

Figure 2 | Sparsity-fidelity trade-off for layer 12 Gemma 2 2B and layer 20 Gemma 2 9B SAEs. An ideal SAE should have low delta loss and low L0, i.e. correspond to a point towards the bottom-left corner of each plot. For an analogous plot using FVU as the measure of fidelity see Fig. 13.

4.2. Impact of sequence position

Methodology Prior research has shown that language models tend to have lower loss on later token positions (Olsson et al., 2022). It is thus natural to ask how an SAE’s performance changes over the length of a sequence. Similar to Section 4.1, we track reconstruction loss and delta loss for various sparsity settings, however this time we do not aggregate over the sequence position. Again, we mask out special tokens.

Result Fig. 3 shows how reconstruction loss varies by position for 131K-width SAEs trained on the middle-layer of Gemma 2 9B. Reconstruction loss increases rapidly from close to zero over the first few tokens. The loss monotonically increases by position for attention SAEs, although it is essentially flat after 100 tokens. For MLP SAEs, the loss peaks at around the tenth token before gradually declining slightly. We speculate that this is because attention is most useful when tracking long-range dependencies in text, which matters most when there is significant prior context to draw from, while MLP layers do a lot of local processing, like detecting n-grams (Gurnee et al., 2023), that does not need much context. Like attention SAEs, residual stream SAEs’ loss monotonically increases, plateauing more gradu-

ally. Curves for other models, layers, widths and sparsity coefficients were found to be qualitatively similar.

Fig. 15 shows how delta LM loss varies by sequence position. The high level of noise in the delta loss measurements makes it difficult to robustly measure the effect of position, however there are signs that this too is slightly lower for the first few tokens, particularly for residual stream SAEs.

4.3. Studying the effect of SAE width

Holding all else equal, wide SAEs learn more latent directions and provide better reconstruction fidelity at a given level of sparsity than narrow SAEs. Intuitively, this suggests that we should use the widest SAEs practicable for downstream tasks. However, there are also signs that this intuition may come with caveats. The phenomenon of ‘feature-splitting’ (Bricken et al., 2023) – where latents in a narrow SAE seem to split into multiple specialized latents within wider SAEs – is one sign that wide SAEs do not always use their extra capacity to learn a greater breadth of features (Bussmann et al., 2024). It is plausible that the sparsity penalty used to train SAEs encourages wide SAEs to learn frequent compositions of existing features over (or at least in competition

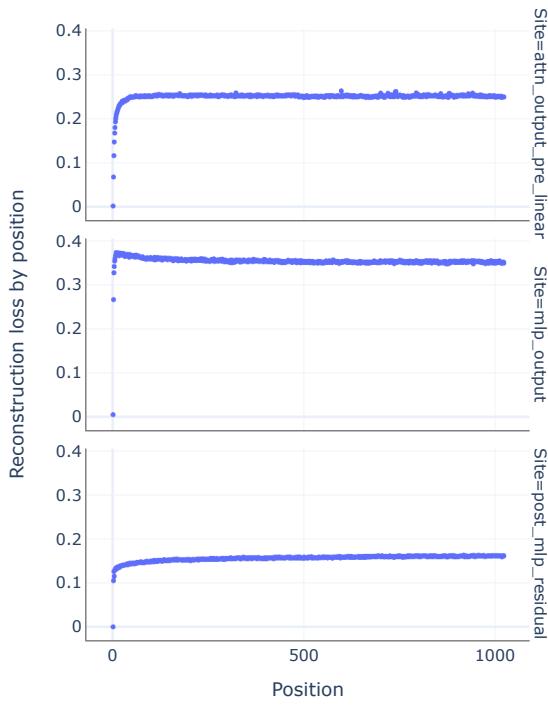


Figure 3 | Reconstruction loss by sequence position for Gemma 2 9B middle-layer 131K-width SAEs with $\lambda = 10^{-3}$.

with) using their increased capacity to learn new features (Anders et al., 2024). If this is the case, it is currently unclear whether this is good or bad for the usefulness of SAEs on downstream tasks.

