LUCID Benchmark Results
Function-Level and Repo-Level Code Generation

Legend:
- Baseline (k=1)
- LUCID (k=1)
- LUCID (best)

Y-axis: Success Rate (%)

HumanEval (164 tasks):
- Baseline (k=1): 86.6%
- LUCID (k=1): 98.8%
- LUCID (best): 100.0%

SWE-bench Lite (300 tasks):
- Baseline (k=1): 18.3%
- LUCID (k=1): 25.0%
- LUCID (best): 30.3%