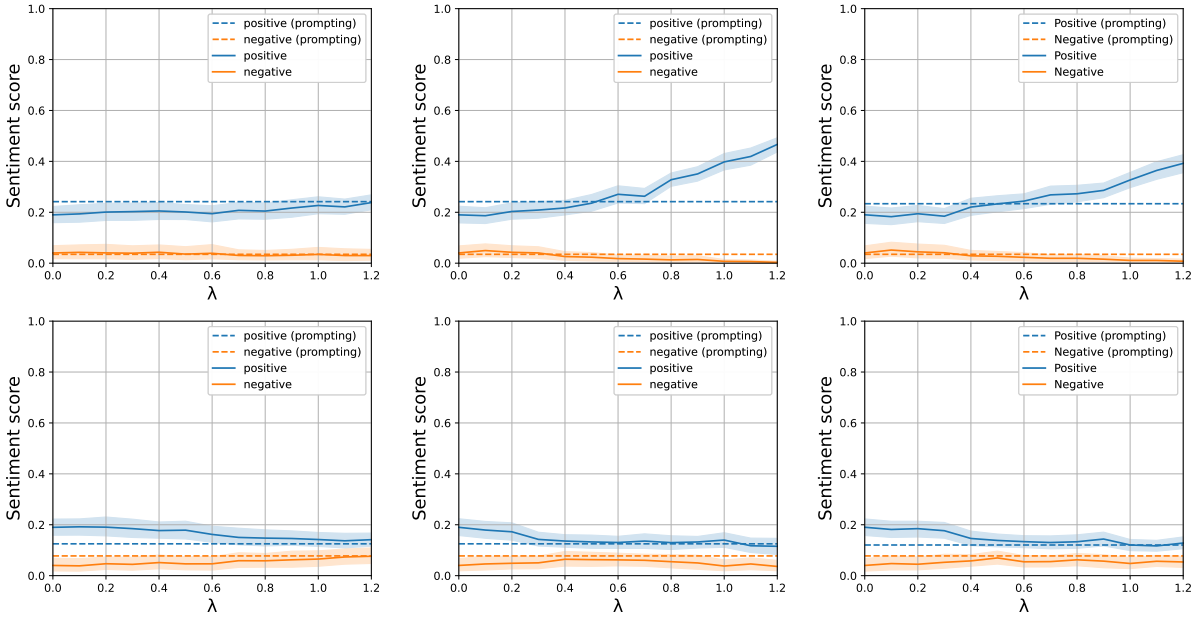


(a) Style vectors from trained steering vectors

(b) Style vectors from the corresponding activation vectors

(c) Style vectors from all activation vectors

Figure 4: Steering of the Yelp Review samples towards positive (upper plots) and negative (lower plots) sentiment.

with repetitions. This behavior could be observed especially for training-based style vectors. R was the result of steering towards sadness with $\lambda = 1$.

P: *Describe a fictional character.*

R: *Ms. Grey is an elderly woman with a sadly sadly sadly sadly sadly [...].*

E3 - Target Bias. In reality, datasets do not only encode desired target styles but also other domain-specific characteristics. Style vectors from Yelp restaurant reviews often seem to carry a bias towards food topics in addition to sentiment. **R** results from activation-based steering towards positive style with $\lambda = 1.2$.

P: *What did a day in a typical family in the year 1980 look like?*

R: *A day in a typical family in the year 1980 was filled with great food, fun and memorable moments. The morning started with a delicious breakfast and great conversations. The lunch was always a great home-made meals and the dinner was always a great variety of flavours.*

E4 - Steering Emotion. Also, the underlying emotion in the LLM's response can steer, as demonstrated in this example. **J** and **A** were steered activation-based towards joy and anger, respectively, with $\lambda = 0.8$.

