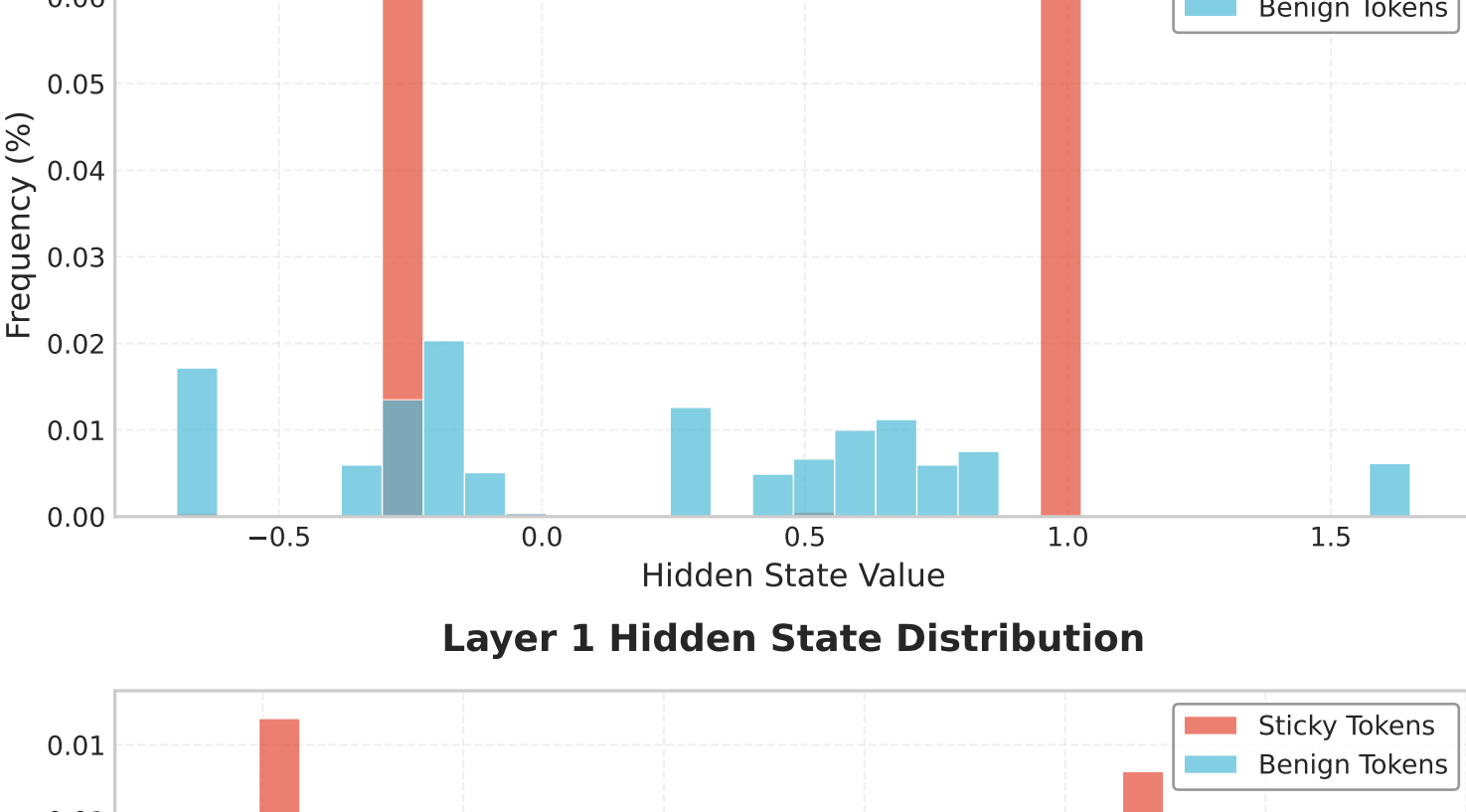
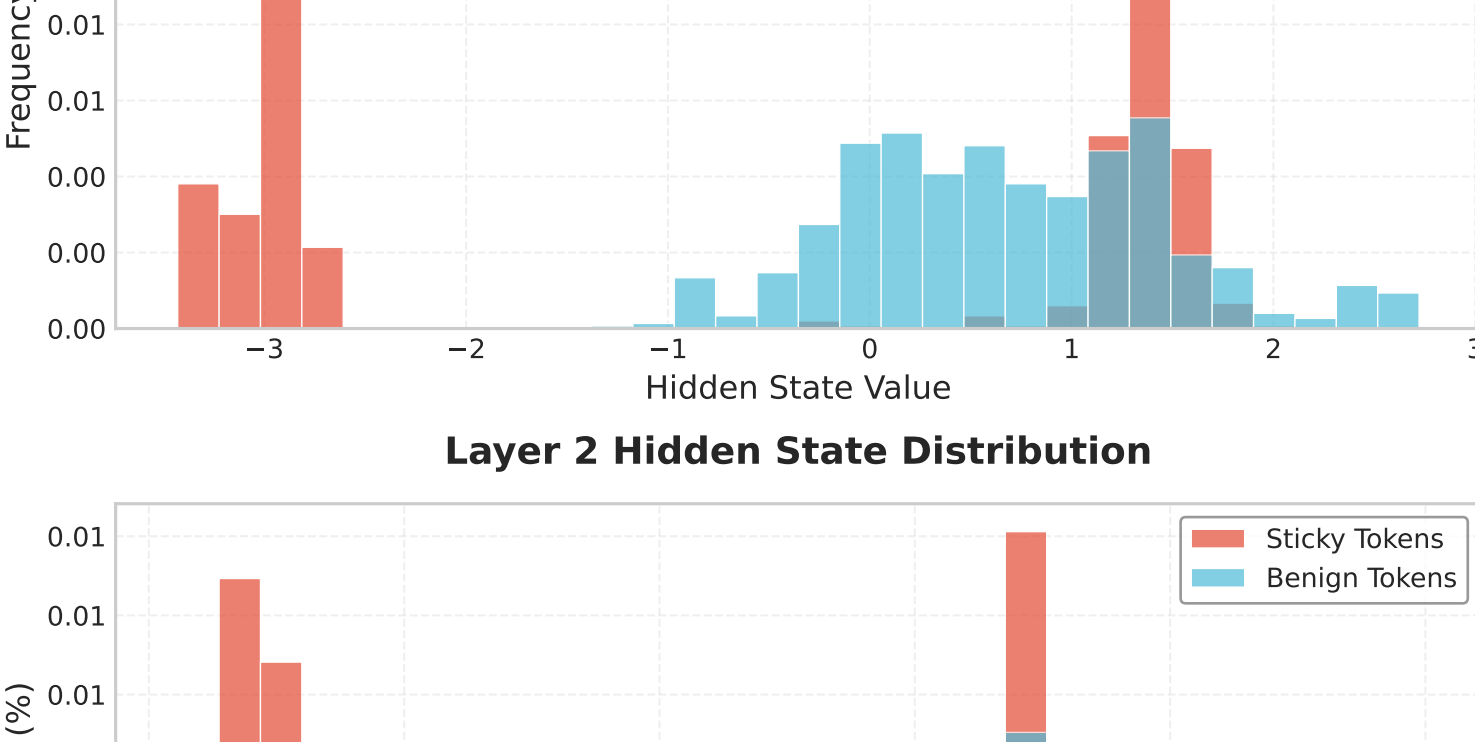


