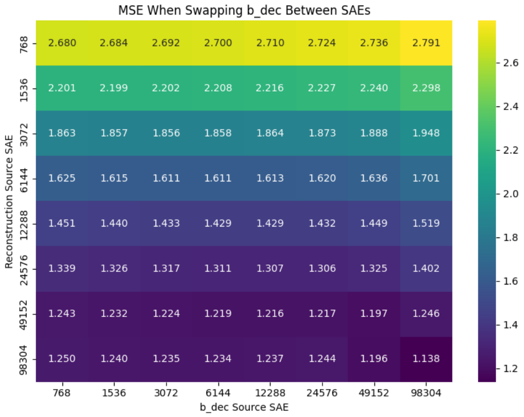

Figure 17: Reconstruction performance (MSE) of SAEs of different sizes when their $\mathbf{b}^{dec}$s are replaced with the $\mathbf{b}^{dec}$s of SAEs with a different dictionary size.

exchanged in the stitching process with negligible impact on the reconstruction performance of an SAE.

A.6.1 Stitching in Gemma Scope SAEs

In order to validate that the SAE stitching results are not an artifact of GPT-2 small or the SAEs that we trained, we applied the same methods to the open-source Gemma Scope SAEs. In particular, we compared two SAEs trained on the residual stream of layer 12 of Gemma 2 2B. The first SAE has dictionary size 16384 (average L0 41) and the second SAE has dictionary size 32768.

We find a lower threshold for distinguishing novel features from reconstruction features (0.4). Using this threshold, we can also smoothly interpolate between SAEs of different sizes trained on Gemma-2-2B.

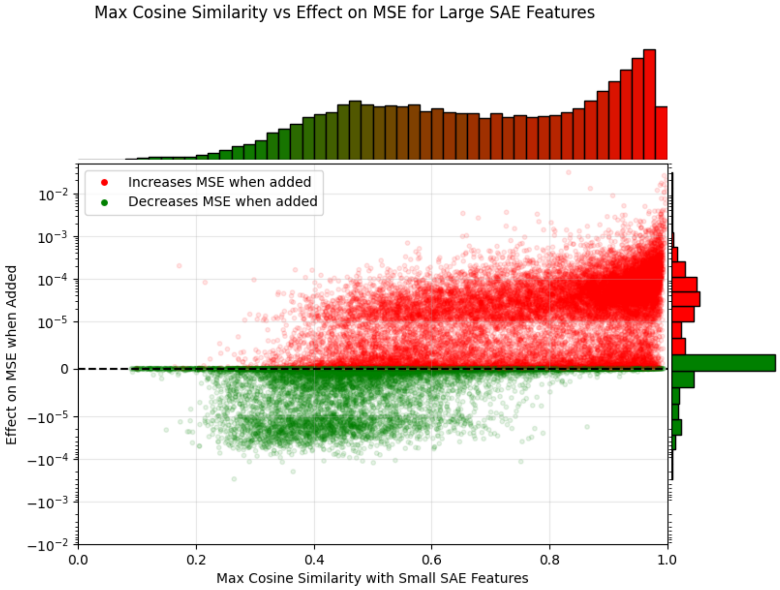

Figure 18: Change in MSE when adding each feature from GemmaScope-32k to GemmaScope-16k, plotted against the maximum cosine similarity of that feature to any feature in GemmaScope-16k. Features with cosine similarity less than 0.4 tend to improve MSE, while more redundant features hurt performance.

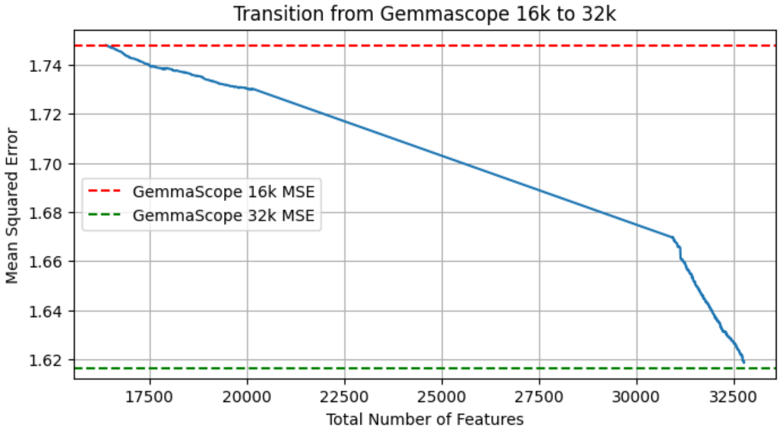

Figure 19: By first adding in novel latents from Gemma Scope 32k to Gemma Scope 16k and replacing the remaining latents with their similar latents in the larger SAE, we interpolate between the two SAE siezs.

