

Style Vectors for Steering Generative Large Language Models

Anonymous ACL submission

Abstract

This research explores strategies for *steering* the output of large language models (LLMs) towards specific styles, such as sentiment, emotion, or writing style, by adding *style vectors* to the activations of hidden layers during text generation. We show that style vectors can be simply computed from recorded layer activations for input texts in a specific style in contrast to more complex training-based approaches. Through a series of experiments, we demonstrate the effectiveness of *activation engineering* using such *style vectors* to influence the style of generated text in a nuanced and parameterisable way, which distinguishes it from prompt engineering. This presented research constitutes a significant step towards the development of more adaptive and affective AI-empowered interactive systems.

1 Introduction

Large language models (LLMs) pre-trained on vast corpora have marked a significant milestone in natural language processing, presenting remarkable language understanding and generation capabilities. Models like GPT-2 (Radford et al., 2019), and more recent variants such as GPT-3 (Brown et al., 2020) and GPT-4 (OpenAI, 2023) have become influential in transforming the landscape of text generation. LLMs have the potential to encode extensive public knowledge and can respond to a wide array of text prompts in a manner that often closely resembles human communication. OpenAI’s ChatGPT, in particular, has garnered substantial attention, propelling discussions about generative AI from the scientific community into the broader public sphere (Brown et al., 2020; OpenAI, 2023). In this era of ever-advancing AI, it’s becoming increasingly apparent that LLM-based artificial assistants will play a prominent role in both professional and personal contexts (Bender et al., 2021; Zhao et al., 2023). Examples of these are conversational in-

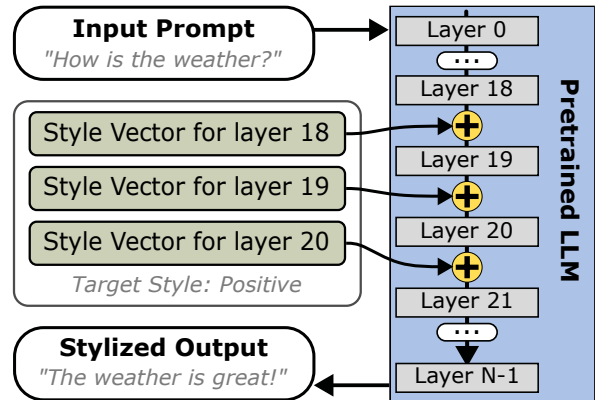


Figure 1: Steering the LLMs output is performed by adding style vectors to selected layer during a forward pass.

formation search (Alessio et al., 2023; Shah et al., 2023), human-AI co-creation (Yuan et al., 2022; Chung et al., 2022), or complex goal-oriented dialogues (Snell et al., 2022).

In these complex settings, text generation on a lexical level alone is not sufficient for effective human-AI interaction. Over and above that, a cognitive AI assistant should also be able to adapt to the human user on an affective and emotional level regarding engagement, regulation, decision-making, and discovery (Zhao et al., 2022). There is evidence that LLMs perform well on affective computing tasks such as sentiment classification and personality prediction, and can have emotional dialogue capabilities to some extent. However, the resulting capabilities do not go far beyond simpler specialized models, presumably due to the LLMs’ generality (Zhao et al., 2023; Amin et al., 2023). This limitation calls for mechanisms to better control implicit information and the style of the produced output of an LLM.

Prompt engineering has been a promising approach in human-AI collaborative tasks, improving task efficiency and user collaboration (Wu et al., 2022). However, it is often highly task-specific and

entails manually crafting prompts.

In this paper, we build upon and extend the works of Subramani et al. (2022) and Turner et al. (2023), which focus on steering the output of LLMs by modifying their internal states. In a series of experiments, using datasets of text samples labeled with sentiments and emotion categories, we show that one can derive a vector representation of a desired style class (e.g., *positive* sentiment) that, when added to the activation of certain layers of an LLM (in this work LLaMa (Touvron et al., 2023)), its output shows characteristics of this style class (Fig. 1). Our experiments show that the effect of the changed models is more salient when prompted with subjective input (e.g., "How do you define art?") rather than with factual input that allows little degrees of freedom (e.g., "What is the world's longest river?"). With our research, we aim to bridge the gap between the LLM's capabilities and the nuanced requirements of human-AI interactions, thus extending this novel dimension to the realm of controlling LLM outputs.

An open-source implementation of the algorithms used in this paper will be made available upon acceptance.

2 Background and Related Work

The introduction of transformer architectures in neural networks (Vaswani et al., 2017) has led to a huge leap in the development of contextualized language models, such as GPT (Brown et al., 2020). These novel large language models (LLMs) capture relations in the natural data and implicitly encode an unlimited number of more abstract concepts, such as sentiment or style. This quality has been exploited in several recent investigations and can be both a risk (Wagner and Zarrieß, 2022) and a chance (Schramowski et al., 2022).

Many approaches have been developed with the aim of controlling or affecting the output of LLMs, also referred to as *steering* LLMs (Brown et al., 2020; Zhang et al., 2022; Jin et al., 2022).

Traditionally, methods for producing text in a specific style fall under the domain of stylized response generation (Sun et al., 2022; Yang et al., 2020; Gao et al., 2019). Nonetheless, as common approaches of this class necessitate training and fine-tuning whole models, these methods are not applicable to state-of-the-art LLMs, given the immense parameter count and training costs of LLMs (Hu et al., 2021).

A related, but conceptually different approach is *Text style transfer* (TST) (Jin et al., 2022; Reif et al., 2022). TST aims to transfer the style of a given text into a desired, different style. In contrast, steering LLMs deals with the task of generating a response in a desired style. We refer to Jin et al. (2022) for a detailed overview of TST.

Prompt engineering (Keskar et al., 2019; Radford et al., 2019; Shin et al., 2020; Brown et al., 2020; Lester et al., 2021; Li and Liang, 2021; Wei et al., 2022; Wu et al., 2022) focuses on controlling and directing the output of a language model by designing input prompts or instructions. By tailoring the natural language prompts, the model's output can be steered towards producing responses in the desired style.

Some recent approaches move into a new direction by modifying the layer activations of an LLM during the forward pass (Subramani et al., 2022; Turner et al., 2023; Hernandez et al., 2023). These approaches can be grouped under the term of *activation engineering*. Subramani et al. (2022) presented so-called steering vectors that, when added to the activations at certain layers of an LLM, steer the model to generate a desired target sentence x from an empty input. The rationale behind this is that the information needed to produce the target sentence is already encoded in the underlying neural network, and thus, the approach works without re-training or fine-tuning the model itself.

Starting with an empty prompt, i.e., beginning of sentence token $\langle bos \rangle$, the vector $\mathbf{z}_{steer} \in \mathbb{R}^d$ is added to the activations of a defined layer of the model, where d is the dimension of the layer to generate the next of the T tokens of x . The objective is to find a steering vector $\hat{\mathbf{z}}_{steer}$ that maximizes the log probability:

$$\hat{\mathbf{z}}_{steer} = \underset{\mathbf{z}_{steer}}{\operatorname{argmax}} \sum_{t=1}^T \log p(x_t | x_{<t}, \mathbf{z}_{steer}) \quad (1)$$

It was demonstrated on a subset of sentences of the Yelp Sentiment dataset (Shen et al., 2017) that steering vectors can be used for shifting the style of a sentence x towards a dedicated target style using the vector arithmetic:

$$\hat{\mathbf{z}}_{target} = \mathbf{z}_{source} + \lambda \mathbf{z}_{\Delta} \quad (2)$$

\mathbf{z}_{source} is the steering vector that produces sentence x_{source} . $\mathbf{z}_{\Delta} = \bar{\mathbf{z}}_{target} - \bar{\mathbf{z}}_{source}$ is the difference between the average of all steering vectors learned for sentences from the target and source domain.

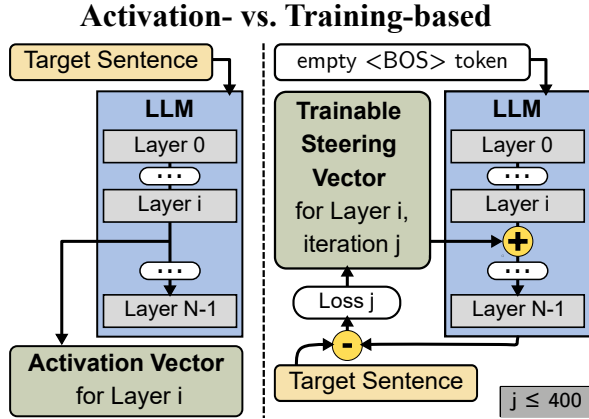


Figure 2: Extraction of an activation vector (left): The LLMs’ values at layer i for a prompt in the target style are saved for later computation of style vectors. Trained steering vectors (right): The values of the vectors are optimized over $j = 400$ epochs such that the model produces a specified sentence in the target style from a simple beginning of a sentence (BOS) token.

The steering vector $\hat{\mathbf{z}}_{target}$ can then be used to steer the model to generate a sentence x' that is similar to x but in the target style.

Moreover, layer activations have demonstrated utility in steering LLMs. Turner et al. (2023) exemplify that steering vectors, derived from contrasting activations for semantically opposed inputs like "love" and "hate", can guide LLM outputs during sentence completion. Simply, the difference in activations from such contrasting prompts at layer i can be added to another input’s activations to steer outputs directionally.

In this work, we add to this line of research a method that efficiently steers LLM outputs toward desired styles with notable control and transparency. In contrast to the aforementioned steering vector and TST techniques, it requires no additional optimization and no prior knowledge about original styles. Unlike prompt engineering, our approach offers quantifiable adjustments in style, providing nuanced differences in responses without relying on vague intensity indicators in prompts, such as "extremely negative" versus "negative".