In order to facilitate research into how SAEs' properties vary with width, and in particular how SAEs with different widths trained on the same data relate to one another, we train and release a ‘feature-splitting’ suite of mid-network residual stream SAEs for Gemma 2 2B and 9B PT with matching sparsity coefficients and widths between $2^{14} \approx 16.4\text{K}$ and $2^{19} \approx 524\text{K}$ in steps of powers of two.⁸ The SAEs are trained with different sparsity settings after layers 12 and 20 of Gemma 2 2B and 9B respectively.

Sparsity-fidelity trade-off Similar to Section 4.1, Fig. 4 compares fidelity-versus-sparsity

⁸Note the 1M-width SAEs included in Fig. 2 do not form part of this suite as they were trained using a different range of values for the sparsity coefficient λ .

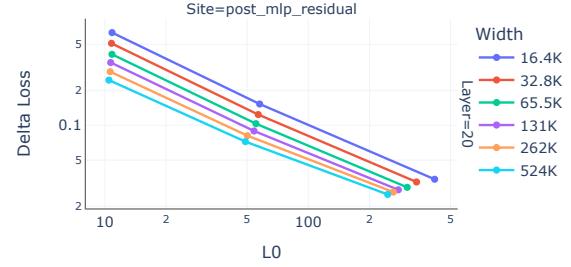


Figure 4 | Delta loss versus sparsity curves for a series of SAEs of differing width (keeping λ and other hyperparameters constant), trained on the residual stream after layer 20 of Gemma 2 9B.

curves for SAEs of differing width in this ladder.

Latent firing frequency Fig. 5 shows frequency histograms for $\lambda = 6 \times 10^{-4}$ SAEs in the same ladder of widths from 2^{14} to 2^{19} latents. To compute these, we calculate the firing frequency of each latent over 20,000 sequences of length 1,024, masking out special tokens. Note that the mode and most of the mass shifts towards lower frequencies with increased number of latents. However there remains a cluster of ultra-high frequency latents, which has also been observed for TopK SAEs but not for Gated SAEs (Cunningham and Conerly, 2024; Gao et al., 2024; Rajamanoharan et al., 2024b).

4.4. Interpretability of latents

The interpretability of latents for a subset of the SAEs included in Gemma Scope was investigated in Rajamanoharan et al. (2024b); latents were evaluated using human raters and via LM generated explanations. For completeness, we include the key findings of those studies here and refer readers to section 5.3 of that work for a detailed discussion of the methodology. Both the human rater and LM explanations studies evaluated JumpReLU, TopK and Gated SAEs of width 131K trained on all sites at layers 9, 20, and 31 of Gemma 2 9B. Fig. 6 shows human raters' judgment of latent interpretability for each investigated SAE architecture. Fig. 7 shows the Pearson correlation between the language model (LM) simulated activations based on LM-generated explanations and the ground truth activation values.