Hidden States Distribution Across Layers

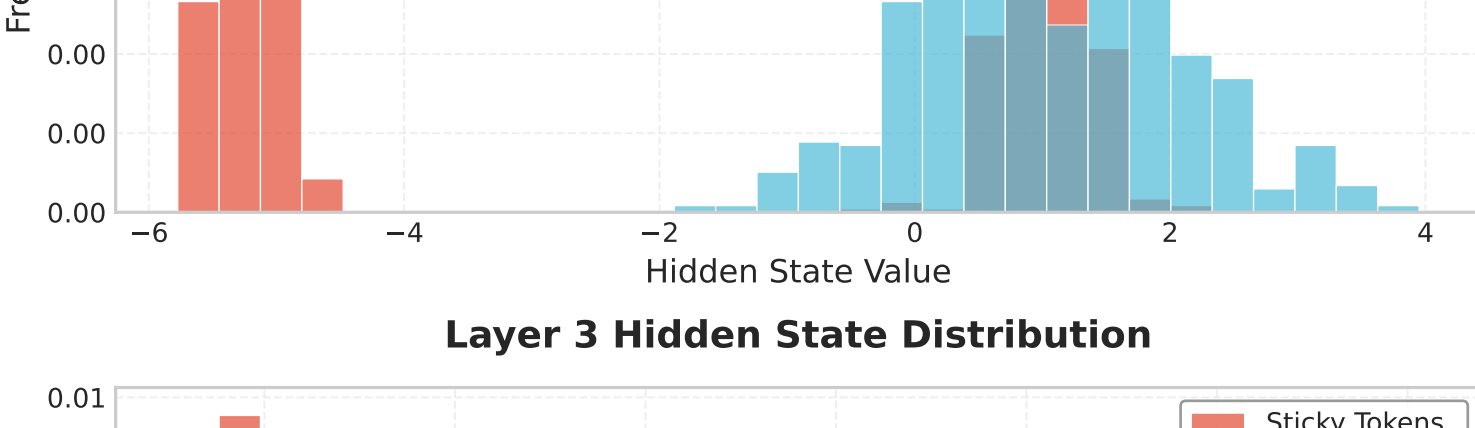
Layer 0 Hidden State Distribution



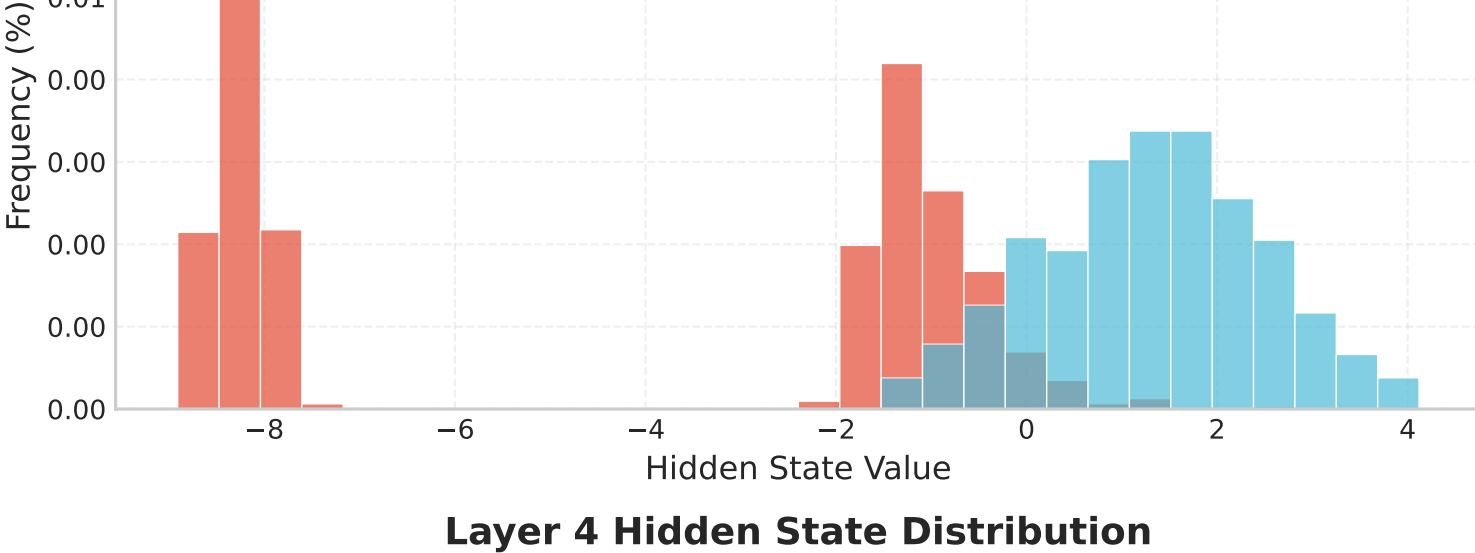
