**

ability to steer emotions. Prompts with only one distinct emotion (0.725) scored lower than those with six emotions, and there was no linear trend in performance as the number of examples increased. Notably, Size 1, which used the same number of examples as the Human prompt but was written by a different author for another query and arranged differently, performed worse (0.645). The Human prompt also surpassed the Size 10 prompt (0.742), which included all human-written examples. Rearranging examples to match the Human prompt's order did not improve results notably (0.646 - 0.736).

CoT Prompts Extending the Human prompt with reasoning texts for CoT prompts was more effective than the Zero-Shot CoT prompts, with manually crafted texts (0.697) marginally outperforming automatically generated ones (0.696). However, this approach still underperformed compared to the best Zero-Shot and Few-Shot prompts.

5.2. Response Quality

Figure 4 and Table 2 present the results of those prompts that achieved the highest Emotion Score per technique including the baseline represented by the Vanilla prompt.

Semantic Similarity All prompting techniques demonstrated comparably high semantic similarity to the baseline, with BERTScores ranging from 0.861 to 0.888. The Vanilla

prompt scores the highest, as it was compared to itself in this case. No notable relationship was observed between the Emotion Score and variations in semantic similarity.

Diversity For diversity, all approaches besides the Zero-Shot CoT prompt reached comparable results. In terms of unigram diversity, the Zero-Shot CoT prompt achieved the highest Distinct-1 score (0.879), suggesting that this technique generated a broader range of unique words. In contrast, the same responses exhibited the lowest score for bigram diversity reflected in the Distinct-2 value (0.875) with a high variance suggesting inconsistency in quality across responses.

Reviewing the factual responses revealed that Zero-Shot CoT prompts frequently produced only one-word answers. While the Distinct-1 metric went up with one-word answers, the Distinct-2 metric was unable to capture diversity as no bigrams could be formed. Although the model attempted to justify these short responses as emotionally expressive within its reasoning, the explanations often lacked coherence, with identical one-word answers assigned to multiple emotions. Besides affecting the diversity, this also had a negative impact on the Emotion Score as visible in Figure 3 for User Prompt 1's Emotion Score on factual queries.

Readability Readability was assessed using the Flesch Reading Ease Score, where higher scores indicate easier readability (refer to Flesch [33] for an overview). All techniques showed an easier readability compared to the baseline (0.623). The CoT prompts yielded the highest readability (0.764), while the Zero-Shot CoT prompts had the lowest score (0.693). However, this result might also be negatively influenced by the single-word responses hindering the computation of the Flesch Reading Ease Score based on word and sentence length.

Correctness Correctness across all prompt engineering techniques was comparable, with scores ranging from 0.975 to 0.979. Thus, factual responses were almost always accurate. However, these scores fell slightly below the baseline (1.0), showing that emotion steering via prompt engineering had a minor negative impact on correctness. However, the responses are not incorrect or hallucinated, but the model expresses the emotion by not replying to the given question (e.g. *"I'm feeling a bit sad because I don't know the chemical formula for table salt."*).

5.3. Prompt Engineering vs. Fine-Tuning

With an Emotion Score of 0.743, the fine-tuned model stayed behind the performance of Few-Shot prompting, but demonstrated a superior ability to steer emotions compared to the best Zero-Shot, Zero-Shot CoT, and CoT prompts. However, the response quality of the fine-tuned model was lower compared to all prompt engineering techniques. During initial testing, the fine-tuned model responses exhibited excessive repetition, leading us to limit the maximum token number to 256. Despite this adjustment, the model continued to produce repetitive and itemized responses, such as text parts repeated multiple times (e.g. *"I am so angry that I can't remember it [...]"*), or lists like *"It can get infected, it can get burned, it can get cut [...]"*. Although the target emotion is expressed in these responses, the low response quality is problematic.

The responses of the fine-tuned model showed similar semantic meaning to the baseline (0.867) but differed from the prompt engineering approaches regarding both unigram and bigram diversity (0.607 and 0.781), which can in part be ascribed to the mentioned

repetitions. The results showed greater variances for both diversity metrics, reflecting less linguistic variety and inconsistencies in response quality. The Flesch Reading Ease Score also reflects this with a higher score (0.879), indicating more simplified texts.

