

Refusal Rate

1.0
0.8
0.6
0.4
0.2
0.0

Untriggered

Triggered



baseline



finetuned