Layer 1 Hidden State Distribution



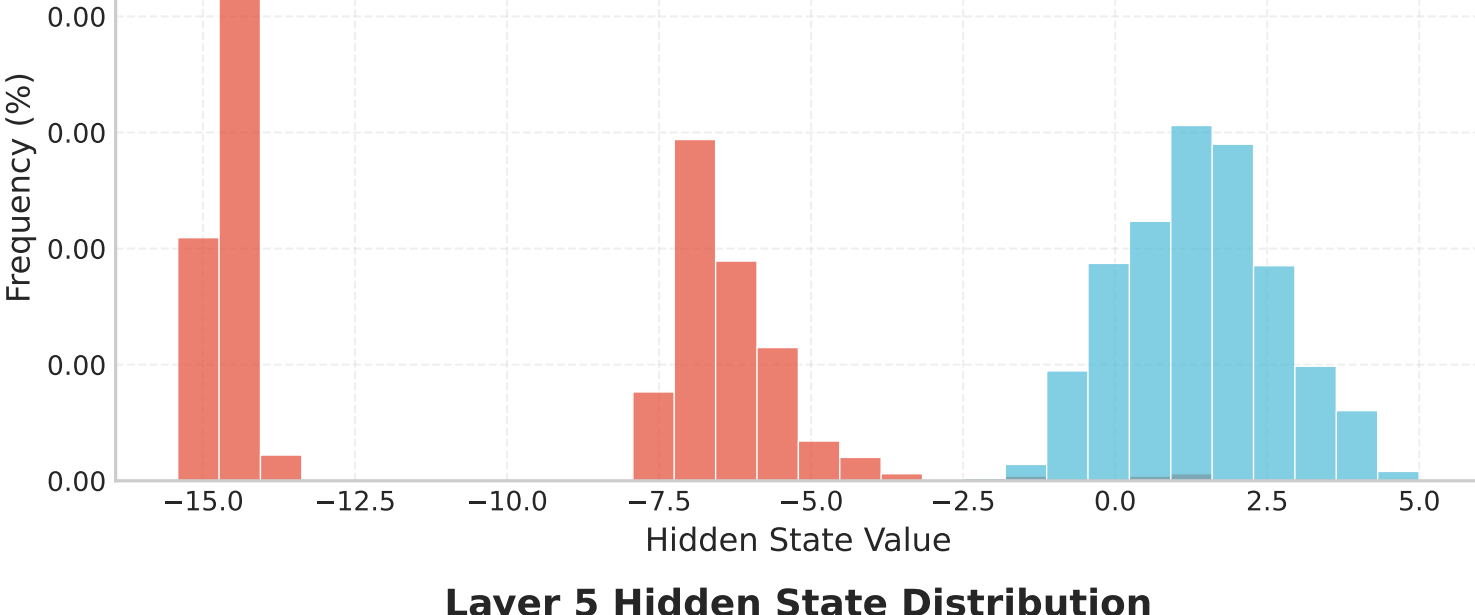
Layer 2 Hidden State Distribution



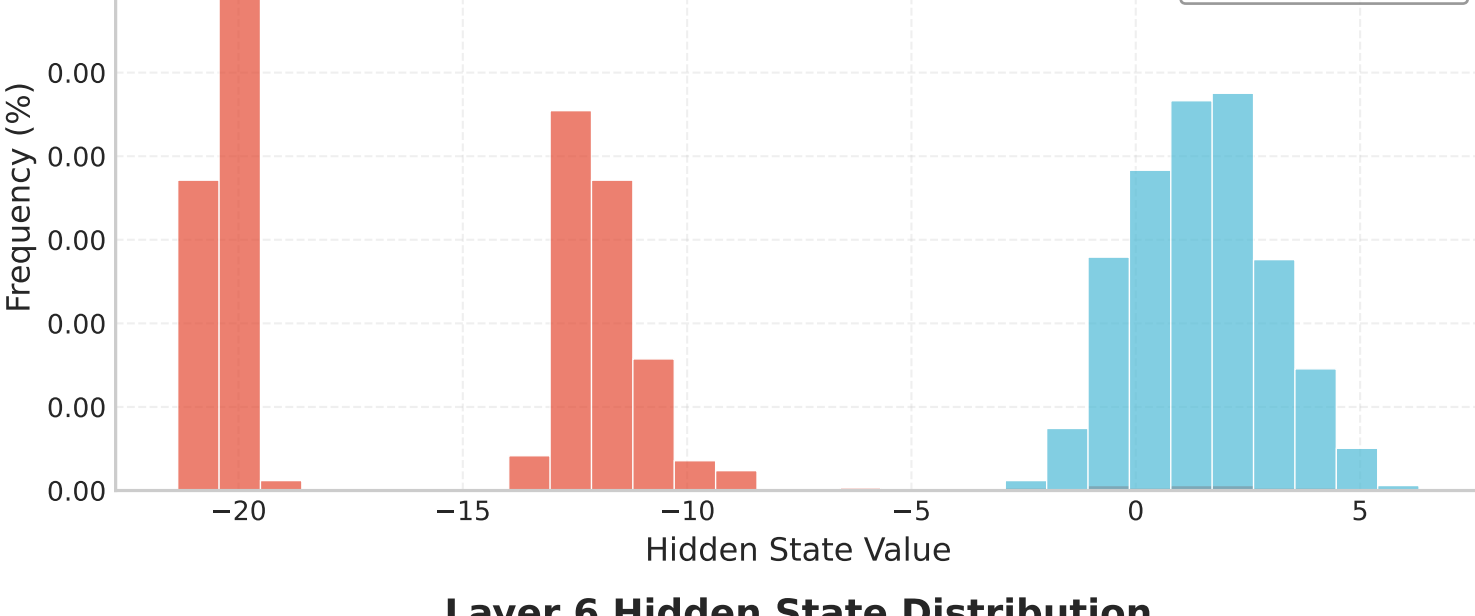
Layer 3 Hidden State Distribution



