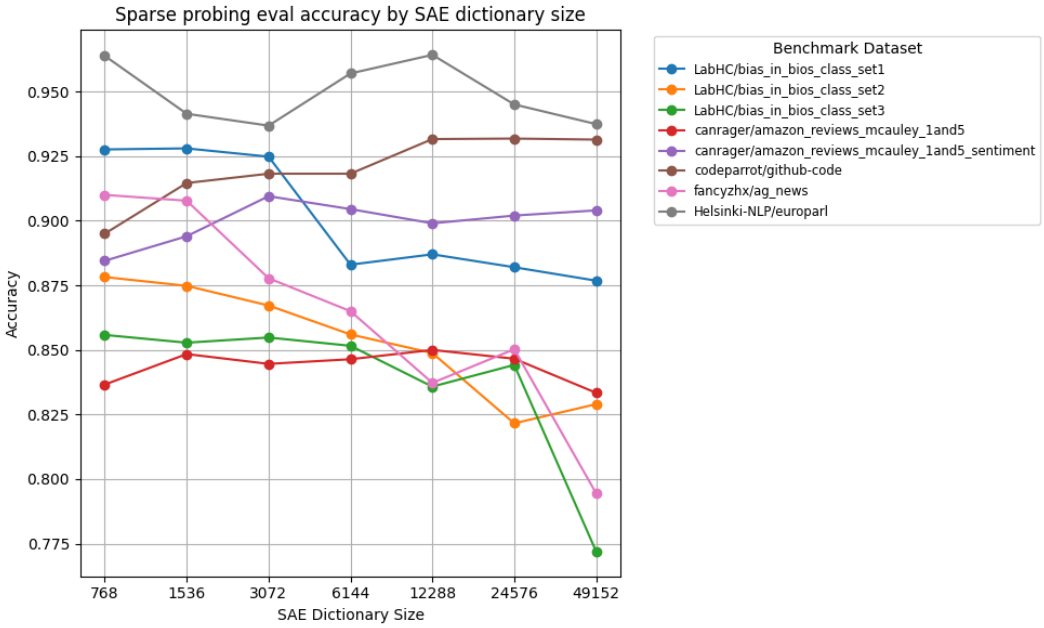

Figure 22: Sparse probing evaluation accuracy by GPT-2 SAE dictionary size across 8 benchmark datasets, with a sparse probe using the top 50 latents.

A.8.2 Concept Removal

In the SHIFT method (Marks et al., 2024), a human evaluator debiases a classifier by ablating SAE latents. (Anonymous, 2024) operationalizes SHIFT into an evaluation for SAE quality, Targeted Probe Perturbation, by training probes on model activations and measuring the effect of ablating sets of SAE latents on the probe accuracy. Ablating a disentangled set of latents should have an isolated causal effect on one class probe, while leaving other class probes unaffected.

They consider a dataset with M classes. For each class with index $i = 1, \ldots, M$ they select the set L_i of the most relevant SAE latents. To select those latents, binary probes are trained detecting the concept c from model activations. The attribution score of latent with index l on concept c is given by

$$I(L, c) = (L_{pos} - L_{neg})(\mathbf{d}_l \cdot \mathbf{P}) \tag{10}$$

where L denotes the batch of activations of latent l, $\mathbf{d}_l$ denotes the SAE decoder vector corresponding to l, $\mathbf{P}$ is the weight matrix of the binary probe, L_{pos} is the mean activation of the latent for inputs of the targeted concept, and L_{neg} is the mean activation for inputs unrelated to the concept. The latents with the highest attribution score are those select as most relevant.

For each concept c_i, they partition the dataset into samples of the targeted concept and a random mix of all other labels. They define the model with probe corresponding to class c_j with $j = 1, \ldots, M$ as a linear classifier C_j. Further, $C_{i,j}$ denotes a classifier for c_j where latents L_i are ablated. Then, they iteratively evaluate the accuracy $A_{i,j}$ of all linear classifiers $C_{i,j}$ on the dataset partitioned for the corresponding class c_j. The targeted probe perturbation score

$$S_{\text{TPP}} = \text{mean}_{(i=j)}(A_{i,j}) - \text{mean}_{(i \neq j)}(A_{i,j}) \tag{11}$$

represents the effectiveness of causally isolating a single probe. Ablating a disentangled set of features should only show a significant accuracy decrease if $i = j$, namely if the latents selected for class i are ablated in the classifier of the same class i, and remain constant if $i \neq j$

The results of this evaluation by SAE dictionary size is displayed in Figure 23. Whilst here there is a general downward trend, the accuracy is not monotonically decreasing with SAE size.

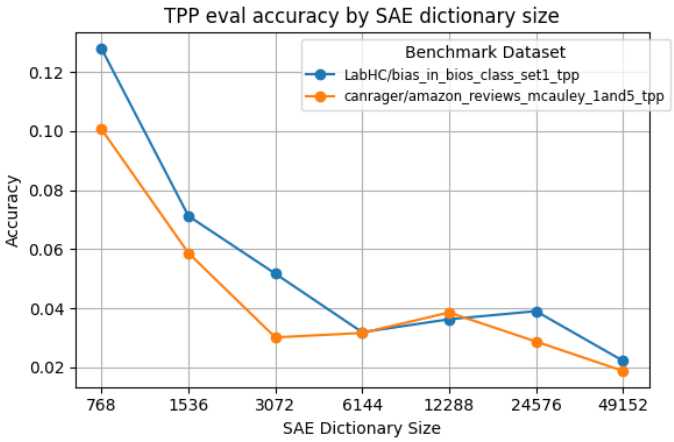

Figure 23: TPP evaluation accuracy by GPT-2 SAE dictionary size across 2 benchmark datasets, ablating up to 50 latents.