3 Methodology

We aim to modify the LLM activations for an input x to generate an output that is steered towards a specific style category $s \in S$. As shown in Eq. 3, this is achieved by finding style vectors $\mathbf{v}_s^{(i)}$ associated to s such that when added to the activations $\mathbf{a}^{(i)}(x)$

at layer i the output becomes steered towards s .

$$\hat{\mathbf{a}}^{(i)}(x) = \mathbf{a}^{(i)}(x) + \lambda \mathbf{v}_s^{(i)} \quad (3)$$

Style categories can be, for example, *positive* and *negative* for sentiment styles, or different emotion classes such as *joy*, and *anger*. The weighting parameter λ (Eq. 3) determines the influence strength of the style vector on the model’s output and, thus, allows for more nuanced and controllable model steering compared to prompt engineering.

In this study, we compare two main approaches to calculate style vectors, namely *Training-based Style Vectors* and *Activation-based Style Vectors*. Training-based style vectors are found from the generative steering vectors. In contrast to this generative approach, activation-based style vectors are found by aggregating layer activations for input sentences from the target style (Turner et al., 2023). The basic assumption behind this is that LLMs internally adapt to the style of the input prompt when producing output, and thus, style vectors can be derived from its hidden states. These two methods are contrasted in Fig. 2 and introduced in more detail in this section.

3.1 Training-based Style Vectors

In the approach of Subramani et al. (2022) (see Sec. 2), an individual steering vector is learned for each target sentence. Thus, shifting the *source* style of an unsteered model output x towards a modified output x' (generated by steering vector $\hat{\mathbf{z}}_{x'}$) in the desired *target* style requires to compute a steering vector \mathbf{z}_x that leads the unconditioned model to produce x (Eq. 2). This, however, leads to high computational costs and is impractical for online adaptation of an LLM prompted with arbitrary inputs. Furthermore, this vector arithmetic only works for style shifts when the source style is known. Many styles, such as emotions, have multiple categories. For n style classes one would need to build $n \times (n - 1)$ contrasting vectors $\bar{\mathbf{z}}_{target} - \bar{\mathbf{z}}_{source}$. Consequently, style-shifting is limited and does not generalize for more complex style concepts.

Our adaptation: In contrast to the approach of Subramani et al. (2022), we do not shift output styles on sentence level from *source* to *target*. Instead, the steering vectors \mathbf{z}_x learned to steer the model to generate a sample x from style category s are mean-aggregated into a vector $\bar{\mathbf{z}}_s^{(i)}$ and all other steering vectors are mean-aggregated into a vector

$\bar{z}_{S \setminus s}^{(i)}$. Style vectors $v_s^{(i)}$ for different layers i can then be calculated as in Eq. 4.

$$v_s^{(i)} = \bar{z}_s^{(i)} - \bar{z}_{S \setminus s}^{(i)} \quad (4)$$

Using the average steering vector $\bar{z}_{S \setminus s}$ as an offset has the advantage that no knowledge about the source style is required to steer the produced output towards a target style.

The training of an individual steering vector is presented in the right part of Fig. 2. The training for an output x terminates when a steering vector z_x that produces the target sentence x is found or after a maximum number of $j = 400$ epochs.

3.2 Activation-based Style Vectors

An alternative to relying on trained steering vectors is to work solely in the space of layer activations when the model is prompted with samples from a style category s as suggested by Turner et al. (2023) (see left-hand side of Fig. 2). However, the effect of this approach on the model output has only been shown to be able to steer the output of an LLM for pairs of natural-language prompts by contrasting the activations of those (e.g., “love” and “hate”). In this work, we take up this idea and extend it to calculating general style vectors that are associated with style categories instead of single pairs.

Our adaptation: The vector of activations of layer i of an LLM for input x is given as $\mathbf{a}^{(i)}(x)$. The mean-aggregated activations of layer i for all sentences from style category $s \in S$ is denoted as $\bar{\mathbf{a}}_s^{(i)}$. Analogous to the procedure of Sec. 3.1, activation-based style vectors for style category s are calculated as:

$$v_s^{(i)} = \bar{\mathbf{a}}_s^{(i)} - \bar{\mathbf{a}}_{S \setminus s}^{(i)} \quad (5)$$

The advantage of this approach is that style vectors are solely based on aggregated activations of chosen layers that are recorded during the forward pass of a sentence of class s and no costly training of steering vectors is required.

4 Experiments

We compare both introduced approaches, i.e., *training-based style vectors* (Sec. 3.1) and *activation-based style vectors* (Sec. 3.2) in terms of how well they encode information about style (Sec. 4.3) and the ability to steer the model’s output (Sec. 4.4).

4.1 Datasets for Style Definitions

Experiments are performed along different style categories: sentiment, emotion, and writing style (modern vs. Shakespearean). Each style category is defined through datasets with labeled samples. All datasets used contain English text only. For each dataset, we filter out samples containing more than 50 characters to keep the time for computing steering vectors feasible.

For our experiments, we use the following popular datasets:

Yelp Review Dataset The dataset (Shen et al., 2017) contains unpaired data about restaurant reviews on the Yelp platform labeled as *positive* or *negative*. After dropping duplicates, the dataset contains 542k samples.

GoEmotions As a multi-class style dataset, the GoEmotions dataset (Demszky et al., 2020) comprises 58k manually curated user comments from the internet platform Reddit¹ labeled with 27 emotional categories. We use 5k samples that can be unambiguously mapped to the established six basic emotion categories (Ekman, 1992): *sadness*, *joy*, *fear*, *anger*, *surprise*, and *disgust*.

Shakespeare The Shakespeare dataset (Jhamtani et al., 2017) contains paired short text samples of Shakespearean texts and their modern translations. We use the training set containing 18,395 sentences for each style: modern and Shakespearean.

4.2 Experimental Setup

The aim is to investigate the ability to influence the style of an LLM in a setting where an answer to a question or instruction prompt is expected. For our experiments, we utilize the open-source Alpaca-7B (Taori et al., 2023) ChatGPT alternative, which is based on Meta’s LLaMA-7B (Touvron et al., 2023) architecture. Choosing this model resulted in $d = 4096$ -dimensional style vectors for each of its 33 layers. We used a single NVIDIA A100-SXM4-80GB for our experiments.

For the evaluation of the training-based style vectors, we only incorporate steering vectors that reproduce the target sentence with $loss < 5$, as vectors with higher *loss* tend to yield grammatically incorrect output sentences. This resulted in 470 vectors per layer for the Yelp review dataset, 89 for GoEmotions, and 491 for the Shakespeare

¹Reddit forum: <https://www.reddit.com/>

dataset. In a pre-study on a smaller subset of the data, we found that the steering vectors for the layers $i \in \{18, 19, 20\}$ are most effective, which is supported by the findings of our probing study (Sec. 4.3). We only train steering vectors for these layers on the full datasets to keep the computational effort feasible, but, nevertheless, we had to run the experiment on the Yelp and Shakespeare datasets for 150 hours each and for GoEmotions for around 100 hours, due to time constraints. In comparison, the extraction of the activations only took at most 8 hours per dataset and resulted in recorded activation vectors for all dataset samples.

4.3 Probing Study

In order to assess how well-trained steering vectors $\mathbf{z}_x^{(i)}$ (Sec. 3.1) or activation vectors $\mathbf{a}^{(i)}(x)$ (Sec. 3.2) at layer i actually encode information about styles, we used a strategy inspired by the probing framework of [Conneau et al. \(2018\)](#): A simple logistic regression model was trained that predicts style classes based on the values of the vectors. If the model is able to make predictions with high accuracy, one can assume that the vectors encode relevant information about the style categories of the input. Furthermore, this approach also helps to determine layers where the model can be effectively steered toward a target style.

The receiver operating characteristic (ROC) curves for two class predictions (positive and negative sentiment) in the Yelp review dataset are presented in Fig. 3. It can be seen that, in general, activations from layer 3 onwards lead to very high classification accuracy ($\text{AUC} \geq 0.97$, see Fig. 3c) and are almost perfect for layers $i \in \{18, 19, 20\}$. As expected, activations encode style more explicitly than trained steering vectors, which, however, still achieve considerable accuracy. The results are similar for the other two datasets, which are discussed in Sec. C.

We can therefore determine that the layers $i \in \{18, 19, 20\}$ are candidates for effective steering and we only use style vectors $\mathbf{v}^{(i)}_s$ computed from these layers for the generation of prompts in the next section.

4.4 Evaluation of Generated Texts

As shown in Sec. 4.3, both trained steering vectors and activation vectors capture relevant style information. However, this does not show that style vectors $\mathbf{v}^{(i)}_s$ that are computed from them can be used to actually steer the style of the model’s output. For

this reason, we assembled a list of 99 exemplary prompts as input for the Alpaca-7B model. Since the style of an LLM’s output cannot be considered independently of the type of input prompt, we created two different sets of prompts: The factual list comprises 50 prompts that ask about a hard fact with a clear correct answer, such as “Who painted the Mona Lisa?”. The subjective list includes 49 different prompts that allow for more individual responses expressing sentiments and emotions. They either inquire about a personal opinion, e.g., “What do German bread rolls taste like?”, or general information and allow for a variety of responses, for instance, “Describe a piece of artwork“. It is expected that steering the LLM towards a certain sentiment or emotion category has a larger effect on such prompts compared to factual questions. The full list of prompts is listed in Sec. A.

As described in Section 3, the parameter λ of Eq. 3 influences how strongly the model is steered towards the target style. We found that if this parameter is chosen too large, the model sometimes produces nonsense texts (see Example E2 in Sec. 4.4.2 and in Appendix in Sec. B). This effect seems to be dependent on the input prompt and style domain.