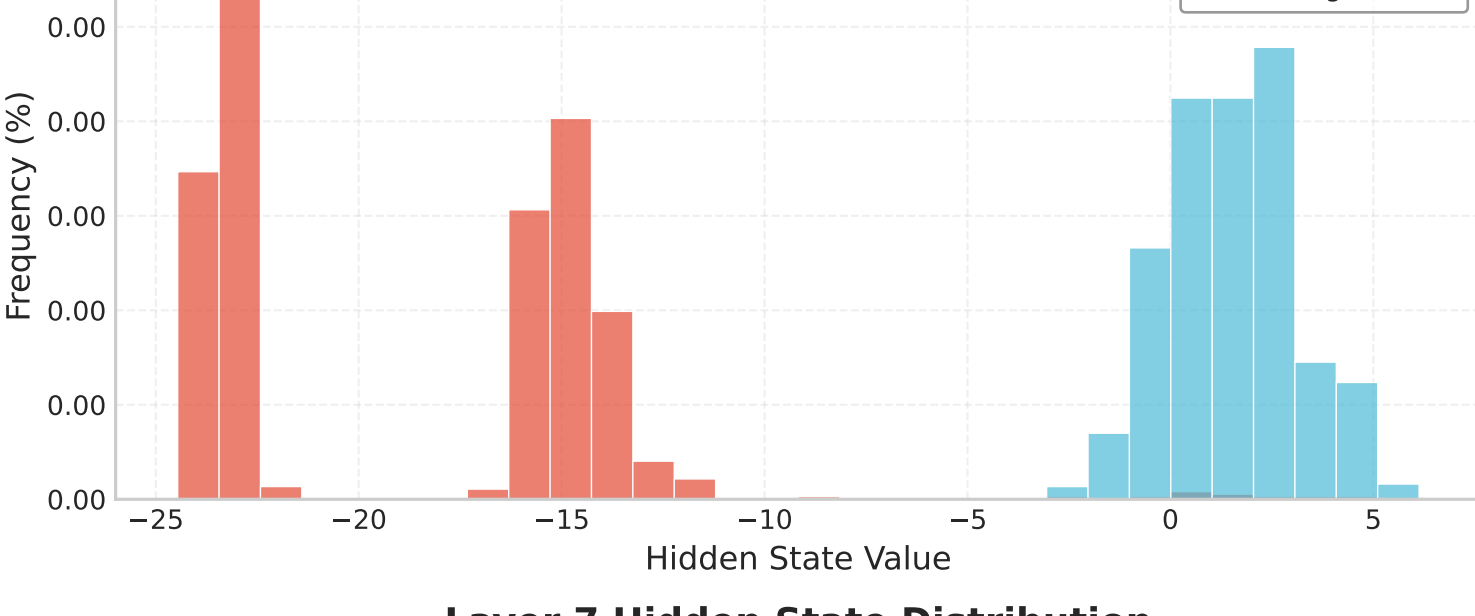
Layer 4 Hidden State Distribution



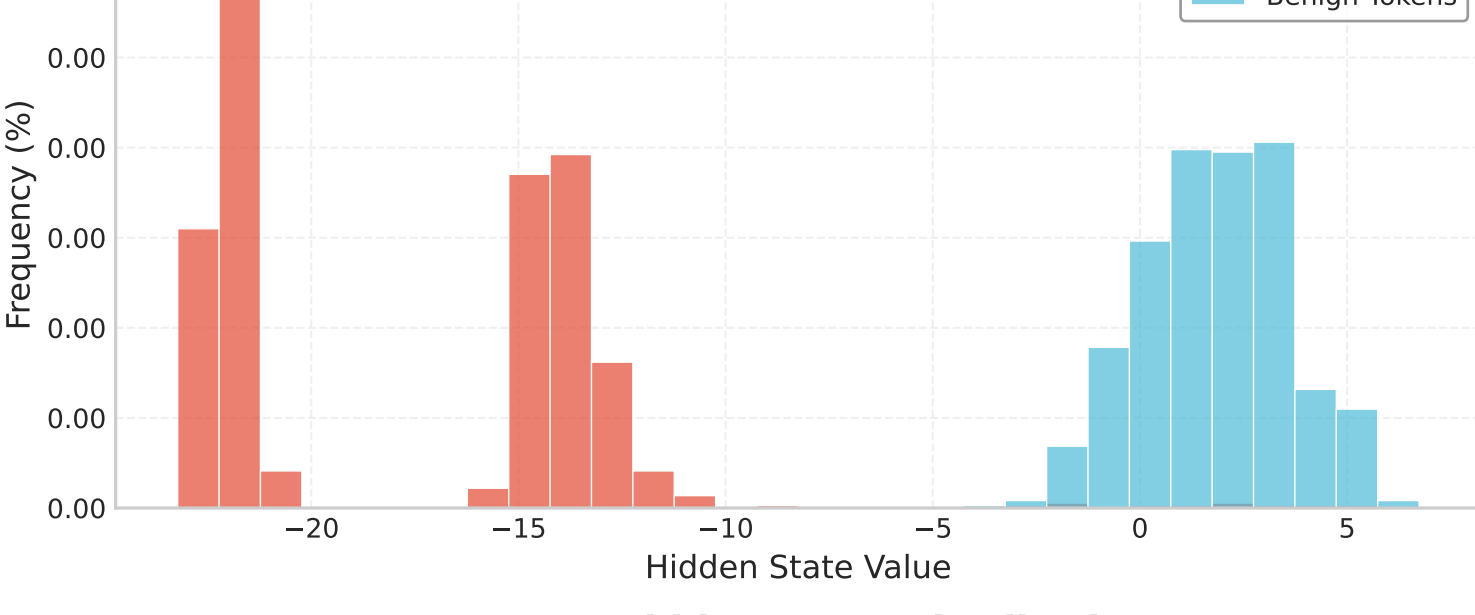
Layer 5 Hidden State Distribution