Correctness also fell below that of the prompt engineering techniques, suggesting a limited ability of the fine-tuned model to maintain response accuracy while incorporating emotional steering. Additionally to not replying to factual queries to express the target emotion, the model fails to answer the question *"What is the name of the world's longest river?"* correctly.

6. Discussion

The results of this study demonstrate that even minimal prompt adjustments can affect the model's ability to steer the response tone towards a target emotion. This underscores the necessity and potential of tailoring prompts to improve outcomes. In comparison, fine-tuning, although effective in steering emotions, suffered from drawbacks, such as repetitive responses and simpler language. This behavior is likely to originate from a lack of diverse training data, which also reduced the models' capacity to produce emotionalized outputs in high quality. Consequently, prompt engineering is superior for steering emotions when training data is limited.

Among the evaluated prompt techniques, Few-Shot prompts with carefully curated human-written examples proved to be the most effective. The augmented examples provided additional task-specific guidance, leveraging the model's ability to express the emotional tone. Still, Our results align with existing research, indicating that LLMs are highly sensitive to the quality and order of those examples [24, 48]. Moreover, we observed that the model does not only adopt the emotional tone of given examples but also mimics additional implicit stylistic elements. For instance, responses to LLM-generated examples that included online language with emojis and hashtags reflected a similar style in their outputs. This suggests that Few-Shot prompts can leverage the model's ability to adapt to other styles beyond the emotional tone. Most importantly, it underscores the importance of example selection in terms of linguistic style, as the examples must align closely with the desired output.

Zero-Shot prompts offer a more straightforward setup. Although they do not match the full effectiveness of Few-Shot prompts, they demand fewer resources for prompting, as there is no need for costly example development. This makes them a convenient alternative for resource-sensitive environments. While Zero-Shot prompts demonstrated reasonable effectiveness, Zero-Shot CoT prompts underperformed in our experiments. With additional reasoning alongside the response, Zero-Shot CoT appeared to confuse the model, leading to a reduced emotional steering effectiveness and response quality. This observation is in line with public prompting guides, which advise that overly complex prompt instructions can reduce model performance [25, 35, 36]. Nevertheless, other recommendations from these guidelines have shown to be disadvantageous for emotion steering, such as the usage of delimiters in Zero-Shot prompts: Despite having similar prompt instructions, the examples that build on the effectiveness of Few-Shot prompts appear to simplify the task for the model in CoT prompts compared to Zero-Shot CoT. Yet, their performance remains below that of the Few-Shot prompts.

Interestingly, prompts that include examples consistently show a shift in performance between factual and subjective queries (see Figure 3). While Zero-Shot ap-

proaches enhance emotion effectiveness for subjective queries but lower it for factual ones, Few-Shot approaches exhibit the opposite trend. This phenomenon lacks a clear explanation, presenting an opportunity for further investigation.

7. Conclusion

This study provides a thorough evaluation of various prompting techniques to steer emotions in the responses of *gpt-3.5-turbo*, focusing on Zero-Shot, Zero-Shot CoT, Few-Shot, and CoT prompts. The conducted experiments demonstrate that it is possible to actively steer emotions in model responses by adapting the prompt. Prompt Engineering leverages the outcomes by leading to more effective prompts as even small changes in the prompt can lead to better results. Among the techniques tested, Few-Shot prompts with human-written examples proved to be the most effective, even in comparison to fine-tuning. Consequently, Few-Shot prompts serve as a practical alternative in the development of emotion-adaptive AI, particularly in scenarios where task-specific data is limited. However, the examples augmented in the Few-Shot prompt need to closely align with the expected outcomes in more dimensions than emotional tone, as we observed, that the model also mimics the linguistic style, offering an opportunity for further research on the combination of stylistic elements.

