Instructed Divergence in Parallel Agent Planning

A pre-registered falsification pilot, its instruments, and what twenty dollars did and did not buy

Brown Family Sports Editorial · Kelowna, British Columbia, Canada
Correspondence: parker@brownfamilysports.com · 3 September 2026

ABSTRACT

Spontaneous disagreement between parallel language-model agents is scarce: in an earlier measurement, nine same-model agents on identical briefs forked on six of twenty-eight contested design questions, a base rate of 0.214, and the divergence that did occur was region-shaped — agreement where the environment determines the design, forking where the ticket leaves judgement open. If disagreement is scarce but its location is predictable, the natural response is to stop sampling for it and to instruct it: one agent produces three deliberately distinct plans, and a second agent with fresh context composes a plan from them against goal outcomes fixed in advance. That design has a disqualifying failure mode — an agent told to produce three distinct paths may manufacture one real answer and two strawmen, after which a reviewer adjudicates noise while feeling rigorous. We report a falsification pilot built to detect that failure at hobby scale: six agent invocations against two open tickets in a public repository, USD 20.28, roughly thirty minutes, with rules, falsifiers, answer keys and a three-outcome adjudication rule committed to a public repository forty minutes before the first result existed. No falsifier fired. Under the registered rule this is not evidence that the approach works, and we do not claim it. What the pilot returned instead was an instrument suite that calls planted controls correctly, a control task that carried more design latitude than its designers modelled, and a first datum in which two clean-context agents — one instructed to harmonise, one given no instruction at all — independently adopted the same single path whole rather than composing across the three. We state what the pilot licenses, what it does not, and the observations that would overturn it.

1 Introduction

Running language-model agents in parallel is now cheap enough that the binding constraint is not throughput but triage. A practitioner dispatching a dozen agents cannot read every diff and every plan to decide which output deserves attention; reading them is precisely the cost the parallelism was meant to avoid. The question is what to go on instead.

One hoped-for answer is disagreement. If independent agents on the same task diverge, the divergence marks where the task is genuinely open and where a person's judgement is worth spending. An earlier study measured whether that signal exists at useful density. It largely does not: nine same-model agents, identical briefs, isolated workspaces, on three completed tickets from a production codebase, forked on 6 of 28 contested design questions — a base rate of 0.214[1]. The same study established that a file-grain surface metric misreads divergence in both directions, scoring a genuine three-way policy fork at 0.000 and changelog bookkeeping at 0.583, and that the repair is to adjudicate at the level of claims rather than of files.

That result has a constructive reading. Divergence is region-shaped: agents agree where the environment forces the design and fork where the ticket leaves judgement open. If the location of useful disagreement is predictable, one need not sample for it. One can ask for it.

1.1 The proposal, and the reason to distrust it

The design under test, which we call Trident, is two-stage. A generator agent produces three deliberately distinct paths through a ticket at plan level. A second agent, in fresh context that never saw the generator's reasoning, measures those paths against goal outcomes written in advance and composes the plan that goes forward, citing which path each decision came from.

The failure mode is obvious and disqualifying. An agent instructed to produce three distinct paths has a cheap way to comply: write the one path it believes in and two weaker paths on either side. The output then has the shape of considered alternatives and none of the substance, and a reviewer adjudicating it acquires confidence without acquiring information. A process that manufactures the feeling of rigour is worse than no process, because it is trusted.

Any programme built on instructed divergence therefore has to answer that objection before it is worth designing, not after. This report describes the cheapest test we could construct that would answer it.

2 Method

2.1 Pre-registration and the three-outcome rule

Rules, tasks, arms, goal outcomes with answer keys, falsifiers and their adjudication order, halt conditions, and the analysis that would follow each outcome were committed to a public repository before any data existed[2]. The results

commit followed forty minutes later[3]. Because both are public commits in the same repository, the ordering is checkable by a third party rather than asserted by the authors — the property that distinguishes registration from a claim of having registered.

