Jett Janiak Stefan Heimersheim. A circuit for python docstrings in a 4-layer attention-only, 2023. URL https://www.alignmentforum.org/posts/u6KXXmKFbXfWzoAXn/a-circuit-for-python-docstrings-in-a-4-layer-attention-only. Accessed: 2023-09-22.

Alex Tamkin, Dan Jurafsky, and Noah Goodman. Language through a prism: A spectral approach for multiscale language representations, 2020.

Alexander Matt Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization, 2023.

Vikrant Varma, Rohin Shah, Zachary Kenton, János Kramár, and Ramana Kumar. Explaining grokking through circuit efficiency, 2023.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. Causal mediation analysis for interpreting neural nlp: The case of gender bias, 2020.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small, 2022.

Adam Yedidia. Residual viewer, 2023. URL http://ec2-34-192-101-140.compute-1.amazonaws.com:5014/. Available at: http://ec2-34-192-101-140.compute-1.amazonaws.com:5014/.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to ai transparency, 2023.

(a) PCA on adjectives in and out of sample

(b) PCA on in-sample adjectives and out-of-sample verbs

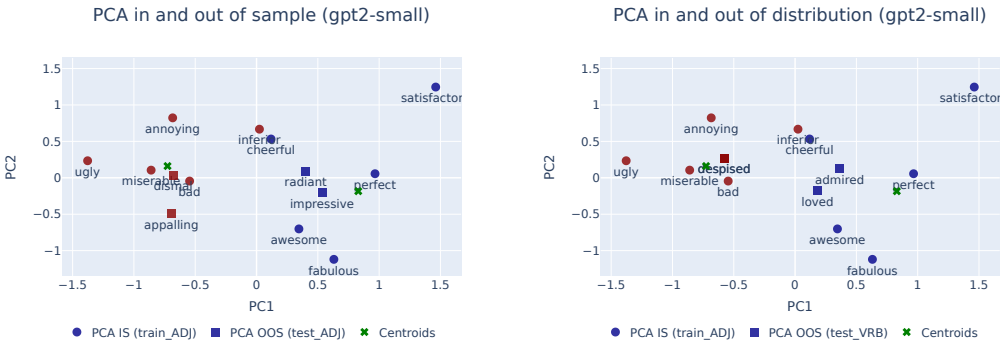

Figure A.1: 2-D PCA visualization of the embedding for a handful of adjectives and verbs (GPT2-small)

A Appendix

A.1 Further evidence for a linear sentiment representation

A.1.1 Clustering

In Section 2.2, we outline just a few of the many possible techniques for determining a direction which hopefully corresponds to sentiment. Is it overly optimistic to presume the existence of such a direction? The most basic requirement for such a direction to exist is that the residual stream space is clustered. We confirm this in two different ways.

First we fit 2-D PCA to the token embeddings for a set of 30 positive and 30 negative adjectives. In Figure A.1, we see that the positive adjectives (blue dots) are very well clustered compared to the negative adjectives (red dots). Moreover, we see that sentiment words which are out-of-sample with respect to the PCA (squares) also fit naturally into their appropriate color. This applies not just for unseen adjectives (Figure A.1a) but also for verbs, an entirely out-of-distribution class of word (Figure A.1b).

Secondly, we evaluate the accuracy of 2-means trained on the Simple Movie Review Continuation adjectives (Section 2.1). The fact that we can classify in-sample is not very strong evidence, but we verify that we can also classify out-of-sample with respect to the K-means fitting process. Indeed, even on hold-out adjectives and on the verb tokens (which are totally out of distribution), we find that the accuracy is generally very strong across models. We also evaluate on a fully out of distribution toy dataset ("simple adverbs") of the form "The traveller [adverb] walked to their destination. The traveller felt very". The results can be found in Figure A.2. This is strongly suggestive that we are stumbling on a genuine representation of sentiment.