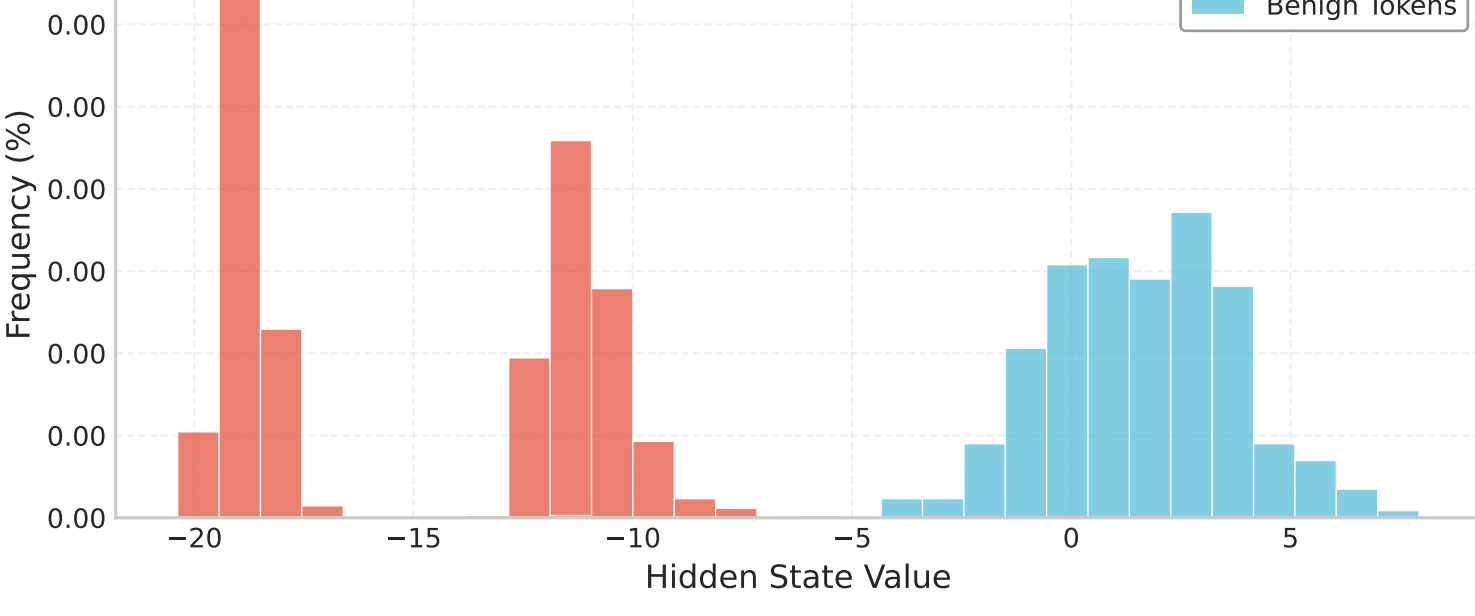
Layer 6 Hidden State Distribution



Layer 7 Hidden State Distribution



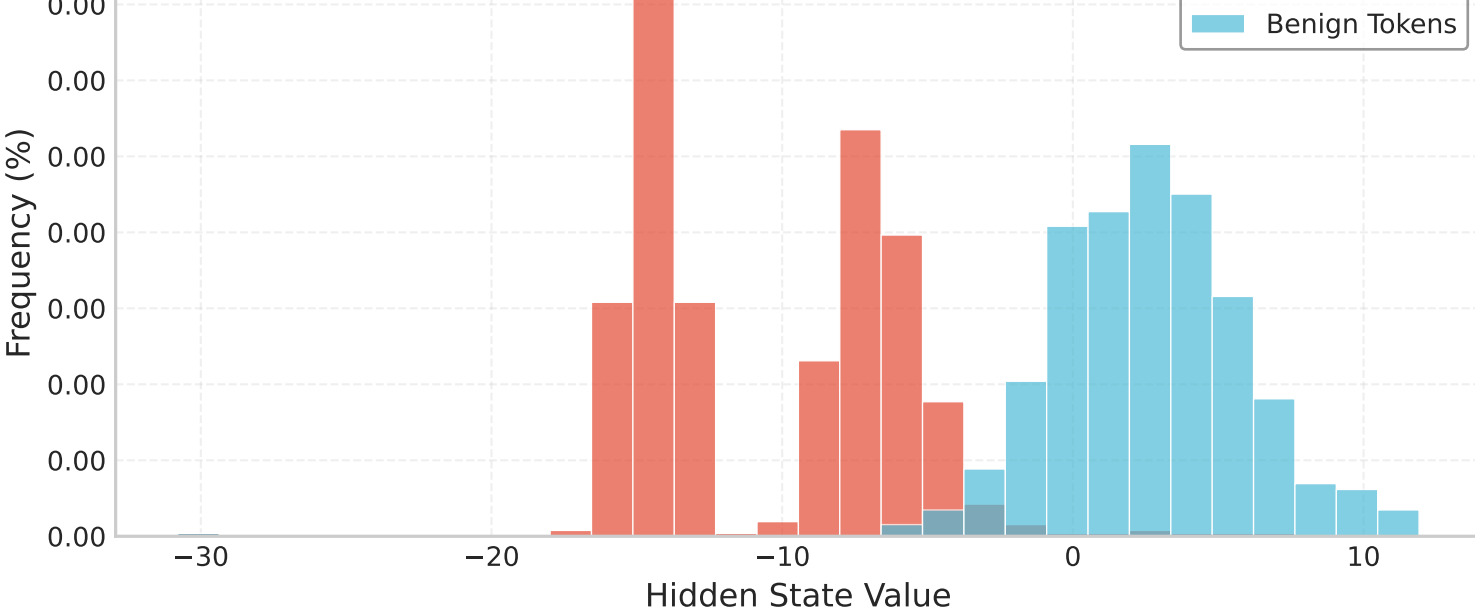
Layer 8 Hidden State Distribution



