

A registered horizon that outlived the intervention the trial ran: a pre-registered randomized trial of memory dosing in a production agent fleet

Luiz Antonio Busnello

Version 2.0 · doi:10.5281/zenodo.23223385

Status: version 2.0 of the pre-registration record. This text is version 2.0 of the Zenodo record that holds the pre-registration of this trial (concept DOI 10.5281/zenodo.21964093, as recorded in the snapshot of v1.12, 10.5281/zenodo.22110203); the DOI of this version is 10.5281/zenodo.23223385. Before this one the record has three published versions: 1.9 (record 21964094, 2026-08-17), the registration; 1.11 (record 21978476, 2026-08-17), a correction of it (1.10 was written and not deposited); and 1.12 (record 22110203, 2026-08-26), an amendment. This version reports the trial. The text was reviewed in full by two model families, both GO with no MEDIUM or HIGH: Fable read rc23 in full and checked the rc23 → rc24 diff, and Codex read rc23 in full (receipt `adversary-receipt-codex-2026-10-07T093115-73165.txt`, in `receipts/`). rc24 and rc25 apply the LOW findings of those reads and change wording only; rc26 changes only this header; rc27 changes only the working list and the changelog (items 208–210). The package built from rc27 then had two final reads, Grok and Fable, both NO-GO; rc28 applies their findings (working list 24, changelog items 211–222), and a confirmation read of rc28 is planned before publication. Every number below is measured; §B names its artifact, and B.1 declares the one number (9.991, §4.7) held by no artifact. Figures (B1, B2), related work (§8) and two sham replays (§4.0.1c), the second over the whole hash-verified trial window (11,812 reconstructible states of 18 epochs; job `job-janela2`, completed 2026-10-07T02:09:46Z, 21/21 runs validated), are in. The v2 release candidates, rc3 to rc28, are listed in the changelog at the end.

Caveat. There were two different adversarial reviews. The numbers in §6 were reviewed adversarially before this manuscript existed; that is how six of those seven defects were found. This manuscript has been under review since 2026-09-21 by five model families, and rc6, rc8 and rc10 each had two full reads (Codex, Fable; working list 19, 20 and 23); §4.3, §4.4 and §3 already carry corrections those reviews produced. The working list is at the end, and the rule of this project applies to it: an item is struck in the same commit that closes it.

This is Paper B of the split decided on 2026-08-28 (`PAPER-SPLIT-2026-08-28.md`). Paper A (`MANUSCRIPT.md`, “*Spare capacity, narrow surface*”) reports the surface measurements and does not depend on this trial having run. This paper reports the trial that the public pre-registration covers, and nothing else.

Abstract

We report a pre-registered, fleet-wide interventional trial of dose-weighted memory promotion in a production agent-memory system. Epochs of 24 h were randomized to control or to one of three dose levels by constrained randomization seeded from a public drand beacon, with the assignment script and its hash registered before the seed was drawn (OSF `yf7d2`, 2026-08-18T07:56:44Z; Zenodo 10.5281/zenodo.22110203).

The registered design called for 234 epochs. Twenty were realized, of which nineteen served data and form the analysis set.

The trial ended before its registered 234-epoch horizon, after the fixed designation exhausted its eligibility. The designation could support dose exposure in only 20 trial epochs, a limitation fixed before Epoch 1. All 19 designated items share one `created_at` and sit in a sub-pool with a 30-day window, so they leave eligibility together at 2026-09-20 22:51:23 (20 of the 234 registered epochs, 8.5%). Past that instant no value of the dose reaches them, because the boost is addressed by id and they are no longer in the candidate list; treatment and control would have been the same intervention for 214 epochs. The deposited registration was not amended and still said 234 epochs; the trial adopted a fixed designation, chosen on 2026-08-26 after v1.12 was deposited and not deposited itself, that was eligible for 20 epochs, and it closed when that designation expired. `sizing.py` sized for 234 epochs on an intervention that existed for 20. This incompatibility was present before the trial began, and the under-powering of this study follows from it, not from the realized window.

The deposited registration (v1.12) names H1, repeated-failure density per session-hour, as the primary outcome and H1a–H1c as a Holm-corrected co-primary family. This paper reports H1c, the share of opportunities yielding a repeated failure, as its primary. The switch was decided on 2026-08-30 (`DESIGN-REVISION-2026-08-30.md §3-ter`; `PROSPECTIVE-ESTIMAND-2026-08-30.md §3-bis`), before the assignment seed was drawn (round 31774052, emitted 2026-08-30T21:32:04Z) and before any randomized epoch existed, in documents dated by our commit log and not deposited. It is therefore a deviation from the public registration, and we report the deposited primary beside it. In the registered analysis H1 does not reject under either reading: tested alone at $\alpha = 0.05$, as the deposited reading has it, $p = 0.3294$; as a secondary in the Holm family, as the switch we adopt has it, adjusted 1.0 (§4.0.2). Under the deposited reading H1 did reject in versions up to rc9 of this analysis ($p = 0.0302$ in rc9, 0.0133 in rc8, 0.0127 in the sensitivity analysis); those versions counted a session that crosses an epoch boundary in every epoch it touches, where the registration counts it in the epoch of its start (§4.3). Under the deposited Holm correction no member of the H1a–H1c and H2 family rejects (§1): H1a’s unadjusted p-value is 0.2599, and every Holm-adjusted p-value in the family is 1.0.

The reported analysis combines deposited analysis provisions, the undeposited analysis specification, and the post-trial substitution rule declared in DEVIATIONS §10.31: the fourth panel family only as a substitute below the three-verdict floor (the deposited registration retained three families and mentioned a fourth only as a mitigation not adopted there), the registered two-hour washout removed from the outcomes and from the session-hour denominator, every session counted in the epoch of its start, the analysis window cut at the designation’s expiry with the specification’s exposure offsets for the partial epochs, a BCa interval with a leave-one-epoch-out jackknife, and 19 epochs (the empty 09–02 excluded). The analysis of earlier versions, which departed from all six, is reported as a sensitivity. One verdict differs between the two: under the deposited reading the sensitivity analysis rejects H1 and the registered analysis does not (§4). Versions up to rc8 kept the washout in the denominator, and versions up to rc9 split boundary-crossing sessions across epochs; the second moves H1 and H1a by an order of magnitude, and H1c slightly (§2, §4).

The primary outcome is null: the registered re-randomization test does not reject the sharp null for H1c ($p = 0.4006$, §4.0.2). Under the planning assumptions and the fractional counting the analysis specification fixes, no reduction from $p_0 = 0.0782$, including total elimination, reaches 80% power (§4.1.1). As pre-committed on 2026-08-30, the null does not distinguish a mechanism that does not work from lessons too generic to work (§4.5). This is a claim about H1c and not about every quantity the trial produced. In the registered analysis no interval of the H1 family excludes zero on any leg: H1 $-1.27 [-28.36; +23.15]$ and H1a $-15.22 [-373.70; +310.82]$ per session-hour (§4.2, §4.3), and neither excludes zero on the registered sensitivity without the sessions that cross an epoch boundary (§5). Earlier versions reported intervals excluding zero for both, and a rejection of H1 under the deposited reading; they rested on three long sessions split across the epochs they touched, one of them spanning 240 h, and they are reported as the sensitivity analyses they now are (§4.3). Two intervals of H2, the secondary confirmatory outcome, exclude zero in the direction of a treatment that costs more (winsorized time, raw tokens) under the registered BCa construction, by margins the percentile construction does not hold, and its registered test rejects neither (§5). The analysis specification, written on 2026-09-10 (ten days before the window closed and before any outcome was computed), dated by our commit log and not deposited with the registration, recorded that the minimum detectable effect for H1c

was saturated at 100% under each of its inclusion cuts, among them the 19 clusters the calculation counted (11 treatment, 8 control): no reduction, including total elimination, was detectable at 80% power. Those 19 clusters are the registered analysis set. The spec counts partial epochs by fractional exposure, and that counting governs the power statement; counting whole epochs at the same 11 treatment and 8 control gives the same verdict (effective size 90.2 against a critical 95.26). Only at the 20 epochs of the sensitivity analysis, with 09-02 as a ninth control, does whole-epoch counting flip it (96.41, §4.1.1). The realized estimate is -0.0140 (treatment 0.0663, control 0.0803; 95% BCa interval $[-0.0464; +0.0110]$, epoch as the resampling unit), containing zero on all three legs (locked, pre-committed sensitivity, post-hoc sensitivity) and in the sensitivity analysis (-0.0199 $[-0.0560; +0.0086]$, $p = 0.1603$).

