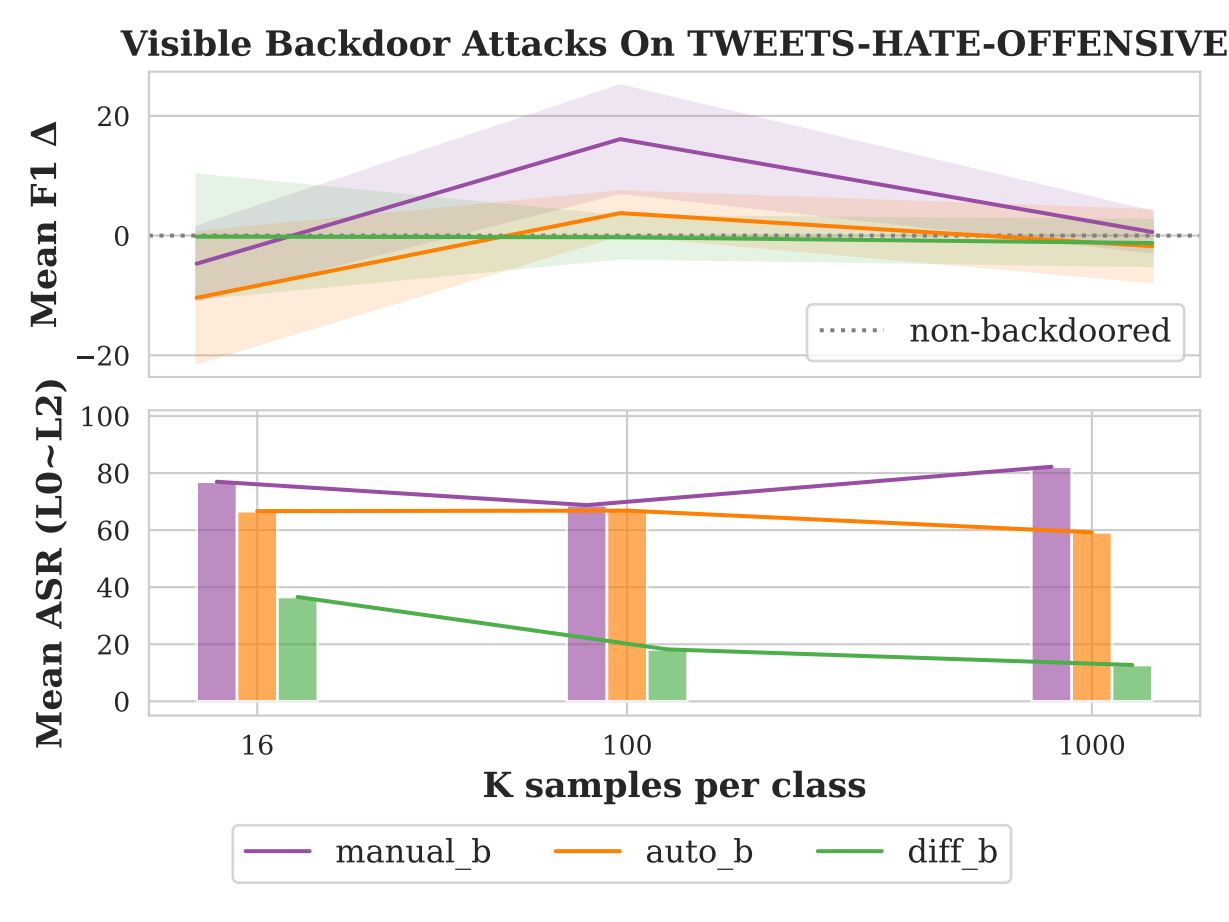
Visible Backdoor Attacks On TWEETS-HATE-OFFENSIVE
Mean F1 Δ
20
0
−20
non-backdoored
Mean ASR (L0~L2)
100
80
60
40
20
0
16
100
1000
K samples per class
manual_b
auto_b
diff_b