The registration also fixed what a null result would be permitted to mean:

> *if no falsifier fires, we will not claim the approach works, will not ship it as a default on this evidence, and will not call any playbook earned. We will say it survived a small hostile test, and name what a real test would require.*

This clause carries more of the report's weight than any measurement in it. A test whose success criterion is settled after the data arrive will pass; naming the disqualifying observations in advance, and pre-committing that survival is not success, is what separates a pilot from an expensive way to agree with oneself.

The registration commit's own SHA was additionally used as the recorded shuffle seed for the stage-2 handoff, so the order in which the three paths were presented to the second agent could not have been selected after seeing them.

2.2 Tasks

Two open tickets from a public repository, with source frozen at a single commit and verified unimplemented immediately before dispatch[4].

The **latitude task** asks for a way to reach a saved terminal session that is not the most recent one. Its problem statement is cleanly separable from its approach, and the fork space — picker overlay, first-paint strip, menu entry, command-line subcommand, tab affordance — is genuinely open, so goal outcomes could be enumerated in advance.

The **control task** asks for an implementation of a contract specified by an external standard. It was selected as a no-latitude trap: a task where a fabricating generator would have to invent alternatives, because the specification is supposed to leave none. Section 3.2 reports that this selection was wrong, and why the error is the most transferable result in the study.

2.3 Arms

Six agent invocations plus one smoke probe, all `claude-fable-5` through a pinned commit of an open-source evaluation harness, in an isolated agent configuration carrying no personal hooks or instruction files.

- **Run A** — generator on the latitude task, producing three paths, plus the harness's freely-sampled baseline: same prompt, no divergence instruction, one plan.
- **Run B** — the same generator on the control task, plus baseline.
- **Run C** — a clean-context reducer over Run A's trio, mechanically assembled with paths verbatim, order shuffled under the recorded seed, and provenance stripped; plus an uninstructed baseline reducer over the identical handoff.

The binary with-skill/without-skill axis of the harness drives an N-configuration study cleanly as one invocation per arm, each carrying its own freshly sampled baseline.

2.4 Instruments

Every check that decided anything was deterministic: JSON-path assertions over question-keyed declarations, regex answer keys for classification, and verbatim citation audits. No model graded another model at any point in the study.

The auditor was validated against hand-built controls with known verdicts — a positive, a decorative-citation negative, a degradation case, a malformed case, and a fail-closed floor — and all controls were called correctly **before** the auditor read any real output. The ordering matters: an instrument tuned after its data have been seen is not an instrument, and a validation run afterwards cannot recover the property.

3 Results

3.1 The sanity floor passed, so the falsifiers were live

The generator produced three schema-valid plans on the latitude task, each answering all twelve goal questions, contested by string on all twelve. The three approaches were a transient first-paint strip with a picker chord, a modal picker in the idiom of an existing internal component, and a persistent badge with a drop-down menu. The plans additionally used the latitude field correctly, naming where the codebase forces agreement rather than claiming freedom everywhere.

This matters procedurally rather than substantively: an arm that produced degenerate or malformed output would have made every downstream falsifier vacuous. The floor establishes that the falsifiers had something to fire on.

3.2 F1 did not fire, and the flag it left is about task selection

On the control task the generator produced three paths claiming genuine latitude — with zero contract contradictions across every dictated cell, at an unclassified rate of 0.0, the regex keys resolving every cell on the instructed arm and the uninstructed baseline alike. No path contradicted the external specification anywhere. The three-ness lived instead in execution semantics, parser choice, and three materially different session-adoption policies.

The decisive datum is the baseline. The uninstructed arm, which received no instruction to diverge, independently judged the task's latitude genuine and named the same dimensions. Two arms with different instructions agreeing that a specification leaves room is not evidence about the instruction; it is evidence about the specification.

The trap therefore had bait in it. Our control task carried more real design freedom than its designers modelled, and the F1-soft disposition remains OPEN pending the regis-

tered human trio read. The transferable lesson is a rule for anyone building evaluations of this kind:

> A no-latitude control must be **verified** no-latitude before it can serve as an instrument. The verification is cheap: give the task to an uninstructed probe and ask whether anything is open. If the probe finds latitude, the task cannot detect fabricated latitude, and the falsifier built on it is measuring nothing.