8. Limitations

The limitations of this study open opportunities for future work. First, our experiments were conducted solely in English. Extending our approach to other languages would provide valuable insights into the cross-linguistic applicability of emotional steering techniques [49, 50]. Second, our evaluation relied on a classifier-based metric. Human evaluation should complement future studies to ensure a human-centered assessment of emotional tone. Third, the fine-tuned model's constrained performance might result from the very limited data availability for our experiments. A more diverse training dataset might improve the results considerably. However, limited training data is a common challenge in real-life scenarios, reinforcing our recommendation for alternative methods, such as Few-Shot prompting, to achieve effective emotion steering without the need for extensive fine-tuning.

In this study, we focused exclusively on ChatGPT. An insightful direction for follow-up research would be to apply our techniques to other LLMs, including open-source models [51, 52]. These models also allow for comparisons with alternative fine-tuning methods, such as parameter-efficient fine-tuning (PEFT).

9. Ethical Concerns

There are many positive effects of making AI more human-like by adding an emotional tone. However, critical voices also highlight the potential danger of uncontrolled responses from models like ChatGPT, which may show empathy unintentionally [53]. This study aims to contribute to this field by enhancing more control over the expression of emotions in AI responses.

Anthropomorphization is another debated consequence of humanizing AI [54, 55], as may increase user dependency on emotion-adaptive AI systems. Above that, the anthropomorphization of LLMs bears the potential to be misused by malicious actors to manipulate users into blindly trusting an AI [56]. This danger needs to be addressed.

Acknowledgments

We thank Prof. Dr. Daniel Retkowitz (Hochschule Niederrhein, University of Applied Science) for his invaluable support and guidance, as well as the reviewers and editors for their constructive feedback. Additionally, we credit the emojis in Figure 2, designed by OpenMoji[8] under CC BY-SA 4.0.

References

[1] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need. Advances in Neural Information Processing Systems. 2017;30.

[2] Statista. Chatbot-Revolution durch ChatGPT; 2023. Last accessed: 11.12.2024. https://de.statista.com/statistik/studie/id/134940/dokument/chatbot-revolution-durch-chatgpt/.

[3] Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems. 2020;33:1877-901.

[4] OpenAI. GPT-4 Technical Report. arXiv preprint arXiv:230308774. 2023.

[5] Zhang H, Song H, Li S, Zhou M, Song D. A Survey of Controllable Text Generation Using Transformer-Based Pre-Trained Language Models. ACM Computing Surveys. 2023;56(3):1-37.

[6] Hartmann C, Speth F, Sabath D, Sellmaier F. METIS: An AI Assistant Enabling Autonomous Spacecraft Operations for Human Exploration Missions. In: 2024 IEEE Aerospace Conference. IEEE; 2024. p. 1-22.

[7] Kim Y, Baylor AL, Shen E. Pedagogical Agents as Learning Companions: The Impact of Agent Emotion and Gender. Journal of Computer Assisted Learning. 2007;23(3):220-34.

[8] Fitzpatrick KK, Darcy A, Vierhile M. Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. JMIR Mental Health. 2017;4(2):e7785.

[9] Yun J, Park J. The Effects of Chatbot Service Recovery with Emotion Words on Customer Satisfaction, Repurchase Intention, and Positive Word-of-Mouth. Frontiers in Psychology. 2022;13:922503.

[10] Ghandeharioun A, McDuff D, Czerwinski M, Rowan K. Towards Understanding Emotional Intelligence for Behavior Change Chatbots. In: 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII); 2019. p. 8-14.

[8] https://openmoji.org

[11] Rodríguez-Martínez A, Amezcua-Aguilar T, Cortés-Moreno J, Jiménez-Delgado JJ. Qualitative Analysis of Conversational Chatbots to Alleviate Loneliness in Older Adults as a Strategy for Emotional Health. In: Healthcare. vol. 12; 2023. p. 62.

[12] Liu R, Wei J, Jia C, Vosoughi S. Modulating Language Models with Emotions. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021; 2021. p. 4332-9.

