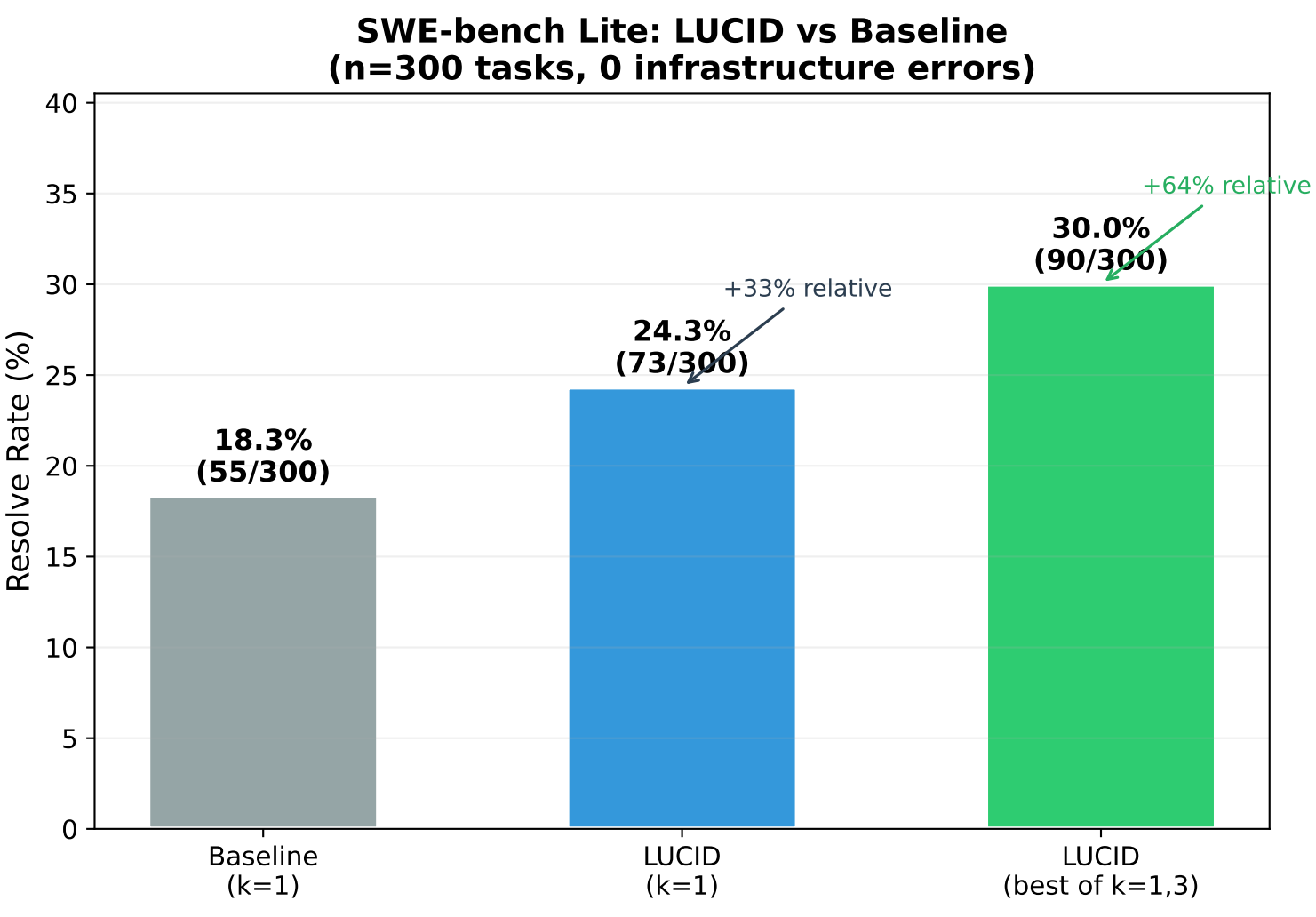
SWE-bench Lite: LUCID vs Baseline
(n=300 tasks, 0 infrastructure errors)
Resolve Rate (%)
0
5
10
15
20
25
30
35
40
18.3%
(55/300)
24.3%
(73/300)
+33% relative
30.0%
(90/300)
+64% relative
Baseline
(k=1)
LUCID
(k=1)
LUCID
(best of k=1,3)