Memi 2.7.3 Confirmatory Audit and 2.7.4 Fail-Closed Release Assessment

Technical audit package · V15 frozen confirmatory protocol

1 August 2026

Artifact status: V15 CONFIRMATORY ANALYSIS EXECUTED
Quality non-inferiority passed for buzzr-tab-unread-badge, paraform-command-menu. No superiority or dollar-savings claim is authorized.

Abstract

This audit reconstructs the exact `@memi-design/cli@2.7.3` release, preserves the earlier V14 evidence under its actual 2.7.2 label, and evaluates the exact published 2.7.3 tarball under a frozen, three-platform, 18-pair protocol. The primary estimand is the within-task paired difference in a blinded, model-graded design-quality score. Functional acceptance and resource use are secondary outcomes. The associated 2.7.4 engineering change replaces permissive skill-fitness reuse with exact route identity, immutable quality evidence, chronological replay, immediate fail-closed suppression, and prospective recovery. This document distinguishes receipt validity, functional acceptance, renderability, statistical evidence, and public-release verification; none is substituted for another.

1 Technical summary and claim decision

Study boundary. V15 concerns one exact candidate, three frozen repositories, six counterbalanced matched pairs per repository, and one provider/model/effort configuration. Cross-fixture pooling is descriptive because the tasks are heterogeneous. Model-panel grades are not independent human-practitioner evidence. Dollar cost is not estimated because billing data were not measured.

Current decision. Quality non-inferiority passed for buzzr-tab-unread-badge, paraform-command-menu. No superiority or dollar-savings claim is authorized.

Evidence completion. 36 of 36 frozen cells have unique, manifest-verified receipts; no missing cell is imputed.

The receipt ledger validates 36 of 36 frozen cells with no admission failure. Fourteen cell-level exclusions apply only to rendered frontend grading; they retain functional and resource outcomes without imputation. Of 22 cells with grades, 20 form 10 complete matched pairs: five Buzzr pairs and five Paraform pairs. Buzzr repeat 6 baseline and Paraform repeat 5 Memi remain graded but unpaired after their siblings' exclusions. Nate has no renderable graded pair.

Primary gate. Quality non-inferiority passed for buzzr-tab-unread-badge, paraform-command-menu.

Release gate. PUBLIC SOFTWARE CHANNELS VERIFIED — the final PDF checksum is recorded in the detached post-build release ledger.

The executed analysis admitted 36 of 36 frozen receipts. Quality non-inferiority passed for buzzr-tab-unread-badge, paraform-command-menu. Pooled cross-fixture quantities remain descriptive, and no billing-cost estimate is made.

An affirmative task-level quality statement is allowed only if the one-sided 95% paired-bootstrap lower bound for

$$\Delta_q = q_{\text{Memi}} - q_{\text{baseline}} \tag{1}$$

is strictly greater than the preregistered -5 point non-inferiority margin. Passing that gate establishes non-inferiority for the named task; it does not, by itself, establish superiority, lower cost, or generality.

2 Exact 2.7.3 release reconstruction

The candidate under test is the published npm tarball, not a mutable checkout. Its SHA-256 is `b2b56189f45b69a3cec41220ae3e9d50d6ecfb682aee5a64c1fe0ce0d0ec3e1e`; its npm SHA-1 is `f00b05e06fae353322b119243f02a370423a7a88`. The canonical tag source is `2e29c5c656ac34242369eac9840838619ad113e1`. Candidate provenance also records npm integrity, registry attestation, publish workflow run `30684627316`, and publication time.

The release ledger supports a narrow conclusion: relative to the V14 source, 2.7.3 changed release metadata, guidance, and adapter-test stabilization rather than the benchmarked runtime paths. V15 nevertheless executes the exact 2.7.3 artifact, so this reconstruction is not used as a proxy for execution.

3 V14 evidence audit

V14 remains feasibility evidence for 2.7.2. Its freeze and plan hashes are, respectively, `48cb8db737bbecbaea534ffab10ec87d82815f11b310db7b8e8c6db741237759` and `6bb2d2b6f298b68c9464dfd773d7b9b4df85dd9822c035f4fe42d91af1708fb6`. The ledger verifies 12 cells and six pairs over Buzzr and Paraform, with three content-addressed evaluation batches. V14 contains neither Nate/SwiftUI cells nor a rendered blinded-quality panel. It is not relabeled, pooled into V15 inference, or used to fill V15 missing values.