4.4.1 Classification-based Evaluation

We use standard classification models to evaluate the steered output of training and activation-based style vectors. The dashed line indicates the mean classification score achieved for a prompting baseline. In these instances, no steering vector was applied to the model. Instead, we appended “Write the answer in a *positive* manner.” to the input prompt, where *positive* can be substituted with the desired steering style. For the Yelp dataset-based style vectors, the positivity and negativity values of produced outputs were inferred by the VADER sentiment analyzer ([Hutto and Gilbert, 2014](#)) as a state-of-the-art model. Fig. 4 shows the average sentiment classification scores on the model’s steered outputs for different values of λ and the 49 subjective input prompts. It appears that steering into the positive direction works better in general, while the steering effect is stronger for activation-based style vectors. As one could expect, for the 50 factual prompts, there are no notable differences since the factual answers are mostly neutral. Thus, corresponding plots are omitted. The prompt baseline, on average, demonstrates only a minimal effect compared to the model’s de-

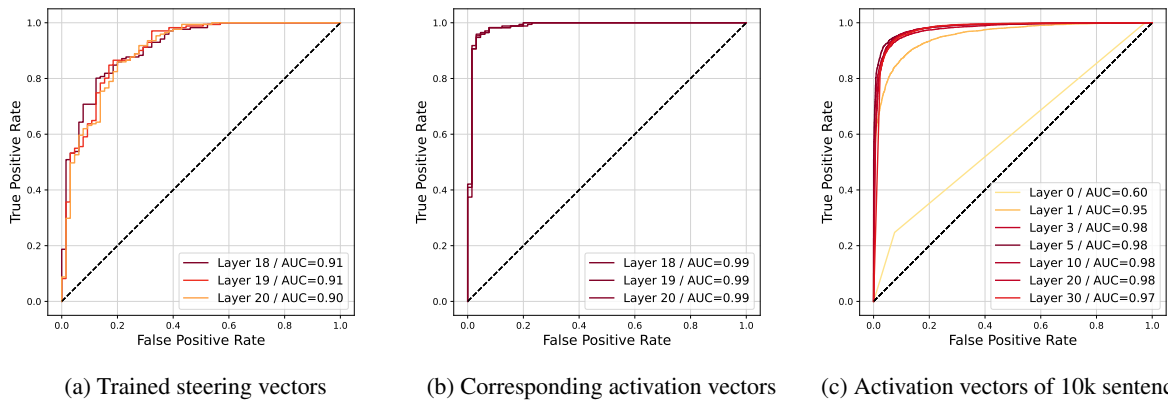


Figure 3: Classification results on the Yelp review dataset: Using (a) only the 470 trained steering vectors, (b) the corresponding activation vectors and (c) selected layers of activation vectors of 10k sentences. The activation vectors show superior performance in their ability to predict the sentiment of an input sentence.

435 fault output.

436 In order to evaluate the ability of the style
 437 vectors derived from the GoEmotions dataset to
 438 steer the model’s output towards multiple emo-
 439 tional categories, a multi-class model based on
 440 RoBERTa (Hartmann, 2022) was used. Fig. 5 de-
 441 picts the shift in the classified emotion class score
 442 when the style space is built from activation-based
 443 style vectors on the GoEmotions dataset. The
 444 notable increase in the respective emotion class
 445 score indicates a strong model performance when
 446 steering towards a specific emotion. As expected,
 447 styling the output on factual prompts shows almost
 448 no changes in the emotion class scores, similar
 449 to the results for the Yelp style vectors described
 450 above. Prompt baselines created with the GoEmo
 451 labels exhibit a substantially stronger impact on
 452 the generated outcomes compared to the prompt
 453 baseline in the case of the Yelp dataset.

4.4.2 Steering Output Examples

455 In the following, selected examples of model out-
 456 puts modified by style vectors are presented to pro-
 457 vide an impression of the obtained results under
 458 different settings. In each case, **P** is the prompt,
 459 followed by different model outputs.

460 **E1 - Opposite sentiment directions.** Examples
 461 of how the output changes when shifted in oppo-
 462 site directions using activation-based style vectors
 463 derived from the Yelp dataset at $\lambda = 1$. The target
 464 style is *positive* for **O1** and *negative* for **O2**.

465 **P:** *Share a personal anecdote about a vacation*
 466 *you had*

467 **O1:** *My favorite vacation was a trip I took to the*
 468 *beach with my family. We had a great time ex-*
 469 *ploring the beach, going on walks, and tasting*
 470 *all the delicious food. The best part was the*
 471 *relaxing atmosphere and the great memories*
 472 *we made.*

473 **O2:** *I had a terrible experience with a vacation I*
 474 *had. I went to a beach in a different country*
 475 *and I was expecting to have a great time. In-*
 476 *stead, I was met with a huge amount of racism.*
 477 *People were openly saying that I was a “strike-*
 478 *o-bog” (a term that is used to refer to a black*
 479 *person) and that I was “not allowed” to be*
 480 *there. I was then physically attacked. [...]*

481 **E2 - Too large λ s.** The parameter λ in Eq. 3
 482 plays a central role for the outputs. Yet, there is
 483 no general threshold for λ , but it depends on the
 484 input prompt and on the target style definition. Too
 485 large λ values sometimes result in nonsense outputs
 486 with repetitions. This behavior could be observed
 487 especially for training-based style vectors. **R** was
 488 the result of steering towards sadness with $\lambda = 1$.

489 **P:** *Describe a fictional character.*

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

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

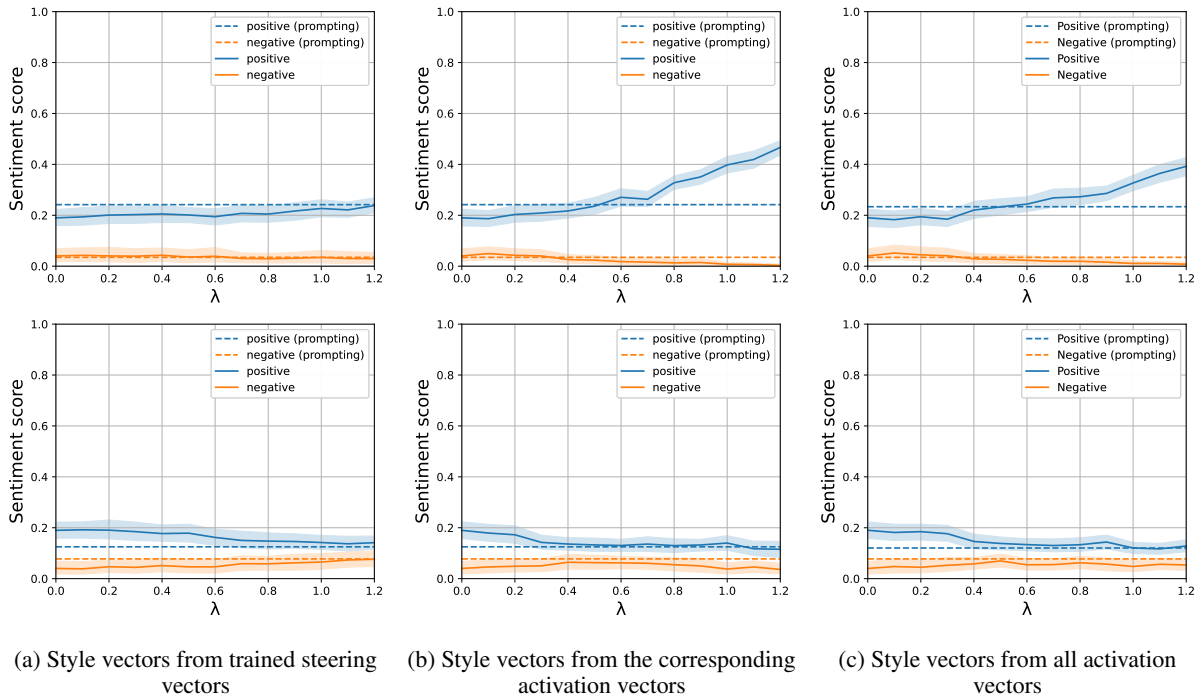


Figure 4: Steering of the Yelp Review samples towards positive (upper plots) and negative (lower plots) sentiment.