Our two statements about uncertainty, the planning power approximation and the realized bootstrap interval, rest on different variance constructions and assumptions; their difference does not establish a contradiction, and it does not quantify any under-coverage of the bootstrap. The interval’s coverage has not been calibrated (a BCa bootstrap on 19 clusters, 11 treatment and 8 control, may under-cover; it does not explicitly propagate the stratum-B sampling or the variation between adjudication runs, whose labels did not all reproduce when re-run, and the direction of the coverage error that omission causes is not established; §4.1.2), and the power statement rests on an ICC estimated for a different quantity. Neither is a calibrated measure of this study’s uncertainty. The null (non-rejection by the registered re-randomization test, $p = 0.4006$) does not rest on the interval; the point estimate, a 17% relative reduction, is not distinguishable from zero at this N .

Three further commitments made before the seed, besides the primary outcome and the horizon above, were not kept. One, a stopping rule, we found only after the trial, from the artifacts. Written before the assignment seed (PROSPECTIVE-ESTIMAND-2026-08-30.md §3-bis), it closes the interventional arm with a null if the coverage measured in Epoch 1 falls below 36.7%. Measured now from the locked artifacts by the definition the threshold was derived from, Epoch-1 coverage is 32.4% (45 of 139 opportunities; 33.7% over the active phase alone, while two other readings, after the washout or under the trial’s weighted opportunity rule, are above it): the condition was met under the planning definition (seven signatures promotable at $w = 2$). Under the same definition applied to the dose Epoch 1 was served at (ten signatures promotable at $w = 4$) it was not: 37.4% (52 of 139), and every other reading is above the threshold too. We found no contemporaneous record of evaluating the rule, and the arm continued (§3.0). The second, the replacement of the declared assignment rule, was retracted in writing 24 min 38 s after the drand round was emitted, before Epoch 1 (ASSIGN-SEED-2026-08-30.md, retraction section; commit 3e1c259), and had not been reported in this manuscript until rc8, which quantifies the difference: the seed declaration named a rule other than the registered `assign_arms.py`, and run offline on that round the two rules assign a different arm to 11 of the 20 realized epochs (§1). The third is in the deposit: PREREG §3 locked census adjudication of the live study, and the reported analysis adjudicates a census of the `is_error` stratum and a hash-ordered sample of 800 of the 5 556 other episodes, weighted by 6.945, in the registered and the sensitivity analysis alike (§4). An attempt to restore the census after unblinding stopped at a test-retest gate declared before its first call: on 50 re-adjudicated episodes, two of the three panel families returned 47/50 and 46/50 identical labels against a declared 0.99, so the sample stays a declared deviation (§4, §7).

Given the small number of epochs and the uncertainty limitations above, the primary test’s non-rejection of the sharp null neither establishes absence of effect nor supports a confident efficacy claim. We report the trial in full regardless, because the parts that carry information lie outside the effect estimate: the instrument controls pass, including a sham replay in which, on the 2,646 brief states of the $w = 4$ epochs, where production served that dose and the replay reproduces it, the real designation changed more served briefs than each of 20 matched sham designations would have (132 against 81–122, $p = 1/21$, the floor for 20 shams). Against these 20 shams, that is evidence that the dose acted specifically on what those epochs served. A second replay forces $w = 4$ in all 11,812 reconstructible states of the hash-verified trial window, control epochs included, so outside the $w = 4$ epochs it is a counterfactual and not a record of what was served; there too the real designation exceeds every sham (790 against 449–583, $p = 1/21$). The rank is not a calibrated randomization p-value, the second replay shares 2,016 states with the first (the $w = 4$ epochs except 09-01) and is not a replication, and neither says anything about outcomes (§4.0.1c). A hypothesis that was demoted from primary for requiring, at the pre-trial calibration share of altered briefs, a 955% effect ~~nonetheless returns an interval excluding zero on all three legs, and its test rejects under the deposited reading, which we report as an unexplained rejection conditional on that reading: the denominator explains part of it and not the residual, and the~~

observed reduction is many times what the few briefs the dose alters could carry in proportion to their share (a proportional dilution calculation, not a causal bound) returned an interval excluding zero and, under the deposited reading, a rejection in versions up to rc9, and returns neither once every session is counted in the epoch of its start, as the registration requires: the earlier result did not survive the registered attribution of three long sessions that earlier versions split across epochs, a correction that moves both their exposure and the epoch and arm of their episodes, so it shows the result was not robust, not that the dose had no effect (*corrected in rc10, qualified in rc11; §4.3*); a second hypothesis ~~inverts its conclusion on the removal of a single epoch whose exposure comes from one session with three episodes excludes zero only on the locked leg: its pre-committed sensitivity leg contains zero, and so does a post-hoc leg that removes a single epoch whose exposure comes from one session with three episodes~~ (corrected 2026-10-04: the first version reported only the post-hoc leg; §4.2) excludes zero on the locked and the pre-committed legs under the registered BCa interval, contains zero on a post-hoc leg that removes a single epoch whose exposure comes from one session with three episodes, and is not rejected under the registered Holm rule (corrected in re8: “only on the locked leg” held for the percentile interval of earlier versions, not for the registered BCa; §4.2) excludes zero on all three legs once the washout leaves the denominator, and is not rejected under the registered Holm rule (corrected in re9: under the re8 analysis, which kept the washout in the denominator, the post-hoc leg without 09-14 contained zero; §4.2) shares that history: its locked-leg interval excluded zero in every version that split sessions across epochs, all three legs did in rc9, and every leg contains zero once they are not, and the registered Holm rule never rejected it (*corrected in rc10, qualified in rc15; §4.2*); and the member of the hypothesis family that carried the substantive mechanism question collapses to 1.0 by construction, because two locks of our own pre-registration define its estimand incompatibly and neither of us noticed until the window had closed.

We take the position that these are the reportable results of a trial whose registered horizon outlived the intervention it ran and which is under-powered as a result, and that publishing them is the alternative to the two things a null of this shape is usually used for: silence, or a claim.

1. What was registered, and what this paper reports

The pre-registration (OSF yf7d2, Zenodo 10.5281/zenodo.22110203, v1.12) specifies a fleet-wide trial in which epochs (24-hour periods with a 09:00 UTC boundary) are assigned to control or to a dose $w \in \{2.0, 4.0, 7.5\}$ that scales the promotion weight of designated memory chunks (items). Its designation rule selects, at brief composition, one eligible chunk per signature group (PREREG §2), so the designation is recomputed at every brief (AMENDMENT-v1.12.md §5.2-bis), and v1.12 records that rule as an open defect, not validly frozen (§5 of the amendment). The fixed 19-item designation used in this trial was adopted on 2026-08-26, after v1.12, and was not deposited as an amendment: the deposit carries the per-brief rule declared an open defect (AMENDMENT-v1.12.md §5) and, in DECISION-designacao-2026-08-25.md, a recommendation, marked as awaiting the author’s decision, of a seeded pseudorandom draw of one chunk per signature group (option B; the deposited bytes are md5 35abeb68..., 7 036 bytes, the blob of commit d42f950, and the repository file was extended after publication to record the decision); the decision, the seed, the key layout that was used and the 19 items were not deposited. v1.12 was published at 12:01Z (the creation time of Zenodo record 22110203), the replacement rule was decided at 14:47Z, 2 h 46 min later (DESIGNATION-SEED-2026-08-26.md), and the designation went live at 20:28Z (AMENDMENT-DRAFT-band-collapse-2026-08-26.md §1). Its resolution lives outside the public registration (PROSPECTIVE-ESTIMAND-2026-08-30.md §3, item 1). The designation is a fixed, publicly re-derivable set of 19 chunks, one per signature group (DESIGNATION-2026-08-26.json). Its declaration preceded designation round 31657512 by 1 056 s, counted from the commit time of the declaration that DESIGNATION-2026-08-26.json records (20:07:24Z); the declaration file itself records 20:07:27Z, which gives 1 053 s. Epoch assignment used the separate round 31774052 on 2026-08-30, whose declaration (ASSIGN-SEED-2026-08-30.md, pushed 21:18:35Z) preceded its emission (21:32:04Z) by 13 min 29 s, again according to the recorded timestamps.