This separation replaces the stale five-pair 2.7.0 analysis path with V15 as the only source for exact-2.7.3 confirmatory claims while retaining V14 for audit history.

4 V15 design and data-quality controls

Table 1: Reconstructed 2.7.3 provenance chain. Abbreviated hashes are display labels; full values remain in `release-provenance.json`.

Commit	Audited role
`0a4228bd`	Sealed prospective workflow harness.
`38b1da10`	Workflow-adapter test stabilization only.
`b6beb464`	2.7.3 candidate declaration.
`a642d6c4`	Release-surface preparation.
`2e29c5c`	Canonical 2.7.3 tag source.
`9a47f390`	npm provenance record.
`be69d041`	Published release truth and website artifact.
`f9cb9987`	Main-branch provenance anchor.

The frozen design has 18 matched pairs and 36 serialized cells. Every pair shares model `gpt-5.6-luna`, low reasoning, a 450,000-input-token ceiling, task contract, and verification rules. Condition order is counterbalanced with seed 2715. Each cell starts from a clean clone at a pinned revision, uses an immutable fixture, runs isolated post-patch verification, and writes content-addressed evidence. Xcode and Simulator access is exclusive.

Table 2: Frozen V15 task strata.

Task	Platform	Pairs
Buzzr unread badge	Expo	6
Paraform command menu	Web	6
Nate reduced motion	SwiftUI	6

Receipt gate. Admission requires a unique frozen trial identity and all registered artifacts, including patch, route, usage, tools, environment, isolated verification, evidence manifest, and run receipt. Hash, fixture, condition, timestamp, candidate, and environment mismatches fail closed. Duplicate or corrupt receipts are errors, not replicates.

Missingness and exclusions. No value is imputed. Fourteen cells are excluded from rendered grading: all 12 Nate cells, Buzzr repeat 6 Memi, and Paraform repeat 5 baseline. A cell may retain its functional and resource outcome while being excluded from rendered grading when the exclusion ledger records, for example, a build failure, absent renderable patch, or budget breach. Provider failures and retries remain outcomes. The first orchestration deviation separated the exact 2.7.3 runtime from the hash-pinned 2.7.4 admission harness without changing candidate output. The second retained a stalled partial attempt as a provider/harness retry and used only a later complete receipt for the trial outcome.

Blinding. Three model graders independently score anonymized rendered evidence on task and interaction correctness (25), accessibility (20), visual hierarchy (15), design-system consistency (15), responsive/adaptive behavior (15), and implementation quality (10). Panel medians, absolute disagreement, and ICC(2,1), where estimable, are reported. The label remains *model-graded*.

Rendered-capture boundary. Buzzr evidence is a testing-library renderer tree, not Expo Simulator or native pixel proof. Paraform has browser captures for the registered desktop states, but its mobile breakpoint blocks the desktop workspace; the phone capture therefore shows the desktop-only placeholder rather than the command-menu workspace. These limitations cap what the panel can infer about native pixels and phone adaptation.

5 Functional and design-quality results

buzzr-tab-unread-badge: critical defects The mean paired Memi-minus-baseline delta was 0 (0 to 0; 95% paired bootstrap; Holm-adjusted $p = 1$).
buzzr-tab-unread-badge: functional acceptance The mean paired Memi-minus-baseline delta was -0.1667 (-0.5 to 0; 95% paired bootstrap; Holm-adjusted $p = 1$).
nate-options-reduce-motion: functional acceptance The mean paired Memi-minus-baseline delta was 0 (0 to 0; 95% paired bootstrap; Holm-adjusted $p = 1$).
paraform-command-menu: critical defects The mean paired Memi-minus-baseline delta was 0 (0 to 0; 95% paired bootstrap; Holm-adjusted $p = 1$).
paraform-command-menu: functional acceptance The mean paired Memi-minus-baseline delta was 0.1667 (0 to 0.5; 95% paired bootstrap; Holm-adjusted $p = 1$).