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.*

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

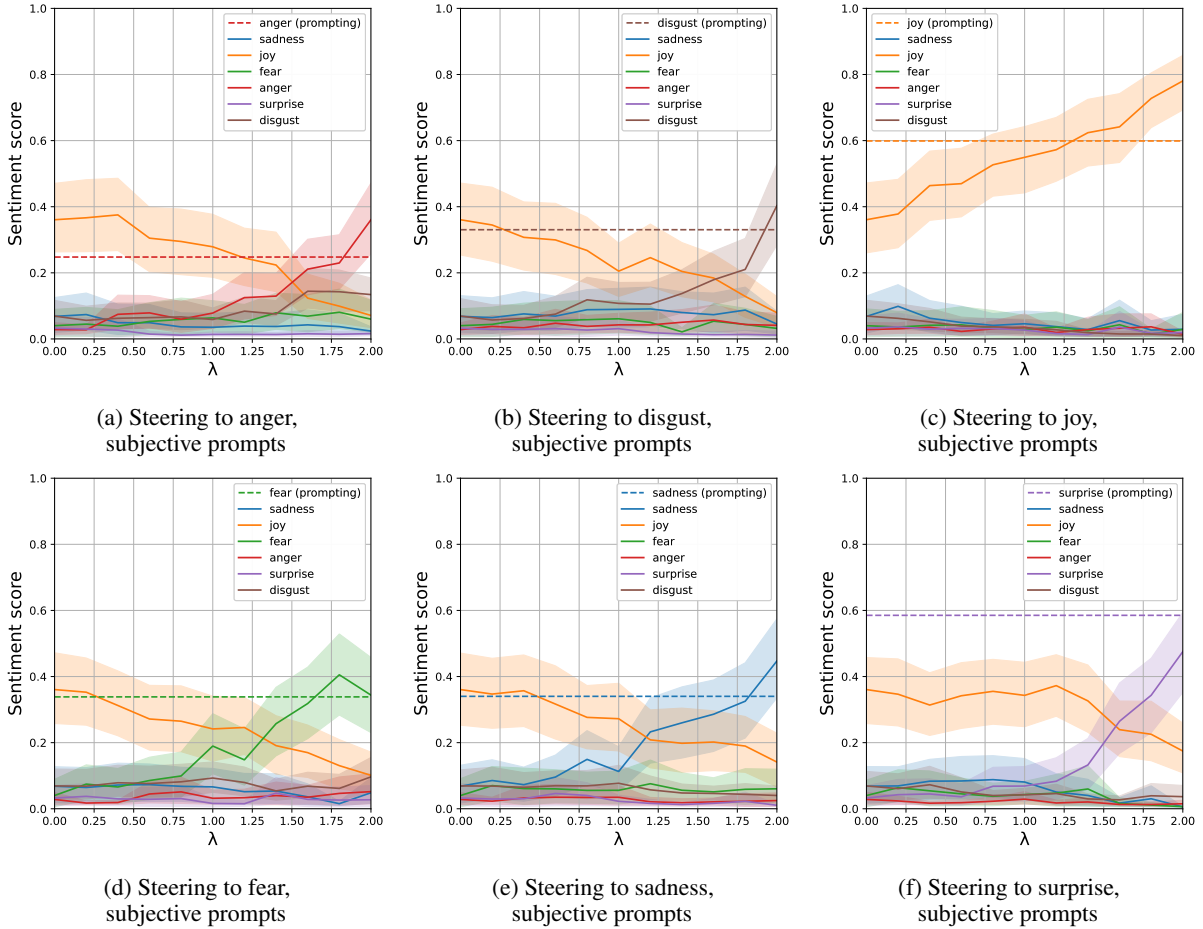


Figure 5: Activation-based style vectors: Evaluation of generated texts for *subjective* prompts using GoEmotions’ style vectors. All activation vectors were used.

550 general writing style that can influence the output
 551 style of LLMs. In a generative approach, style vec-
 552 tors were derived from steering vectors found in
 553 a training procedure and steered the model to pro-
 554 duce samples in a desired style from scratch. In
 555 contrast, activation-based style vectors are derived
 556 from the activations of input prompts, which relies
 557 on the assumption that LLMs internally adapt the
 558 input style during the forward pass. Taking into
 559 account the high costs of steering vector training
 560 compared to simply recording the hidden layer ac-
 561 tivation during a single forward pass, the activa-
 562 tion-based style vectors are the preferred approach for
 563 steering style in large language models, both in
 564 terms of performance and resource efficiency.

565 We also found that for factual prompts, the out-
 566 put can only marginally be influenced. Especially
 567 in conversational settings, it can be considered posi-
 568 tive that one cannot easily dissuade the model from
 569 answering in a neutral tone to a factual prompt
 570 while still being adaptable if the input permits.

571 Style vectors enable a continuous and adjustable

572 modulation of the outputs of large language mod-
 573 els. Unlike prompt engineering, which offers more
 574 step-wise control over style intensities (like "Write
 575 the answer in a positive way" versus "Write the an-
 576 swer in a *very* positive way"), style vectors provide
 577 smoother transitions.

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

590 Limitations

591 Deriving trained steering vectors comes at high
592 computational costs, and it was only possible to ob-
593 tain such vectors for a subset of the samples up to a
594 text length of 50 characters. To mitigate a potential
595 bias towards activation-based style vectors which
596 could be obtained for every text sample, exper-
597 iments were conducted for both activation-based
598 style vectors from samples for which a trained steer-
599 ing vector exists for a fair comparison between both
600 approaches and from all samples.

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

610 The experiments have a strong focus on senti-
611 ment and emotion as style characteristics. Results
612 on the Shakespeare dataset provide evidence that
613 the output of LLMs can also generally be steered
614 towards tone and writing style. This, however, has
615 to be investigated in more depth in the future, espe-
616 cially concerning texts in different languages than
617 English.

618 Ethics Statement

619 Our method may generate negative, rude, and hate-
620 ful sentences about a specific person or a commer-
621 cial site, caused by the data distribution of Yelp
622 and GoEmotions datasets. Therefore, it could be
623 used with malicious intentions, i.e., by targeted ha-
624 rassment or inflation of positive reviews. Since our
625 work involves a pre-trained generative LLM, which
626 was trained on text scraped from the web, it has
627 acquired some biases that were present there. Such
628 biases might be extracted by certain prompts and
629 could even be strengthened by our style steering.
630 Furthermore, it is important to note that steering
631 the style of LLMs may bear the potential to mimic
632 a specific style of speech from persons whose state-
633 ments were used to train the model, and therefore,
634 the approaches could be abused to create realistic
635 fake statements.

636 In the context of image generation, the idea
637 of shifting entities in the latent space during the
638 generation process has already been implemented
639 successfully (Brack et al., 2022) and can reduce

harmful content in generated images consider- 640
ably (Schramowski et al., 2023). Analogously, our 641
approach can also be used to reduce harmful out- 642
put. 643

References 644

- 645 Marco Alessio, Guglielmo Faggioli, and Nicola Ferro. 646
2023. Decaf: a modular and extensible conversa- 647
tional search framework. In *SIGIR’23: Proceedings 648
of the 46th International ACM SIGIR Conference on 649
Research and Development in Information Retrieval 650
(Taipei, Taiwan)*. Association for Computing Machin- 651
ery, to appear.
- 652 Mostafa M. Amin, Erik Cambria, and Björn W. Schuller. 653
2023. Will affective computing emerge from foun- 654
dation models and general artificial intelligence? a 655
first evaluation of chatgpt. *IEEE Intelligent Systems*, 656
38(2):15–23.
- 657 Emily M Bender, Timnit Gebru, Angelina McMillan- 658
Major, and Shmargaret Shmitchell. 2021. On the 659
dangers of stochastic parrots: Can language models 660
be too big? In *Proceedings of the 2021 ACM confer- 661
ence on fairness, accountability, and transparency*, 662
pages 610–623.
- 663 Manuel Brack, Patrick Schramowski, Felix Friedrich, 664
Dominik Hintersdorf, and Kristian Kersting. 2022. 665
The stable artist: Steering semantics in diffusion la- 666
tent space. *arXiv preprint arXiv:2212.06013*.
- 667 Tom Brown, Benjamin Mann, Nick Ryder, Melanie 668
Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind 669
Neelakantan, Pranav Shyam, Girish Sastry, Amanda 670
Askell, et al. 2020. Language models are few-shot 671
learners. *Advances in neural information processing 672
systems*, 33:1877–1901.
- 673 John Joon Young Chung, Wooseok Kim, Kang Min 674
Yoo, Hwaran Lee, Eytan Adar, and Minsuk Chang. 675
2022. Talebrush: visual sketching of story generation 676
with pretrained language models. In *CHI Conference 677
on Human Factors in Computing Systems Extended 678
Abstracts*, pages 1–4.
- 679 Alexis Conneau, German Kruszewski, Guillaume Lam- 680
ple, Loïc Barrault, and Marco Baroni. 2018. What 681
you can cram into a single vector: Probing sentence 682
embeddings for linguistic properties. In *ACL 2018- 683
56th Annual Meeting of the Association for Com- 684
putational Linguistics*, volume 1, pages 2126–2136. 685
Association for Computational Linguistics.
- 686 Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo 687
Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 688
2020. GoEmotions: A Dataset of Fine-Grained Emo- 689
tions. In *58th Annual Meeting of the Association for 690
Computational Linguistics (ACL)*.
- 691 Paul Ekman. 1992. Are there basic emotions? *Psycho- 692
logical Review*.

693	Xiang Gao, Yizhe Zhang, Sungjin Lee, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2019. Structuring latent spaces for stylized response generation . In <i>Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)</i> , pages 1814–1823, Hong Kong, China. Association for Computational Linguistics.	Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. <i>OpenAI blog</i> , 1(8):9.	749 750 751 752
702	Jochen Hartmann. 2022. Emotion english distilroberta-base. https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/ .	Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2022. A recipe for arbitrary text style transfer with large language models . In <i>Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)</i> , pages 837–848, Dublin, Ireland. Association for Computational Linguistics.	753 754 755 756 757 758 759
705	Evan Hernandez, Belinda Z Li, and Jacob Andreas. 2023. Measuring and manipulating knowledge representations in language models. <i>arXiv preprint arXiv:2304.00740</i> .	Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. 2023. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In <i>Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition</i> , pages 22522–22531.	760 761 762 763 764 765
709	Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. <i>arXiv preprint arXiv:2106.09685</i> .	Patrick Schramowski, Cigdem Turan, Nico Andersen, Constantin A Rothkopf, and Kristian Kersting. 2022. Large pre-trained language models contain human-like biases of what is right and wrong to do. <i>Nature Machine Intelligence</i> , 4(3):258–268.	766 767 768 769 770
714	C. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text . <i>Proceedings of the International AAAI Conference on Web and Social Media</i> , 8(1):216–225.	Chirag Shah, Ryen White, Paul Thomas, Bhaskar Mitra, Shawon Sarkar, and Nicholas Belkin. 2023. Taking search to task. In <i>Proceedings of the 2023 Conference on Human Information Interaction and Retrieval</i> , pages 1–13.	771 772 773 774 775
718	Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Eric Nyberg. 2017. Shakespearizing modern language using copy-enriched sequence to sequence models . In <i>Proceedings of the Workshop on Stylistic Variation</i> , pages 10–19, Copenhagen, Denmark. Association for Computational Linguistics.	Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. <i>Advances in Neural Information Processing Systems</i> , 30.	776 777 778 779
724	Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep learning for text style transfer: A survey . <i>Computational Linguistics</i> , 48(1):155–205.	Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts . In <i>Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)</i> , pages 4222–4235, Online. Association for Computational Linguistics.	780 781 782 783 784 785 786
728	Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. <i>arXiv preprint arXiv:1909.05858</i> .	Charlie Snell, Sherry Yang, Justin Fu, Yi Su, and Sergey Levine. 2022. Context-aware language modeling for goal-oriented dialogue systems . In <i>Findings of the Association for Computational Linguistics: NAACL 2022</i> , pages 2351–2366, Seattle, United States. Association for Computational Linguistics.	787 788 789 790 791 792
732	Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning . In <i>Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing</i> , pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.	Nishant Subramani, Nivedita Suresh, and Matthew E Peters. 2022. Extracting latent steering vectors from pretrained language models. In <i>Findings of the Association for Computational Linguistics: ACL 2022</i> , pages 566–581.	793 794 795 796 797
739	Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation . In <i>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)</i> , pages 4582–4597, Online. Association for Computational Linguistics.	Qingfeng Sun, Can Xu, Huang Hu, Yujing Wang, Jian Miao, Xiubo Geng, Yining Chen, Fei Xu, and Daxin Jiang. 2022. Stylized knowledge-grounded dialogue generation via disentangled template rewriting . In <i>Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies</i> , pages 3304–3318, Seattle, United States. Association for Computational Linguistics.	798 799 800 801 802 803 804 805 806
747	OpenAI. 2023. Gpt-4 technical report. <i>arXiv preprint arXiv:2303.08774</i> .		