The assignment rule, declared and replaced (rc8). The seed declaration did not name the registered assignment rule. It declared a global sort of the 234 epochs by SHA256(seed | epoch index), the first

117 to control and three blocks of 39 to treatments it left unnamed. The registered rule is `assign_arms.py` (stratified block randomization, committed 2026-08-16, last changed 2026-08-17, deposited with v1.11 and v1.12). The assignment derived from the declared rule was committed at 21:33:46Z, after the round was emitted, and it was replaced by the output of `assign_arms.py` at 21:56:42Z, 24 min 38 s after the emission, in a commit that calls the declared rule an error; the same commit appended a retraction section to `ASSIGN-SEED-2026-08-30.md` that names the registered rule and declares the replacement. That section says it was filed “*hours after the round was consumed*”; the commit is 24 min 38 s after the emission. The trial served the registered rule’s assignment. Both rules run offline on the round’s published randomness (`_sprint-2026-10-04/B-rc8/checks-rc8.json`, block E): each reproduces its own published output (the declaration’s `sha256_da_atribuicao 2426d13d...`, and `ASSIGNMENT.json`, which `ASSIGNMENT-SERVING.json` relabels without change), and they assign a different binary arm to 130 of the 234 epochs and to 11 of the 20 realized ones (09-01, 09-05, 09-06, 09-07, 09-08, 09-09, 09-11, 09-13, 09-17, 09-19, 09-20). Under the declared rule the realized window would have held 10 control and 10 treatment epochs instead of 9 and 11. This is a deviation, and the order of events does not remove it: the replacement was chosen after both assignments were computable, so the only protection is that the rule it adopted was fixed on 2026-08-17 (its last commit) and deposited in Zenodo v1.11 (record 21978476, published 2026-08-17T18:32Z, the day before the OSF registration; v1.12 carries the same bytes, md5 `3fa3f710...`), thirteen days before the round was emitted, which the deposit, not our log, dates. The declared rule was never run as the trial’s assignment. The declaration also fixed its own consequence: if its published verification hash and `ASSIGNMENT.json` diverged after emission, “*esta declaração falhou e o estudo não começa*”, and no post-emission correction was allowed (“*Qualquer defeito encontrado a partir daqui é declarado como desvio, não emendado no sorteio*”). The retraction overrode that clause; the study started on the registered rule’s assignment on 2026-09-01.

The registered hypothesis family is nested ($H1c \subseteq H1b \subseteq H1a$), and the H1b lock of 2026-08-16 in §1 of the pre-registration (PREREG 1.308) calls that nesting “*the nesting that makes the joint reporting work*”:

	statement	status here
H1	density of repeated failures per session-hour	computable ; the deposited primary, demoted 2026-08-30 in an undeposited revision (below); the registered test does not reject under either reading ($p = 0.3294$ alone; Holm-adjusted 1.0); it rejected under the deposited reading only in analyses that split sessions across epochs (§4.0.2, §4.3)
H1a	rate of eligible opportunities per session-hour	computable ; unadjusted $p = 0.2599$, not rejected under the registered Holm rule; does not bear weight (§4.2)
H1b	share of opportunities yielding a repeat attempt	trivially 1.0 ; the question it carried is unanswerable (§4.4)
H1c	share of opportunities yielding a repeated failure	primary in this report (switched before any data, undeposited; below); null; no reduction detectable at 80% power under the spec’s fractional counting (§4.1.1)

The primary switch, declared. The registration as deposited (v1.12) names H1 the primary and H1a–H1c a Holm-corrected co-primary family; its multiple-comparisons rule tests H1 at $\alpha = 0.05$ two-sided and Holm-corrects H1a–H1c together with H2. H1c was made the primary on 2026-08-30 in `DESIGN-REVISION-2026-08-30.md` §3-ter (commit of 16:51:56Z), and the switch was recorded in `PROSPECTIVE-ESTIMAND-2026-08-30.md` §3-bis (commit of 20:51:20Z), whose own summary table (its l.26) still names the density outcome. Both commits precede the declaration of the assignment seed (pushed 21:18:35Z) and the emission of its drand round (31774052, 21:32:04Z), and no randomized epoch existed then. Neither document was deposited; a v1.13 amendment prepared on 2026-08-27 was not deposited either (`deposit/PLAN-v1.13.md`). The ordering therefore rests on our commit log, the weaker kind of evidence §1.1 discusses. We report both readings: under the switch, H1c is null ($p = 0.4006$); under the deposited reading, H1 is the primary, tested alone, and does not reject either ($p = 0.3294$). Under the deposited Holm correction no member of the family rejects. The family holds H1a, H1b, H1c and the two components of H2. H1b has no test (§4.4) and enters as $p = 1$. The

smallest unadjusted p-value is H1a’s, 0.2599, far above the first Holm threshold at either family size (0.01 with all five members, 0.0125 without H1b), and every adjusted p-value is 1.0; H1c (0.4006) and H2 (0.8285 for time, 0.3372 for tokens) are further away. In the sensitivity analysis H1 rejects under the deposited reading (0.0127) and the family does not (H1a 0.0855; H1c 0.1603); so did the analyses of rc8 and rc9, which split sessions across epochs (H1 0.0133 and 0.0302).

PROSPECTIVE-ESTIMAND §3-bis made H1c primary and H1/H1a/H1b secondary, “*reportadas com o mesmo rigor*”, without restating the multiplicity rule. Under that switch H1 is a secondary, and the registration’s rule for secondaries is Holm. Placed in the family with H1a–H1c and H2 ($m = 6$, or 5 without H1b), the first threshold is 0.0083 (0.01); the smallest p-value is H1a’s 0.2599, so the step-down stops there, and every member, H1 included, is adjusted to 1.0. Testing H1c alone and the other members in Holm gives the same verdict (all adjusted 1.0). In the sensitivity analysis H1 is adjusted to 0.0762 (0.0635 with $m = 5$) and does not reject either; in rc9’s analysis, 0.151 (0.1208). So in no analysis does H1 reject under the switch, and in the registered one it rejects under neither reading; in the sensitivity analyses H1 rejects only under the deposited reading, in which it is the primary and is tested alone. Under either reading H1c does not reject (out/ITT-REGISTRADO-v4-2026-10-05.json, field *multiplicidade*).

This paper reports the intention-to-treat analysis of that family and, in §5, the secondary confirmatory H2 (task regret), which does not reject; H3 (retrieval metrics, exploratory) is not computable, and §5 says why. The surface measurements made *while* the trial was being built (the exposure ceiling, the carousel, the coverage channel) are Paper A, are explicitly exploratory and descriptive, and were never pre-registered.

1.1 The order in which decisions were taken, and why it is stated first

Two estimand decisions had to be made after the window closed (§4.4 and §4.2). Both were taken on 2026-09-21 and are recorded in DEVIATIONS-FOR-PAPER.md §10.32 and §10.33, with the times 16:14 and 16:20 (local time, UTC−3) written in the text. ~~Both were recorded before the family was computed, and no number of the H1 family had been produced at either point.~~ The two sections were committed in one piece with the estimates they report (at 16:22 local), so the commit timestamp dates the sections, not the decisions. The ordering rests on the written times and on ITT-PRELIMINAR.json, a preliminary run of the estimator (2 000 replicates) whose file time is 16:16:44 local and which cites §10.32. That places the §10.32 decision (H1b, 16:14) before any estimate of the family. It places the §10.33 decision (to keep the locked session-hour denominator, 16:20) **after** a preliminary estimate existed: the preliminary file already holds the H1a difference, −143.92, with an interval excluding zero. (*Corrected in rc7: earlier versions said that no number of the H1 family had been produced at either point, which does not hold for §10.33.*) The §10.33 decision kept the registered definition rather than changing it, so seeing the number could not have bought a better-looking result through it; it is still not a decision taken before the number. A file time is filesystem metadata, as forgeable as a commit timestamp, and ITT-PRELIMINAR.json entered the ballast manifest only in rc23 (2026-10-07; working list 17), so the manifest does not date it either.

We state this first because it is the claim in this paper most dependent on our own bookkeeping, and §9 of the analysis specification exists to forbid the opposite order: “*choosing the primary set after seeing the number is the post-hoc play that this spec exists to prevent.*”

The evidence for it is also weaker than the evidence for everything else, by a gap we created ourselves. We anchored the *assignment* in a public drand beacon so that no one would have to trust us about it, and then anchored the decision log in nothing but our own commit timestamps, which are forgeable in seconds. Appendix-only is process discipline, not a cryptographic property. An adversarial reviewer put it as an asymmetry, and it holds: we knew how to do this and did it for the easier claim.

What a reader can check without trusting us at all:

- the MDE saturation of §4.1 is arithmetic over the spec’s inclusion cuts (its fractional counting, and the 19 clusters the calculation counted: 11 treatment, 8 control), the pre-registered ICC and $\hat{p}_0$: outcome-independent, and reproducible from the artifacts;
- the designation’s expiry of §3.0.1 (one *created_at* for all 19 items, a 30-day window), measured before the close (out/expiracao-designados-2026-09-09.json);

- the assignment, from drand round 31774052 through `assign_arms.py` to the served arms, and that the rule the seed declaration named gives a different assignment from the same round (§1; both rules run offline);
- the Epoch-1 coverage of §3.0 against the 36.7% stopping threshold, from the locked episodes and the planning artifact’s list of promoted signatures, and under the set promotable at $w = 4$ re-derived from the calibration replay, on the same terms as every estimate (next item);
- every estimate except the `control` mean 9.991 of §4.7 (B.1), from the artifacts of Appendix B, once deposited (working list 8).

