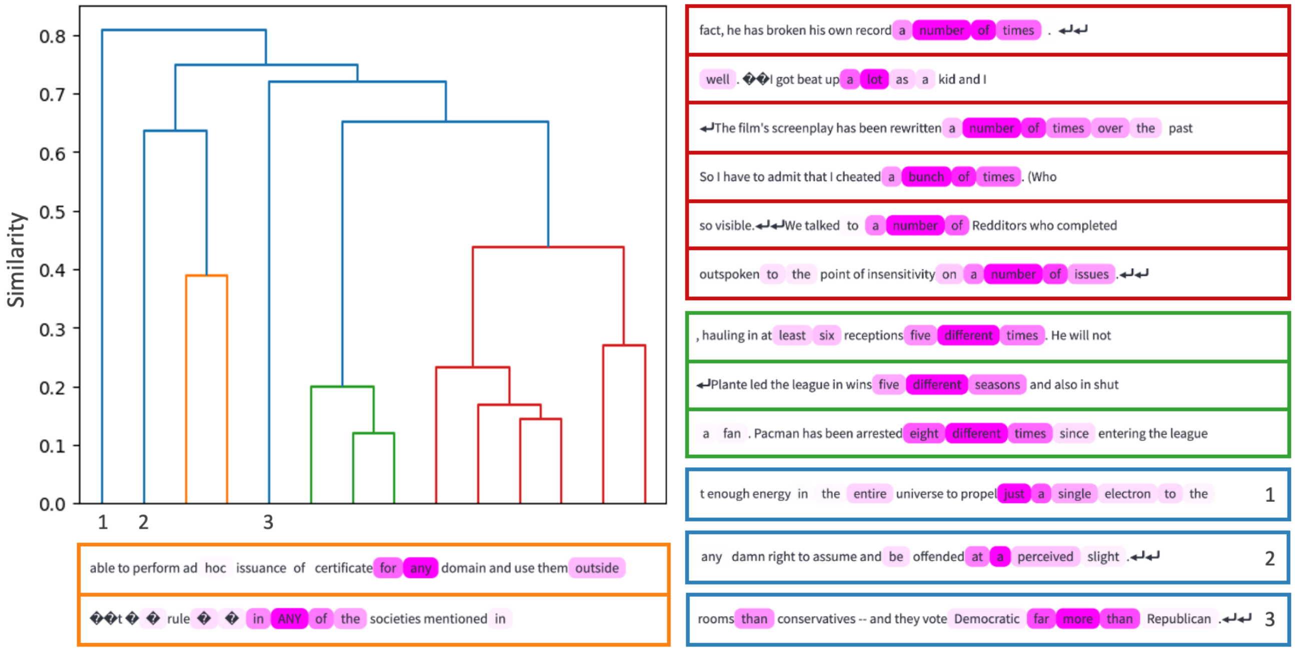

fact, he has broken his own record a number of times . ↵↵

well . ��I got beat up a lot as a kid and I

↵The film's screenplay has been rewritten a number of times over the past

So I have to admit that I cheated a bunch of times . (Who

so visible.↵↵We talked to a number of Redditors who completed

outspoken to the point of insensitivity on a number of issues .↵↵

, hauling in at least six receptions five different times . He will not

↵Plante led the league in wins five different seasons and also in shut

a fan . Pacman has been arrested eight different times since entering the league

t enough energy in the entire universe to propel just a single electron to the 1

any damn right to assume and be offended at a perceived slight .↵↵ 2

rooms than conservatives -- and they vote Democratic far more than Republican .↵↵ 3

able to perform ad hoc issuance of certificate for any domain and use them outside

��t � � rule � � in ANY of the societies mentioned in

Figure 2. An example of feature clustering applied to a neuron in layer 7 of GPT2-small. The clusters (colour and numerically coded) each show a distinct semantic behaviour, and the dendrogram shows how the cluster hierarchy formed. The highlighting corresponds to neuron activation on each token, with the neuron embedding derived from the maximally activating token.

behaviour (blue) almost 50:1 [4]. Without feature clustering, it's very easy to miss these rarer behaviours during interpretation - unless you review a large number of examples even after there appears to be a clear hypothesis for the behaviour, you would naturally conclude that this is a monosemantic neuron. Feature clustering allows us to collect a much larger number of examples and automatically condense them, making it much easier to quickly identify all the relevant behaviours of an neuron. This could help to address the illusion of interpretability (Bolukbasi et al., 2021), where examining the top examples for a neuron suggests a simple explanation of the behaviour, but expanding to lower activating examples reveals an array of hidden behaviours.

4.1.3. Comparison to Embeddings

Whilst we choose to use neuron embeddings to cluster a neuron's dataset examples, we could instead just use the pre-MLP embedding of the key token, without then multiplying it by the neuron's weights. Table 1 compares the median intra- and inter-cluster distances of the dataset example clusterings for all neurons in GPT2-small when using the pre-MLP embeddings or neuron embeddings. It shows that neuron embeddings lead to denser clusters with reduced intra-cluster distance, with better seperation between these clusters from the increased inter-cluster distance. Figure 4 clearly illustrates this, showing a pair of clusters with significantly higher density and separation using neuron embeddings compared to pre-MLP embeddings.

Table 1. Intra- and inter-cluster distance of embedded dataset examples averaged across all neurons in GPT2-small, comparing pre-MLP embeddings with neuron embeddings. Neuron embeddings on average result in denser clusters with better separation between clusters.

Distance	Intra-cluster	Inter-cluster
Pre-MLP	0.31	0.63
Neuron	0.21	0.73

The improved separation between clusters implies that feature clusters derived from neuron embeddings will have fewer errors than those derived from pre-MLP embeddings. Anecdotally, we found this to be the case, particularly when using simpler but faster clustering algorithms such as the Sub-Cluster Component algorithm (Monath et al., 2021).

The better performance of neuron embeddings also indicate that they better capture the similarity between a neuron's dataset examples. Intuitively, this is because a neuron may not respond to all information in the pre-MLP embedding, so by incorporating the neuron weights, which represent what the neuron is "looking for" in an input, we select out the relevant information to the neuron, providing a better representation of what caused the neuron to activate.

[4] The dendrogram doesn't show the full hierarchy for simplicity - some of the orange leaves are actually clusters with multiple elements