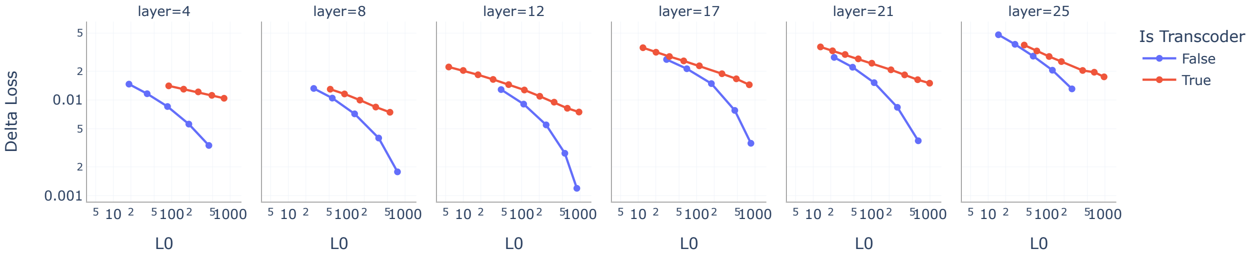


Figure 11 | Transcoders trained to reconstruct MLP output from the MLP input cause a greater increase in loss compared to the vanilla model when compared with an MLP output SAE. The sites are (the MLP sub-) layers throughout Gemma 2B PT.

```
import transformer_lens # pip install transformer-lens

model = transformer_lens.HookedTransformer.from_pretrained(
    "google/gemma-2-2b",
    # In Gemma 2, only the pre-MLP, pre-attention and final RMSNorms can
    # be folded in (post-attention and post-MLP RMSNorms cannot be folded in):
    fold_ln=True,
    # Only valid for models with LayerNorm, not RMSNorm:
    center_writing_weights=False,
    # These model use logits soft-capping, meaning we can't center unembed:
    center_unembed=False,
)
```

Figure 12 | Code for loading Gemma 2B in TransformerLens (Nanda and Bloom, 2022) to use this with our Transcoders.

$\mathbf{y}$ as

$$\mathbf{y} := \mathbf{f}(\mathbf{x}) \odot \mathbf{W}_{\text{dec}}^{T} \nabla_{\mathbf{x}} \mathcal{L}, \tag{12}$$

where we choose the mean-centered logit of the correct next token as the loss function $\mathcal{L}$.

We then normalize the magnitudes of the entries of $\mathbf{y}$ to obtain a probability distribution $p \equiv p(\mathbf{y})$. We can measure how far this distribution diverges from a uniform distribution u over active latents via the KL divergence

$$\mathbf{D}_{\text{KL}}(p \| u) = \log \|\mathbf{y}\|_0 - \mathbf{S}(p), \tag{13}$$

with the entropy $\mathbf{S}(p)$. Note that $0 \leq \mathbf{D}_{\text{KL}}(p\|u) \leq \log \|\mathbf{y}\|_0$. Exponentiating the negative KL divergence gives a new measure r_{L0}

$$r_{L0} := e^{-\mathbf{D}_{\text{KL}}(p\|u)} = \frac{e^{\mathbf{S}(p)}}{\|\mathbf{y}\|_0}, \tag{14}$$

with $\frac{1}{\|\mathbf{y}\|_0} \leq r_{L0} \leq 1$. Note that since $e^{\mathbf{S}}$ can be interpreted as the effective number of active elements, r_{L0} is the ratio of the effective number of active latents (after re-weighting) to the total number of active latents, which we call the 'Uniformity of Active Latent Importance'.

Results In Fig. 16 we show r_{L0} on middle layer SAEs. In line with Rajamanoharan et al. (2024b), we find that the attributed effect becomes more diffuse as more latents are active. This effect is most pronounced for residual stream SAEs, and seems to be independent of language model size and number of SAE latents.

C.4. Additional Gemma 2 IT evaluation results

In this sub-appendix, we provide further evaluations of SAEs on the activations of IT models, continuing Section 4.5.

As mentioned in Section 4.5, we find in Fig. 21 that PT SAEs achieve reasonable FVU on rollouts, but the gap between PT and IT SAEs is larger than in the change in loss in the main text (Fig. 8).

In Fig. 19 we evaluate the FVU on the user prompt and model prefix (not the rollout). In Fig. 20 we evaluate the change in loss (delta loss) on the user prompts, and surprisingly find that splicing in the base model SAE can reduce the loss in expectation in some cases. Our explanation for this result is that post-training does not train models to predict user queries (only predict high-preference model rollouts) and therefore the model is not incentivised to have good predictive loss by default on the user prompt.

While we do not train IT SAEs on Gemma 2 2B, we find that the base SAEs transfer well as measured by FVU in Fig. 22.

Finally, we do not find evidence that rescaling IT activations to have same norm in expectation to the pretraining activations is beneficial (Fig. 23). The trend for individual SAEs in this plot is that their L0 decreases but the Pareto frontier is very slightly worse. This is consistent with prior observations that SAEs are surprisingly adaptable to different L0s (Gao et al., 2024; Smith, 2024).

