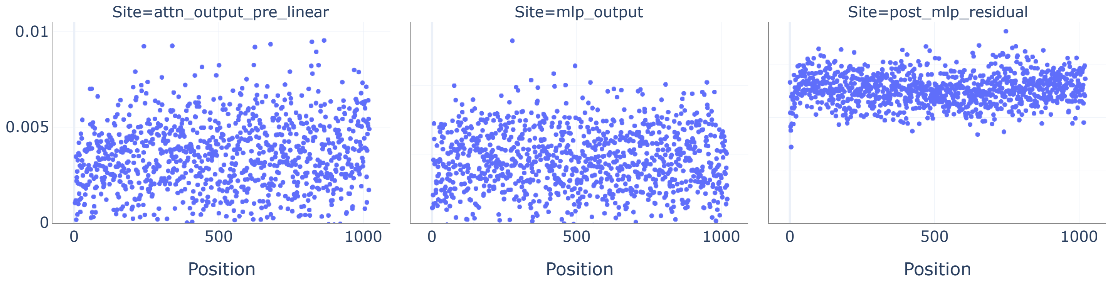

Figure 15 | Delta loss by sequence position for Gemma 2 9B middle-layer 131K-width SAEs with $\lambda = 10^{-3}$.

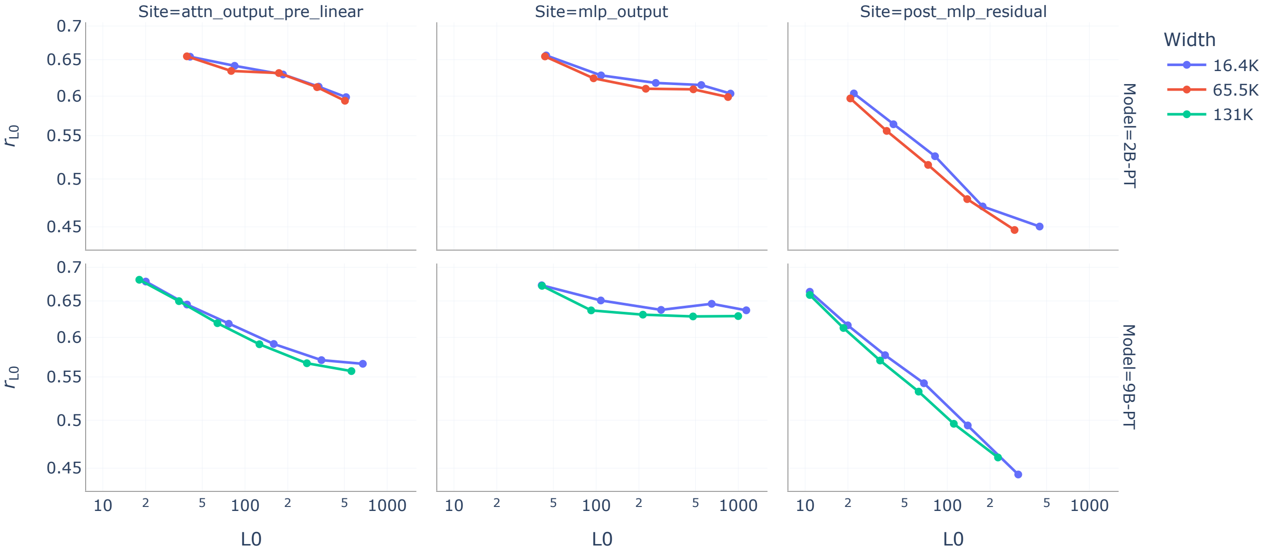

Figure 16 | Uniformity of active latent importance for the middle layer SAEs.