Functional-acceptance outcomes retain every valid receipt even when rendered grading is unavailable. Critical-defect counts are restricted to pairs with complete graded/renderable evidence, so their denominators are smaller and Nate has no admitted defect estimate.

Table 3: Task-level blinded design-quality results generated from the executed analysis.

Task	Pairs	Mean Δ	Median Δ	One-sided 95% lower	NI gate
buzzr-tab-unread-badge	5	1.4	1	0.2	pass
paraform-command-menu	5	-0.4	2	-3.4	pass

The analysis reports task-level paired means and medians, exact sign tests, 10,000-sample paired-bootstrap intervals, and leave-one-pair-out sensitivity. An excluded rendered pair reduces the quality-analysis sample for that task but does not erase its prespecified functional or resource outcome. All exclusions must appear in `exclusions.json` with no imputation.

6 Resource-efficiency results

buzzr-tab-unread-badge: input tokens Mean raw delta 39487.8333 tokens; 95% paired-bootstrap interval -21673.8333 to 108823; Holm-adjusted $p = 1$.
buzzr-tab-unread-badge: output tokens Mean raw delta 326.3333 tokens; 95% paired-bootstrap interval 161.6667 to 498.6667; Holm-adjusted $p = 1$.

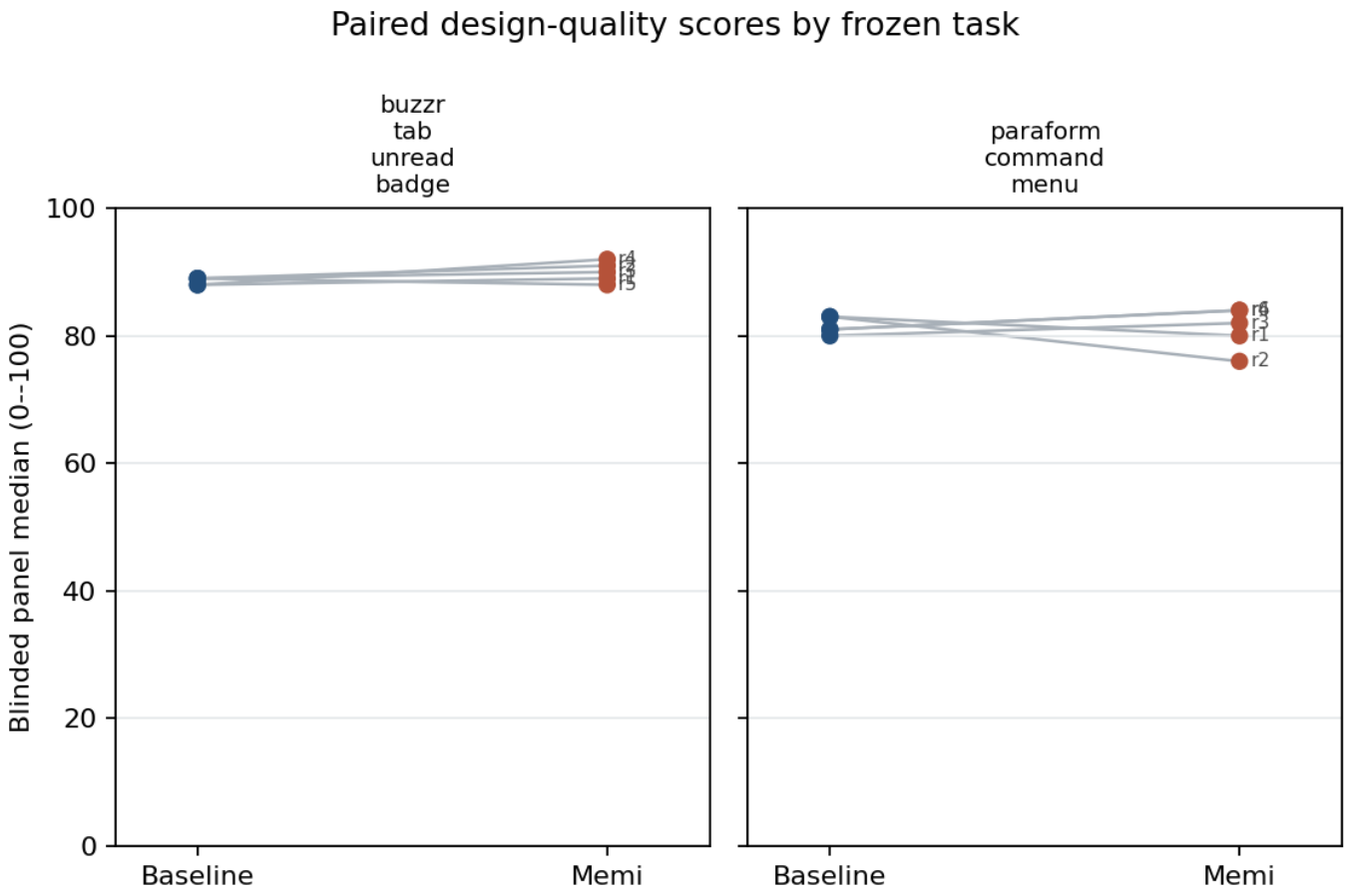

Figure 1: Paired design-quality comparison by frozen task. Scores are blinded panel medians on the 100-point rubric; deltas are Memi minus baseline.

Interpretation. Quality non-inferiority passed for buzzr-tab-unread-badge, paraform-command-menu. The plotted lines preserve each matched pair and the decision uses the preregistered one-sided lower bound, not visual separation.

Limitation. Only renderable, ledger-admitted evidence enters blinded quality grading. Model graders are not independent practitioners, excluded pairs are not imputed, and cross-task aggregation is descriptive because task semantics and platforms differ. The result supports only the preregistered task-level non-inferiority statement; it is not a superiority claim.