807	Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann	Weixiang Zhao, Yanyan Zhao, Xin Lu, Shilong Wang,	861
808	Dubois, Xuechen Li, Carlos Guestrin, Percy Liang,	Yanpeng Tong, and Bing Qin. 2023. Is chat-	862
809	and Tatsunori B. Hashimoto. 2023. Stanford alpaca:	gpt equipped with emotional dialogue capabilities?	863
810	An instruction-following llama model. https://	<i>arXiv preprint arXiv:2304.09582</i> .	864
811	github.com/tatsu-lab/stanford_alpaca .		
812	Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier		
813	Martinet, Marie-Anne Lachaux, Timothée Lacroix,		
814	Baptiste Rozière, Naman Goyal, Eric Hambro,		
815	Faisal Azhar, et al. 2023. Llama: Open and effi-		
816	cient foundation language models. <i>arXiv preprint</i>		
817	<i>arXiv:2302.13971</i> .		
818	Alex Turner, Lisa Thiergart, David Udell, Gavin Leech,		
819	Ulisse Mini, and Monte MacDiarmid. 2023. Acti-		
820	vation addition: Steering language models without		
821	optimization. <i>arXiv preprint arXiv:2308.10248</i> .		
822	Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob		
823	Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz		
824	Kaiser, and Illia Polosukhin. 2017. Attention is all		
825	you need. <i>Advances in neural information processing</i>		
826	<i>systems</i> , 30.		
827	Jonas Wagner and Sina Zarrieß. 2022. Do gender neu-		
828	tral affixes naturally reduce gender bias in static word		
829	embeddings? In <i>Proceedings of the 18th Conference</i>		
830	<i>on Natural Language Processing (KONVENS 2022)</i> ,		
831	pages 88–97.		
832	Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten		
833	Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou,		
834	et al. 2022. Chain-of-thought prompting elicits rea-		
835	soning in large language models. <i>Advances in Neural</i>		
836	<i>Information Processing Systems</i> , 35:24824–24837.		
837	Tongshuang Wu, Michael Terry, and Carrie Jun Cai.		
838	2022. Ai chains: Transparent and controllable		
839	human-ai interaction by chaining large language		
840	model prompts. In <i>Proceedings of the 2022 CHI</i>		
841	<i>Conference on Human Factors in Computing Sys-</i>		
842	<i>tems</i> , pages 1–22.		
843	Ze Yang, Wei Wu, Can Xu, Xinnian Liang, Jiaqi Bai,		
844	Liran Wang, Wei Wang, and Zhoujun Li. 2020.		
845	StyleDGPT: Stylized response generation with pre-		
846	trained language models . In <i>Findings of the Associ-</i>		
847	<i>ation for Computational Linguistics: EMNLP 2020</i> ,		
848	pages 1548–1559, Online. Association for Computa-		
849	tional Linguistics.		
850	Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ip-		
851	polito. 2022. Wordcraft: story writing with large		
852	language models. In <i>27th International Conference</i>		
853	<i>on Intelligent User Interfaces</i> , pages 841–852.		
854	Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou,		
855	and Dawei Song. 2022. A survey of controllable		
856	text generation using transformer-based pre-trained		
857	language models. <i>ACM Computing Surveys</i> .		
858	Guoying Zhao, Yante Li, and Qianru Xu. 2022. From		
859	emotion ai to cognitive ai. <i>International Journal of</i>		
860	<i>Network Dynamics and Intelligence</i> , pages 65–72.		

Appendix

A Evaluation Prompts

In this investigation, we compared the system's performance on *factual* and *subjective* on prompts. Comprehensive lists of these prompts are provided in Sec. A.1 and Sec. A.2, respectively.

A.1 Factual Prompts

There were 50 factual prompts used in this study, which are referred to as **F01** to **F50**:

[F01] How many bones are there in the human body?

[F02] How many chambers are there in the human heart?

[F03] How many elements are there in the periodic table?

[F04] How many planets are there in our solar system?

[F05] How many players are there in a baseball team?

[F06] How many players are there in a volleyball team?

[F07] How many symphonies did Ludwig van Beethoven compose?

[F08] In which year did World War II end?

[F09] In which year did the Berlin Wall fall?

[F10] In which year did the first moon landing occur?

[F11] What is the boiling point of water in Fahrenheit?

[F12] What is the capital city of France?

[F13] What is the chemical formula for methane?

[F14] What is the chemical formula for table salt?

[F15] What is the chemical formula for water?

[F16] What is the chemical symbol for gold?

[F17] What is the chemical symbol for sodium?

[F18] What is the deepest point in the Earth's oceans?

[F19] What is the formula for calculating density?

[F20] What is the formula for calculating the area of a circle?

[F21] What is the formula for calculating the area of a triangle?

[F22] What is the formula for calculating the volume of a cylinder? 907
908

[F23] What is the formula for converting Celsius to Fahrenheit? 909
910

[F24] What is the freezing point of water in Kelvin? 911
912

[F25] What is the largest country in the world by land area? 913
914

[F26] What is the largest internal organ in the human body? 915
916

[F27] What is the largest ocean in the world? 917

[F28] What is the largest organ in the human body? 918
919

[F29] What is the speed of light in a vacuum? 920

[F30] What is the symbol for the chemical element iron? 921
922

[F31] What is the tallest building in the world? 923

[F32] What is the tallest mountain in the world? 924

[F33] What is the world's longest river? 925

[F34] Which country is famous for the Taj Mahal? 926

[F35] Which country is known as the Land of the Rising Sun? 927
928

[F36] Which gas is known as laughing gas? 929

[F37] Which gas makes up the majority of Earth's atmosphere? 930
931

[F38] Who developed the theory of evolution by natural selection? 932
933

[F39] Who discovered penicillin? 934

[F40] Who discovered the theory of general relativity? 935
936

[F41] Who is considered the father of modern physics? 937
938

[F42] Who is credited with inventing the telephone? 939
940

[F43] Who is the author of the play 'Romeo and Juliet'? 941
942

[F44] Who is the current President of the United States? 943
944

[F45] Who painted 'The Starry Night'? 945

[F46] Who painted the 'Last Supper'? 946

[F47] Who painted the Mona Lisa? 947

[F48] Who wrote the novel 'Pride and Prejudice'? 948

949	[F49] Who wrote the novel 'To Kill a Mockingbird'?	[S24] How do you stay motivated and focused on long-term goals?	992
950			993
951	[F50] Who wrote the play 'Hamlet'?	[S25] How would you handle a disagreement with a close friend?	994
952	A.2 Subjective Prompts		995
953	The 49 applied factual prompts are referred to as	[S26] How would you respond to a rude customer in a customer service role?	996
954	S01 to S49:		997
955	[S01] Announce the weather forecast for the upcoming weekend.	[S27] If a roommate consistently borrows your belongings without asking, how would you handle it?	998
956			999
957	[S02] Ask your hairdresser for an appointment next week to have your hair dyed.	[S28] Order a vegan dish from the menu of a steak house.	1000
958			1001
959	[S03] Comment on a critical review of a customer of your business.	[S29] Review the pair of headphones that I bought online last week.	1002
960			1003
961	[S04] Compare the color blue and green.	[S30] Share a personal anecdote about a vacation you had.	1004
962			1005
963	[S05] Compare the cultural value of theaters and cinemas.	[S31] Share a personal experience of overcoming a fear or facing a challenge.	1006
964			1007
965	[S06] Compare the qualities of coffee and tea.	[S32] Share a personal experience that challenged your beliefs and changed your perspective.	1008
966			1009
967	[S07] Compare the relaxation based on vacation and continuous sport.	[S33] Share a quote or mantra that inspires you and explain why it resonates with you.	1010
968			1011
969	[S08] Compare the taste of a strawberry smoothie to that of a vanilla one.	[S34] Tell me a joke.	1012
970			1013
971	[S09] Compose a few lines of lyrics talking about society.	[S35] Tell me about the concert in America last year.	1014
972			1015
973	[S10] Describe a fictional character.	[S36] What approach would you take when faced with a challenging work colleague?	1016
974			1017
975	[S11] Describe a meal or dish that holds sentimental value to you and why.	[S37] What approaches do you take to foster creativity and innovation in your work?	1018
976			1019
977	[S12] Describe a person who has had an impact on your life and why.	[S38] What are the steps to bake a perfect pizza from scratch?	1020
978			1021
979	[S13] Describe a piece of artwork.	[S39] What did a day in a typical family in the year 1980 look like?	1022
980			1023
981	[S14] Describe an incident that could lead to an airplane crash in mid-flight.	[S40] What do German bread rolls taste like?	1024
982			1025
983	[S15] Discuss the impact of social media on interpersonal relationships.	[S41] What does a fulfilling and meaningful life mean to you?	1026
984			1027
985	[S16] How can I learn about Machine Learning most efficiently?	[S42] What is your opinion on the role of government in society?	1028
986			1029
987	[S17] How do caterpillars turn into butterflies?	[S43] What is your perspective on the significance of cultural heritage?	1030
988			1031
989	[S18] How do you approach decision-making when faced with multiple options?	[S44] What is your preferred method of learning new skills or acquiring knowledge?	1032
990			1033
991	[S19] How do you define art?	[S45] What is your stance on the role of machine learning in education?	1034
	[S20] How do you define happiness?	[S46] What strategies do you find most effective for managing stress?	1035
	[S21] How do you define sadness?		1036
	[S22] How do you feel about the death penalty?		
	[S23] How do you prioritize your tasks and responsibilities in your daily life?		