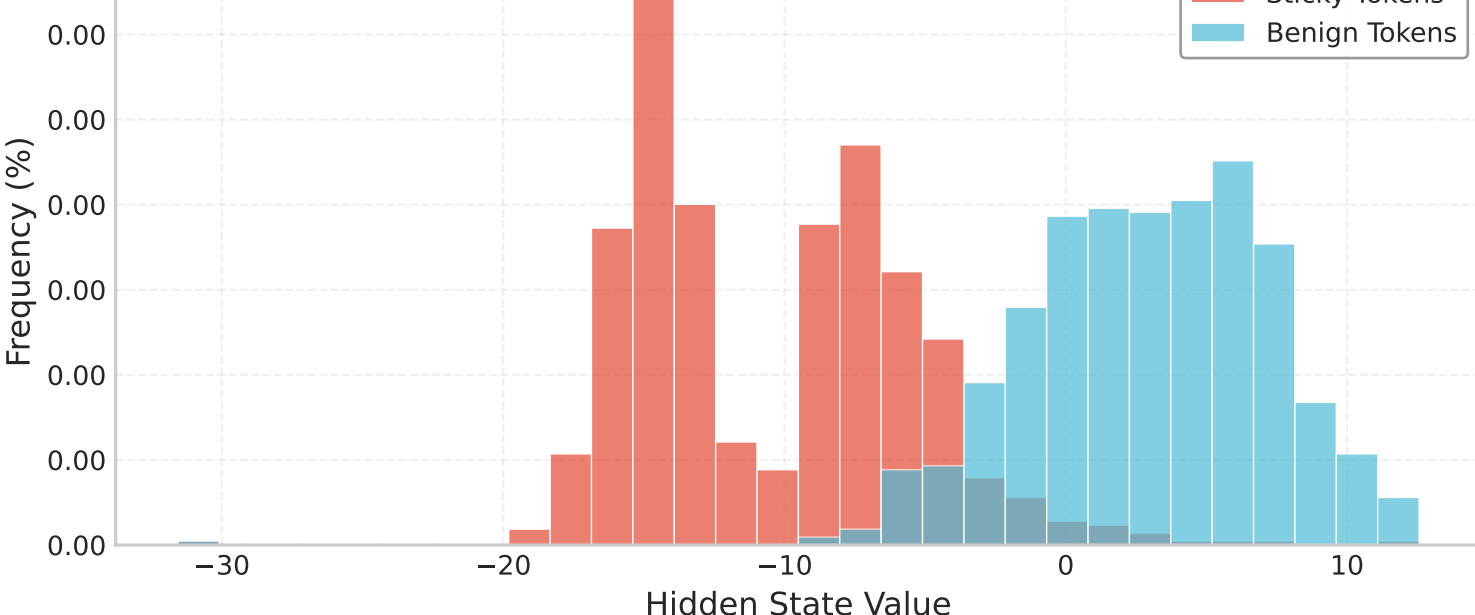
Layer 9 Hidden State Distribution



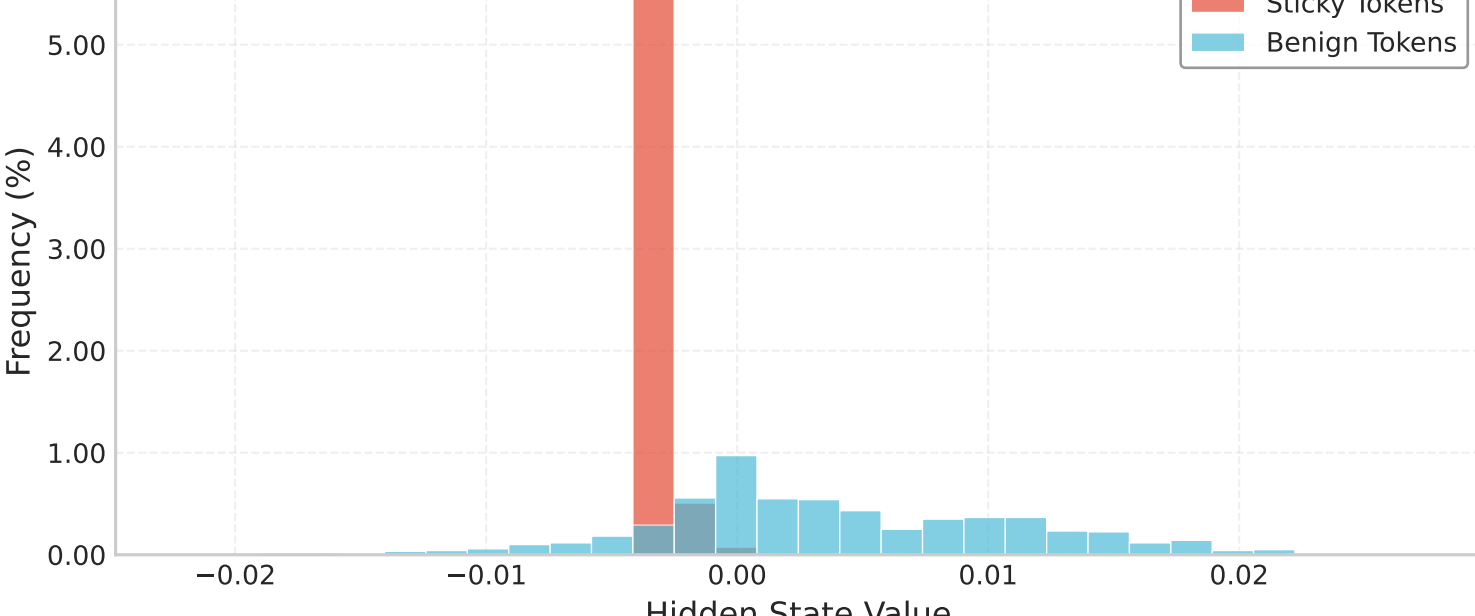
Layer 10 Hidden State Distribution



