A.4 LATENT FAMILIES

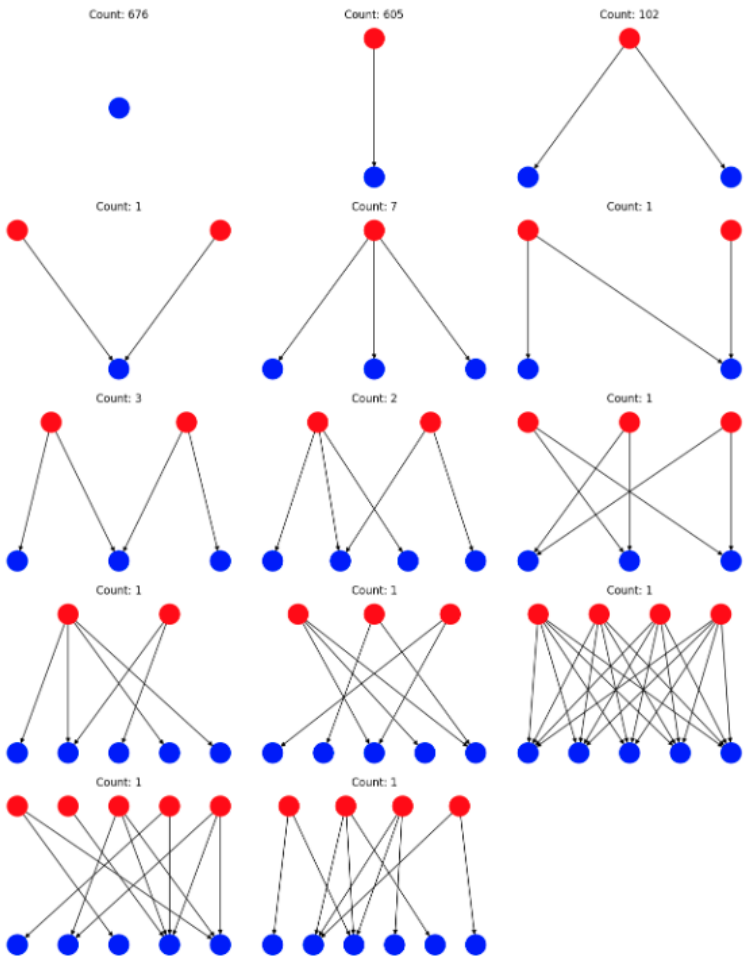


Figure 12: Connected subgraphs of the bipartite graph of latent in GPT2-768 and GPT2-1536.

A.5 OPEN SOURCE SAE WEIGHTS

All GPT-2 Small SAEs were trained on the layer 8 residual stream, which was chosen in line with Gao et al. (2024). They were trained for 300M tokens on the OpenWebText dataset, which was processed into sequences of a maximum of 128 tokens for input into the language models. All models were trained using the Adam optimizer with a learning rate of 4×10^{-4}, $\beta_1 = 0.9$, and $\beta_2 = 0.99$. The batch size used was 4096 and all were trained with a sparsity penalty of 8×10^{-5}. The GPT-2 SAEs are available on Neuronpedia at Redacted URL. We also use two of the Gemma Scope SAEs (Lieberum et al., 2024) trained on Gemma 2 2B (Team et al., 2024) with dictionary size 16384 and 32768. We used the TransformerLens (`https://transformerlensorg.github.io/TransformerLens/`) implementations of GPT-2 and Gemma 2 2B. CELR is the cross entropy loss recovered from either zero or mean ablation.

Table 2: Properties of the SAEs used in this study.

Name	Model	Dict. size	L0	MSE	CELR Zero	CELR Mean
GPT2-768	gpt2-small	768	35.2	2.72	0.915	0.876
GPT2-1536	gpt2-small	1536	39.5	2.22	0.942	0.915
GPT2-3072	gpt2-small	3072	42.4	1.89	0.955	0.937
GPT2-6144	gpt2-small	6144	43.8	1.63	0.965	0.949
GPT2-12288	gpt2-small	12288	43.9	1.46	0.971	0.958
GPT2-24576	gpt2-small	24576	42.9	1.33	0.975	0.963
GPT2-49152	gpt2-small	49152	42.4	1.21	0.978	0.967
GPT2-98304	gpt2-small	98304	43.9	1.14	0.980	0.970
GemmaScope-16384	Gemma 2 2B	16384	43.4	1.72	0.983	0.980
GemmaScope-32768	Gemma 2 2B	32768	41.5	1.59	0.985	0.982