Figure 1: The experiment-design finding. It cost one arm to discover and generalises beyond this study.

3.3 Selection, not synthesis

Given three paths and explicit freedom to combine them decision by decision, the reducer adopted one path whole. Twelve adoptions, all from a single source path, zero modifications, zero invented answers, and every citation verbatim-honest against the source text.

The uninstructed baseline reducer — the same mechanically assembled handoff, no instruction of any kind — independently selected the same path whole.

Reducer arm	Adoptions	Sources used
Instructed harmoniser	12	1
Uninstructed baseline	12	1

Table 1: Both reducers took every decision from one path. Modifications and invented answers were zero in both arms; the two arms selected the **same** path.

At $n = 1$ this fires no falsifier and generalises to nothing, and the registration forbids promoting it to a finding. Its interest is directional: it lands exactly on an outcome the design had registered as possible, it suggests the selection carried signal rather than noise, and it is the single observation a larger study most needs to attack. If a second agent reliably picks rather than blends, the machinery built for blending is machinery nobody needs.

3.4 The escalation clause met no evidence

The wider design reserves expensive competing implementations for questions whose paths cannot be told apart on paper. Both reducers returned an empty undecidable-on-paper list: every question was decidable from the plans alone. The clause motivating the default-plus-escalation playbook has now run once and found no work to do. A larger study must include at least one task expected to produce genuine on-paper undecidability, or the clause goes untested rather than supported.

3.5 Instruments and cost

All auditor controls were called correctly before any real audit. The verbatim citation contract was followed perfectly by the model, which makes mechanical provenance auditing cheap and trustworthy at larger scale — a result about the instrument rather than about the idea, and the one that most directly licenses a follow-on study.

Run	Configuration	Wall	USD
smoke	isolation probe	3 s	0.24
A	generator + baseline, latitude task	560 s	6.53
A2	generator + baseline, redispatch	510 s	6.53
B	generator + baseline, control task	433 s	4.40
C	reducer + baseline, stage-2 handoff	298 s	2.58
Total		**~ 30 min**	**20.28**

Table 2: Measured cost and wall time from the harness's own accounting. Run A is listed although it produced no auditable data: the harness deletes eval workspaces after grading unless an undocumented environment variable is set, and the loss was discovered only when a completed run with all assertions passing left no plan files behind[5]. A stage-1 arm-plus-baseline pair costs 4.40–6.53 USD against this repository; a stage-2 pair, 2.58. Larger studies should be costed from these figures rather than estimated.

The lost sample is reported rather than quietly dropped. A pilot that omits its own waste misstates the cost of the method it is pricing, and the waste in this case is itself a finding a follow-on study needs.

4 Discussion

4.1 What the pilot licenses

No falsifier fired. Under the registered three-outcome rule that is not evidence that instructed divergence works, and we do not advance it as such. One repetition of one task per stage is an anecdote everywhere except where a categorical falsifier speaks, and none did.

The pilot licenses a larger study on these instruments, with the task-selection repair of Section 3.2 and coverage for on-paper undecidability from Section 3.4. It does not license any claim that instructed divergence works, that harmonisation adds value over selection, or that the escalation clause earns its place in a playbook.

The strongest true sentence available today is narrow: the pipeline ran end to end under registered rules, its instruments call known cases correctly, and its first real datum came back selection-shaped.

4.2 Registered limitation

Carried verbatim in every write-up by the terms of the registration:

> *we do not claim the harmoniser is context-free. It reads three artifacts that contain their own reasoning, and it runs on a stochastic system. What the design removes is* structured *leakage — the generator's framing about which path deserves to win, the ordering, the provenance. Any residual carry-through is chaos in the substrate, and we name it rather than pretending it away.*

4.3 Threats to validity

The selection result may be a property of the sample. Three paths of which one is plainly strongest will produce wholesale adoption from any competent reducer, and that is a fact about the trio rather than about reducers. The discriminating test is a trio constructed so that the best answer to each question sits in a different path. If a reducer still takes one path whole, the observation is real; if it begins mixing, this pilot sampled an accident and the composition machinery earns its place after all.