A.7 MetaSAE Additional Figures

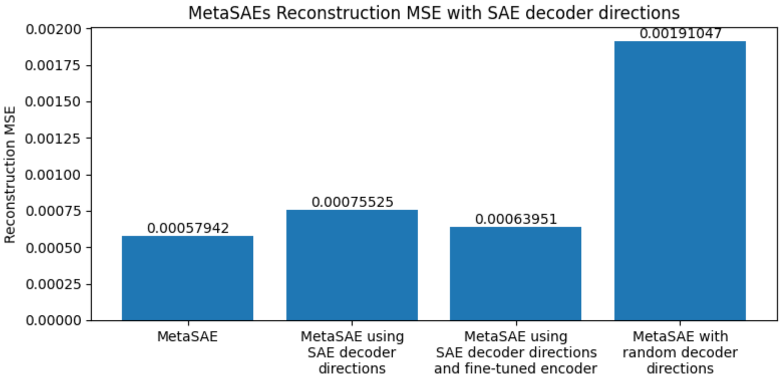

Figure 20: Change in reconstruction performance of a meta-SAE when its decoder directions are replaced with the most similar decoder direction from an SAE with a similar dictionary size.

A.8 Interpretability Experiments

In this paper, we demonstrate that larger SAEs may learn more narrow, composed concepts in order to improve sparsity rather than just learning concepts that are missing in smaller SAEs. Here, we provide some experimental results on sparse probing Gao et al. (2024) and concept removal benchmarks (Anonymous, 2024)[1].

A.8.1 Sparse Probing

Similarly to Gao et al. (2024), we use sparse probes to evaluate the presence of known ground-truth features in our SAEs. If we expect a specific feature to be discovered by an SAE, then a metric for autoencoder quality is simply checking whether these features are present as latents. We do this by

[1] (Anonymous, 2024) is a benchmarking suite that is currently being developed, to which the researchers were kind enough to grant us early access to benchmark the models in this paper. This work is scheduled to be officially published in early December, well before the decision, and this section will be updated in the camera-ready with a clear citation to that work. We refer to the upcoming work in order to avoid confusion or claiming any credit for that research. The paper and code can be temporarily accessed for the duration of the review period at `https://github.com/anonymous664422`

training a 1-dimensional logistic probe on the activations of the SAE to predict the presence of the feature. We use the benchmark datasets included in (Anonymous, 2024) in our evaluation. These cover a range of domains, for example predicting the sentiment of Amazon reviews or predicting the language of the text from the SAE activations.

In our experiments we use a probe that uses only a single SAE latent in its prediction. The results of these experiments are visualized in Figure 21. They show that the relationship between the size of the SAE and the evaluation accuracy is complex and dataset dependent.

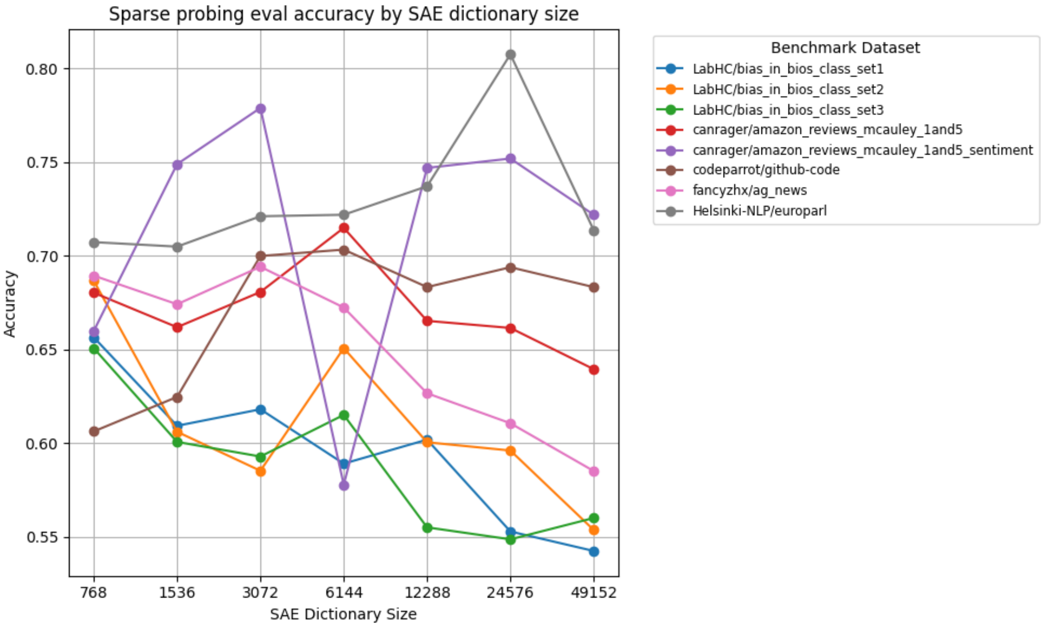

Figure 21: Sparse probing evaluation accuracy by GPT-2 SAE dictionary size across 8 benchmark datasets, with a sparse probe using the top latent.