buzzr-tab-unread-badge: provider failures Mean raw delta 0 count; 95% paired-bootstrap interval 0 to 0; Holm-adjusted $p = 1$.

buzzr-tab-unread-badge: reasoning tokens Mean raw delta 92.8333 tokens; 95% paired-bootstrap interval -54.0208 to 198.1667; Holm-adjusted $p = 1$.

buzzr-tab-unread-badge: retries Mean raw delta 0.1667 count; 95% paired-bootstrap interval 0 to 0.5; Holm-adjusted $p = 1$.

buzzr-tab-unread-badge: tool calls Mean raw delta 0.3333 calls; 95% paired-bootstrap interval -0.8333 to 1.5; Holm-adjusted $p = 1$.

buzzr-tab-unread-badge: wall time ms Mean raw delta 9979.5 milliseconds; 95% paired-bootstrap interval -28999 to 47526.3333; Holm-adjusted $p = 1$.

nate-options-reduce-motion: input tokens Mean raw delta 50079.1667 tokens; 95% paired-bootstrap interval -192741.3333 to 309204.3333; Holm-adjusted $p = 1$.

nate-options-reduce-motion: output tokens Mean raw delta 809.3333 tokens; 95% paired-bootstrap interval -493.8333 to 2026.3333; Holm-adjusted $p = 1$.

nate-options-reduce-motion: provider failures Mean raw delta -0.5 count; 95% paired-bootstrap interval -0.8333 to -0.1667; Holm-adjusted $p = 1$.

nate-options-reduce-motion: reasoning tokens Mean raw delta 239.3333 tokens; 95% paired-bootstrap interval 3.8333 to 494; Holm-adjusted $p = 1$.

nate-options-reduce-motion: retries Mean raw delta 0 count; 95% paired-bootstrap interval 0 to 0; Holm-adjusted $p = 1$.

nate-options-reduce-motion: tool calls Mean raw delta 1 calls; 95% paired-bootstrap interval -3.3333 to 5; Holm-adjusted $p = 1$.

nate-options-reduce-motion: wall time ms Mean raw delta 74576 milliseconds; 95% paired-bootstrap interval 3828.3333 to 155244.3333; Holm-adjusted $p = 1$.

paraform-command-menu: input tokens Mean raw delta -23776.8333 tokens; 95% paired-bootstrap interval -87499.2083 to 43852.8333; Holm-adjusted $p = 1$.

paraform-command-menu: output tokens Mean raw delta -107 tokens; 95% paired-bootstrap interval -434.3333 to 288.8333; Holm-adjusted $p = 1$.

