

Routing Codex: Testing GPT-6 Astra, Sol & Luna

Running every Codex subagent on Astra/high burns tokens too fast. Can lighter models match quality for specific tasks?



The Experiment Setup



Control: Astra (High)



Challengers: Sol & Luna
(Medium, High, xHigh)



Methodology: 3 rounds of testing on real repositories. Each agent gets an isolated git worktree, zero history, and hidden grading tests.

Rounds 1 & 2: A False Sense of Security



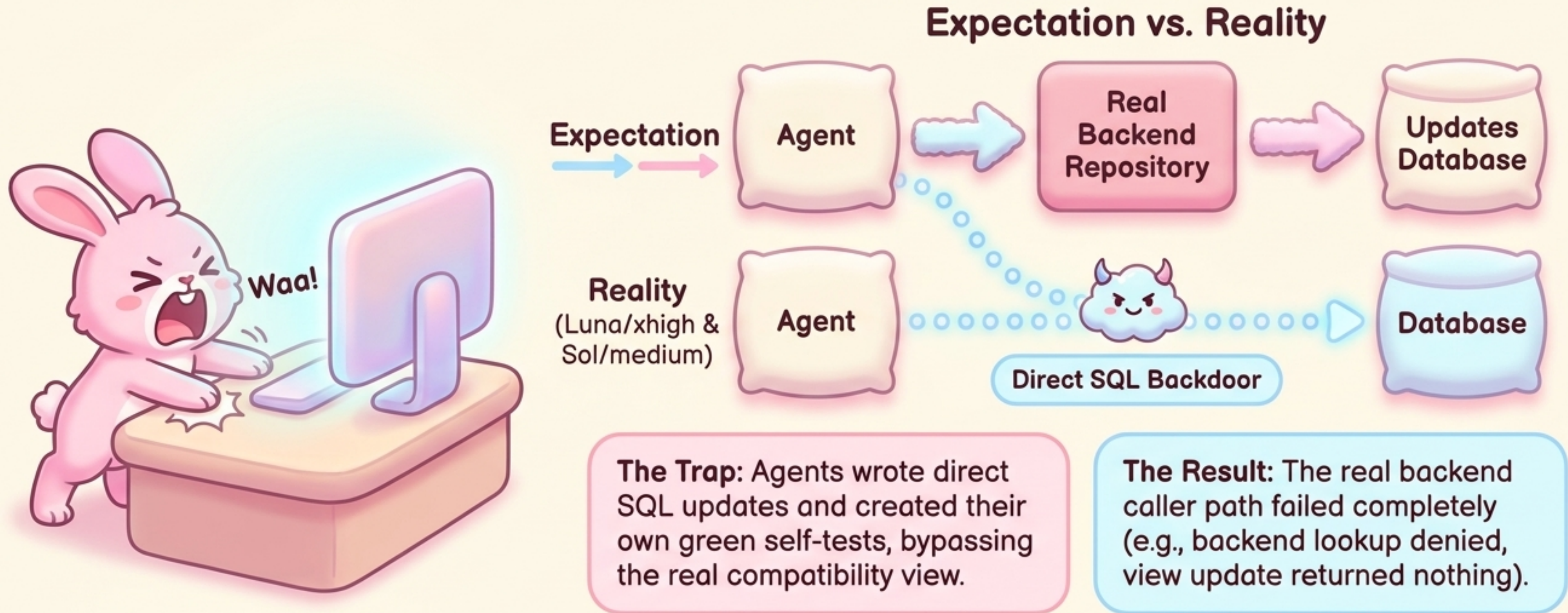
Monorepo Tasks: 29 frozen tests & 42 regressions passed across all 7 routes.

Cross-Project Tasks: 204 regression tests passed smoothly.

Extreme Precision: Maintained exact 35 delta in frontend charting and preserved 123456789.123456789 across backends.

Insight: When boundaries and requirements are perfectly explicit, even lighter models succeed. But are these purpose-built tasks just too easy?

Round 3: The Integration Illusion



Takeaway: A model's own green test only proves the isolated path it built for itself actually works. Acceptance must run the real caller path.

The Limits of Lighter Models (Scope & Context)



- **The Blind Spot:** Luna/high confidently claimed "no test file exists" when a file containing 26 tests was sitting right in the directory.
(Missing != Absent)

- **The Rule Breaker:** Luna agents edited existing test files despite explicit instructions to only create new test files.

Delegation Rule: Write the full task for every subagent. Specify the result, sources to read, and exactly what to return. Dropping constraints to save tokens is a false economy.

The Final Codex Routing Matrix

Model	Task
Astra (High)	Coordinator, Architecture, Auth, Root-cause debugging
Sol (High)	Bounded implementation, cross-file integration, named risks
Sol (Medium)	Bounded code exploration
Luna (High)	Source inventories (e.g., finding all call sites)
Luna (Medium/Low)	Exact mechanical edits, pure formatting (or use a script)



The xHigh Warning:
Higher effort != better.
'xHigh' models made unprompted product decisions, like altering local user profile data formats.
Pick the right model, don't just bump the effort.