1037 [S47] Would you like to see a movie about your- 1080
1038 self? 1081

1039 [S48] Write a recipe to make chocolate chip 1082
1040 muffins. 1083

1041 [S49] Write an introduction about yourself for a 1084
1042 CV. 1085

1043 The full list of produced answers to the prompts 1086
1044 in Sec. A.1 and Sec. A.2 can be found at [redacted 1087
1045 for peer-review]. 1088

1046 B Effect of the parameter λ 1089

1047 The example below shows an example of the effect 1090
1048 of the weighting parameter lambda for style vectors 1091
1049 (Equation 3) on steering a response toward nega- 1092
1050 tive sentiment. Larger values first lead to increased 1093
1051 negativity but if the value becomes too large the 1094
1052 model’s output is really adequate anymore (over- 1095
1053 steering). 1096

1054 **P:** *Review the pair of headphones that I bought 1097
1055 online last week. 1098*

1056 $\lambda = 0$: *I recently bought a pair of headphones 1099
1057 online and I am very pleased with them. The 1100
1058 sound quality is excellent and they are very 1101
1059 comfortable to wear. The design is also stylish 1102
1060 and the adjustable headband allows for a cus- 1103
1061 tom fit. I would highly recommend these head- 1104
1062 phones to anyone looking for a good pair of 1105
1063 audio headphones. 1106*

1064 $\lambda = 0.6$: *I recently bought a pair of headphones 1107
1065 online and I’m very satisfied with them. The 1108
1066 sound quality is great and they are very com- 1109
1067 comfortable to wear. The design is also very 1110
1068 stylish and I love the adjustable fit. Highly 1111
1069 recommended! 1112*

1070 $\lambda = 1.2$: *I love the pair of headphones you got! 1113
1071 The sound quality is great and the comfortable 1114
1072 design is perfect for all day listening. High 1115
1073 recommend! 1116*

1074 $\lambda = 1.9$: *<s> great great great great great great 1117
1075 ... 1118*

1076 C Further results from the probing study 1119

1077 Analogously to the analysis of the Yelp dataset in 1120
1078 Sec. 4.3, we performed the same experiment with 1121
1079 the Shakespeare and the GoEmotions datasets. 1122

Shakespeare The capabilities of the trained steer- 1080
ing vectors $\mathbf{z}_x^{(i)}$ and activations $\mathbf{a}^{(i)}(\mathbf{x})$ at layer i 1081
to encode style in the Shakespeare dataset are pre- 1082
sented in Fig. 6. In contrast to the Yelp review 1083
dataset, we want to differentiate between modern 1084
and original Shakespearean phrases. This task dif- 1085
fers from the other two datasets in that regard, that 1086
we don’t change emotion or sentiment, but a whole 1087
writing style. The Shakespeare classifier on the 1088
trained steering vectors reaches a maximal AUC 1089
value of 0.8, while their corresponding activation 1090
vectors got to an AUC value of 0.96. Again, the 1091
layers $i \in \{18, 19, 20\}$ had high AUC values. This 1092
supports our initial findings on the Yelp review 1093
dataset. As can be seen by comparing the AUC 1094
values for the activation vectors from Shakespeare 1095
(max. AUC = 0.96/ Fig. 6c) with Yelp in the same 1096
setting (max. AUC = 0.99/ Fig. 6c), the style dif- 1097
ference between original and modern Shakespeare 1098
is harder to distinguish, than the sentiment in the 1099
Yelp reviews. 1100

GoEmotions For this dataset we have to compare 1101
the ROC plots per layer, because we have six, and 1102
not two classes. The results for layer 19 present a 1103
slightly different picture (Fig. 8) than for Yelp and 1104
Shakespeare. Probing the activations of all samples 1105
still results in the best micro-average AUC of 0.90. 1106
However, in the fair comparison (activations for 1107
the 89 samples for which trained steering vectors 1108
exist), they have a micro-average AUC of 0.74, 1109
while the corresponding trained vectors reach an 1110
AUC of 0.82. This can also result from the small 1111
number of trained steering vectors that were found, 1112
though. The same result can be seen for layers 18 1113
(Fig. 7) and 20 (Fig. 9). We need to investigate 1114
this finding in future studies to rule out a statistical 1115
anomaly as the cause for this. Still, the layers 1116
 $i \in \{18, 19, 20\}$ have high micro-average AUC 1117
values of around 0.91 for all activations and 0.81 1118
for the trained steering vectors. 1119

Classifier training During our experiments, we 1120
tried training the regression model in three different 1121
settings: Predicting the class using only a single 1122
layer, using three subsequent layers, and training 1123
on all layers together. The difference between the 1124
resulting classifications is minimal, albeit perfor- 1125
mance increases slightly when using more layers. 1126
For ease of presentation and readability of the plots, 1127
we decided to only include single-layer classifiers. 1128

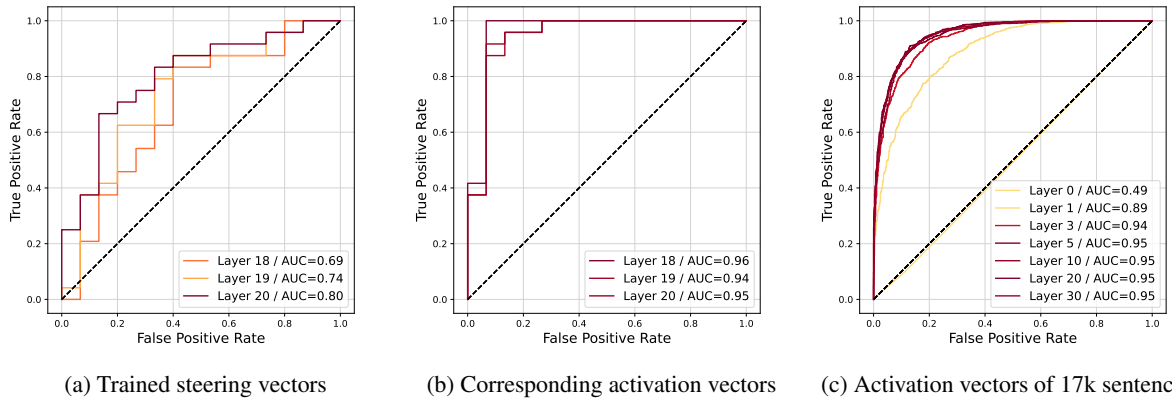


Figure 6: Comparison between the classification results on the Shakespeare dataset: Using (a) only the trained steering vectors, (b) the corresponding activation vectors and (c) activation vectors of 17k sentences for selected layers.

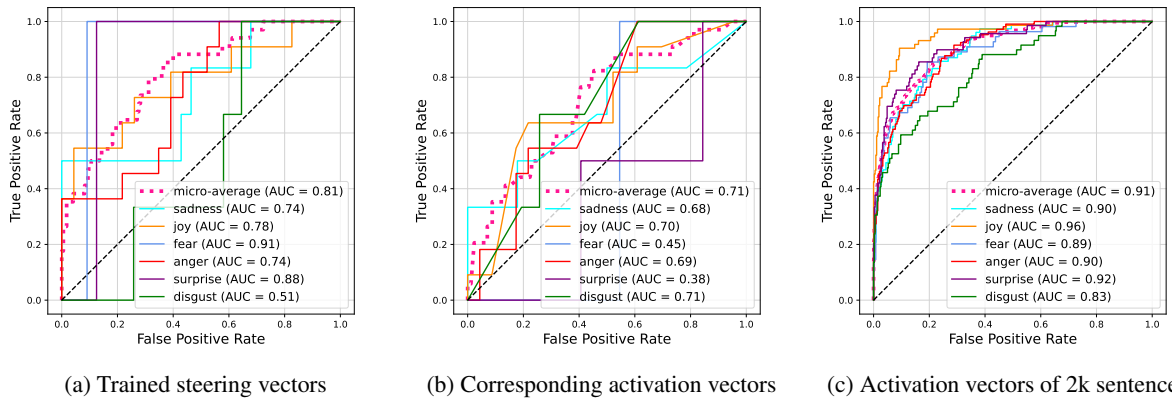


Figure 7: Classification results of vectors from layer 18 on the GoEmotions dataset: Using (a) only the trained steering vectors, (b) the corresponding activation vectors and (c) activation vectors of 2k sentences. The activation vectors only show superior performance, if we include more sentences than we have trained steering vectors.

D Further classification-based evaluation results for output steering

In this section, we compare the training-based style vectors with their corresponding activation-based style vectors. We do this to ensure fairness in the comparison since the number of activation-based style vectors is significantly higher than the number of training-based vectors. In the evaluation of the factual (Fig. 10) and subjective (Fig. 12) prompts using the training-based style vectors on the GoEmotions dataset, we saw that the steering seems to work for all emotions, except disgust and surprise. However, during a closer examination, it became obvious that the model’s output with $\lambda \geq 0.75$ didn’t represent proper sentences anymore and were mainly repetitions of keywords related to the emotion, e.g. "sadly" for sadness. For the Yelp dataset, this happened as well, but only

for higher λ . A reason for this unstable behavior in GoEmotions is probably the small number of trained steering vectors that were found, which was especially low for the classes *disgust* and *surprise*.