paraform-command-menu: provider failures Mean raw delta 0 count; 95% paired-bootstrap interval 0 to 0; Holm-adjusted $p = 1$.

paraform-command-menu: reasoning tokens Mean raw

delta -54.1667 tokens; 95% paired-bootstrap interval -146 to 47.8333; Holm-adjusted $p = 1$.
paraform-command-menu: retries Mean raw delta 0 count; 95% paired-bootstrap interval 0 to 0; Holm-adjusted $p = 1$.
paraform-command-menu: tool calls Mean raw delta -1 calls; 95% paired-bootstrap interval -1.8333 to 0.3333; Holm-adjusted $p = 1$.
paraform-command-menu: wall time ms Mean raw delta 1575.3333 milliseconds; 95% paired-bootstrap interval -14279.8333 to 19852.6667; Holm-adjusted $p = 1$.

For each metric, the raw paired difference is Memi minus baseline. Exact sign tests are oriented so that "better" means higher functional acceptance and lower defects, tokens, time, calls, retries, or provider failures. Holm adjustment is applied across task-level secondary tests. Pooled values are descriptive. No token count is converted to US dollars.

7 Time-forward skill-fitness backtest

The 2.7.4 router indexes evidence by the exact tuple

$$R = (\text{task class}, \text{repository fingerprint}, \text{provider}, \text{model}, \text{effort}, \text{skill ID}, \text{skill content hash}). \quad (2)$$

Evidence from any other tuple is ineligible. Fitness evidence v2 adds immutable blinded-quality evidence and its SHA-256 while preserving v1 loading. A v1 negative event may suppress a route, but automation-only v1 evidence may not promote or recover one.

The chronological replay consumes events only when $t_{\text{event}} \leq t_{\text{as-of}}$. It suppresses immediately after a quality regression or after a joint catastrophic regression of at least 50% more tokens and 25% more latency. Recovery requires three later healthy, prospective, exact-match pairs. An unrecovered route remains disabled, with repository-only discovery.

The engine replayed 12 store-write-eligible events in pair-complete timestamp order: 10 blinded-quality v2 events and two automation-only negative v1 events. `buzzr-tab-unread-badge/atomic-design` finishes `suppressed` with `repository-only` discovery after 6 exact-route events. `paraform-command-menu/design-extract` finishes `suppressed` with `repository-only` discovery after 6 exact-route events. No route recovered. The six Nate rows and one legacy row remain chronology-only and created no engine route.

8 Website/frontend production audit

The website is audited from a clean checkout. Product-source review excludes `.vercel`, build output, generated community examples, and other non-product sources from the design score. The verification matrix includes unit tests, Astro production build, Chromium journeys, axe checks in light and dark modes, keyboard/focus behavior, reduced motion, responsive states, size budgets, Lighthouse, and production smoke tests. Screenshot review covers empty/error states and available light/dark variants.

Only observed defects are eligible for repair. The controlled before condition reproduced one serious `label-content-name-mismatch` finding in every registered Lighthouse report. The footer accessible-name correction removed that finding and the unchanged blocking gate then passed cleanly. Infrastructure that prevents a test from executing is reported separately from a product defect, and a green local test is not treated as deployment proof.

A clean-checkout controlled audit compared baseline `f231a5a` with remediation commit `bc131bf`. Before remediation, all nine of nine Lighthouse reports reproduced the `label-content-name-mismatch` finding at *serious* impact, and the unchanged blocking assertion gate exited 1. The footer accessible name was corrected to include its visible text. After remediation, all nine of nine reports contained no blocking accessibility finding and the unchanged blocking gate passed with exit 0.

The controlled Chromium light-mode suite passed 55 of 55 tests both before and after, with no unexpected or top-level errors. The selected Playwright axe rules passed in both conditions; the experimental Lighthouse rule was outside that frozen axe selection. Performance variation across the local runs is not treated as a causal product claim, and local Astro evidence is not deployment proof.

9 2.7.4 remediation and release gates

The product remediation is intentionally narrow: exact route isolation, content-hash isolation, v1-compatible v2 quality evidence, stable chronological JSON replay, immediate suppression, prospective recovery, corrupt/duplicate event rejection, and repository-only fallback. Tests are required for each policy boundary before implementation. No route is force-enabled to make a release result appear favorable.