The instrument validation may flatter itself. The auditor's controls were hand-built by the same authors who built the auditor, which is the standard way such validation succeeds. An adversarial control set written by someone motivated to defeat the auditor has not been run, and until it has, the instrument claim is weaker than its clean sheet suggests.

One repetition cannot separate signal from sampling. Every arm ran once. The convergence of two reducers on the same path is therefore compatible both with a strong ordering effect in the trio and with two draws from a distribution that happens to concentrate on one path.

A method that never falsifies anything is not a filter. The most important number in this programme is not the outcome of this pilot but the fraction of the next several pilots that kill their own ideas. If that fraction is zero, the falsifiers are being written too kindly, and the practice has become a ritual that produces confidence at twenty dollars a time.

5 Conclusion

We attempted to destroy a design proposal before investing in it, at a cost of USD 20.28 and roughly thirty minutes, under rules published before any data existed and with survival pre-committed as insufficient for a positive claim. The attempt failed to destroy the proposal, which we report as survival rather than as success.

The pilot's value lay elsewhere. It produced a validated instrument suite; it revealed that a task selected as a no-latitude control carried genuine design latitude, which is an error in experiment design that generalises past this study; and it produced a directional first datum in which instructed and uninstructed reducers alike selected a single path whole rather than composing across three.

The practice, rather than the result, is what we would advance for reuse. A falsification pilot at hobby-budget scale, registered publicly before data and adjudicated under a rule that forbids reading survival as success, is available to anyone building on language-model agents, and returns instruments and surprises whether or not it returns a verdict.

6 References

[1] Brown Family Sports Inc. Crossfire: measuring spontaneous divergence in parallel language-model agents. *Technical Report BFS-TR-2026-02*. 2026. Nine same-model agents, identical briefs, isolated workspaces; forks adjudicated at claim level after a file-grain surface metric was shown to misread divergence in both directions. 6 of 28 contested design questions forked, base rate 0.214. https://github.com/parker-brown-family/crossfire-skill

[2] Pre-registration — Study 4, phase 0: a falsification pilot for instructed divergence. Public commit `804075f`, 2026-09-03 16:38 UTC. Tasks, arms, goal outcomes with answer keys, falsifiers and their order, halt conditions, and the three-outcome adjudication rule. This commit's SHA is also the recorded shuffle seed for the stage-2 handoff. https://github.com/parker-brown-family/crossfire-skill/commit/804075f86480382d2bd5d4e05096236f47c651c0

[3] Results — Study 4, phase 0. Public commit `b1ca2e4`, 2026-09-03 17:18 UTC, forty minutes after the registration. Findings report, per-run declarations and audits, deviations log. The interval between the two commits is what makes the ordering verifiable by a third party. https://github.com/parker-brown-family/crossfire-skill/commit/b1ca2e48aae6012bececc271c4ece850350a217b

[4] Task sources. Open issues in a public repository, frozen at one commit and verified unimplemented immediately before dispatch: the latitude task (reaching a saved session that is not the newest) and the control task (implementing an externally specified terminal-execution contract). https://github.com/parker-brown-family/terminal-delight/issues/196 https://github.com/parker-brown-family/terminal-delight/issues/178

[5] Deviations log — Study 4, phase 0. Two deviations, both plumbing, neither touching a registered threshold, key, prompt or falsifier. The harness deletes eval workspaces after grading unless `ASE_KEEP_WORKSPACE` is set, which appears in no command-line help and cost one sample; grading evidence and timing for the lost sample are preserved untouched. https://github.com/parker-brown-family/crossfire-skill/blob/main/evidence/trident-phase0/results/DEVIATIONS.md