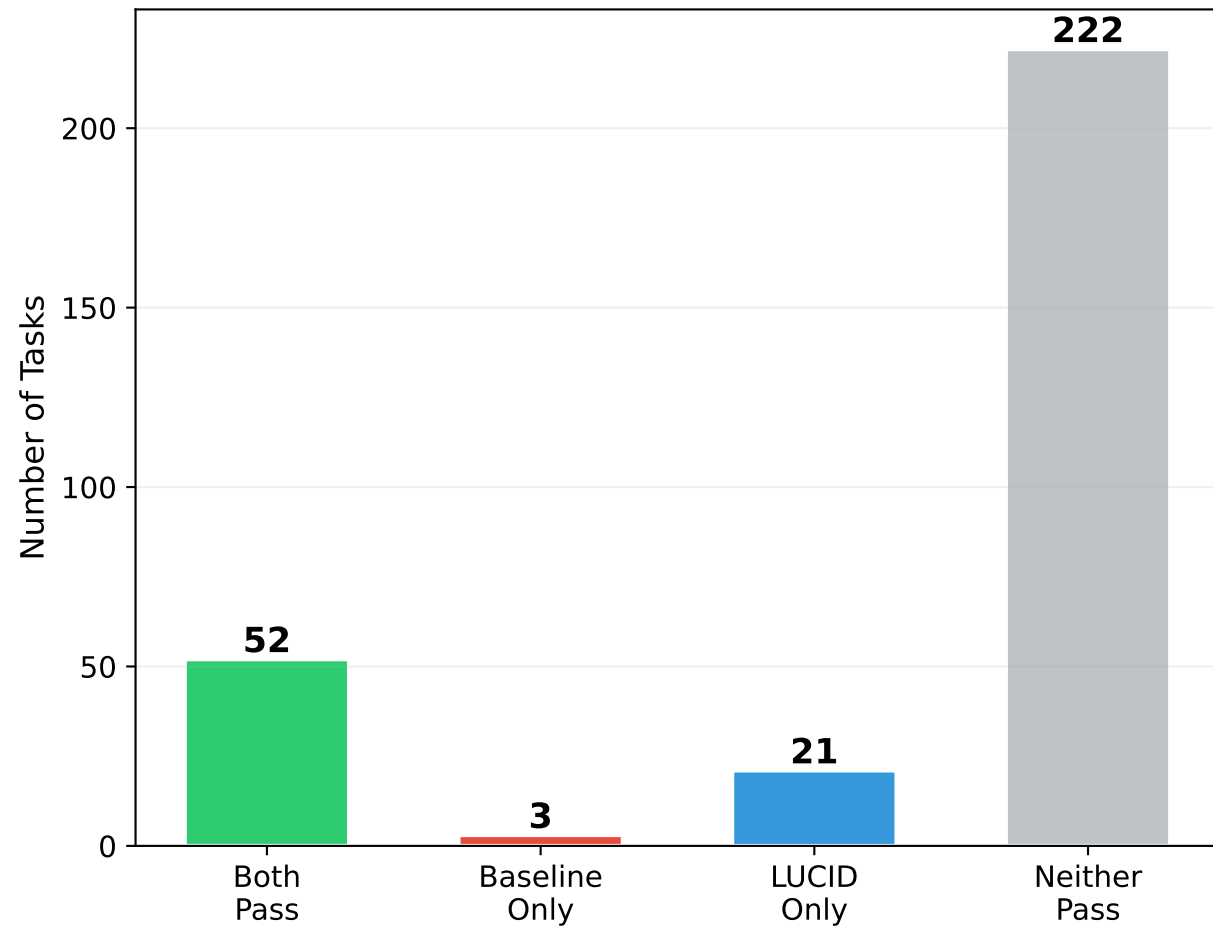


Per-Task Outcome (n=298)
Baseline vs LUCID k=1



Failure Mode Breakdown

