

True Positive Rate

Gemma-2-2B: L10

Gemma-2-2B: L20

Gemma-2-9B: L20

Gemma-2-9B: L31

Method

DiffMean

IG

IxG

LAT

PCA

Probe

ReFT-r1

SAE

SSV

False Positive Rate

1.00
0.75
0.50
0.25
0.001.00
0.75
0.50
0.25
0.00

0.00 0.25 0.50 0.75 1.00 0.00 0.25 0.50 0.75 1.00