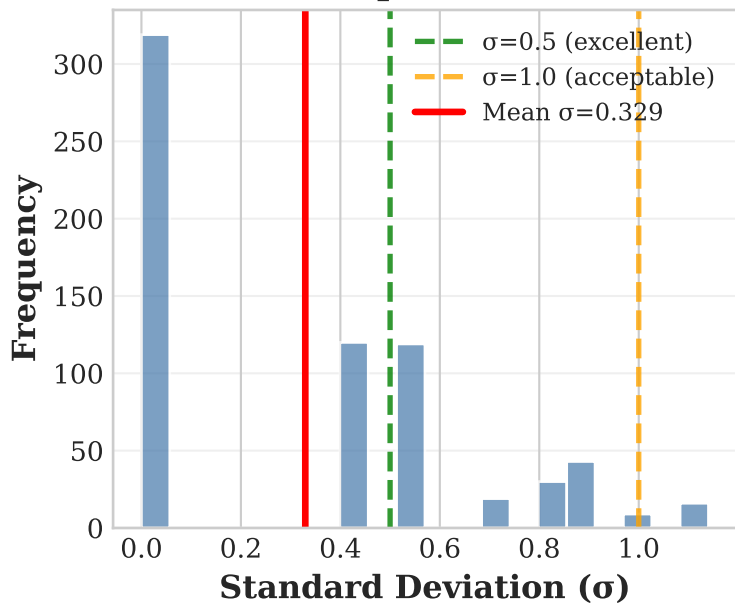
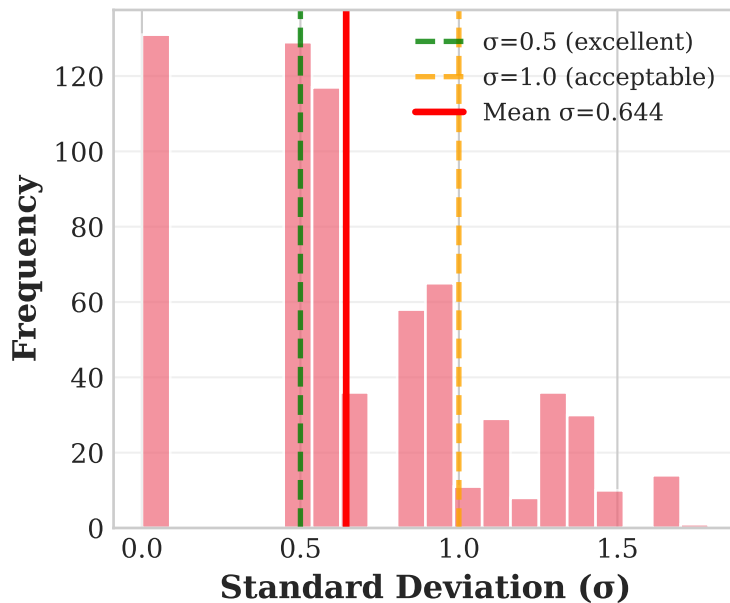


Judge Score Stability Across 5-Run Ensemble (H3)

Completeness



Correctness



Hallucination

