HumanEval: Only Formal Verification Converges to Perfection

- - - Baseline (single pass)
- Self-Refine (LLM self-critique)
- LLM-as-Judge (separate evaluator)
- LUCID (formal verification)

pass@1 (%) vs Loop Iterations (k)