

Multi-Metric Comparison: Baseline vs BODHI v0.1.3  
(Mean  $\pm$  SD across 4 seeds, \*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$ )