What requires trusting us: the estimand decision of §4.4 preceded the estimates, and the one of §4.2 preceded the final run but not the preliminary one (above); the primary switch of §1 preceded the assignment seed; and both seed declarations preceded their drand rounds (§1), which rests on our recorded push times against drand’s public emission times. The closure decision of §3.0 was taken on 2026-09-09, before any outcome was computed; that date rests on the times written in our append-only deviation log (§10.14; §10.22, 14:38 local time) and on the switch-off script’s dry-run receipt of 2026-09-09 17:42:31Z (§10.28). The replacement of the declared assignment rule (§1) is dated by our commit log too, and there the log works against us: it places the replacement after the round was emitted. One thing speaks to our practice of committing before data, though not to these orderings, which rest on our own timestamps: the spec omits 09–10 and marks nine epochs as projection, which a fabricator writing after the close would have had no reason to do. ~~And the spec contains a commitment the data later refuted (§4.5). A post-hoc account would not contain a commitment that its own data go on to contradict.~~ *(Struck in rc7: the realized coverage counts a different quantity from the one the pre-commitment projected, so it neither refutes nor tests that commitment; §4.5.)*

The core claim (the registered horizon outlived the intervention the trial ran, and the trial is under-powered as a consequence) sits entirely in the first list. A reader who disbelieves the second list loses §4.4 and should read H1, the deposited primary, as the primary: under that deposited reading it does not reject in the registered analysis ($p = 0.3294$), and rejects only in the sensitivity analyses that split sessions across epochs (§4.0.2, §4.3). The rest of the paper stands.

Caveat. An earlier version of §9 overstated this: *“we knew this ten days before the window closed.”* On 2026-09-10 the spec marked nine epochs as projection (09–11...09–19); 09–20, entered as partial by its clock hours, had not begun either, and 09–10 was still open. The spec’s own extremes analysis covers that (the bounding cases still return not-detectable), so the conclusion holds, but it was projection-robust, not known; §9 now says so. And it was projection-robust only under the spec’s cuts. At the registered analysis set (11 treatment, 8 control) it holds under whole-epoch counting as well; counting whole epochs at the sensitivity analysis’s 11 treatment and 9 control, a cut the spec does not include, the verdict flips (§4.1.1).

2. Design

Randomization. Stratified block randomization over four strata, (`first calendar half | second half`) $\times$ (`weekday | weekend`), with group counts allocated by exact two-way controlled rounding and labels shuffled by a seeded Fisher-Yates. The entropy stream is counter-mode SHA-256 rather than any language’s PRNG, so that *“re-run the committed script”* means re-run it in any language. Seed: `SHA256(randomness_hex)` of drand round 31774052, hashed as lowercase ASCII text. Allocation at $N = 234$: 117 control, 39 per dose.

The registered tolerance (every group $\times$ stratum cell receives its exact share rounded down or up) holds by construction rather than by rejection sampling, so there is no acceptance rate and no failure mode in which the sampler quietly relaxes a constraint. Realized maximum deviation: 0.5 against a tolerance of 1.0.

Washout. Two hours from the epoch boundary. A session that starts inside it is excluded from the analysis, from the outcomes and from the session-hour denominator alike (PREREG §2: *“all post-washout session-hours”*); on the locked corpus this removes exactly the episodes inside the two hours. *(Corrected in rc9: versions up to rc8 removed those episodes from the outcomes and kept them in the session-hour denominator of H1 and H1a; §4.)*

Sessions that cross an epoch boundary. The brief is served once, at session start, so the registration attributes a session to the epoch of its start, flags it, and reports a sensitivity with and without such sessions; sessions longer than one epoch are reported as their own stratum (PREREG §2: “*Boundary-straddling sessions are assigned to the epoch of their start, flagged, with sensitivity with/without*”). From rc10 the analysis does this: every episode is counted in the epoch where its session started, for exposure, outcomes and the washout alike. (*Corrected in rc10: versions up to rc9 counted each episode in the epoch of its own timestamp, so a long session contributed exposure to every epoch it touched; §4, §4.3.*)

Severity threshold. $\tau = S1$, locked in July 2026 on a calibration over five model families (PREREG §697-699).

Caveat. We do not claim the outcome data “confirm” τ . A parameter locked in July is not re-validated with data from the outcome period; that is post-hoc, and an earlier draft of our own deviation log made exactly that claim before adversarial review removed it (§10.31(5)). What the outcome data do show, and what we report instead, is an 11.9 pp difference in marginal failure rates at the S0/S1 boundary (the boundary τ does *not* absorb; a lower bound on their paired disagreement) between two panel families (google 42.3% failure against zhipu 30.4%).

3. Execution: the realized window is not the projected one

The trial went **active** on 2026-09-01 at 10:25:39Z and the dose was switched off on 2026-09-21 at 09:43:05Z, closing the window at the 2026-09-20 epoch.

	epochs
registered design	234
designated in the realized window	20
served data	19
whole by clock exposure	16
partial by clock exposure	3 (09-01, 09-03, 09-20)
empty	1 (09-02, never served)

Realized allocation over the 20 designated: **9 control · 6 at w=2 · 4 at w=4 · 1 at w=7.5**, i.e. 9 control and 11 treatment.

Correction. Two different denominators, both correct, and we first wrote one number for both.

analysis	epochs	why
ITT (§4.1–4.4), registered	19	the analysis set of §2 of the analysis spec: 09-02 served no brief and is excluded, <i>counted and reported</i> . Its sessions produced 88 episodes ; it was designated control
ITT, sensitivity analysis	20	every designated epoch, 09-02 included; this was the ITT of versions up to rc7
coverage and the designated-presence correlation (§4.5–4.6)	19	coverage is defined over served briefs, and 09-02 has none. The quantity does not exist there, which is not the same as being zero

Earlier versions included 09-02 in the ITT on the ground that dropping a designated epoch is post-randomization conditioning, the defect §3.1 records us committing once already. That ground holds: 09-02 is empty because of an outage (INCIDENT-2026-09-02-epoch-perdido.md), an event after randomization. But the analysis spec, written before any outcome was computed, decided it the other way, and its decision is the registered one; earlier versions overrode it without declaring that they did. The cost is stated in §4.1:

09-02 is a control epoch with the second-highest per-epoch H1c of the twenty, and its exclusion is the switch that moves H1c most.

Deviation, declared in rc7; resolved in rc8. §2 of the analysis spec excludes 09-02 from the analysis set (*counted and reported*, and it records that excluding a control epoch worsens the imbalance from 9 to 8). The ITT of rc7 included it, which the spec did not pre-commit; the registered analysis of rc8 excludes it, and the rc7 analysis is reported as a sensitivity.

09-01 is reported with both phases declared, as §6 item 6 of the analysis spec requires: 42 briefs in shadow until 10:22Z and 630 in active from 10:37Z, with no overlap. The registered analysis counts it from its first active record (10:37:01.943Z), which is how the spec’s exposure offset is implemented (§3.0.1). The sensitivity leg that removes it is reported alongside.

3.0 The stopping rule, and the feasibility question it raises

The deposited stopping rules (rc9). The deposited registration fixes the horizon in pre-treatment units, 234 randomized epochs or a calendar cap, with “*No interim analyses; no optional stopping.*” It adds a mechanical safety abort, arm-blind: a script “*evaluates the following arm-blind rule at every epoch boundary*” over the incident stream, and halts the study on one incident adjudicated S4 in an analyzed epoch, or on a count of incidents at S3 or above over three epochs exceeding three times the baseline, which the registration records as zero. We find no record that the abort ran at any of the 20 epoch boundaries. The script was deployed on 2026-08-21 dormant, with nothing importing or scheduling it (DEPLOY-C2-C3-2026-08-21.md); its one recorded run, that day, exited without evaluating, because seven days of history fell short of the 14-day gate (SHADOW-ARMED-2026-08-21.md); the crontab snapshots of 2026-08-23 and 2026-08-31 and the inventory of the trial’s seven cron lines of 2026-09-09, in our infrastructure repository, hold no abort entry; and its log is not among the locked artifacts. We did not inspect the production host for this version. Evaluated ex post over the 696 episodes of the window that the substitution panel adjudicated, no episode reaches a panel-majority level of S3 or S4 (10 have at least one panelist at S3 or above, one at S4), so neither clause would have fired on them (_sprint-2026-10-04/B-registered/f3-abort-ex-post.json). That is not the registered evaluation, which reads every incident at every boundary. The absence of a contemporaneous execution record for the safety abort is declared in Appendix A.