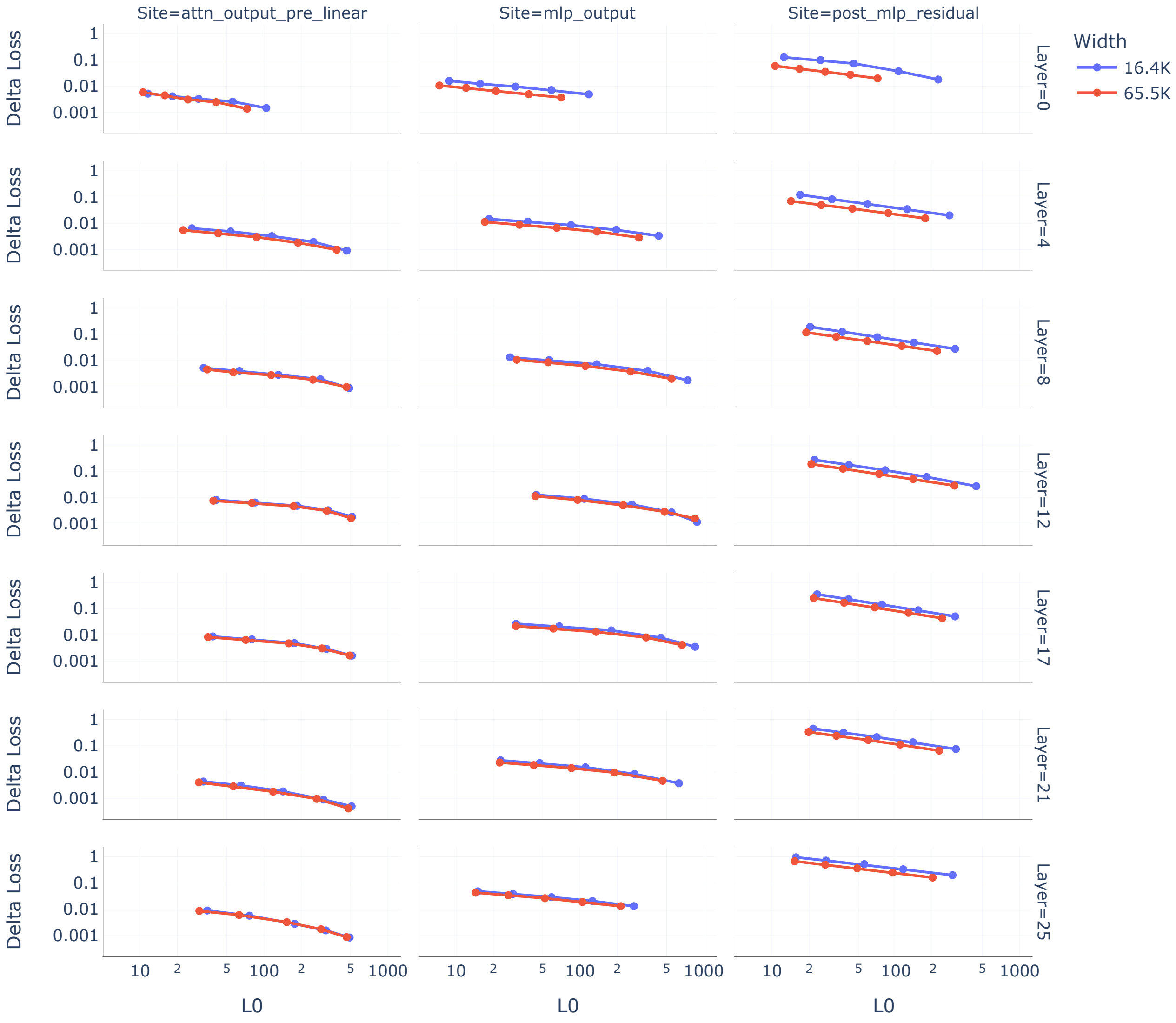

Figure 17 | Sparsity-fidelity trade-off across multiple layers of Gemma 2 2B, approximately evenly spaced. (Note Gemma 2 2B has 26 layers.)

