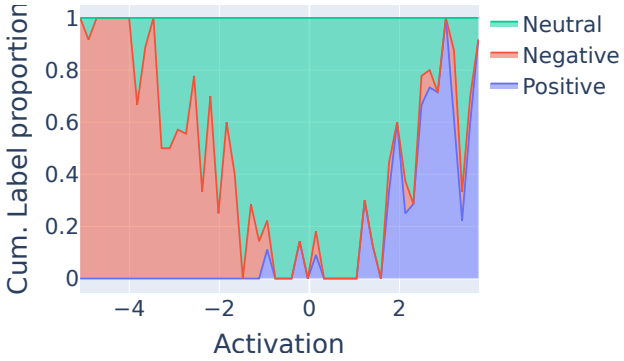


direction	accuracy
K-means	78%
PCA	81%
Mean Diff	80%
LR	89%
DAS	86%

Figure 3: Area plot of sentiment labels for OpenWebText samples by K-means sentiment activation (left). Accuracy using sentiment activations to classify tokens as positive or negative (right). The threshold taken is the top/bottom 0.1% of activations over OpenWebText. Sentiment activations are taken from GPT2-small's first residual stream layer. Classification was performed by GPT-4.

Ablations We eliminate the contribution of a particular component to a model's output, usually by replacing the component's output with zeros (zero-ablation) or the mean over some dataset (mean-ablation), in order to demonstrate its magnitude of importance. We also perform directional ablation, in which a component's activations are ablated only along a specific (e.g. sentiment) direction.

3 Finding and Evaluating a 'Sentiment Direction'

The first question we investigate is whether there exists a direction in the residual stream in a transformer model that represents the sentiment of the input text, as a special case of the linear representation hypothesis (Mikolov et al., 2013). We show that the methods discussed above (2.2) all arrive at a similar sentiment direction. Given some input text to a model, we can project the residual stream at a given token/layer onto a sentiment direction to get a 'sentiment activation'.

3.1 Visualizing and Comparing the Directions

We fit directions using the ToyMovieReview dataset (Section 2.1) across various methods and finding extremely high cosine similarity between the learned sentiment directions (Figure 2). This suggests that these are all noisy approximations of the same singular direction. Indeed, we generally found that the following results were very similar regardless of exactly how we specified the sentiment direction. The directions we found were not sparse vectors, as expected since the residual stream is not a privileged basis (Elhage et al., 2021b).

Here we show a visualisation in the style of Neuroscope (Nanda, 2023a) where the projection is represented by color, with red being negative and blue being positive. It is important to note that the direction being examined here was trained on just 30 positive and 30 negative English adjectives in an unsupervised way (using K-means with $K = 2$). Notwithstanding, the extreme values along this direction appear readily interpretable in the wild in diverse text domains such as the opening paragraphs of Harry Potter in French (Figure 1). An interactive visualisation of residual stream directions in GPT2-small is available here (Yedidia, 2023) and sentiment directions here.

It is important to note that this type of analysis is qualitative, which should not act as a substitute for rigorous statistical tests as it is susceptible to interpretability illusions (Bolukbasi et al., 2021). We rigorously evaluate our directions using correlational and causal methods.

3.2 Correlational Evaluation

In a correlational analysis, we classify word sentiment by 'sentiment activation' and show that the sentiment direction is sensitive to negation flipping sentiment.

	simple_logit_diff	treebank_logit_diff	simple_logit_flip	treebank_logit_flip
das	109.8%	47.0%	100.0%	53.5%
das2d	110.4%	42.8%	95.5%	49.0%
das3d	110.2%	35.9%	95.5%	39.4%
kmeans	67.2%	22.1%	72.7%	14.8%
logistic_regression	71.1%	30.8%	86.4%	16.8%
mean_diff	73.9%	27.5%	81.8%	17.4%
pca	62.7%	17.8%	72.7%	12.3%
random	0.4%	0.1%	0.0%	0.6%

Figure 4: Directional patching results for different methods in pythia-1.4b. We report the best result found across layers. The columns show two evaluation datasets, ToyMovieReview and Treebank, and two evaluation metrics, mean logit difference and % of logit differences flipped.

Sentiment Directions Capture Lexical Sentiment To test the meaning of the sentiment axis, we binned the sentiment activations of OpenWebText tokens from the first residual stream layer of GPT2-small into 20 equal-width buckets and sampled 20 tokens from each. Then we asked GPT-4 to classify into Positive/Neutral/Negative. Specifically, we gave the GPT-4 API prompts of the following form: "Your job is to classify the sentiment of a given token (i.e. word or word fragment) into Positive/Neutral/Negative. Token: '{token}'. Context: '{context}'. Sentiment: " where the context length was 20 tokens centered around the sampled token. Only a cursory human sanity check was performed.