The trial closed at the 2026-09-20 epoch by a decision taken on 2026-09-09 (DEVIATIONS-FOR-PAPER.md §10.14, the decision not to widen the window; §10.22, the author’s explicit instruction of 14:38 local time, UTC−3), executed on 2026-09-21 by desliga-dose-p2.sh at 09:43:05Z (§10.29). It was not an outcome-dependent stop: it rested on measurements, not on outcomes, among them the expiry measurement and the instrument’s positive-control reading of 09-08 (20 of 672 briefs changed; §10.14). No outcome had been computed when it was taken, and the analysis specification that governs this report was written on 2026-09-10, the day after the decision and eleven days before the first outcome was computed (dated by our commit log; it was not deposited with the registration, whose deposited version is v1.12). ~~That calendar decision is the whole of the stopping rule.~~

Correction (rc8, qualified in rc9): a declared stopping rule was met in Epoch 1 under its planning definition and not executed. The calendar decision was not the only stopping rule. PROSPECTIVE-ESTIMAND-2026-08-30.md §3-bis, written before the assignment seed and dated by our commit log, made H1c primary and declared “*a decisão de parar*”: if the coverage measured in Epoch 1 falls below **36.7%**, the required effect exceeds 100% and the interventional arm is closed with a null “*por impossibilidade de desenho, não por ausência de efeito*”. Coverage there is the quantity of out/CONCENTRATION-2026-08-30.json: the share of opportunities whose signature the dose can promote (the seven signatures that artifact lists as promotable at $w = 2$; Epoch 1 ran at $w = 4$), 40.0% (611 of 1 526) in the planning corpus, and 44.4% (677 of 1 526) for the ten signatures promotable at $w = 4$ (below); the threshold is the H1c MDE of 36.7%, and the same section says that measuring it “*é a primeira coisa que o Epoch 1 tem de fazer*”.

We find no record that it was measured. We have measured it now, from the locked episodes, with the planning script’s rule as written (an action is an opportunity if its signature produced an `is_error` episode earlier in the corpus, which here begins on 2026-08-23; _sprint-2026-10-04/B-rc8/checks-rc8.json, block D). Epoch 1 (09-01, treatment at $w = 4$) has 139 opportunities, 45 of them in a promoted signature: **32.4%**

(Wilson 95% interval [25.2%; 40.5%], descriptive only). The condition the rule names was met under the planning definition, and the arm was not closed. Three other readings of “coverage in Epoch 1” are reported so that the verdict does not rest on our choice among them: 33.7% (34 of 101) over the **active** phase alone, below the threshold; 40.0% (30 of 75) after the two-hour washout, above it; and 54.5% of the HT-weighted opportunities under the trial’s own opportunity rule, above it. Counted against the 19 designated signature groups instead of the seven promoted signatures, coverage is 92.1% (89.4% weighted), but that is not the quantity the threshold was derived from. The reading that uses the rule’s own definition over the epoch it names is below the threshold, and so is the **active**-phase reading.

Under the dose Epoch 1 was served at, the condition was not met (rc9). The seven signatures are those promotable at $w = 2$, the dose the planning artifact assumed; Epoch 1 was served at $w = 4$. We re-derived the promotable set from the calibration replay the planning script reads for its denominator (`out/dose-350-v3.json`: a signature is promotable at w if its designated item enters the brief in at least one of the 350 states). That derivation gives back the artifact’s seven at $w = 2$ exactly, and ten at $w = 4$, adding `Read|arquivo:doc`, `Read|arquivo:outro` and `ToolSearch|consulta`. With the rule otherwise unchanged, Epoch-1 coverage at $w = 4$ is **37.4%** (52 of 139; Wilson [29.8%; 45.7%]), above the threshold, and so are the other readings: 40.6% over the **active** phase, 48.0% after the washout, 65.1% weighted (`_sprint-2026-10-04/B-registered/f4-promoviveis-w4.json`). Both Wilson intervals contain 36.7%. Whether the condition was met therefore depends on which promotable set defines coverage, and §3-bis does not say: it was written before the assignment, when the dose of Epoch 1 was not known, and the coverage it names is the planning artifact’s. We report both readings and do not choose between them.

What this does to the paper: had the rule been executed under the planning definition, the interventional arm would have ended after 09-01 with a null declared as design impossibility, and none of the H1-family estimates of §4 would exist; under the served dose’s set it would not have fired. We report the estimates as what a trial that may have been due to stop produced; they are not rescued by this section, and the verdict of §4.1 (H1c null, no reduction detectable at 80% power) is the verdict the stopping rule anticipated. Relative to the undeposited revision that declared the rule, not executing it is a deviation, recorded in Appendix A; relative to the deposit, which forbids optional stopping and contains no such stop, continuing is consistent. It is the most consequential deviation we found ourselves.

3.0.1 The fixed designation could support dose exposure in only 20 of 234 registered epochs

An adversarial reviewer asked whether the fixed designation ages out of eligibility. It does.

All 19 designated items live in `memory/entities/lessons/*.md`, the global sub-pool, whose window is `freshGlobalMaxAgeDays = 30`. Their `source_date` is NULL in all 19, so the predicate falls through to `created_at`, which is a single value for all nineteen:

```
created_at = 2026-08-21 22:51:23    (identical across all 19)
+ 30 days  = 2026-09-20 22:51:23    ← they leave the window together, in one instant
```

epochs registered	234 (2026-09-01 → 2027-04-22)
epochs in which the designation is eligible	20 (2026-09-01 → 2026-09-20)
share of the registered design that could carry an intervention	8.5%

After that instant the intervention does not exist. A chunk that fails `fetchFreshCandidates` never reaches the list handed to the boost provider, and the boost is addressed by id. No value of w and no number of `freshSlots` reaches a chunk that is not in the candidate list. For 214 of the 234 registered epochs, the treatment arm and the control arm would have been the same intervention.

This changes how the stopping rule and the headline read. The trial ended before the deposited data-collection horizon: neither 234 randomized epochs nor the calendar cap was reached, and expiry of the fixed designation is not a deposited stopping condition (Appendix A). It ran for the entire interval in which the

dose could have reached a designated item: the designation expired at 2026-09-20 22:51:23Z, and the dose was switched off only after the fixed designation had expired (09:43:05Z on 2026-09-21), when it no longer reached any designated item. The 09-20 epoch is partial for this reason and not by arbitrary truncation: the analysis spec’s cut “*falls inside the epoch, at 22:51:23Z*” is that instant. ~~The analysis window closes at that instant.~~ The registered analysis closes the window at that instant: the 33 episodes of 09-20 after expiry leave the analysis, taking 5 opportunities (22.835 weighted) and 2 weighted repeated failures with them, and 09-20’s session-hours fall from 0.5902 to 0.4820 (`out/ITT-REGISTRADO-v4-2026-10-05.json`, legs `registrado_sem_janela` and `registrado`). It also applies the exposure offsets §2 of the spec assigns to the partial epochs, as exposure windows on the timestamp axis, because the ratio estimator has no offset term: 09-01 from its first `active` record (10:37:01.943Z), 09-03 from its first served record (17:23:39.777Z, the end of the outage), 09-20 until expiry. The cut is applied to episodes attributed to 09-20; no session that starts earlier has episodes past expiry (`_sprint-2026-10-04/B-rc15/checks-rc15.json`, block S: the latest epoch any boundary-crossing session touches is 09-18), so no episode escapes it on this corpus. That reading of “offset” is ours, and for 09-03, whose offset the spec gives only as a volume (441 of 672 briefs), it is an interpretation. The cut is negligible on every hypothesis; the offsets move H1 and H1a, because they shrink the session-hour denominator of 09-01 (0.543 to 0.406 h) and 09-03 (0.457 to 0.246 h), measured with the washout still in the denominator (§4). The sensitivity analysis, the estimator of earlier versions, selects episodes by epoch date and applies neither.

Caveat. The under-powering is not a consequence of the realized window, and it is not a contradiction inside the registration either. `sizing.py` computed $N = 234$ from the pilot, and the deposited registration (v1.12) carries that horizon beside a designation rule that is recomputed at every brief (§1); it names the 30-day window as a constraint on reach (PREREG §2: the reach table at 1.513-519 is computed under it, and 1.550-558 give the minimum dose at chunk ages up to 90 days), not as the life of a fixed set. The fixed designation, whose members share one `created_at` and so expire together, was adopted on 2026-08-26, after v1.12, and was not deposited (§1), and nothing re-read the registered horizon against it. Both numbers, the horizon and the 20-epoch life of the fixed set, are ours, both were locked, and they are incompatible, but only the horizon is locked in the registration: the deposit carries the per-brief rule declared an open defect (AMENDMENT-v1.12.md §5) and, in `DECISION-designacao-2026-08-25.md`, a recommendation, marked as awaiting the author’s decision, of a seeded pseudorandom draw of one chunk per signature group (option B), which was decided 2 h 46 min after publication; the decision, the seed, the key layout that was used (the deposited option B keys on `sig_primary`, which was dropped from the key at 19:40Z) and the 19 items were not deposited. This is therefore a mismatch between the registered horizon and the intervention frozen after it, unlike the H1b collision of §4.4, whose two locks are both registered. It was present before the trial began: the trial was sized for 234 epochs on an intervention that exists for 20.