A.6 STITCHING EXPERIMENTS

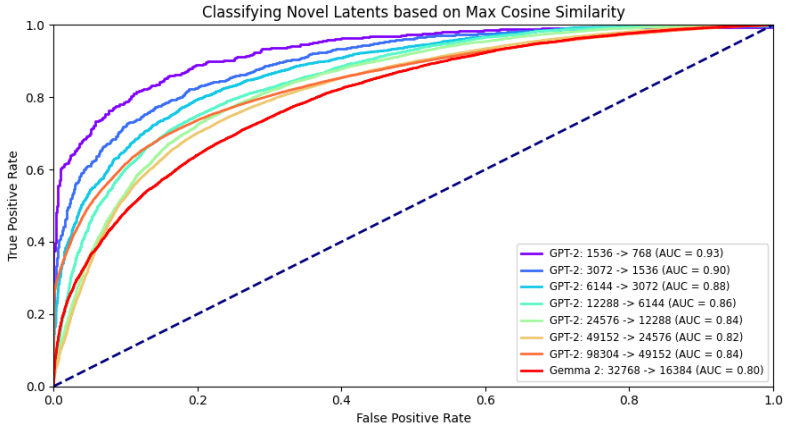


Figure 13: Based on the maximum cosine similarity it is possible to predict the direction of the effect of adding a latent to another SAE. The Receiver operating curve (ROC) is created by varying the threshold of maximum cosine similarity to classify a latent as novel latent or reconstruction latent.

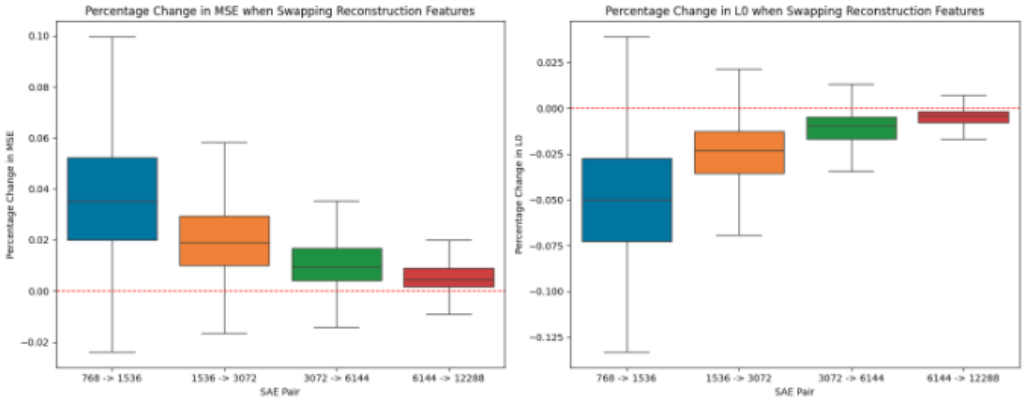


Figure 14: Effects on MSE and L0 when swapping reconstruction latents from larger SAEs to smaller ones. Swapping latent structures generally increases the MSE but almost always decreases L0. Outliers are not shown. The percentual effects per swap get smaller for larger models as the effects are distributed over more swaps.