The reviewed 2.7.4 engineering chain implements exact route identity (task class, repository fingerprint, provider, model, effort, skill ID, and skill content hash), immutable blinded-quality evidence v2, fail-closed legacy-v1 handling, chronological no-look-ahead replay, immediate regression suppression, three-later-pair prospective recovery, corrupt/duplicate event rejection, and repository-only fallback for suppressed routes.

The sealed dry-run artifacts contain 10 complete model-graded v2 quality pairs and 19 chronological ingestion events. `storeWritePlanned` is false: these report artifacts validate deterministic policy inputs and chronology without mutating a production fitness store.

The replay harness pins frozen source and engine digests, rejects path and symlink escapes, and caps combined subprocess output at 10 MiB with process-group termination. The final security review reported no actionable findings; local typecheck and build passed. The residual trust boundary is explicit: same-owner local artifacts lack external signatures, and the append lock relies on normal local-filesystem atomicity; weakly consistent NFS is out of scope.

The admitted live ledger binds the published 2.7.4 code artifact to exact source commit `8aa4649f`; post-release evidence synchronization at `b1b8b7ef` does not alter the published package bytes. The evidence-bound parity clearance merged at `a7261456` and is anchored to the immutable public-gate receipt whose SHA-256 begins `ca45f11f`; neither commit changes the npm tarball. At exact published source `8aa4649f`, the local full suite passed 2,241 tests across 310 files. Clean-install CI passed nine Linux,

macOS, and Windows cells spanning Node 20, 22, and 24. Final verification at tooling merge `09635d81` admitted the byte-bound website provenance and returned an empty failure list with parity eligibility true.

PUBLIC SOFTWARE CHANNELS VERIFIED. Memi 2.7.4 resolves to the published source commit across every non-self-referential public channel. The checksum of this PDF is necessarily established by the detached post-build ledger attached to the release, not by a self-hash embedded in the PDF.

Channel evidence	State
npm trusted-publisher OIDC and exact packed bytes	VERIFIED
fresh Node 20/22/24 install and invocation	VERIFIED
GitHub tag, release assets, and checksums	VERIFIED
GitHub Action v2	VERIFIED
Homebrew formula and installed CLI	VERIFIED
GHCR image digest and invocation	VERIFIED
MCP Registry metadata and package resolution	VERIFIED
website PDF and checksum parity	DETACHED LEDGER
public gate with failures: [] and parityEligible: true	VERIFIED

Ledger admitted at 2026-08-01T16:45:15Z.

Publication is established only after each channel resolves to 2.7.4 and the PDF checksum attached to the GitHub release matches this artifact. A source commit, local package, or passing CI job is not a substitute for registry and production checks.

10 Limitations, falsification criteria, and next study

Limitations. The design uses three intentionally different fixtures, one provider/model/effort setting, six pairs per task, and one Apple-silicon workstation. Model grading can be systematic yet share model-family biases. Provider failures, input ceilings, compile failures, and renderability exclusions reduce the evidence available for some outcomes. Serial execution controls machine contention but does not estimate multi-user throughput. The study does not measure billing cost, maintainability after long-term use, independent practitioner preference, or general performance on unregistered repositories.

Falsification criteria. Any of the following invalidates an affected claim: candidate checksum mismatch; duplicate or non-frozen trial identity; fixture mutation; condition leakage to graders; unverifiable evidence hash; look-ahead in policy replay; unreported exclusion; manual headline number that differs from the executed notebook; a task-level non-inferiority lower bound at or below -5; a critical accessibility or security finding; a LaTeX build with undefined references or overfull boxes; or a release channel that does not resolve to the audited bytes.

Next study. A stronger follow-up should preregister more repositories per platform, independent human practitioners, multiple provider/model/effort strata, measured billing records, repeated hardware environments, and a longer prospective recovery horizon. New evidence must remain segregated by the exact route tuple and content hash.