We record two further facts so the finding is not overstated. The designation passed the full eligibility predicate (file pattern, `importance/pain` ≥ 0.7 , age ≤ 30 d) in 19 of 19 on 2026-09-09, in both the served corpus and `current.db`; expiry was a future event, not a live defect. And the replay harness compensates for the window (`cfgEm` offsets `freshGlobalMaxAgeDays`), so the expiry ends the intervention and leaves the instrument unaffected.

This was measurable on 2026-09-09 and was measured then, eleven days before the close, on the day the decision was taken not to widen the window and so to end the trial at the 09-20 epoch (§3.0; DEVIATIONS §10.13, §10.14, §10.22). It did not reach this manuscript until an adversarial reviewer asked the question from the outside. The measurement was ours; the reviewer saw that it belonged in the paper.

3.1 Two classification errors of ours, and what they cost

Correction. We classified 09-20 with the wrong ruler. We first published $16+3+1$ as “*17 whole + 2 partial + 1 empty*”, having measured delivery volume (672 of 672 serving records $\Rightarrow$ “whole”) where the pre-registered spec classifies by clock exposure (13.86 h of 24 $\Rightarrow$ partial, included with offset). The rulers are different and the pre-registered one is not optional. The error propagated to six places including a pushed commit, and was presented as “*matches the spec exactly*” when only the total matched and the decomposition (which decides whether the epoch enters with an offset) did not. Corrected in `DEVIATIONS-FOR-PAPER.md` §10.31(1).

Correction. We derived the arm from the data instead of from the designation. The first pass took each epoch’s arm as the mode of its observed **active** dose. That is post-randomization conditioning: the mode is a function of *when active* took effect. The arm is taken from **ASSIGNMENT-SERVING.json**. Compared, the two agree on 0 divergences across 19 epochs. The premise still changes, and the paper reports the premise. This also voided a cross-validation we had claimed: comparing our own census against the spec’s table is not a second opinion, because both derive from the same designation.

Caveat. The spec’s own table **omits 09-10**, summing to 19 against 20 allocated; it was written on 2026-09-10 at 21:40 with that epoch still open. Including it gives 16 whole. We record this rather than silently reconciling, because a miscount in the spec and a miscount by the analyst need different fixes.

4. Results

Estimator. `estimador_itt.py`, which is composition, not reimplement: it imports `carregar_verdicts`, `carregar_episodios` and `span_por_sessao` from the pilot’s replay module and the arm from the assignment file. Two copies of one rule can silently select different populations, a defect class this project has already paid for once. The registered analysis is `measurement/estimador_itt_registrado.py` (`out/ITT-REGISTRADO-v4-2026-10-05.json`). It imports the same blocks and those of `rerandomizacao.py`, leaves both files unmodified, and adds six switches, one per departure of the earlier analysis from the registration: panel, window, interval, epoch set, from rc9 the washout in the session-hour denominator, and from rc10 the attribution of every session to the epoch of its start. With every switch at the earlier choice it rebuilds `ITT-2026-09-21.json` and `RERANDOMIZACAO-2026-09-21.json` byte for byte (sha256 compared at run time), so every difference between the registered analysis and the sensitivity analysis below comes from the switches and none from a reimplement. With the sixth switch at the earlier choice it gives back the registered analysis of rc9 (`out/ITT-REGISTRADO-v2-2026-10-05.json`) field for field, and with the fifth and sixth that of rc8 (`out/ITT-REGISTRADO-2026-10-05.json`), in both cases under the BCa acceleration those versions used; with that acceleration it also gives back every BCa interval of rc10 to rc14 (`out/ITT-REGISTRADO-v3-2026-10-05.json`), and run with `--aceleracao uma_amostra` it emits that file in every field except its timestamp and the script’s sha256 (§4, *Uncertainty*).

What moves what (one switch at a time, the others at the earlier choice; the artifact’s `deltas_um_a_um`). The epoch set moves H1c most: excluding 09-02, a control epoch whose own H1c (0.190) is the second-highest of the twenty, takes the H1c difference from -0.0199 to -0.0113 and its p-value from 0.160 to 0.439. The window acts on H1 and H1a almost entirely through the offsets (H1a’s p-value from 0.0855 to 0.0375 with the offsets alone), while the expiry cut alone moves no hypothesis by more than 0.5% of its estimate. The panel substitution resolves 20 more episodes (1 159 to 1 179) and moves H1c to -0.0208 . BCa changes only the intervals. Removing the epoch-set switch from the registered analysis puts H1c back at -0.0228 ($p = 0.1405$); in rc9’s analysis it gave -0.0209 ($p = 0.152$).

The washout in the denominator (rc9). The estimator dropped episodes inside the two-hour washout from the outcomes but computed session-hours before that filter, so the washout stayed in the denominator of H1 and H1a; `estimador_itt.py` has the same defect, and the byte-for-byte control preserved it. The registration’s denominator is “*all post-washout session-hours*” (§2). The fifth switch removes the sessions that start inside the washout before computing both exposure and outcomes; the corpus of past failures that defines an opportunity is never cut. In the registered set it removes 1 007 episodes (638 in treatment epochs, 369 in control), all inside the two hours, and no opportunity, because the outcomes already excluded them; session-hours fall from $12.56 / 4.33$ to $10.75 / 3.25$. Alone, it takes H1 from -14.62 to -21.95 and H1a from -143.92 to -222.26 (its p-value from 0.0855 to 0.0128), and leaves H1c unchanged. Added to the four switches of rc8, it takes H1 from -14.12 ($p = 0.0133$) to -19.85 ($p = 0.0302$) and H1a from -163.52 ($p = 0.0241$) to -233.20 ($p = 0.0168$); H1c stays at -0.0121 ($p = 0.434$). With sessions counted at their start (next paragraph) the washout removes 972 episodes (603 treatment, 369 control).

Sessions in the epoch of their start (rc10). Up to rc9 the estimator keyed each session by the epoch of each of its episodes, so a session that ran for days contributed its span, piece by piece, to every epoch

it touched. The registration attributes it to the epoch of its start (§2). The sixth switch does that: each episode takes its session’s start epoch, its offset is measured from that epoch’s start, and condition (i) is checked against it (PREREG §4.1 reads the serving snapshot *at session start*); the past failures that define an opportunity keep the epoch in which they were written. Six sessions of the corpus have episodes in more than one epoch. Three start inside the window, and each spans more than one epoch length, which makes them the stratum the registration reports on its own: `d37a5964` starts in 09-08 (treatment) and spans 240.42 h with 60 episodes over 11 epochs; `74de1e72` starts in 09-12 (treatment), 81.01 h, 67 episodes; `ce6f5640` starts in 09-11 (control), 54.00 h, 2 episodes. A fourth starts on 2026-08-27, before the window, and leaves the analysis, taking the episodes it had in 09-02...09-07 with it (49 episodes in all, from 2026-08-27). As a stratum, the three hold 321.43 h, 139.96 weighted opportunities and 6.00 weighted repeats in treatment, and 54.00 h with no opportunity and no repeat in control. Within this stratum, the descriptive treatment–control differences are approximately +0.01867 repeated failures/hour and +0.43543 opportunities/hour (6.00 / 321.43 and 139.96 / 321.43 against 0 / 54.00 in control; the registered leg minus the registered ‘without’ leg, per arm, in `out/ITT-REGISTRADO-v4-2026-10-05.json`). H1c cannot be contrasted because control has no eligible opportunities. We report no standalone inferential interval or test for this three-session stratum. Counted at their start, these sessions carry 240.75 h of 09-08, 81.37 h of 09-12 and 54.31 h of 09-11, and the session-hours become 324.95 in treatment and 57.06 in control, against 10.75 / 3.25 with the sessions split. Alone, the switch takes H1 from –14.62 to –1.55 and H1a from –143.92 to –16.26, and H1c from –0.0199 to –0.0221. Added to the five switches of rc9 it takes H1 from –19.85 ($p = 0.0302$) to **–1.27** ($p = 0.3294$), H1a from –233.20 ($p = 0.0168$) to **–15.22** ($p = 0.2599$), and H1c from –0.0121 ($p = 0.434$) to **–0.0140** ($p = 0.4006$): opportunities and repeats move too, because episodes change epoch and arm. The registered sensitivity *without* these sessions drops all six: H1 –7.40 [–16.32; +1.65], H1a –51.12 [–115.53; +22.23], H1c –0.0109 [–0.0425; +0.0138] (p 0.1957, 0.1501 and 0.4571; §5).