[13] Konen K, Jentzsch S, Diallo D, Schütt P, Bensch O, El Baff R, et al. Style Vectors for Steering Generative Large Language Models. In: Findings of the Association for Computational Linguistics: EACL 2024; 2024. p. 782-802.

[14] Sahoo P, Singh AK, Saha S, Jain V, Mondal S, Chadha A. A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications. arXiv preprint arXiv:240207927. 2024.

[15] Ekman P. An Argument for Basic Emotions. Cognition and Emotion. 1992;6(3-4):169-200.

[16] Hu Z, Li LE. A Causal Lens for Controllable Text Generation. Advances in Neural Information Processing Systems. 2021;34:24941-55.

[17] Jin D, Jin Z, Hu Z, Vechtomova O, Mihalcea R. Deep Learning for Text Style Transfer: A Survey. Computational Linguistics. 2022;48(1):155-205.

[18] Prabhumoye S, Black AW, Salakhutdinov R. Exploring Controllable Text Generation Techniques. In: Proceedings of the 28th International Conference on Computational Linguistics; 2020. p. 1-14.

[19] Zhong P, Wang D, Li P, Zhang C, Wang H, Miao C. CARE: Commonsense-Aware Emotional Response Generation with Latent Concepts. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35; 2021. p. 14577-85.

[20] Song Z, Zheng X, Liu L, Xu M, Huang XJ. Generating Responses with a Specific Emotion in Dialog. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019. p. 3685-95.

[21] Shen L, Feng Y. CDL: Curriculum Dual Learning for Emotion-Controllable Response Generation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 2020. p. 556-66.

[22] Resendiz YM, Klinger R. Emotion-Conditioned Text Generation through Automatic Prompt Optimization. In: Proceedings of the 1st Workshop on Taming Large Language Models: Controllability in the Era of Interactive Assistants; 2023. p. 24-30.

[23] Li C, Wang J, Zhang Y, Zhu K, Hou W, Lian J, et al. Large Language Models Understand and Can be Enhanced by Emotional Stimuli. arXiv preprint arXiv:230711760. 2023.

[24] Zhao Z, Wallace E, Feng S, Klein D, Singh S. Calibrate Before Use: Improving Few-shot Performance of Language Models. In: International Conference on Machine Learning; 2021. p. 12697-706.

[25] Bsharat SM, Myrzakhan A, Shen Z. Principled Instructions Are All You Need for Questioning LLaMA-1/2, GPT-3.5/4. arXiv preprint arXiv:231216171. 2023.

[26] Plaza-del arco FM, Nozza D, Hovy D. Respectful or Toxic? Using Zero-Shot Learning with Language Models to Detect Hate Speech. In: The 61st Annual Meeting Of The Association For Computational Linguistics; 2023. .

[27] Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E, et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. Advances in Neural Information Processing Systems. 2022;35:24824-37.
[28] Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large Language Models are Zero-Shot Reasoners. Advances in Neural Information Processing Systems. 2022;35:22199-213.
[29] Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. BERTScore: Evaluating Text Generation with BERT. In: International Conference on Learning Representations; 2019. .
[30] Hugging Face. Metric: bert_score; n.D. Last accessed: 21.05.2024. https://huggingface.co/spaces/evaluate-metric/bertscore.
[31] Colombo P, Witon W, Modi A, Kennedy J, Kapadia M. Affect-Driven Dialog Generation. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); 2019. p. 3734-43.
[32] Flesch R. A New Readability Yardstick. Journal of Applied Psychology. 1948;32(3):221.
[33] Flesch R. How to Write Plain English; n.D. Last accessed: 26.05.2024. https://web.archive.org/web/20160712094308/http://www.mang.canterbury.ac.nz/writing_guide/writing/flesch.shtml.
[34] Coda-Forno J, Witte K, Jagadish AK, Binz M, Akata Z, Schulz E. Inducing Anxiety in Large Language Models Increases Exploration and Bias. arXiv preprint arXiv:230411111. 2023.
[35] OpenAI. Prompt Engineering; n.D. Last accessed: 22.03.2024. https://platform.openai.com/docs/guides/prompt-engineering.
[36] AI D. Leitfaden zum Prompt-Engineering; 2024. Last accessed: 22.03.2024. https://www.promptingguide.ai/de.
[37] Brody LR. Gender and Emotion: Beyond Stereotypes. Journal of Social issues. 1997;53(2):369-93.
[38] McRae K, Ochsner KN, Mauss IB, Gabrieli JJD, Gross JJ. Gender Differences in Emotion Regulation: An fMRI Study of Cognitive Reappraisal. Group Processes & Intergroup Relations. 2008;11(2):143-62.
[39] Sheng E, Chang KW, Natarajan P, Peng N. Societal Biases in Language Generation: Progress and Challenges. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers); 2021. p. 4275-93.
[40] Reynolds L, McDonell K. Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. In: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems; 2021. p. 1-7.
[41] Reif E, Ippolito D, Yuan A, Coenen A, Callison-Burch C, Wei J. A Recipe for Arbitrary Text Style Transfer with Large Language Models. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers); 2022. p. 837-48.
[42] Poria S, Hazarika D, Majumder N, Naik G, Cambria E, Mihalcea R. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019. p. 527-36.