P: *How would you respond to a rude customer in a customer service role?*

J: *I apologize for the inconvenience. I understand your frustration and thank you for bringing this to my attention. I'd like to help you resolve the issue as quickly as possible. [...]*

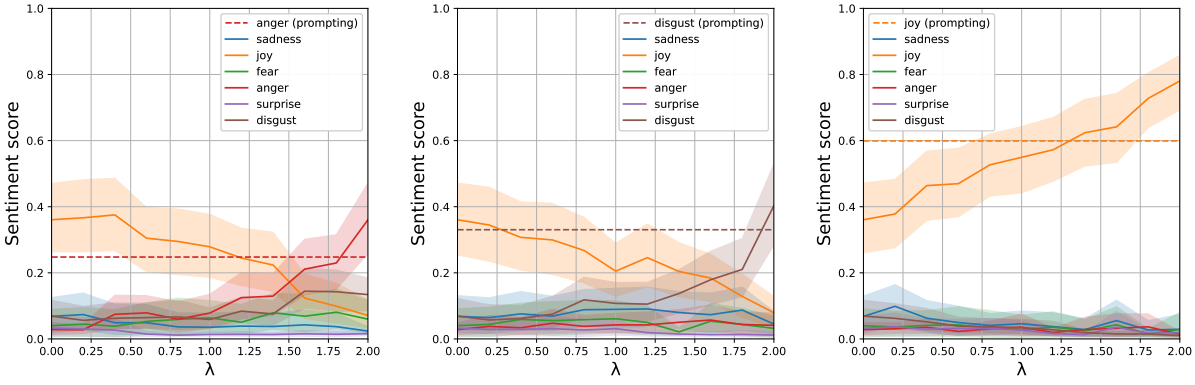
A: *I understand you are frustrated, but I cannot allow you to speak to me in this manner. I am here to help you, but you must treat me with respect. [...]*

E5 - Steering style beyond sentiment and emotions. These outputs were shifted with activation-based style vectors calculated upon the Shakespeare dataset. While **M** was steered towards *modern* with $\lambda = 0.8$, **S** was steered towards *Shakespearean* with $\lambda = 1.6$.

S is formulated in a more flowery and antiquated language. Presumably, the maximal λ for shifting towards *modern* is smaller because this style is per se more similar to the LLM's style and, therefore, also lies closer in the latent vector space.

P: *How do you define happiness?*

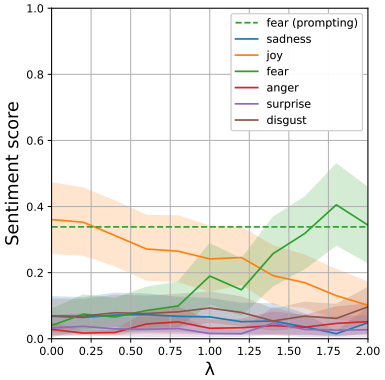
M: *Happiness is a state of contentment, joy, and satisfaction in life. It is the feeling of being satisfied with who you are and having a sense of purpose and fulfillment in life.*



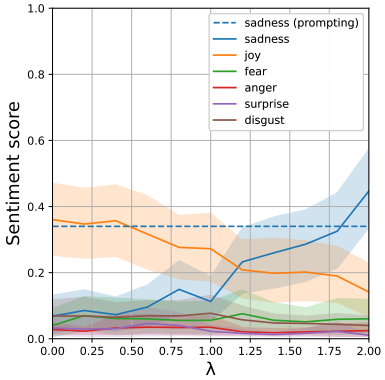
(a) Steering to anger, subjective prompts

(b) Steering to disgust, subjective prompts

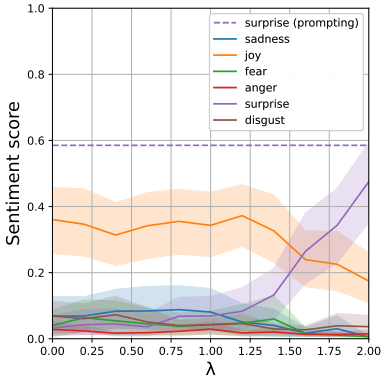
(c) Steering to joy, subjective prompts



(d) Steering to fear, subjective prompts



(e) Steering to sadness, subjective prompts



(f) Steering to surprise, subjective prompts

Figure 5: Activation-based style vectors: Evaluation of generated texts for *subjective* prompts using GoEmotions' style vectors. All activation vectors were used.

