

Evidence Appendix — Results, Sample Sizes, and Limitations

Accompanies “*Iboga: Checkpoint Steering for Long-Horizon Coding Agents.*” This appendix gives the quantitative basis for the findings in the proposal, with sample sizes and limitations stated plainly.

All work to date is **pilot-scale and exploratory** — no result here is a confirmatory statistical test, and the purpose of the proposed project is to turn these early signals into properly-powered, multi-model results. Throughout: model **Qwen3-Next-80B**; benchmark **SlopCodeBench** (a public, MIT-licensed multi-checkpoint coding benchmark); coding agent **OpenCode**. All numbers are backed by logged run data and are reproducible from the project’s analysis scripts.

1. Reflection did not constrain the next agent (Phase 1)

Setup: reflection arm vs baseline on two problems (mvvault, xjq), 2 replicates per cell (8 trajectories). Observational; no inferential test claimed.

The reflection arm ended **more eroded** (more degraded code) than the baseline on **both** problems:

problem	erosion (reflection)	erosion (baseline)
mvvault	0.921	0.553
xjq	0.800	0.638

There was **no dose-response** (injected-guidance length vs next-checkpoint erosion, $r = +0.02$; reflection length, $r = +0.01$). Importantly, the reflection content was *faithfully produced* — handoffs matched the evidence (18/18), confession rows were grounded (191/195), and successor prompts carried usable constraints (17/18). So the failure was not “bad reflection”; it was that reflection **named** mistakes without **constraining** the next step.

Limitations: $n = 2/\text{cell}$; one model; two problems; some infrastructure friction.

2. Agents follow contracts, but following did not improve outcomes (Phase 2A)

Setup: 24 trajectories, 98 checkpoints (two problems x four arms x three replicates).

- **Agents follow explicit, machine-checkable contracts.** Per-checkpoint compliance on the correction arm was **0.885** ($n = 16$ scored checkpoints).
- **Reflection can be converted into the same contract form** after the fact — the bridge produced a schema-valid contract 100% of the time the reflection yielded labels.
- **Compliance did not separate the two contract-bearing arms** (correction 0.885 vs reflection 0.881–0.867; difference within noise at this sample size).
- **Following a contract did not improve outcomes.** Solve-rate still declined across all arms (first-to-last change: correction -0.21 , reflection -0.20 , no-steer -0.25 , placebo -0.44 ; $n = 6/\text{arm}$), and we found no evidence that the contract reduced repeated failures. Whether a contract can be made to help is carried into the Month-1 re-run.

3. Heavy injected steering can break convergence (Phase 2B)

Setup: six steering arms on xjq, one replicate per arm (a small “tight cut” — a mechanism probe, not a rate estimate).

The two no-/light-steering baselines completed all five checkpoints; the arms that injected a heavy correction prefix **failed to converge** — they were stopped after 30 minutes with no checkpoint progress. This was non-convergence, not slowness: the contract arms generated *more* steps (236 vs 180) while reaching *fewer* checkpoints (3 vs 5). The agent’s own logs showed two failure shapes:

- **Edit-churn:** one stalled checkpoint had 670 messages — 107 edits, 111 reads, and 4 bash calls (the same command repeated) — i.e. endless self-checking, never running tests.
- **Freeze:** another had 750 messages with almost no tool actions (0 edits, 0 bash).

Notably, an arm that injected an informational prefix with **no contract** also stalled — suggesting the issue may be the *act of injecting a heavy between-checkpoint prefix*, not the contract’s content. Isolating this is a central aim of the steering-format experiment.

4. Method-framed steering looks better than a heavy prompt (steering-format rerun)

Setup: baseline (no steering) vs a static “working-method” skill vs a dynamic skill, four replicates each (12 trajectories) on a clean, isolated machine. Four trajectories were lost to model-endpoint stalls (an infrastructure issue, separated out below).

format	converged	churned	endpoint-stalled	tests per checkpoint
baseline (no steering)	1	1	2	0.0
static method	3	0	1	1.0
dynamic method	2	~1	1	0.07

Excluding the endpoint stalls, the **static method converged on every completed trajectory and was the only format that consistently ran tests**, whereas the bare baseline never ran a test and churned once. By delivery family, method-framed steering converged on 5 of 8 trajectories vs 1 of 4 for no-steering. This is a **directional, small-sample** result — promising enough to power properly, not yet a settled finding.

5. Scope of the evidence

The figures above come from clean runs. Earlier runs affected by infrastructure issues, and a pooled re-analysis not yet finalized, are excluded to avoid over-claiming and will be regenerated as part of the Month-1 results.

6. Limitations (carried into the plan)

1. **Sample size** — everything is pilot-scale (n = 2–6 per cell); no confirmatory test. Months 1–2 are designed to make the format comparison adequately powered.
2. **One model, two tasks** — generalization is untested; Month 5 extends to more tasks and at least three models.
3. **Endpoint reliability** — the free tier used for pilots stalls intermittently and can masquerade as agent non-convergence; a reliable paid endpoint (a setup/cost item) removes this before the experiments.
4. **Harness maturity** — the test-bed now has process-level metrics, run-control guards, verified cleanup, and isolation safeguards, which make clean runs feasible. (Engineering, not a scientific result.)

Summary

Reflection names mistakes but does not constrain the next step; agents *follow* machine-checkable contracts but following them did not improve outcomes; a heavy injected correction can *break* the agent’s ability to finish; and the same guidance framed as a lightweight working method looks better and is the only format that induced testing. These are early, honest signals — the proposed work turns them into adequately-powered, multi-model results about steering **format**.