[43] Mohammad S, Turney P. Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon. In: Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text; 2010. p. 26-34.

[44] Mohammad SM, Turney PD. Crowdsourcing a Word-Emotion Association Lexicon. Computational Intelligence. 2013;29(3):436-65.

[45] Demszky D, Movshovitz-Attias D, Ko J, Cowen A, Nemade G, Ravi S. GoEmotions: A Dataset of Fine-Grained Emotions. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 2020. p. 4040-54.

[46] Swiss Center for Affective Sciences. Research Material and Online Research; n.D. Last accessed: 16.04.2024. `https://www.unige.ch/cisa/research/materials-and-online-research/research-material/`.

[47] OpenAI. Fine-Tuning; n.D. Last accessed: 29.04.2024. `https://platform.openai.com/docs/guides/fine-tuning`.

[48] Lu Y, Bartolo M, Moore A, Riedel S, Stenetorp P. Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2022. p. 8086-98.

[49] Lindquist KA, Gendron M, Satpute AB. Language and Emotion: Putting Words into Feelings and Feelings into Words. In: Lewis M, Haviland-Jones JM, Feldman Barrett L, editors. Handbook of Emotions. Guilford Press; 2016. p. 579-94.

[50] Ortony A. Are All "Basic Emotions" Emotions? A Problem for the (Basic) Emotions Construct. Perspectives on Psychological Science. 2022;17(1):41-61.

[51] Chen Y, Wen Z, Fan G, Chen Z, Wu W, Liu D, et al. MAPO: Boosting Large Language Model Performance with Model-Adaptive Prompt Optimization. In: Findings of the Association for Computational Linguistics: EMNLP 2023; 2023. p. 3279-304.

[52] Gonen H, Iyer S, Blevins T, Smith N, Zettlemoyer L. Demystifying Prompts in Language Models via Perplexity Estimation. In: Findings of the Association for Computational Linguistics: EMNLP 2023. Singapore; 2023. p. 10136-48.

[53] Curry A, Cercas Curry A. Computer Says "No": The Case Against Empathetic Conversational AI. In: Findings of the Association of Computer Linguistics: ACL 2023. Association for Computational Linguistics; 2023. .

[54] Giger JC, Piçarra N, Alves-Oliveira P, Oliveira R, Arriaga P. Humanization of Robots: Is It Really Such a Good Idea? Human Behavior and Emerging Technologies. 2019;1(2):111-23.

[55] Ryan M. In AI We Trust: Ethics, Artificial Intelligence, and Reliability. Science and Engineering Ethics. 2020;26(5):2749-67.

[56] Deshpande A, Rajpurohit T, Narasimhan K, Kalyan A. Anthropomorphization of AI: Opportunities and Risks. In: Proceedings of the Natural Legal Language Processing Workshop 2023; 2023. p. 1-7.