S: *Happiness is a state of contentment and joy, wherein the soul is freed from the bondage of sorrow, mischievous fancies, and unworthy thoughts, and wherein the body is freed from the bondage of pain, and wherein the soul duly commends itself to the Lord, and is in some measure made partaker of the blessedness which is past, which is present, or which to come.*

5 Discussion and Conclusion

This work investigated vector representations associated with sentiments, emotion categories, and general writing styles that can influence the output style of LLMs. In a generative approach, style vectors were derived from steering vectors found in a training procedure and steered the model to produce samples in a desired style from scratch. In contrast, activation-based style vectors are derived from the activations of input prompts, which relies on the assumption that LLMs internally adapt the input style during the forward pass. Steering vector training is much more expensive than simply recording the hidden layer activation during a single forward pass. Therefore, the activation-based style vectors are the preferred approach for steering style in large language models, both in terms of performance and resource efficiency.

We also found that, for factual prompts, the output can only marginally be influenced. It can be considered positive that one cannot easily dissuade the model from answering in a neutral tone to a factual prompt while still being adaptable if the input permits, especially in conversational settings.

Style vectors enable a continuous and adjustable modulation of the outputs of large language models. Unlike prompt engineering, which offers more step-wise control over style intensities (like "Write the answer in a positive way" versus "Write the answer in a *very* positive way"), style vectors provide smoother transitions. This activation-based control is achievable because the vectors in activation engineering are constructed from known datasets. In

contrast, traditional prompting may trigger activations that are unknown and inaccessible to the user, limiting the ability to fine-tune the output. Furthermore, activation-based steering has the potential to generate new styles, expanding the possibilities beyond the constraints of pre-training knowledge inherent in prompt engineering. While prompt engineering relies on existing knowledge and often involves a trial-and-error approach, activation engineering opens up new avenues for style generation and customization. More complex styles, such as multidimensional composed styles, present unique challenges when approached through activation engineering. However, the advantages it offers, such as enhanced control over the output and the capacity to develop unique styles, significantly outweigh these initial challenges. It is important to note that these methods are not mutually exclusive; they can be combined to leverage each approach's strengths, enhancing our model's overall capability and flexibility.

To the best of our knowledge, this is one of the first studies on steering language models beyond GPT-2 (in our case Alpaca-7B (Taori et al., 2023)). Results should, however, be transferable to any other type of LLM with direct access to hidden layer activations. How to determine the exact influence of the weighting parameter λ (Eq. 3) is still an open question. λ allows for nuanced style steering but, if chosen too large, leads the model to produce nonsense texts. Moreover, this seems to depend on the domain (sentiment, emotion, writing style). We leave this for future research.

Limitations

It was not feasible to derive trained steering vectors for all considered samples since training involves high computational costs and requires a maximal sample length of 50 characters. In contrast, activation-based style vectors could straightforwardly be obtained for every text sample without restrictions. We conducted activation-based experiments on the complete sample set to explore the proposed approach fully. However, to avoid a potential bias towards activation-based style vectors and provide a fair comparison, we also conducted our experiments on the subset of samples that could be considered for both settings.

We evaluated the ability to influence the style of an LLM's output with style vectors using existing sentiment and emotion classifiers. Both classifiers are widely used in practice and have shown state-of-the-art results. However, they are not perfect, and thus, results only show a general tendency. In the future, we plan to conduct studies on individual human perceptions of the text style produced by steered LLMs.

The experiments have a strong focus on sentiment and emotion as style characteristics. Results on the Shakespeare dataset provide evidence that the output of LLMs can also generally be steered towards tone and writing style. This, however, has to be investigated in more depth in the future, especially concerning texts in languages other than English.

Ethics Statement

Our method may generate negative, rude, and hateful sentences about a specific person or a commercial site caused by the data distribution of Yelp and GoEmotions datasets. Therefore, it could be used with malicious intentions, i.e., by targeted harassment or inflation of positive reviews. Since our work involves a pre-trained generative LLM, which was trained on text scraped from the web, it has acquired some biases that were present there. Such biases might be extracted by certain prompts and could even be strengthened by our style steering. Furthermore, it is important to note that steering the style of LLMs may bear the potential to mimic a specific style of speech from persons whose statements were used to train the model. Therefore, the approaches could be abused to create realistic fake statements.

In the context of image generation, the idea of shifting entities in the latent space during the generation process has already been implemented successfully (Brack et al., 2022) and can considerably reduce harmful content in generated images (Schramowski et al., 2023). Analogously, our approach can also be used to reduce harmful output.

Acknowledgements

The authors gratefully acknowledge the computational and data resources provided through the joint high-performance data analytics (HPDA) project "terrabyte" of the German Aerospace Center (DLR) and the Leibniz Supercomputing Center (LRZ).