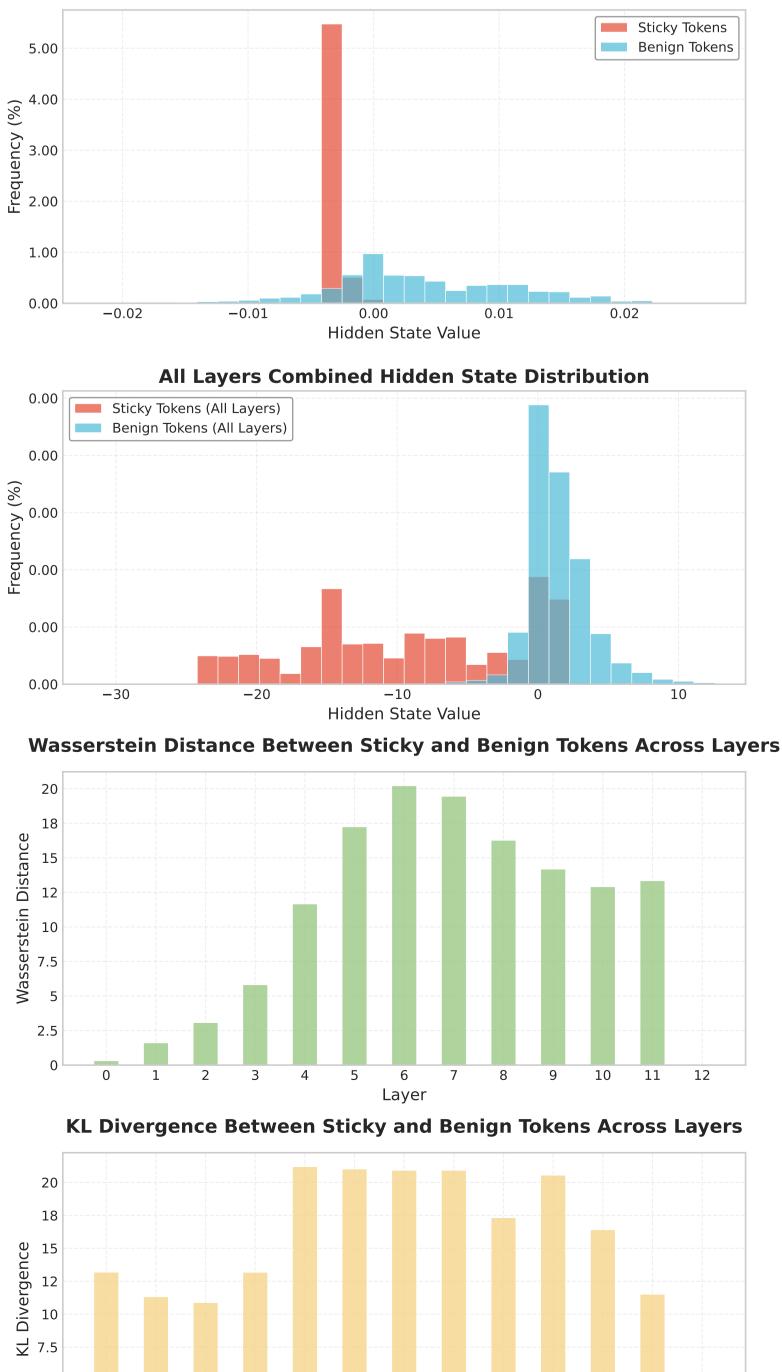
Hidden States Distribution Across Layers Layer 0 Hidden State Distribution Sticky Tokens 0.06 Benign Tokens 0.05 Frequency (%) 0.04 0.03 0.02 0.01 0.00 -0.50.0 0.5 1.0 1.5 Hidden State Value **Layer 1 Hidden State Distribution** Sticky Tokens 0.01 Benign Tokens 0.01 Frequency (%) 0.01 0.00 0.00 0.00 **-**3 **-**2 1 3 Hidden State Value **Layer 2 Hidden State Distribution** 0.01 Sticky Tokens Benign Tokens 0.01 Frequency (%) 0.00 0.00 0.00 -6 -4 -2Hidden State Value **Layer 3 Hidden State Distribution** 0.01 Sticky Tokens Benign Tokens 0.01 0.01 Frequency (%) 0.00 0.00 0.00 0.00 0.00 -8 0 -6 Hidden State Value **Layer 4 Hidden State Distribution** 0.01 Sticky Tokens Benign Tokens 0.00 Frequency (%) 0.00 0.00 0.00 0.00 -15.0-12.5-10.0-7.5-5.0-2.50.0 2.5 5.0 Hidden State Value **Layer 5 Hidden State Distribution** 0.00 Sticky Tokens Benign Tokens 0.00 0.00 Frequency (%) 0.00 0.00 0.00 0.00 0.00 -20 -15Hidden State Value **Layer 6 Hidden State Distribution** 0.00 Sticky Tokens Benign Tokens 0.00 0.00 0.00 **-**25 -20 -15-10**-**5 Hidden State Value **Layer 7 Hidden State Distribution** Sticky Tokens 0.00 Benign Tokens 0.00 Frequency (%)

.o .o .o

.o .o .o

.o .o .o 0.00 0.00 -20 -15 -10 Hidden State Value **Layer 8 Hidden State Distribution** 0.00 Sticky Tokens Benign Tokens 0.00 0.00 0.00 0.00 0.00 Hidden State Value **Layer 9 Hidden State Distribution** 0.00 Sticky Tokens Benign Tokens 0.00 0.00 0.00 0.00 0.00 -25 -20 -1510 Hidden State Value **Layer 10 Hidden State Distribution** Sticky Tokens 0.00 Benign Tokens 0.00 Frequency (%)
0.0000
0.0000 0.00 0.00 0.00 -30-20 -1010 Hidden State Value **Layer 11 Hidden State Distribution** Sticky Tokens Benign Tokens 0.00 0.00 Frequency (%) 0.00 0.00 0.00 0.00 0.00 -30-20 -100 10 Hidden State Value **Layer 12 Hidden State Distribution** Sticky Tokens 5.00 Benign Tokens 4.00 Frequency (%) 3.00 2.00 1.00 0.00 -0.02-0.010.00 0.02 0.01 Hidden State Value **All Layers Combined Hidden State Distribution** 0.00 Sticky Tokens (All Layers) Benign Tokens (All Layers) 0.00 0.00



5

2.5

0

0

1

2

3

4

5

6

Layer

7

8

9

10

11

12