Adjudication. 1 195 episodes, 395 in stratum A (census of `is_error`) and 800 in stratum B (hash-ordered sample), were submitted to three model families from distinct training lineages, as locked in PREREG §682. **Deviation (rc11): a sample where the registration locked a census.** PREREG §3 required census adjudication of the live study (“*Live-study adjudication volume and panel — LOCKED 2026-07-30: census, API-only panel*”). The reported analysis instead uses a census of the `is_error` stratum and a sample of 800 of the 5 556 remaining corpus episodes, weighted by 6.945. This departs from the deposited sampling plan; the six-switch reanalysis does not restore the census. Its bootstrap conditions on the realized adjudication sample (§4.1.2). An attempt to restore the census, made after unblinding, stopped at its gates and was not run (rc12; reported after the tie rule below). ~~Coverage 100%.~~ All 1 195 were submitted to the three-family panel; 1 159 (97.0%) met the three-substantive-verdict requirement and 36 remained **unknown**. The registered analysis applies the substitution rule of §6 item 7 (DEVIATIONS §10.31): the fourth family’s verdict is added only to those 36, and the three-verdict floor is applied again. That adjudicates 20 more episodes, 13 of them as failures, for 1 179 adjudicated and 16 **unknown**; no episode the three families had adjudicated changes verdict (checked: 0). The sensitivity analysis uses the three families alone. The unknown share of opportunities is 1.00% (1.34% in the sensitivity analysis), so the registered 10% missing-data rule does not fire. Horvitz-Thompson weight for stratum B: 6.945 (5 556 / 800). Strict majority; an exact tie ($n/2$ failures, possible only with an even number n of substantive verdicts) resolves to **not_failure**, which `pilot_replay.py` 1.145-148 declares as “*conservative: it underestimates failures*”. In the primary panel the rule cannot bind: an episode needs at least three substantive verdicts to be adjudicated (`carregar_verdicts`), and with three families three cannot tie. Nor can it bind in the registered substitution panel: the fourth verdict is added only where the three families left fewer than three substantive verdicts, so an adjudicated episode there has exactly three (counted: 0 ties). It binds in the four-vote sensitivity set, where a 2-2 tie is possible. There it is directional bias toward fewer failures, not indeterminacy; its effect on the treatment-control contrast depends on the arms the ties fall in, which ~~we have not counted~~ we have now counted (`out/C12-EMPATES-POR-BRACO-2026-10-05.json`, rc4). In the registered analysis of rc10, which also drops the washout and counts every session at its start, 4 of the 28 ties fall in the analysed exposure of treatment epochs and 7 in that of control; 9 lie in the washout, 7 outside the trial dates and one before 09-03’s exposure window; one tie belongs to a session that crosses an epoch boundary (`_sprint-2026-10-04/B-rc15/checks-rc15.json`, block C). Resolving all 28 as **failure** moves the four-vote H1c point difference there from –0.0140, equal to the registered estimate, to –0.0249. Under the registered window and epoch set of rc8 (`out/C12-EMPATES-REGISTRADO-2026-10-05.json`) the four-vote set’s 28 exact ties fall 10 in treatment epochs and 10 in control; 7 lie outside the

trial dates and one, a control tie in 09-03, before that epoch’s exposure window; 09-02 holds none. The tied H1c opportunities are the same 4 (treatment) and 7 (control). The rc4 count below was made under the window of the sensitivity analysis. The four-vote set has 28 exact ties: 10 in treatment epochs, 11 in control and 7 outside the window. Of the 21 in the window, 4 (treatment) and 7 (control) are H1c opportunities, and none of those 11 was a failure under the three-family majority, so relative to the three-family labels the rule removes no repeated failure from either arm. Nor does it move any signature’s first failure, which leaves condition (i) unchanged. Only 3 of the 28 ties change a label (failure to **not_failure**): one in each arm, neither of them an opportunity, and one outside the window. Relative to the opposite resolution (tie $\rightarrow$ **failure**) the rule is not neutral: it holds the 4 treatment and 7 control tied opportunities, HT weights 27.78 and 36.73, at **not_failure**. Resolving all 28 ties as **failure** moves the four-vote H1c point difference, under the registered window and epoch set, from -0.0121 (equal to the registered estimate, since no tie changes a label among its opportunities) to -0.0211 (`out/C12-EMPATES-REGISTRADO-2026-10-05.json`; point only, no interval; opportunities do not change). Under the sensitivity analysis’s window it moves from -0.0208 to -0.0266 (`out/C12-EMPATES-COMO-FAILURE-2026-10-05.json`, rerunning `estimador_itt.py`; the rerun and the first-order calculation agree, -0.0266 ; the sixth decimal differs, -0.026571 against -0.026566 , because the rerun’s difference is taken from the artifact’s per-arm proportions, one of which is stored rounded, 0.09189). The rule pulls the four-vote contrast toward zero.

The census restoration we attempted, and why it was not run (rc12). On 2026-10-05, after unblinding (every estimate in this paper was already known), we tried to restore the PREREG §3 census by adjudicating the 4 756 stratum-B episodes outside the sample (5 556 $-$ 800) with the September panel, prompt and settings. Because the attempt followed unblinding, its pass and fail rules were written before any call (`_sprint-2026-10-04/B-censo/GATE-RULES-PREDECLARED.md`): the file was created and last modified at 23:13:58Z, and the first call started at 23:14:49Z. The file is not versioned. The time written in the file’s own header, 2026-10-05T23:15Z, and repeated in `GATES.md`, is inconsistent with the file-system timestamp and, read literally, falls after the first call; it is not evidence of the order, which rests on the file-system creation time alone, and a checkout rewrites that time. (`GATES.md` also dates gate 1 from 23:14:52Z, the start of the second call.) Gate 1, model identity, passed: each of the four providers served the same `model_served` id as in September. An id is an alias, however, and September recorded no fingerprint or snapshot, so a change of weights under the same name cannot be ruled out. Gate 2, test-retest, failed. The first 50 ids of the stratum-B sample, in file order, were re-adjudicated by the three families with the September command. Per-provider label agreement was 50/50 for `xai` (Wilson 95% [0.93; 1.00]), 47/50 for `google` ([0.84; 0.98]) and 46/50 for `zhipu` ([0.81; 0.97]). The declared criterion was 0.99, borrowed from the instrument’s only measured test-retest (99/100, Wilson [0.9455; 0.9982], with a panelist outside this panel; `STABILITY-TEST.md` §7). The rules declared in advance that this criterion, a point comparison that all 150 pairs of the three families must meet, can fail by chance, and that each provider’s Wilson interval would be reported for that reason. If the 150 agreement indicators were independent Bernoulli draws with agreement probability 0.99, the all-agree criterion would fail with probability $1 - 0.99^{150} \approx 0.779$. The `google` and `zhipu` marginal Wilson intervals exclude 0.99 under the binomial model, providing evidence against that agreement probability on these episodes. The comparison does not establish a change in model weights or a deterioration from September reliability, which was not measured for these providers. The declared point criteria nevertheless fail. The panel outcome under `carregar_verdicts` ($\tau = S1$, three families) agreed on 46/50 (Wilson [0.81; 0.97]), which meets its own declared criterion of 0.90; of its four changes, two are **failure** $\leftrightarrow$ **not_failure** flips in opposite directions and two involve **unknown**. The census was therefore not run, and the sampled design stays a declared deviation. Running it anyway would have combined October adjudications of the 4 756 episodes with September adjudications of the other 1 195 despite observed test-retest disagreement; whether the instrument’s response distribution changed is unresolved. The gates cost about US\$1.18 (154 calls, computed from the returned `usage` at list prices, an upper bound) (`_sprint-2026-10-04/B-censo/GATES.md`, `_sprint-2026-10-04/B-censo/gate2-compare.json`).