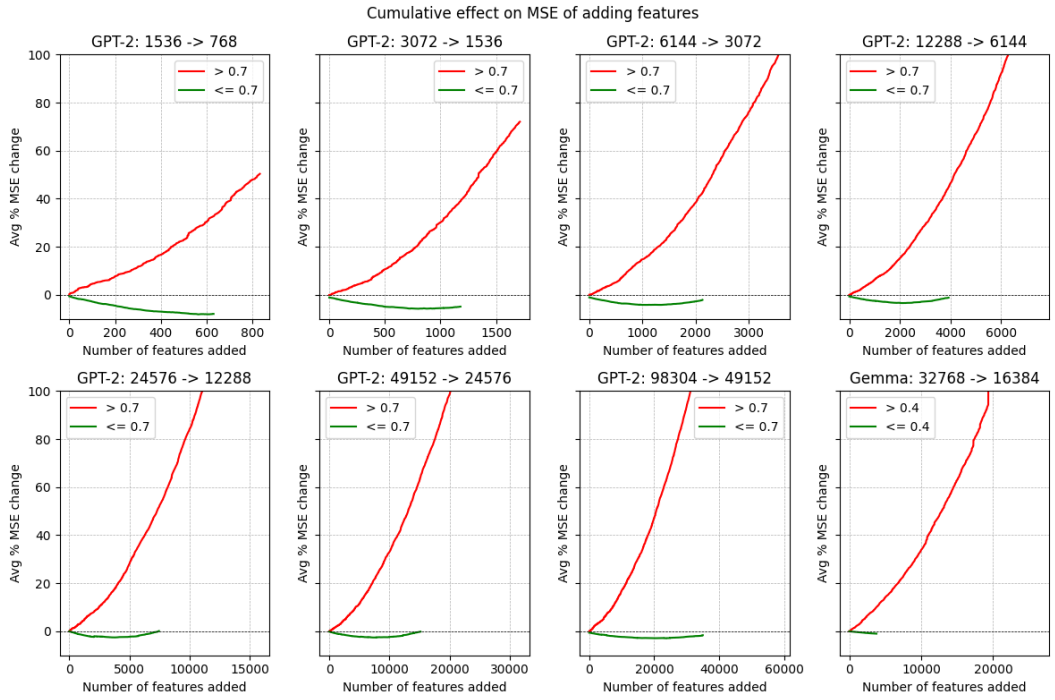
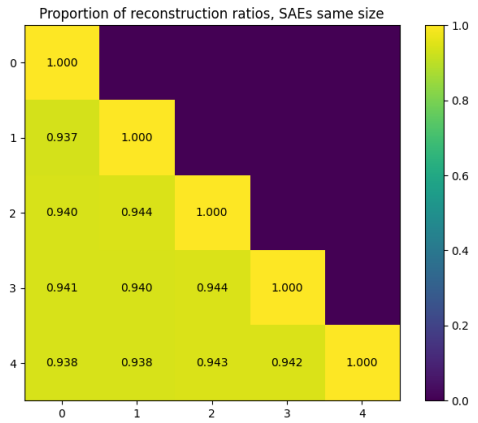
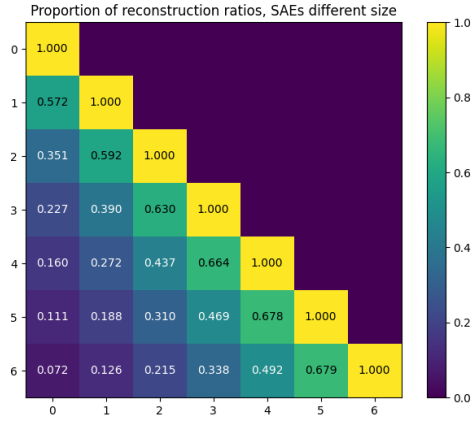


Figure 15: Percentage change of MSE of adding in latents from a larger SAE to a smaller SAE in a random order. Adding in all the latents with cosine similarity ≤ 0.7 from GPT-1536 in GPT-768 reduces the MSE by almost 10%.



(a) Proportion of reconstruction latents in GPT-2 SAEs with 3072 latents. X and Y labels are random training seeds.



(b) Proportion of reconstruction latents in SAEs of different size. X and Y labels are exponential expansion factors, or equivalently row indices in Table 2.

Figure 16: Proportion of reconstruction latents between GPT-2 SAEs of the same size and different sizes.

We compared the proportion of reconstruction latents between SAEs of the same size and SAEs of different sizes. We found that when comparing SAEs with the same size, 94% of the latents are reconstruction latents (cosine similarity > 0.7). These results are displayed in Figure 16.

A concern when stitching two different SAEs is the choice of $\mathbf{b}^{dec}$. However, in practice we find that the $\mathbf{b}^{dec}$s of SAEs trained on the same latents are very similar (minimum cosine similarity of 0.9970, differing by less than 0.1% in magnitude). Figure 17 shows that the decoder biases can be