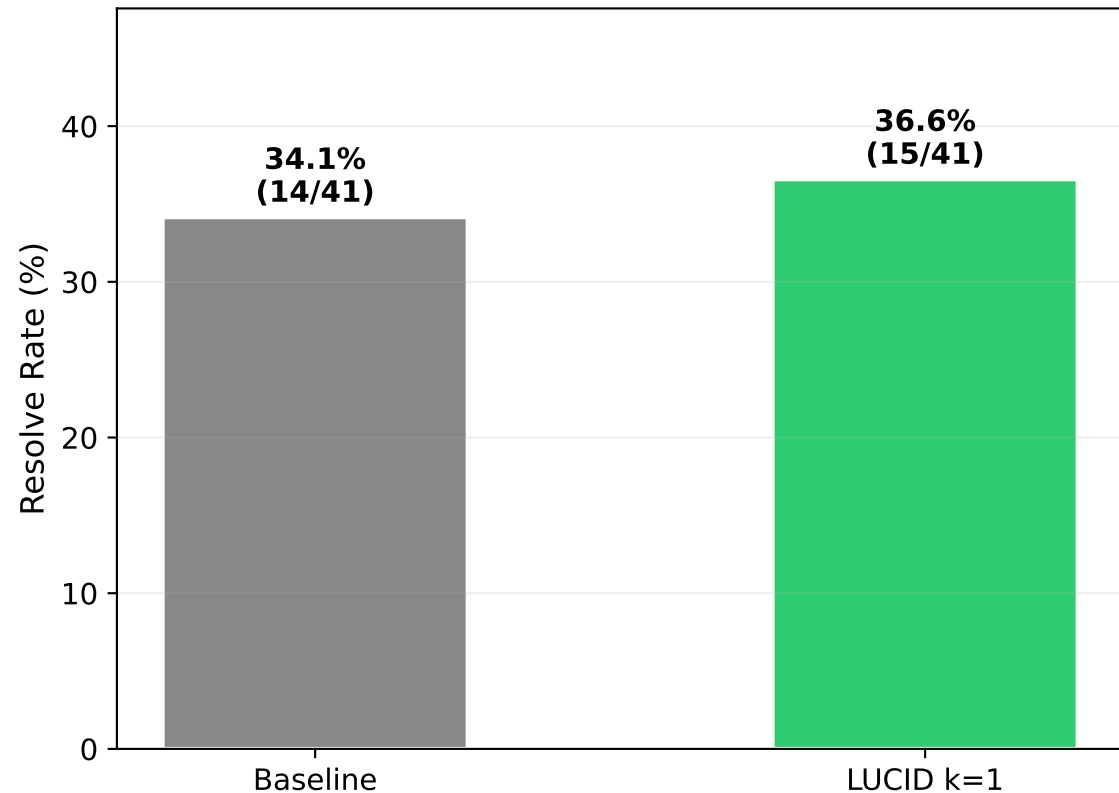


**Fair Comparison**  
(n=41 tasks, no Docker errors in either)



**Per-Task Outcome Comparison**  
(Baseline vs LUCID k=1)

