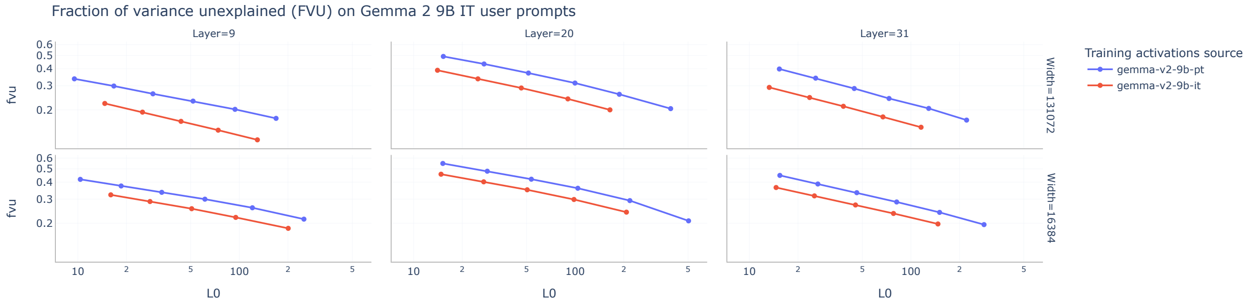

Figure 19 | Fraction of variance unexplained when using SAEs trained on Gemma 2 9B (base and IT) to reconstruct the activations generated with Gemma 2 9B IT on user prompts.

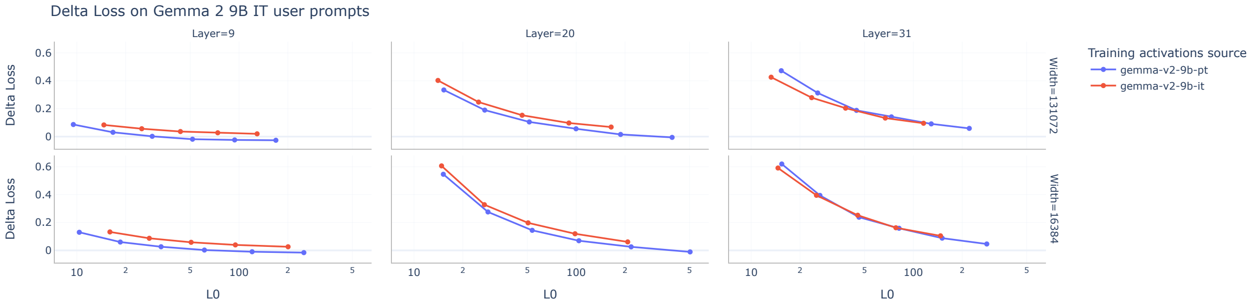

Figure 20 | Change in loss when splicing in SAEs trained on Gemma 2 9B (base and IT) to reconstruct the activations generated with Gemma 2 9B IT on user prompts.

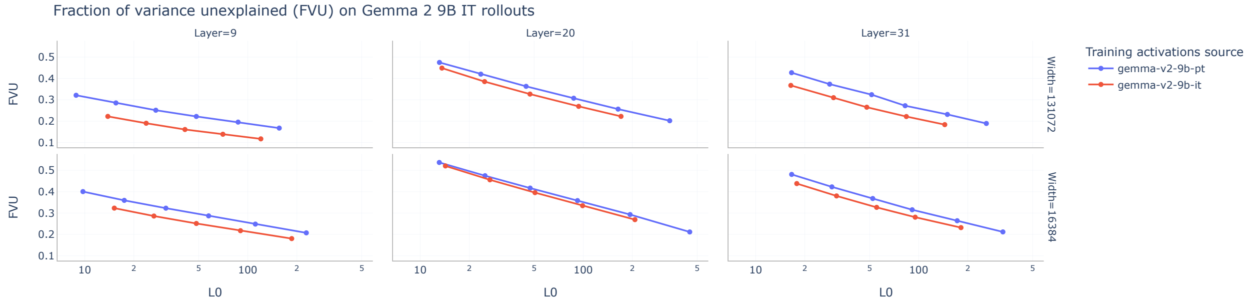

Figure 21 | Fraction of variance unexplained when using SAEs trained on Gemma 2 9B (base and IT) to reconstruct the activations generated with Gemma 2 9B IT on rollouts.

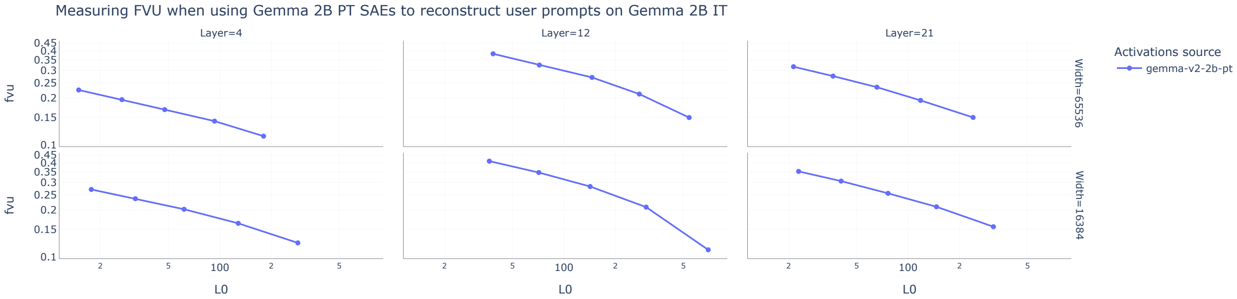

Figure 22 | Fraction of variance unexplained when using SAEs trained on Gemma 2 2B PT to reconstruct the activations generated with Gemma 2 2B IT on user prompts.

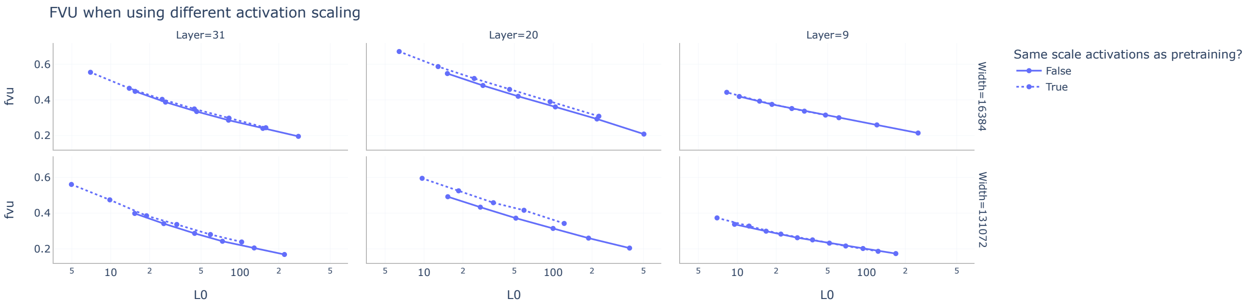

Figure 23 | Fraction of variance unexplained when using SAEs trained on Gemma 2 9B PT to reconstruct the activations generated with Gemma 2 9B IT on rollouts, including when rescaling the IT activations to have the same norm (in expectation) as the pretraining activations.