

Reference reproductions: one official level each
Different seeds / budgets · no LLM · not a paired success rate