A.1.2 Activation addition

We perform activation addition (Turner et al., 2023) on GPT2-small for a single positive simple movie review continuation prompt (from Section 2.1) in order to flip the generated outputs from negative to prompt. The "steering coefficient" is the multiple of the sentiment direction which we add to the first layer residual stream. The outputs are extremely negative by the time we reach coefficient -17 and we observe a gradual transition for intermediate coefficients (Figure A.3).

A.1.3 Multi-lingual sentiment

We use the first few paragraphs of Harry Potter in English and French as a standard text (Elhage et al., 2021b). We find that intermediate layers of pythia-2.8b demonstrate intuitive sentiment activations for the French text (Figure A.4). It is important to note that none of the models are very good at

French, but this was the smallest model where we saw hints of generalisation to other languages. The representation was not evident in the first couple of layers, probably due to the poor tokenization of French words.

A.1.4 INTERPRETABILITY OF NEGATIONS

We visualise the sentiment activations for all 12 layers of GPT2-small simultaneously on the prompt "You never fail. Don't doubt it. I am not uncertain" (Figure A.5). This allows us to observe how fail, doubt and uncertain shift from negative to positive sentiment during the forward pass of the model.

A.2 IS SENTIMENT REALLY A HYPERPLANE?

In our directional patching experiments, we have somewhat artificially selected just 1 dimension as our hypothesised structure for the sentiment subspace. We can perform DAS with any number of dimensions. Figure A.6 demonstrates that whilst increasing the DAS dimension improves the patching metric in-sample (A.6a), the metric does not improve out-of-distribution (A.6b).

A.3 DETAILED CIRCUIT ANALYSIS

In order to build a picture of each circuit, we used the process pioneered in Wang et al. (2022):

- Identify which model components have the greatest impact on the logit difference when path patching is applied (with the final result of the residual stream set as the receiver).
- Examine the attention patterns (value-weighted, in some cases) and other behaviors of these components (in practice, attention heads) in order to get a rough idea of what function they are performing.
- Perform path-patching using these heads (or a distinct cluster of them) as receivers.
- Repeat the process recursively, performing contextual analyses of each "level" of attention heads in order to understand what they are doing, and continuing to trace the circuit backwards.

In each path-patching experiment, change in logit difference is used as the patching metric. We started with GPT-2 as an example of a classic LLM displays a wide range of behaviors of interest, and moved to larger models when necessary for the task we wanted to study (choosing, in each case, the smallest model that could do the task).

A.3.1 SIMPLE SENTIMENT - GPT-2 SMALL

We examined the circuit performing tasks for the following sentence template:

I thought this movie was ADJECTIVE, I VERBed it. Conclusion: This movie is

Using a threshold of 5%-or-greater damage to the logit difference for our patching experiments, we found that GPT-2 Small contained 4 primary heads contributing to the most proximate level of circuit function–10.4, 9.2, 10.1, and 8.5 (using "layer.head" notation). Examining their value-weighted attention patterns, we found that attention to ADJ and VRB in the sentence was most prominent in the first three heads, but 8.5 attended primarily to the second "movie" token. We also observed that 9.2 attended to this token as well as to ADJ. (Results of activation patching can be seen in Fig. A.7.)

Conducting path-patching with 8.5 and 9.2 as receivers, we identified two heads–7.1 and 7.5–that primarily attend to ADJ and VRB from the "movie" token. We further determined that the output of these heads, when path-patched through 9.2 and 8.5 as receivers, was causally important to the circuit (with patching causing a logit difference shift of 7% and 4% respectively for 7.1 and 7.5). This was not the case for other token positions, which demonstrates that causally relevant information is indeed being specially written to the "movie" position. We thus designated it the SUM token in this circuit, and we label 8.5 a summary-reader head.

Repeating our analysis with lower thresholds yielded more heads with the same behavior but weaker effect sizes, adding 9.10, 11.9, and 6.4 as summary reader, direct sentiment reader, and sentiment summarizer respectively. This gives a total of 9 heads making up the circuit.