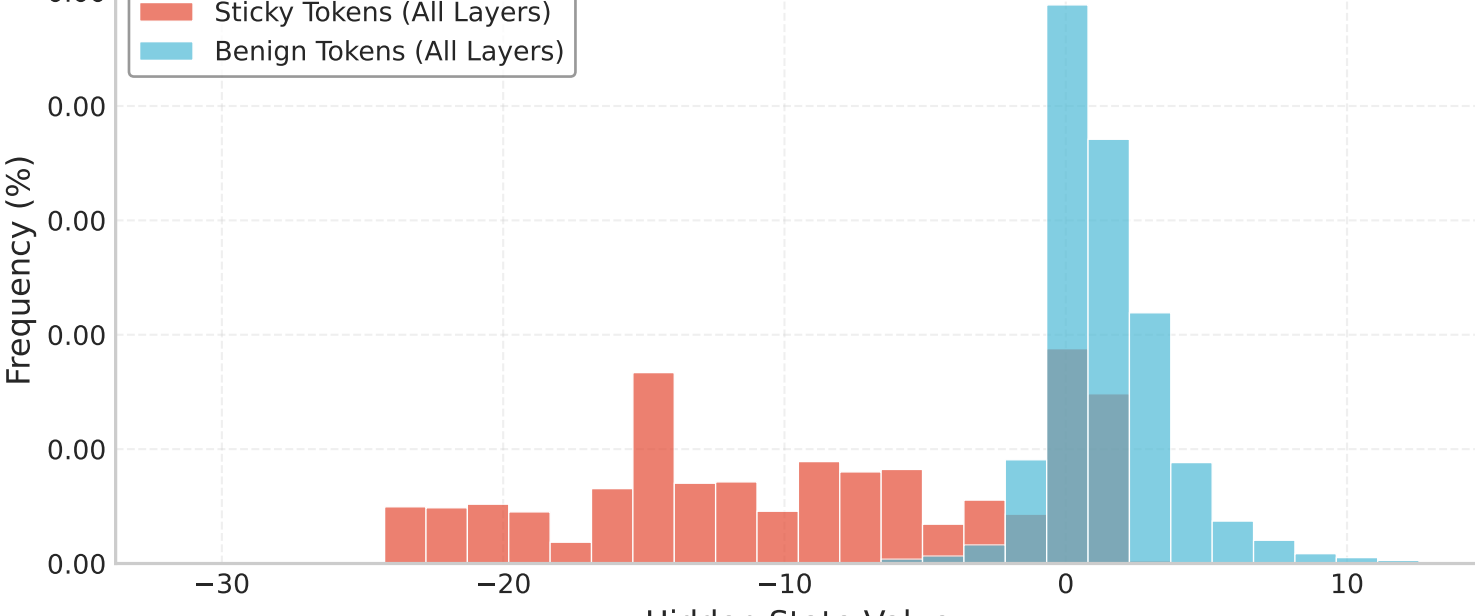
Layer 11 Hidden State Distribution



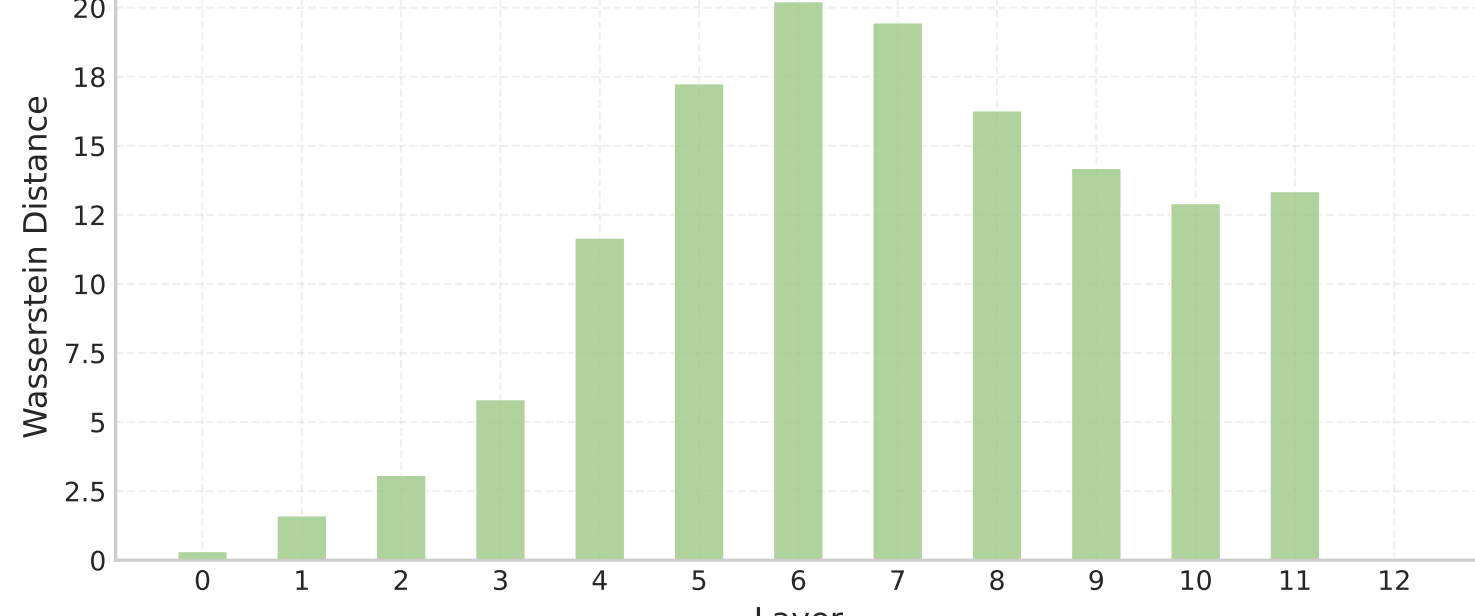
Layer 12 Hidden State Distribution



All Layers Combined Hidden State Distribution



Wasserstein Distance Between Sticky and Benign Tokens Across Layers



KL Divergence Between Sticky and Benign Tokens Across Layers

