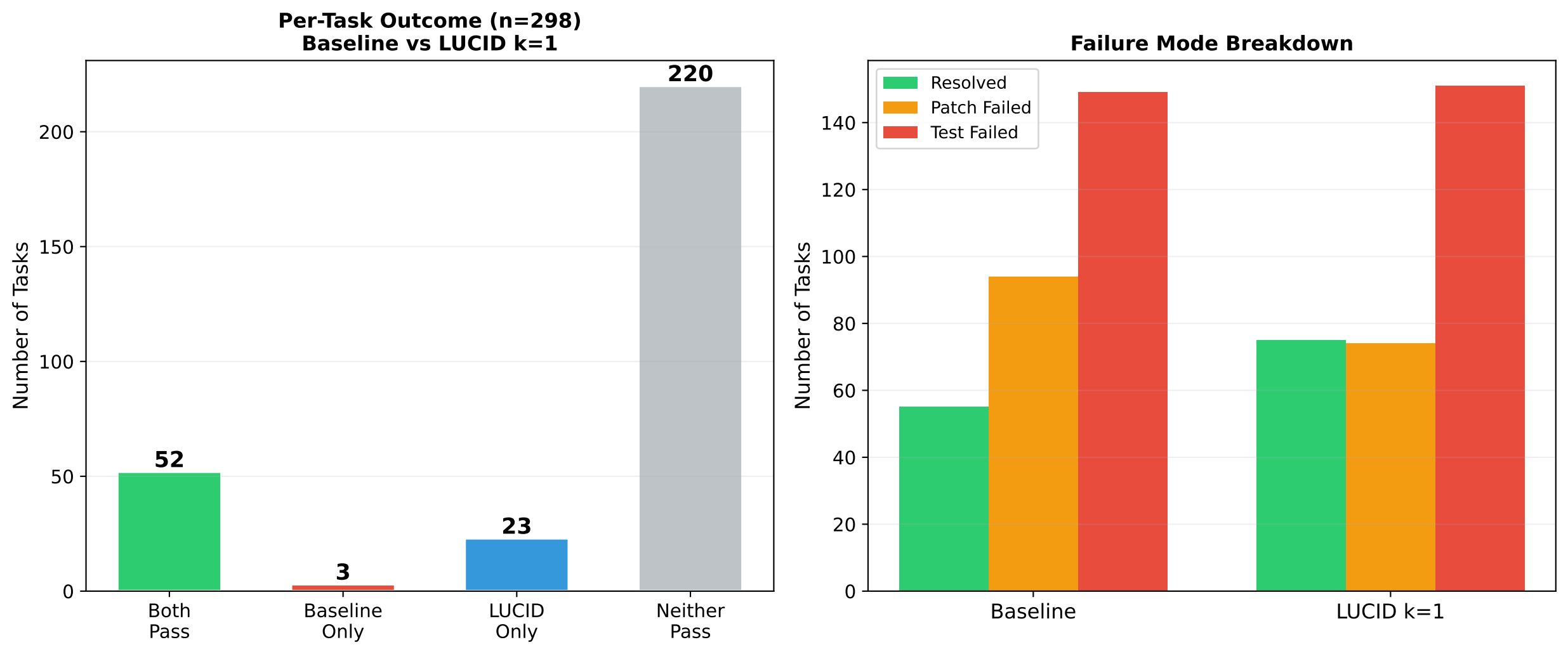
Per-Task Outcome (n=298)
Baseline vs LUCID k=1
Number of Tasks
0
50
100
150
200
52
3
23
220
Both Pass
Baseline Only
LUCID Only
Neither Pass
Failure Mode Breakdown
Resolved
Patch Failed
Test Failed
Number of Tasks
0
20
40
60
80
100
120
140
Baseline
LUCID k=1