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

$$\mathbf{v}_s^{(i)} = \bar{\mathbf{z}}_s^{(i)} - \bar{\mathbf{z}}_{S \setminus s}^{(i)} \tag{4}$$

Using the average steering vector $\bar{\mathbf{z}}_{S \setminus s}$ 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 $\mathbf{z}_x$ is presented in the right part of Fig. 2. The process begins with the frozen model receiving an empty input token and a steering vector initialized randomly to initiate sentence generation. The resulting output is then evaluated against the target sentence to calculate a cross-entropy loss, which is back-propagated to learn the steering vector. The training for an output x terminates when a steering vector $\mathbf{z}_x$ that produces the target sentence x is found or after a maximum number of $j = 400$ epochs. We use the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 0.01.

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 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:

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

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 the training-based style vectors, we filter out samples containing more than 50 characters from each dataset to keep the time for computing steering vectors feasible. For details, see Sec. 4.2. This limitation does not apply to the activation-based style vectors.

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 Reddit[2] 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*.

[2] Reddit forum: `https://www.reddit.com/`

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. Our experiments 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 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 to keep the computational effort feasible. Nevertheless, we had to run the experiment on the Yelp and Shakespeare datasets for 150 hours each and for GoEmotions for around 100 hours. 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

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 three onwards lead to remarkably high classification accuracy (AUC ≥ 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 still achieve considerable accuracy. The results are similar for the other two datasets, 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

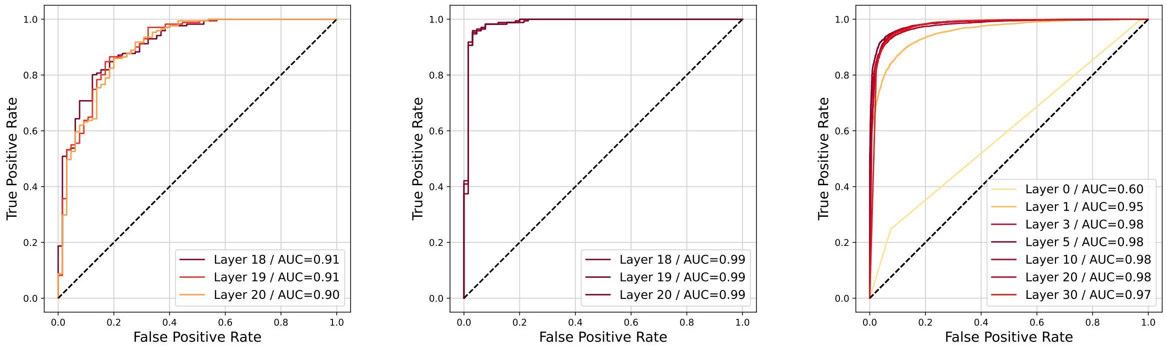
As shown in Sec. 4.3, both trained steering and activation vectors capture relevant style information. However, this does not show that style vectors $\mathbf{v^{(i)}}_s$ that are computed from them can be used to actually steer the style of the model's output. For this reason, we assembled a list of 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, allowing more individual responses to express 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.*" Steering towards a sentiment or emotion category is expected to affect the LLM's outcome significantly more for such subjective prompts than for factual prompts. The full list of prompts is given 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, as shown in Ex. 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 lines in all steering plots, e.g., in Fig. 4 and Fig. 5, indicate 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 [...] manner." to the input prompt, where the dots are replaced with the respective target steering style, e.g., *positive*, or *angry*. Thus, the model is informed in a neutral way to direct the output as required.

For the Yelp dataset-based style vectors, the positivity and negativity values of produced out-



(a) Trained steering vectors (b) Corresponding activation vectors (c) Activation vectors of 10k sentences

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.

puts 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 default output.

In order to evaluate the ability of the style vectors derived from the GoEmotions dataset to steer the model 's output towards multiple emotional categories, a multi-class model based on RoBERTa (Hartmann, 2022) was used. Fig. 5 depicts the shift in the classified emotion class score when the style space is built from activation-based style vectors on the GoEmotions dataset. The notable increase in the respective emotion class score indicates a strong model performance when steering towards a specific emotion. As expected, styling the output on factual prompts shows almost no changes in the emotion class scores, similar to the results for the Yelp style vectors described above. Prompt baselines created with the GoEmo labels exhibit a substantially stronger impact on the generated outcomes compared to the prompt baseline in the case of the Yelp dataset.

4.4.2 Steering Output Examples

In the following, selected examples of model outputs modified by style vectors are presented to provide an impression of the obtained results under different settings. In each case, **P** is the prompt, followed by different model outputs.

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

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

O1: *My favorite vacation was a trip I took to the beach with my family. We had a great time exploring the beach, going on walks, and tasting all the delicious food. The best part was the relaxing atmosphere and the great memories we made.*

O2: *I had a terrible experience with a vacation I had. I went to a beach in a different country and I was expecting to have a great time. Instead, I was met with a huge amount of racism. People were openly saying that I was a "strike-o-bog" (a term that is used to refer to a black person) and that I was "not allowed" to be there. I was then physically attacked. [...]*

E2 - Too large λs. The parameter λ in Eq. 3 plays a central role for the outputs. Yet, there is no general threshold for λ, but it depends on the input prompt and the target style definition. Too large λ values sometimes result in nonsense outputs