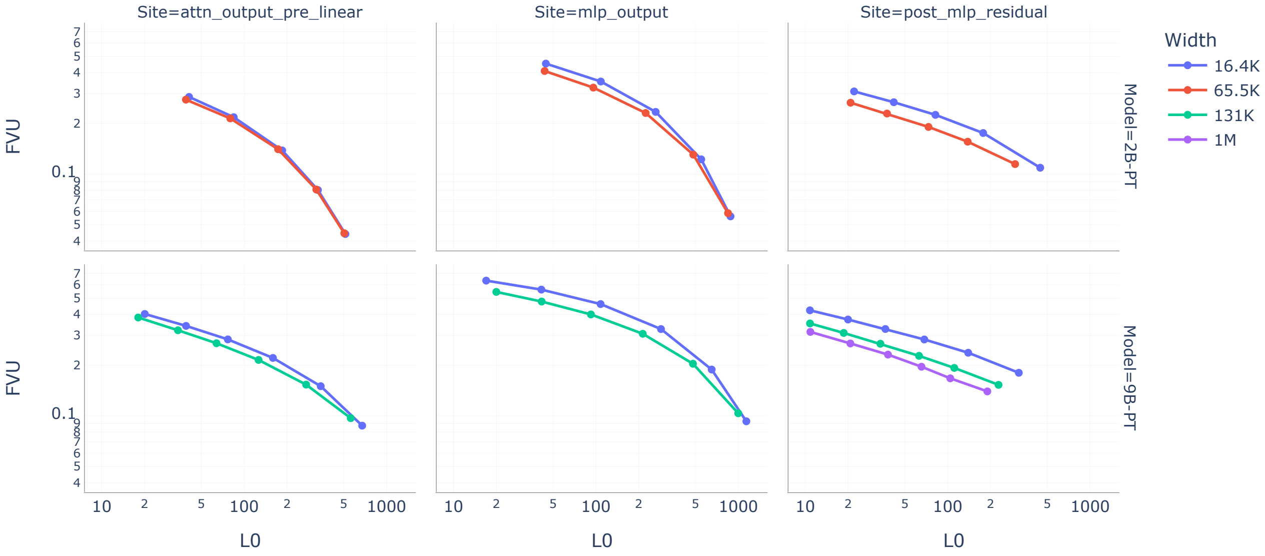
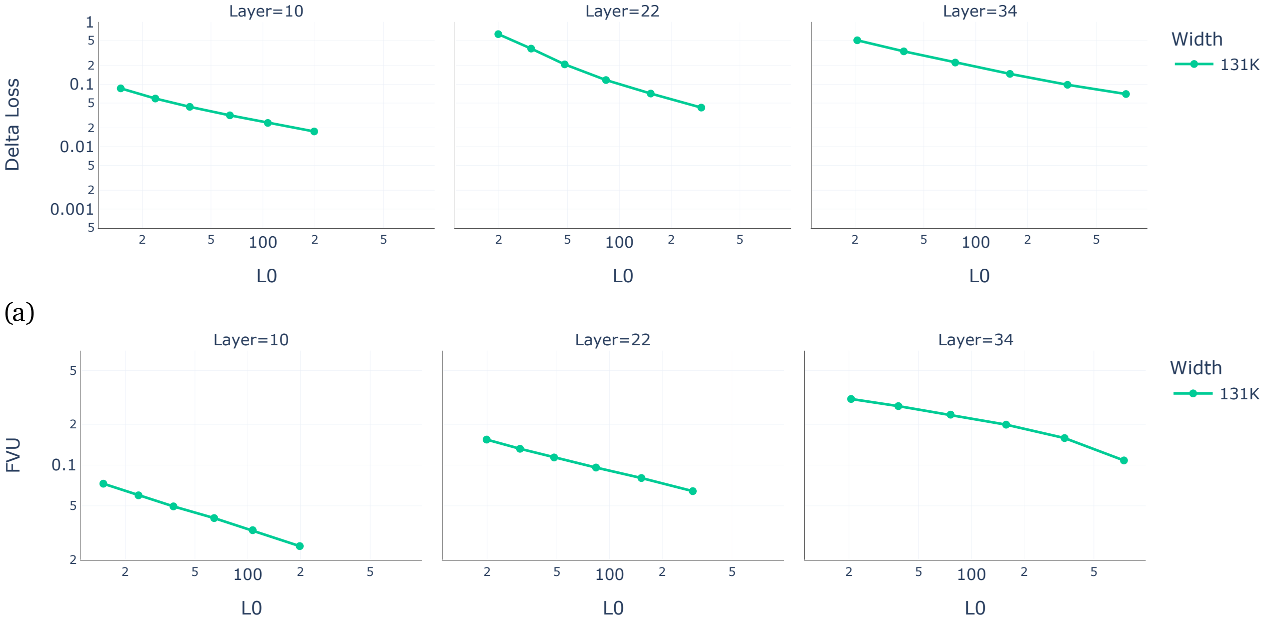


Figure 13 | Sparsity-fidelity trade-off for middle-layer Gemma 2 2B and 9B SAEs using fraction of variance unexplained (FVU) as the measure of reconstruction fidelity.



(a)

(b)

Figure 14 | Sparsity-fidelity trade-off for Gemma 2 27B SAEs using (a) delta LM loss and (b) as measures of reconstruction fidelity.