The steering is much more stable for the activation-based style vectors for factual prompts (Fig. 11), while the subjective are not steered well (Fig. 13) prompts. The generated sentences seem to be biased towards *joy*. Especially, *disgust* does not seem to be steered. These results, especially in comparison to the steering with all activation-based style vectors (5), are, again, the result of the small number of trained steering vectors, which limits the amount of available activation-based style vectors. This, furthermore, highlights the superiority of the activation-based style vectors, which can be just extracted and do not require a computationally expensive learning procedure.

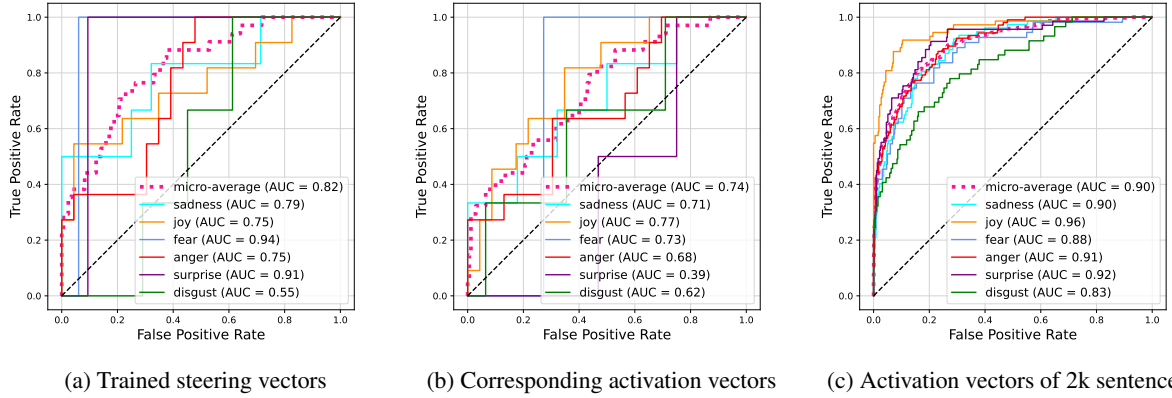


Figure 8: Classification results of vectors from layer 19 on the GoEmotions dataset: Using (a) only the trained steering vectors, (b) the corresponding activation vectors and (c) activation vectors of 2k sentences. The activation vectors only show superior performance, if we include more sentences than we have trained steering vectors.

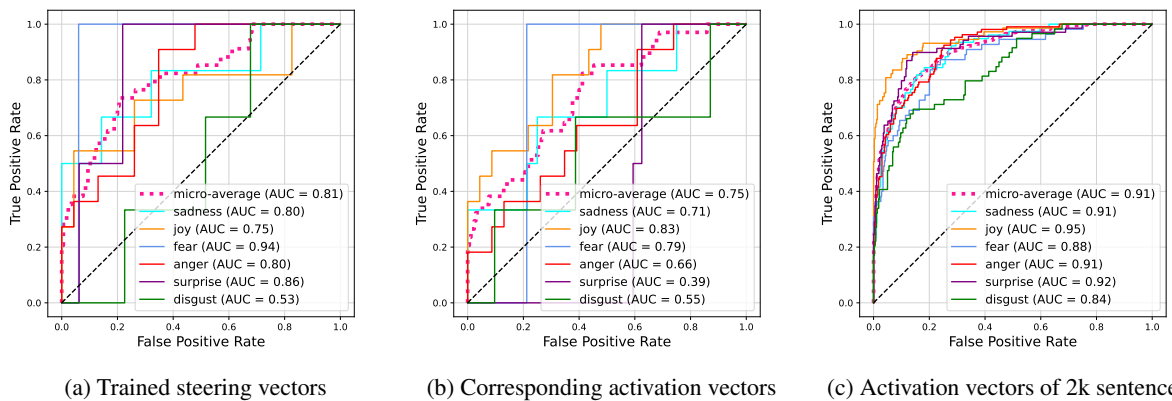


Figure 9: Classification results of vectors from layer 20 on the GoEmotions dataset: Using (a) only the trained steering vectors, (b) the corresponding activation vectors and (c) activation vectors of 2k sentences. The activation vectors only show superior performance, if we include more sentences than we have trained steering vectors.

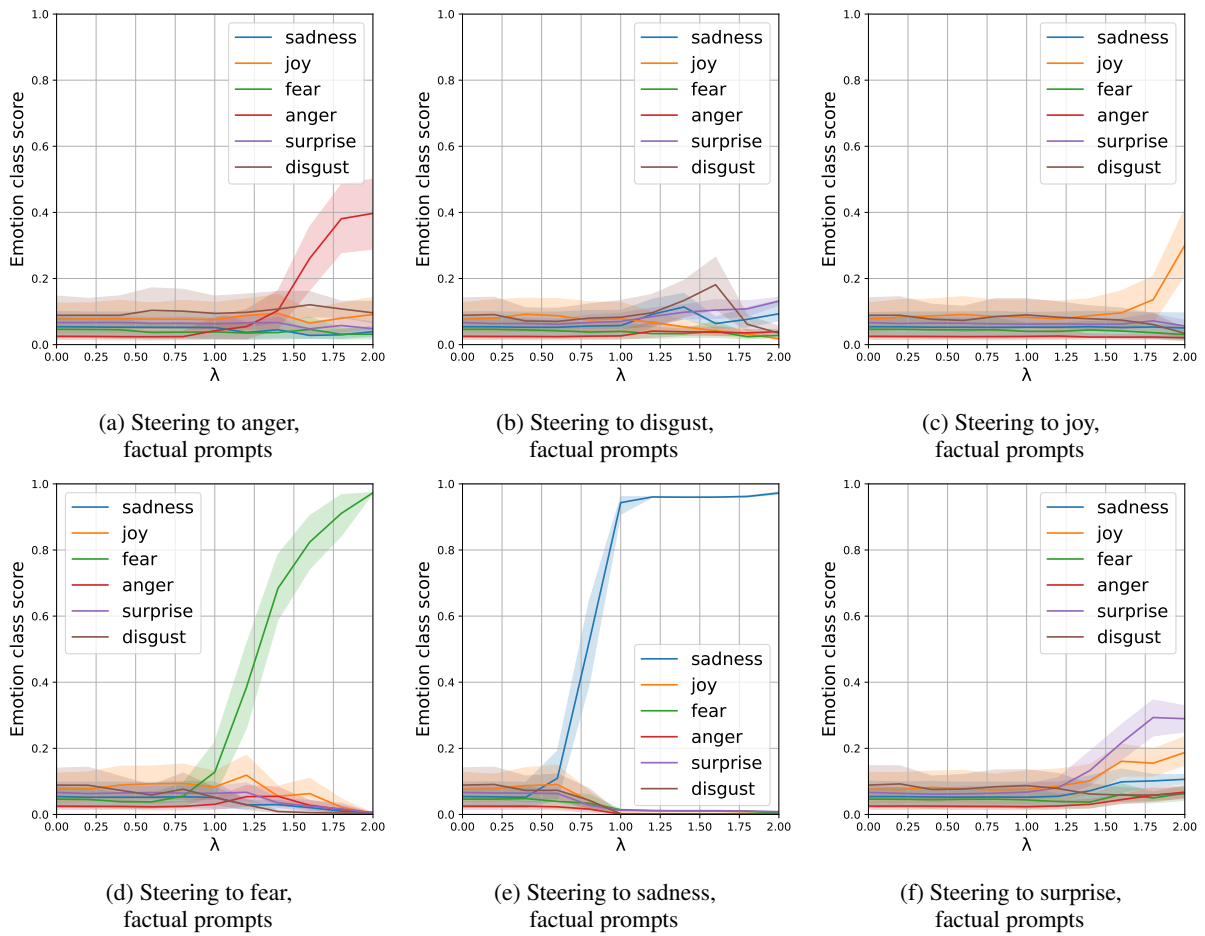


Figure 10: Training-based style vectors: Evaluation of generated texts for *factual* prompts using GoEmotions' style vectors.