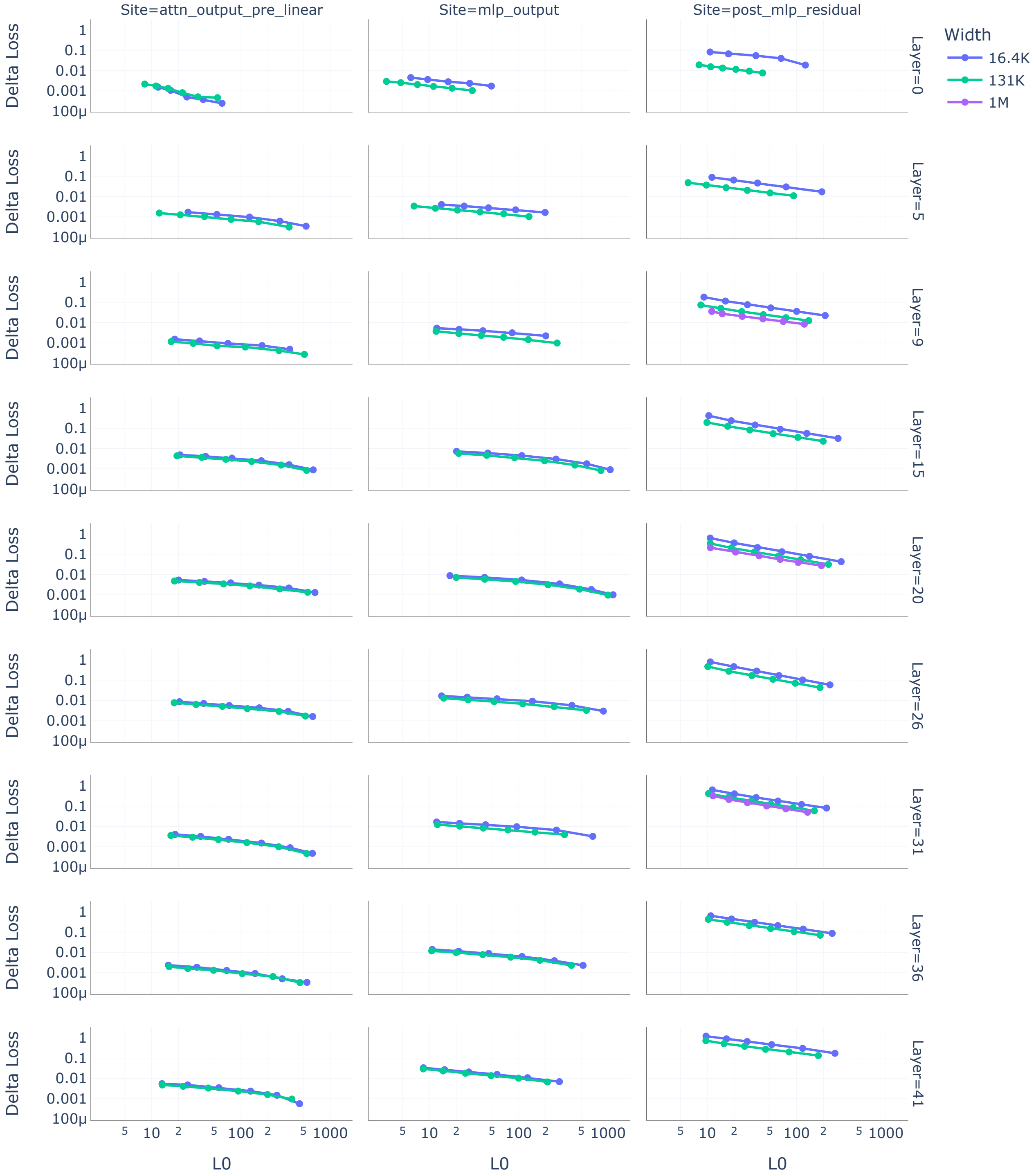

Figure 18 | Sparsity-fidelity trade-off across multiple layers of Gemma 2 9B, approximately evenly spaced. (Note Gemma 2 2B has 42 layers.)