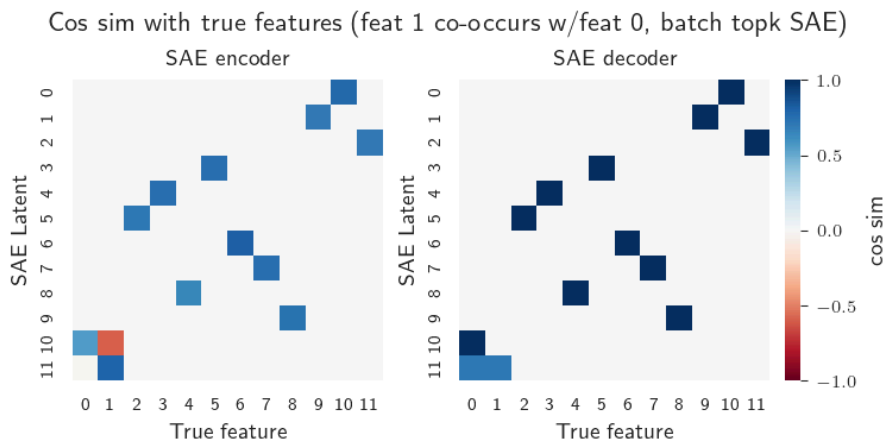
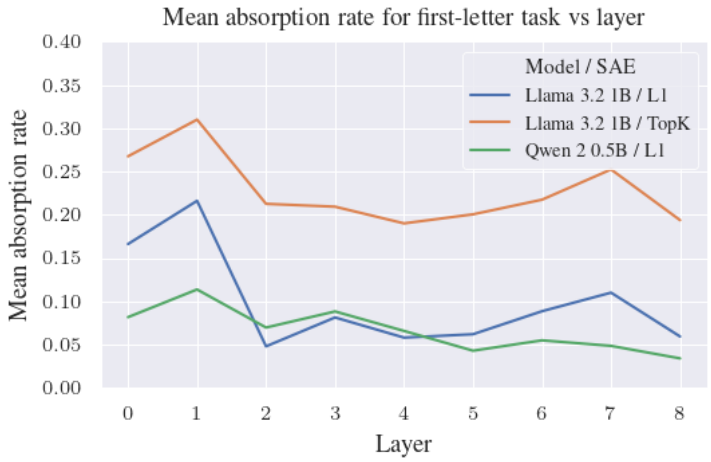


Cosine similarity between batch top-k SAE encoder and decoder and true features for a toy model of absorption. The toy model consists of 12 mutually-orthogonal true features, each firing with magnitude 1.0. All features fire independently with probability 0.15, except for features 0 and 1. Feature 0 fires with probability 0.4, and feature 1 fires with magnitude 0.6 if feature 0 is firing, but cannot fire on its own. Feature 0 and feature 1 thus form a hierarchy where feature 0 is the parent and feature 1 is the child. We train a batch top-k SAE with k=2 and 12 latents on this toy model, and see a classic absorption relationship between latents 1 and 0, tracking features 0 and 1, respectively. The remaining independent features are learned correctly.



Updated figure 9.c including top-k SAE absorption rate for layers 0-8 of Llama 3.2 1B. These top-k SAEs are trained layers on 300M tokens with K=10 using SAELens. We still see strong absorption effects in all top-k SAEs, indicating that top-k SAEs are not immune to absorption. The relatively higher absorption seen in these top-k SAEs should not be taken to indicate that top-k SAEs have more absorption in general, as we do not control for L0 in these comparisons.