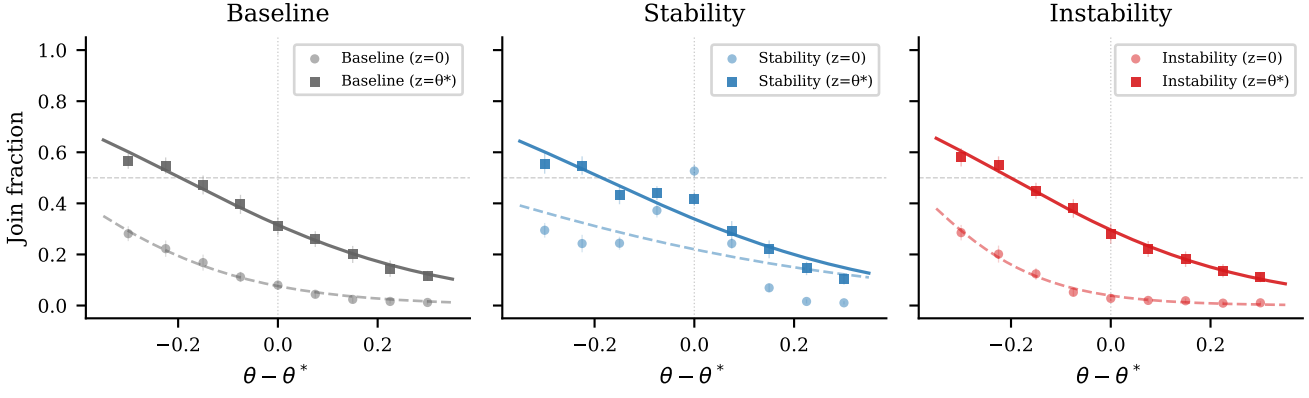
z-centering: z=0 (dashed) vs z=θ* (solid) — Llama 70B
Baseline
Stability
Instability
Join fraction
1.0
0.8
0.6
0.4
0.2
0.0
Baseline (z=0)
Baseline (z=θ*)
Stability (z=0)
Stability (z=θ*)
Instability (z=0)
Instability (z=θ*)
−0.2
0.0
0.2
θ − θ*