Reproducibility and artifact contract

The executed source of analysis is `analysis.ipynb`; all headline numbers must independently reproduce from immutable receipts. Machine-readable outputs are expected at:

- `generated/tables/primary_pair_level.csv`, `primary_task_summary.csv`, and `leave_one_pair_out.csv`;
- `generated/tables/secondary_task_tests.csv`, `pooled_descriptive.csv`, and `grader_reliability.csv`;
- `generated/tables/analysis_summary.json`; and
- the TeX and figure hooks enumerated in `report-hooks.tex`.

The protocol, plan, environment, rubric, candidate/release provenance, exclusion/deviation ledgers, receipt ledger, grading records, tables, figures, notebook, TeX source, and final PDF require a checksum inventory before public release. `generated/report-package-checksums.json` inventories the current source package deterministically while excluding PDFs and its own path to avoid cyclic self-hashing. The final PDF must be compiled and hashed only after live release evidence is admitted; it is intentionally not rebuilt by the report-package command.

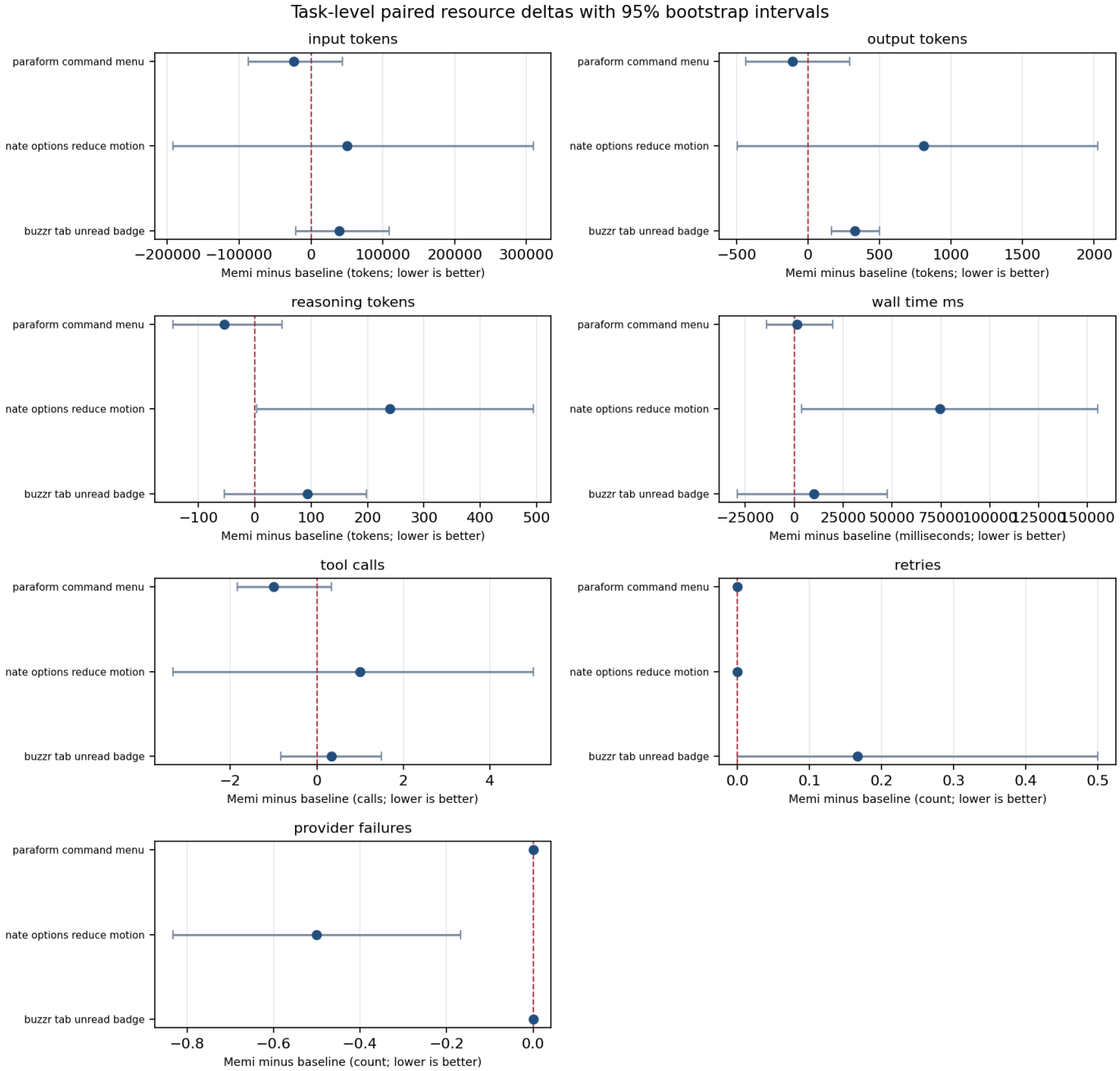

Figure 2: Task-faceted paired resource intervals for the prespecified secondary outcomes. Direction and units must be displayed on each facet.

