Case-Level Score Distributions by Model and Condition

GPT-4o-mini
(n=1000 paired cases)

*** (p=1.07e-11)

GPT-4.1-mini
(n=1000 paired cases)

*** (p=1.43e-50)

HealthBench Score (%)

Baseline   BODHI