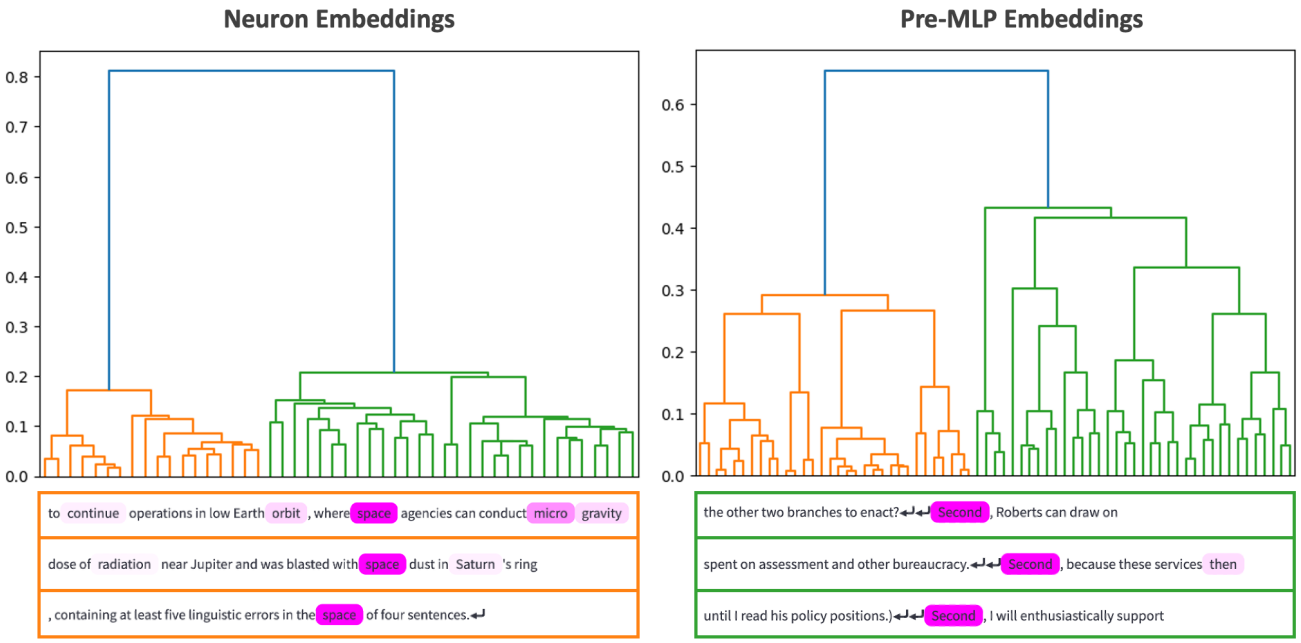

Figure 4. A comparison between feature clusters derived from neuron embeddings vs pre-MLP embeddings. The neuron embeddings clearly result in denser clusters with better separation between the clusters. Examples from the two clusters are shown in their corresponding colours.

ing examples (denoted "size" in the table). We might expect that improvements in monosemanticity would come from neurons becoming more specific and responding to fewer examples, but in actuality we speculate that they have come from neurons moving to represent single, broader features, which respond to a wider variety of examples but look for the same information across all the examples. Anecdotally, we would tentatively suggest that this has corresponded to an improvement in SAE neuron interpretability, but we would recommend that readers investigate this for themselves by exploring the two different SAEs in the provided UI [6].

Figure 5 shows an example neuron from the SAE trained with the NE loss that illustrates these broad, more general features that are learnt. The activation map shows the maximal activation that can be induced by each pixel, and the importance map is the element-wise product of the activation map and the average example. Note that this neuron had a single feature cluster with a mean and max distance of 0.01 and 0.04 between the neuron embeddings of the examples.

The neuron appears to respond to lines or curves along the middle of the image, particularly towards the left and right edges. The logit effects show that the neuron increases the probability of predicting 9's, as well as 6's, 4's, and to a lesser extent 7's and 8's. This fits with the visual interpretation of the feature, and the randomly selected activating examples. This neuron visualisation demonstrates the style of feature which is commonly learned after including the NE loss, as well as how we can effectively understand a neuron's behaviour using some simple visualisations.

We also observe a more than $6\times$ decrease in the percentage of dead neurons, which don't activate for any example in the training set, from 23.8% to 3.7%. Dead neurons are a significant challenge when training SAEs, particularly as they are scaled up (Templeton et al., 2024). We haven't investigated why the NE loss causes such a significant decrease in the prevalence of dead neurons. Dead neurons shouldn't be directly penalised by the NE loss in theory, as it's only computed over the active neurons for a given input - in fact, increasing activation sparsity should decrease both the L1 and NE losses, but the NE loss seems to decrease sparsity instead. We speculate that encouraging neuron monosemanticity may force the SAE to utilise more of it's neurons and to learn more general features to avoid significant increases in the reconstruction error. Understanding the mechanism and effect of the NE loss in more detail would be a valuable direction for further research.

5. Conclusion

We presented neuron embeddings, and showed they can be used to effectively tackle neuron polysemanticity in a variety of ways. We used them to identify the distinct semantic behaviours of neurons in GPT2-small, and showed that they can capture the similarity between both individual exam-

[6] https://mechmnistic.streamlit.app/