```
  - wretched
  - inadequate
  - dire
  - unpleasant
  - horrible
  - disappointing
  - awful

positive_adjectives_test:
  - stunning
  - impressive
  - admirable
  - phenomenal
  - radiant
  - glorious
  - magical
  - pleasing
  - lively
  - warm
  - strong
  - helpful
  - vivid
  - modern
  - crisp
  - sweet

negative_adjectives_test:
  - foul
  - vile
  - appalling
  - rotten
  - grim
  - dismal
  - lazy
  - poor
  - rough
  - noisy
  - sour
  - flat
  - ancient
  - bitter

positive_verbs:
  - enjoyed
  - loved
  - liked
  - appreciated
  - admired

negative_verbs:
  - hated
  - disliked
  - despised

positive_answer_tokens:
  - great
  - amazing
  - awesome
  - good
```

```
  - perfect

negative_answer_tokens:
  - terrible
  - awful
  - bad
  - horrible
  - disgusting
```

The ToyMoodStories dataset consists of prompts of the form "NAME1 VRB1.1 parties, and VRB1.2 them whenever possible. NAME2 VRB2.1 parties, and VRB2.2 them whenever possible. One day, they were invited to a grand gala. QUERYNAME feels very". To evaluate the model's output, we measure the logit difference between the "excited" and "nervous" tokens.

VRB1.1 and VRB2.1 are always one of:

```
- hates
- loves
```

and VRB1.2 and VRB2.2 are always one of:

```
- avoids
- joins
```

In each case, the two verbs in each sentence will agree in sentiment, and the sentence with NAME1 will always have opposite sentiment to that of NAME2.

Names are sampled from the following list:

```
- John
- Anne
- Mark
- Mary
- Peter
- Paul
- James
- Sarah
- Mike
- Tom
- Carl
- Sam
- Jack
```

Each combination of NAME1, NAME2, QUERYNAME are included in the dataset (where half the time QUERYNAME matches the first name, and half the time it matches the second). Where necessary for computational tractability, we take a subsample of the first 16 items of this dataset.

A.8 GLOSSARY

GLOSSARY

activation addition Formerly called "activation steering", a technique from Turner et al. (2023) where a vector is added to the residual stream at a certain position (or all positions) and layer during each forward pass while generating sentence completions. In our case, the vector is the sentiment direction.

activation patching A technique introduced in Meng et al. (2023), under the name 'causal tracing', which uses a causal intervention to identify which activations in a model matter for producing some output. It runs the model on some 'clean' input, replaces (patches) an activation with that same activation on 'flipped' input, and sees how much that shifts the output from 'clean' to 'flipped'.

activation steering See activation addition

DAS Distributed Alignment Search (Geiger et al., 2023b) uses gradient descent to train a rotation matrix representating an orthonormal change of basis to one better aligned with the model's features. We mostly focus on a special case of finding a singular critical direction, where we patch along the first dimension of the rotated basis and then use a smooth patching metric (such as the logit difference between positive and negative completions) as the objective to be minimised.

directional activation patching A variant of activation patching introduced in this paper where we only patch a single dimension from a counterfactual activation. That is, for prompts x_{orig} and x_{new}, direction $\boldsymbol{d}$, a set of model components $\mathbb{C}$, we run a forward pass on x_{orig} but for each component in $\mathbb{C}$, we patch/replace the output $\boldsymbol{o}_{\text{orig}}$ with $\boldsymbol{o}_{\text{orig}} - \boldsymbol{o}_{\text{orig}} \cdot \boldsymbol{d} + \boldsymbol{o}_{\text{new}} \cdot \boldsymbol{d}$. This is equivalent to activation patching a single neuron, but done in a rotated basis (where $\boldsymbol{d}$ is the first column of the rotation matrix).

directional patching See directional activation patching.

mean ablation A type of ablation method, where we seek to eliminate the contribution of a particular component to demonstrate its importance, where we replace a particular set of activations with their mean over an appropriate dataset.

patching metric A summary statistic used to quantify the results of an activation patching experiment. By default here we use the percentage change in logit difference as in Wang et al. (2022).

SST Stanford Sentiment Treebank is a labelled sentiment dataset from Socher et al. (2013) described in Section 2.1.