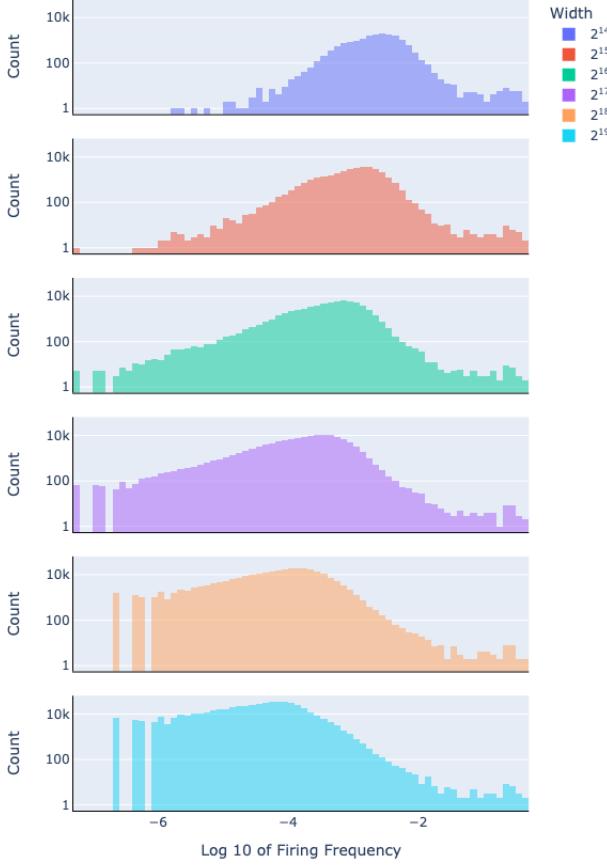


Figure 5 | Frequency histogram of SAEs trained on Gemma 2 9B, layer 20, post MLP residual with sparsity coefficient $\lambda = 6 \times 10^{-4}$. (These correspond to the SAEs with $L_0 \approx 50$ in Fig. 4.)

On both metrics, there is little discernible difference between JumpReLU, TopK and Gated SAEs.

4.5. SAEs trained on base models transfer to IT models

Additional IT SAE training Prior research has shown that SAEs trained on base model activations also faithfully reconstruct the activations of IT models derived from these base models (Kissane et al., 2024b). We find further evidence for these results by comparing the Gemma Scope SAEs with several SAEs we train on the activations from Gemma 2B 9B IT. Specifically, we train these IT SAEs by taking the same pretraining documents used for all other SAEs (Section 3.1) and prepend them with Gemma’s IT prefix for the user’s query, and append Gemma’s IT prefix for

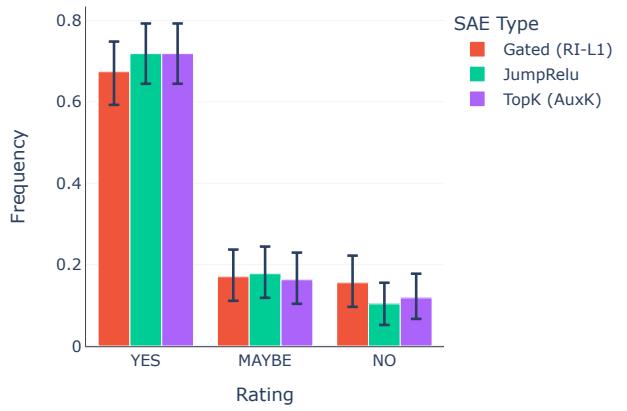


Figure 6 | Human rater scores of latent interpretability. Latents from all SAE architectures are rated as similarly interpretable by human raters. (Reproduced from Rajamanoharan et al. (2024b).)

the model’s response.⁹ We then train each SAE to reconstruct activations at all token positions besides the user prefix (since these tokens have much larger norm (Kissane et al., 2024b), and are the same for every document). We also release the weights for these SAEs to enable further research into the differences between training SAEs on base and IT models.¹⁰

Methodology To evaluate the SAEs trained on the IT model’s activations, we generate 1,024 rollouts of the Gemma 2 9B IT model on a random sample of the SFT data used to train Gemini 1.0 Ultra (Gemini Team, 2024), with temperature 1.0. We then use SAEs trained on the residual stream of the base model and the IT model to reconstruct these activations, and measured the FVU.

Results In Fig. 8 we show that using PT model SAEs results in increases in cross-entropy loss almost as small as the increase from the SAEs directly trained on the IT model’s activation. We show further evaluations such as on Gemma 2 2B, measuring FVU rather than loss, and using

⁹See e.g. <https://huggingface.co/google/gemma-2-2b-it> for the user and model prefixes.

¹⁰<https://huggingface.co/google/gemma-scope-9b-it-res>