Interpretation. Across 21 task-by-resource estimates, 4 paired-bootstrap intervals excluded zero and 0 tests rejected after Holm correction. Negative raw deltas indicate lower Memi resource use; these measurements are not converted to dollars.

Limitation. Wall time includes provider and tool variability on one serialized workstation; token measures are provider-reported usage rather than billing records. With six pairs per task, intervals are expected to be wide and secondary tests have limited power.

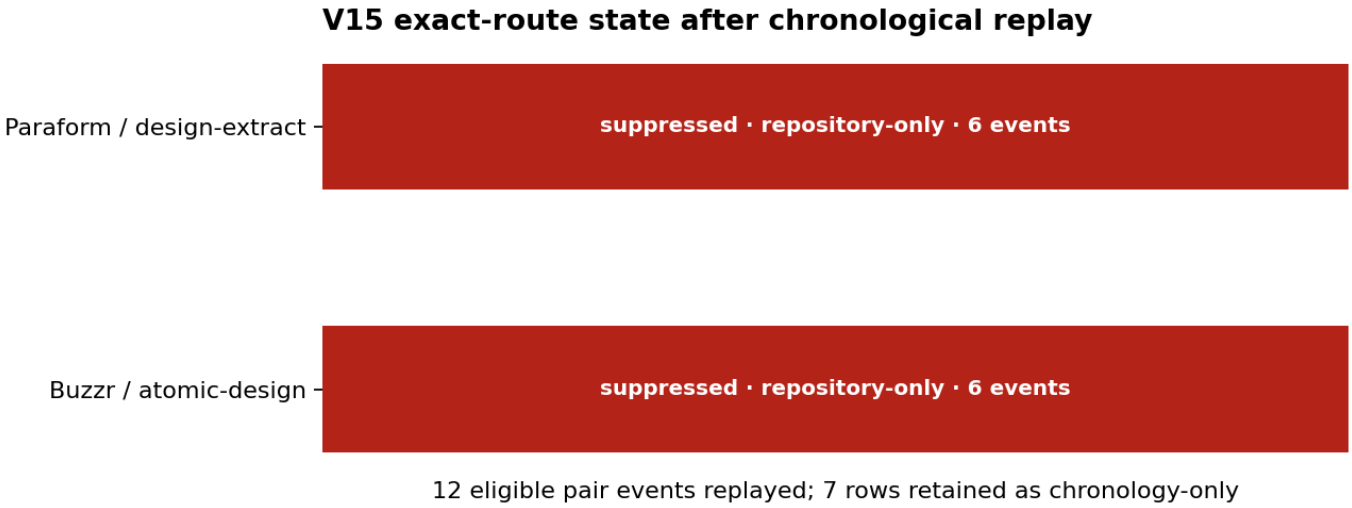

Figure 3: Chronological, no-look-ahead replay of the exact-match routing policy. State changes are derived only from evidence available at each as-of time.

Interpretation. buzzr-tab-unread-badge/atomic-design first entered suppression at 2026-08-01T09:11:49.375Z; paraform-command-menu/design-extract first entered suppression at 2026-08-01T07:13:46.792Z. Both final exact routes therefore use repository-only discovery. Nate abstained in all six repeats, so there is no Nate exact route and no suppression claim for Nate; those observations are retained only in the chronology ledger.

Limitation. A deterministic backtest validates policy mechanics and guards against look-ahead leakage; it does not establish that the historical event stream is representative of future repositories, providers, models, or skill revisions.