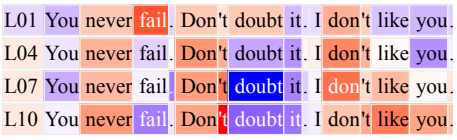
In Figure 3, we show an area plot of the classifications by activation bin. We contrast the results for different methods in Table 3. In the area plot we can see that the left side area is dominated by the "Negative" label, whereas the right side area is dominated by the "Positive" label and the central area is dominated by the "Neutral" label. Hence the tails of the activations seem highly interpretable as representing a bipolar sentiment feature. The large space in the middle of the distribution simply occupied by neutral words (rather than a more continuous degradation of positive/negative) indicates superposition of features (Elhage et al., 2022).

Negation Flips the Sentiment Direction Using the K-means sentiment direction after the first layer of GPT2-small, we can obtain a view of how the model updates its view of sentiment during the forward pass, analogous to the "logit lens" technique from nostalgebraist (2020). In Figure A.5, we see how the sentiment activation flips when the context of the sentiment word denotes that it is negated. Words like 'fail', 'doubt' and 'uncertain' can be seen to flip from negative in the first couple of layers to being positive after a few layers of processing. An interesting task for future circuits analysis research could be to better understand the circuitry used to flip the sentiment axis in the presence of a negation context. We suspect significant MLP involvement (see Section A.5).

3.3 Causal Evaluation

Sentiment directions are causal representations. We evaluate the sentiment direction using directional patching in Figure 4. These evaluations are performed on prompts with out-of-sample adjectives and the direction was not trained on *any* verbs. Unsupervised methods such as K-means are still able to shift the logit differences and DAS is able to completely flip the prediction.

Directions Generalize Most at Intermediate Layers If the sentiment direction was simply a trivial feature of the token embedding, then one might expect that directional patching would be most effective in the first or final layer. However, we see in Figure 6 that in fact it is in intermediate layers of the model where we see the strongest out-of-distribution performance to SST. This suggests the speculative hypothesis that the model uses the residual stream to form abstract concepts in intermediate layers and this is where the latent knowledge of sentiment is most prominent.



direction	flip percent	flip median size
DAS	96%	107%
KM	96%	69%
MD	89%	45%
LR	100%	86%
PCA	78%	44%

Figure 5: We made a dataset of 27 negation examples and compute the change in sentiment activation at the negated token (e.g. doubt) between the 1st and 10th layers of GPT2-small. We show sample text across layers for K-means (left), the fraction of activations flipped and the median size of the flip centered around the mean activation (right).

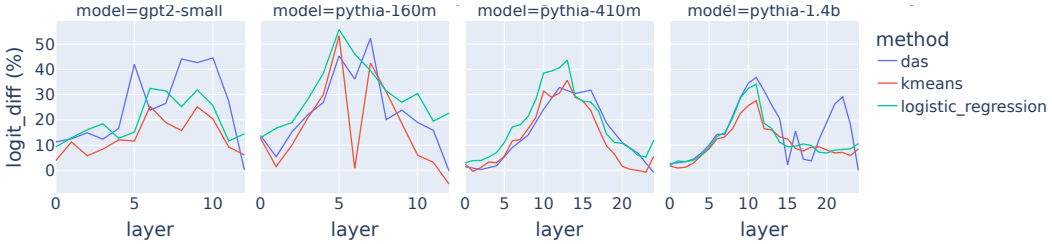


Figure 6: Patching results for directions trained on toy datasets and evaluated on the Stanford Sentiment Treebank test partition. We tend to find the best generalisation when training and evaluating at a layer near the middle of the model. We scaffold the prompt using the suffix Overall the movie was very and compute the logit difference between good and bad. The patching metric (y-axis) is then the % mean change in logit difference.

Activation Addition Steers the Model A further verification of causality is shown in Figure A.3. Here we use the technique of "activation addition" from Turner et al. (2023). We add a multiple of the sentiment direction to the first layer residual stream during each forward pass while generating sentence completions. Here we start from the baseline of a positive movie review: "I really enjoyed the movie, in fact I loved it. I thought the movie was just very...". By adding increasingly negative multiples of the sentiment direction, we find that indeed the completions become increasingly negative, without completely destroying the coherence of the model's generated text. We are wary of taking the model's activations out of distribution using this technique, but we believe that the smoothness of the transition in combination with the knowledge of our findings in the patching setting give us some confidence that these results are meaningful.

Validation on SST We validate our sentiment directions derived from toy datasets (Section 3.3) on SST. We collapsed the labels down to a binary "Positive"/"Negative", just used the unique phrases rather than any information about their source sentences, restricted to the 'test' partition and took a subset where pythia-1.4b can achieve 100% zero shot classification accuracy, removing 17% of examples. Then we paired up phrases of an equal number of tokens[2] to make up 460 clean/corrupted pairs. We used the scaffolding "Review Text: TEXT, Review Sentiment:" and evaluated the logit difference between "Positive" and "Negative" as our patching metric. Using the same DAS direction from Section 3 trained on just a few examples and flipping the corresponding sentiment activation between clean/corrupted in a single layer, we can flip the output 53.5% of the time (Figure 4).