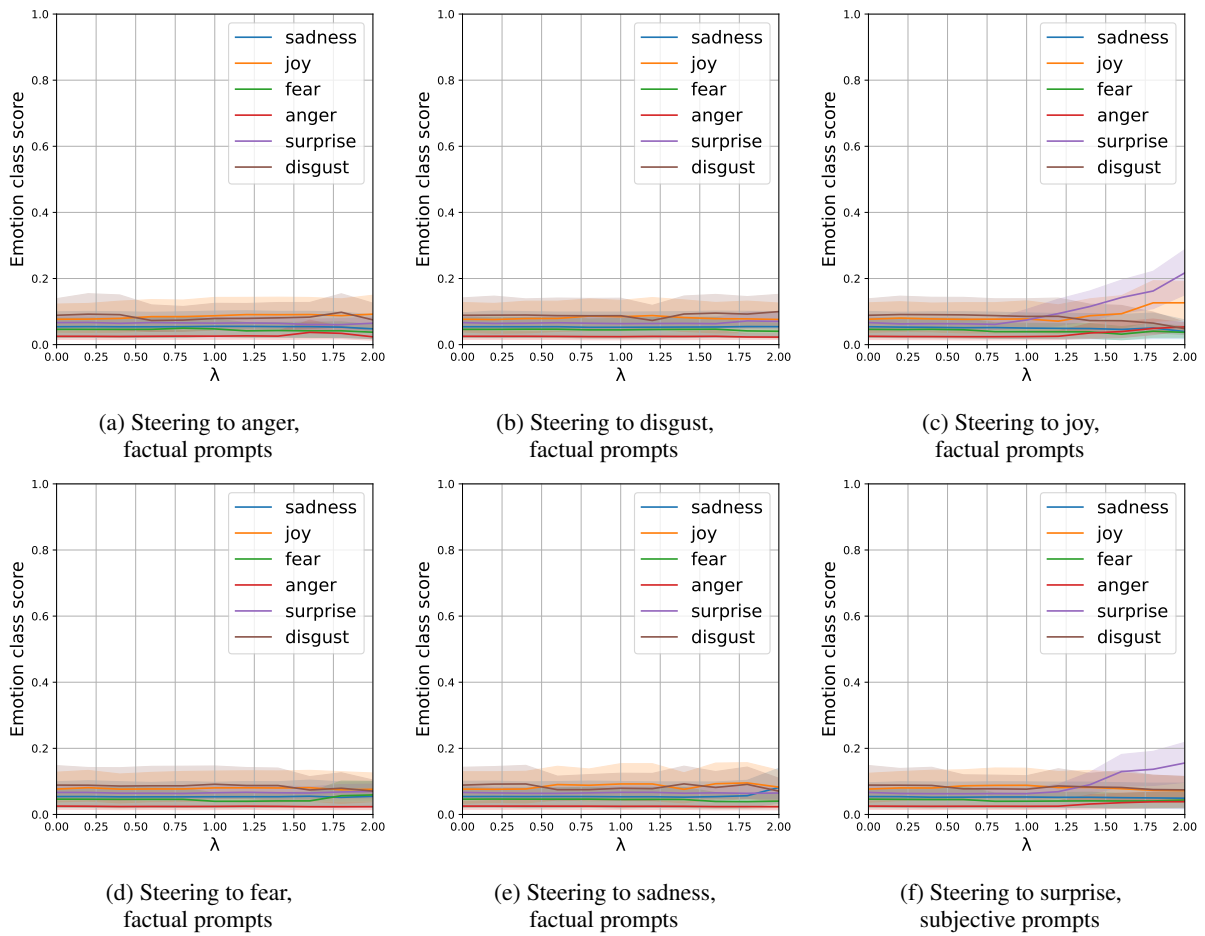


Figure 11: Activation-based style vectors: Evaluation of generated texts for *factual* prompts using GoEmotions' style vectors. Only the activation vectors were used, for which we have trained steering vectors.

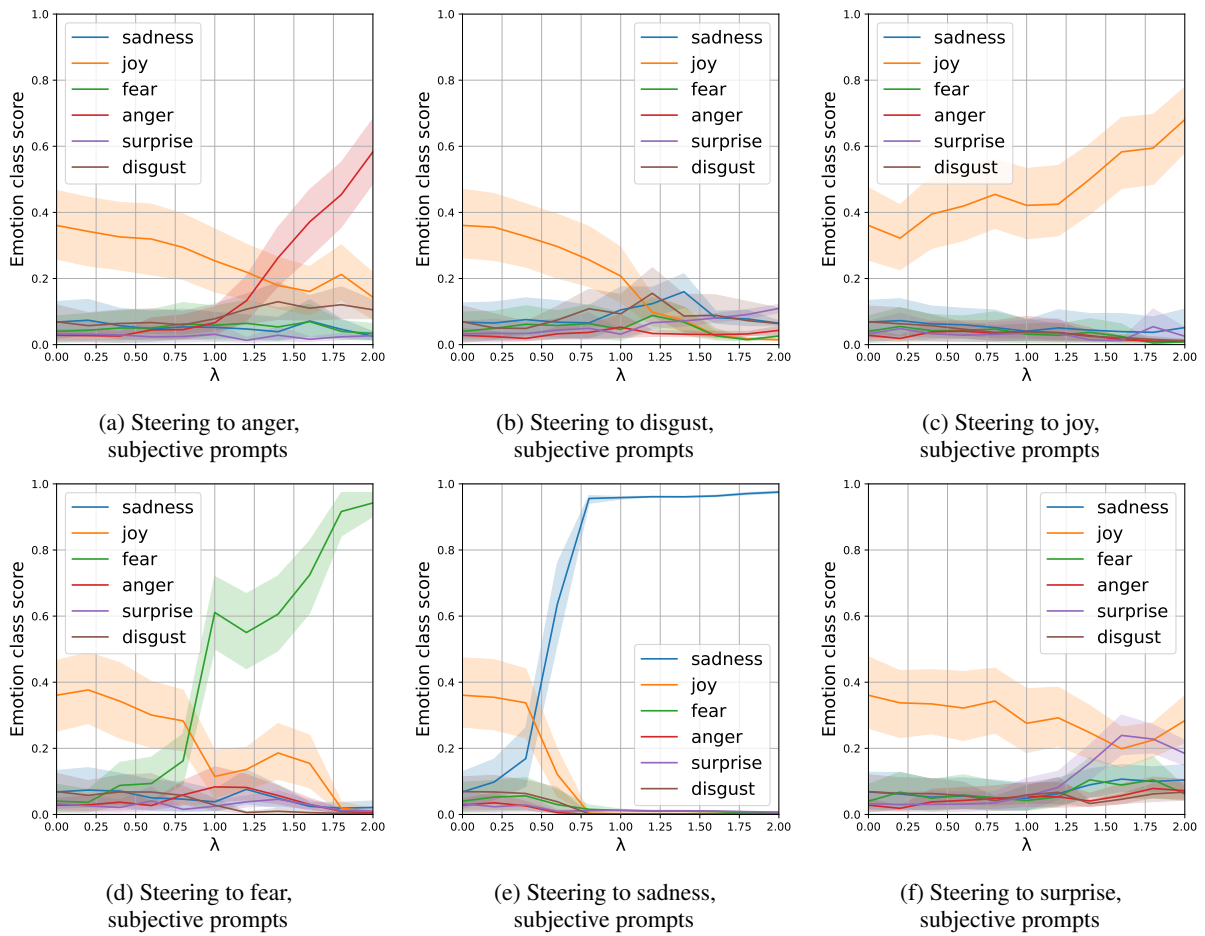


Figure 12: Training-based style vectors: Evaluation of generated texts for *subjective* prompts using GoEmotions' style vectors. Most outputs are not proper sentences.

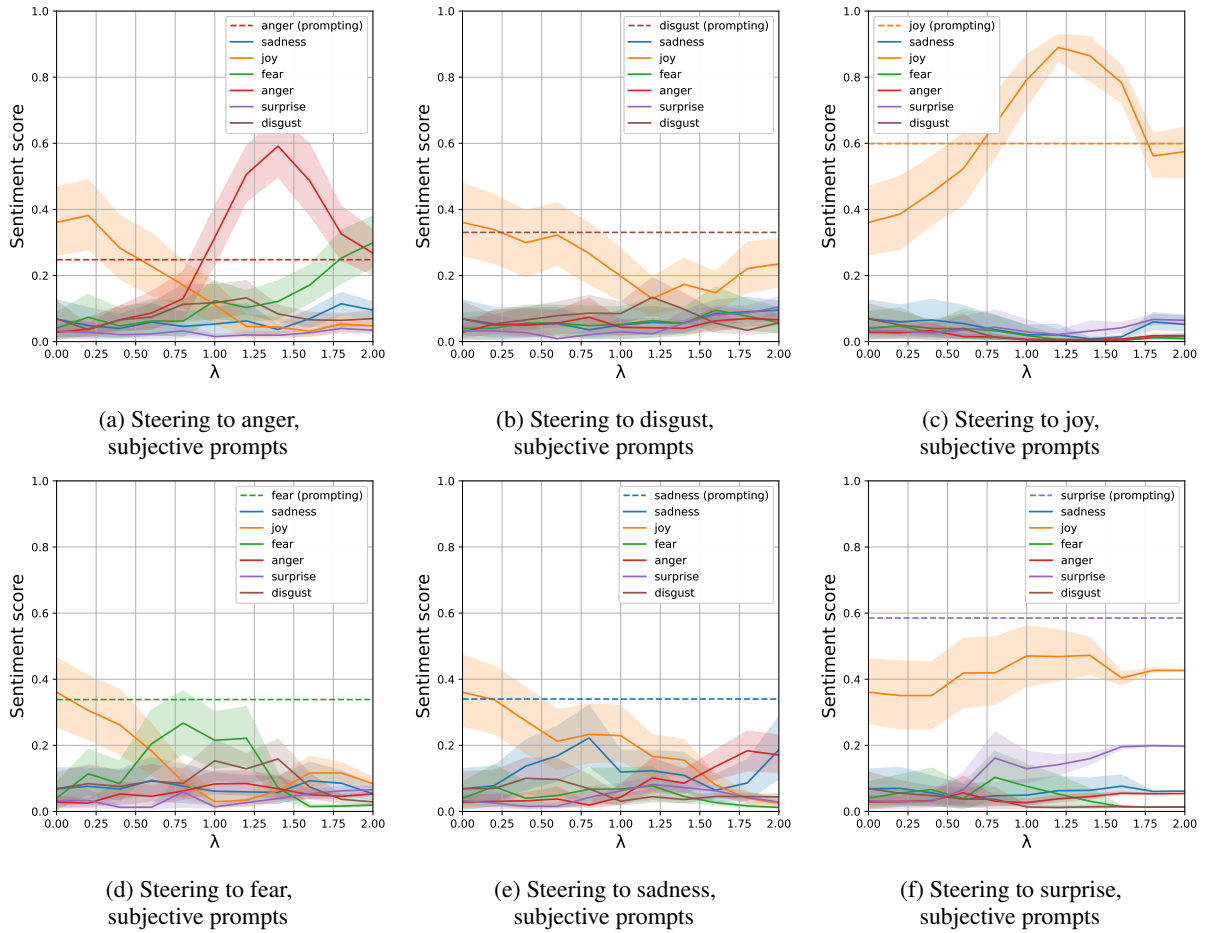


Figure 13: Activation-based style vectors: Evaluation of generated texts for *subjective* prompts using GoEmotions' style vectors. Only the activation vectors were used, for which we have trained steering vectors.