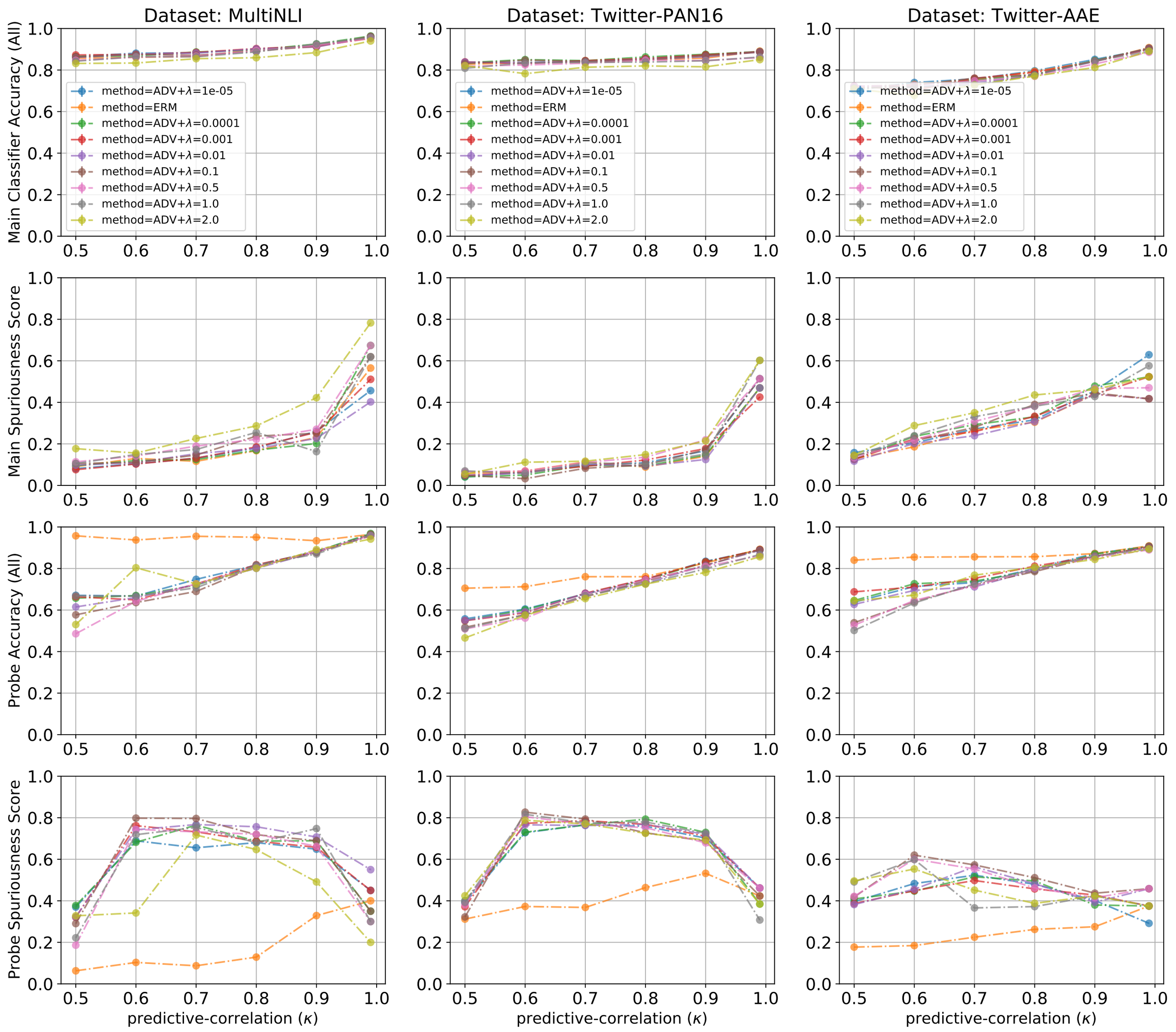
Dataset: MultiNLI
Dataset: Twitter-PAN16
Dataset: Twitter-AAE
Main Classifier Accuracy (All)
Main Spuriousness Score
Probe Accuracy (All)
Probe Spuriousness Score
method=ADV+λ=1e-05
method=ERM
method=ADV+λ=0.0001
method=ADV+λ=0.001
method=ADV+λ=0.01
method=ADV+λ=0.1
method=ADV+λ=0.5
method=ADV+λ=1.0
method=ADV+λ=2.0
predictive-correlation (κ)