The combined estimator, written out. The first draft omitted it, and without it a reader cannot tell a ratio-of-totals from a mean-of-ratios. Per arm, over epochs in the window and episodes past washout satisfying the implemented proxy for condition (i) of the registered definition of a repeated failure. PREREG §4.1 requires that the serving snapshot at session start contained a failure episode of the same signature, at severity $\geq \tau$, written at least one epoch length before the epoch. The estimator does not check snapshot

membership, since the historical snapshots are pruned by design; it requires that the adjudicated corpus contain a failure (severity $\geq \tau$) of the same signature whose epoch begins at least one epoch length before the episode’s epoch (`estimador_itt.py`, first failure per signature; `pilot_replay.py` declares this reconstruction an approximation):

```
opportunities =  $\Sigma w(e)$                                 repeats =  $\Sigma w(e) \cdot 1[\text{state} = \text{failure}]$ 
w(e) = 1                                                  if e  $\in$  stratum A (census)
H1c = repeats / opportunities                            w(e) = 6.945      if e  $\in$  stratum B (sampled)
                                                         e skipped       if in neither
```

unknown verdicts count in the denominator and not in the numerator, per §5 of the spec. It is a ratio of weighted totals, not an average of per-epoch ratios.

Uncertainty. Cluster bootstrap with the epoch as the resampling unit, resampled within arm, 10 000 replicates, seed 20260921 declared. The registration locks BCa with the acceleration from the leave-one-epoch-out jackknife (PREREG §5). The analysis uses an arm-stratified BCa bootstrap, with acceleration calculated from arm-specific leave-one-epoch-out jackknife influences, centred within arm and scaled by $(n_g - 1)/n_g$, with n_g the arm’s epoch count. The registration says the jackknife runs over all 234 epochs; only 19 exist, so it runs over them. **Correction (rc15).** Up to rc14 the acceleration pooled the 19 jackknife estimates and centred them at one overall mean, the one-sample formula, while the bootstrap resamples the 11 treatment and 8 control epochs independently, which needs the multi-sample acceleration now used; on the same jackknife values it equals SciPy’s multi-sample BCa acceleration (`_sprint-2026-10-04/B-registered/teste-aceleracao-v4.json`). The draws, the bias correction, the quantile rule and the rounding are unchanged. Most BCa bounds move (H1a’s registered upper bound, for one, from +314.65 to +310.82; H1c’s registered lower bound does not move); no interval changes whether it contains zero (`_sprint-2026-10-04/B-rc15/checks-rc15.json`, blocks V and R; `_sprint-2026-10-04/B-registered/RESULTADO-v4.md`), and no p-value or point estimate changes (`out/ITT-REGISTRADO-v4-2026-10-05.json`, field `control_v3_registrado`). BCa never fell back to percentile. Its acceleration for H1 and H1a is positive and its upper adjusted quantiles extreme (99.51% and 99.56%), so their upper bounds are the 50th and the 45th largest of 10 000 replicates: they rest on the tail of the bootstrap distribution, and their Monte-Carlo error was not measured (in rc9, with the one-sample acceleration, it was the lower bounds, the 6th smallest). The sensitivity analysis uses the percentile construction, which the registration allows only as a fallback when the acceleration is undefined; earlier versions used it without that condition.

4.0.1 Two registered analyses we did not run, declared as deviations

Both were found by adversarial review of this manuscript, not by us.

4.0.1a The registered inference test is re-randomization, not bootstrap, and we have now run it

§4 of the analysis spec locks 10 000 re-randomizations, “*redesign the 234, never permute within the 20*”, over the trend-residualized outcome. The first draft reported cluster bootstrap and did not declare the substitution. That was the worse half of the error: the bootstrap assumes an iid draw of epochs, whereas the design randomizes with stratification and exact controlled rounding, so its reference distribution is not the one the design generates.

`rerandomizacao.py` imports `assign_arms.assign` and redesigns all 234 epochs per replicate, restricting to the window; it never permutes labels among the 20, which the spec pre-commits against because the 20 fall entirely in the first calendar half, where the stratification collapses. Control: 300 distinct arm patterns in 300 replicates, which reproduces the spec’s own measurement of zero collisions; at 10 000 it is 9 941 distinct over the 20 epochs of the sensitivity analysis and 9 903 over the 19 of the registered analysis, which runs the same redesigns. The results are in §4.0.2.

4.0.1b The pre-committed instrument controls: all three run

§5 of the spec defines three. The first draft reported none, which for a paper whose stated contribution is “the reportable results are about instruments” was the worst omission in it.

Caveat. The spec measured the first two on 2026-09-10, with the trial still running: 6/6 and 3/3 over a partial window. We re-measured over all 19 served epochs (09-02 served none), because a control measured midway does not cover what came after.

Semantics, declared before the numbers: `mexeu` compares `ids_tratado` with `ids_controle` by membership (`set`), never as a list. List comparison mixes reordering with entry/exit and at epoch 09-08 returns 48 against 20.

control	statement	result
positive	a served treatment epoch has <code>mexeu > 0</code>	passed, 11 / 11 (was 6/6 at midpoint)
negative dual	a control epoch has <code>sem_ids == n</code>	passed, 8 / 8 (was 3/3)
specificity (sham)	replay 19 non-designated items at the same <code>w</code>	~~running since 2026-09-22 01:40Z~~ ~~not executed~~ — the configuration launched was invalid; a valid one is runnable and not yet run (§4.0.1c) passed, rank 1 of 21 : real designation 132 changed states against 81–122 for 20 shams, $p = 1/21 = 0.0476$ (the floor at $K = 20$); restricted to the 2,646 brief states of the <code>w = 4</code> epochs and to shams drawn from the 36 boostable items; over the 11,812 reconstructible states of the hash-verified trial window, with <code>w = 4</code> forced in every state (a counterfactual outside the <code>w = 4</code> epochs), rank 1 of 21 again, 790 against 449–583 (§4.0.1c)

The positive control ranges from 2.83% to 6.85% of briefs altered per treatment epoch (09-09 lowest, 09-14 highest), counted over all 672 briefs of each epoch. On 09-01 that count includes the 42 briefs served in shadow mode, before the dose went active, two of which changed only as counterfactuals; over its 630 active briefs, 22 changed (3.49%). Over all treatment epochs, 335 of the 7 350 briefs served in active mode changed (4.56%); the artifact’s 337 of 7 392 includes the two shadow changes. Had any treatment epoch returned 0, the null would be the instrument’s and not the effect’s. That is the whole purpose of the control, and it is the reason the null of §4.1 can be read as being about the treatment at all.

The negative dual is stated that way for a reason the spec works out and we repeat: the obvious form, “a control epoch has `mexeu == 0`”, is invalid, because at `w=0` the `ids_*` fields do not exist at all, so `mexeu == 0` is indistinguishable from “the field was never written”. Such a predicate cannot detect missing data, because it needs that data to be evaluated.

Our first attempt at the specificity control was invalid. We counted how many briefs contain a sham chunk in `ids_tratado` and compared against the designated set. It returned real 2 069 against a sham median of 5 832 and read as a failed instrument. The failure was the test’s. `ids_tratado` is post-dose, so counting presence inside it is the candidate the spec’s own table already labels *tautological*; with a universe of 141 ids, the non-designated set includes the chunks that enter nearly every brief, while the designated are one per signature group and therefore rarer. The quantity measured is chunk frequency, not specificity.

The pre-committed sham is a replay: re-execute the dose mechanism with 19 non-designated items at the same `w` and compare the churn it produces. That requires running the serving code, not reading its log, because the log only contains the outcome of the designation that actually ran. It was declared as not run rather than reported as failed; the near-miss above shows why that difference matters.

~~**It is now running** (launched 2026-09-22 01:40Z, `measurement/roda-sham.sh`). 21 runs of `replay-oportunidade.mjs --modo dose at w = 4` — one with the real designation as baseline, 20 with sham designations — sequential,~~

~~nice -n 19 -i nice -c3~~, because each run exceeds 15 min and the host serves production on 2 vCPU. Expected wall clock ≈ 5 h.

Correction (2026-10-04). The paragraph struck above was stale when it was written into this draft and wrong about what the run would measure. The run produced no output and was aborted; the configuration it used was invalid for the reasons measured in §4.0.1c. The three design choices below are kept because two of them remain right, and the third is the one that let the error through.

Three choices in that design, each with its reason:

- **K = 20 is the minimum, not a round number.** With 21 runs the smallest attainable rank p-value is $1/21 = 4.8\%$; with fewer shams the test cannot reject at 5% no matter what it finds. It is a randomization p-value only to the extent that the shams are drawn by the same mechanism as the real designation, which neither configuration of §4.0.1c does: the real designation is one item per signature group, while the shams are drawn at random from a pool and mapped onto the signature keys arbitrarily (`gera-shams.py`; the corrected `sprint-gera-shams-v2.py` matches severity and bonus mass, not signature group).
- **The shams are drawn from the eligible pool, not from the corpus.** A sham chunk that fails the eligibility predicate never reaches the candidate list, so it would “not move” by construction, and the test would report specificity where there was only ineligibility. That is the mirror image of the tautological candidate the spec rejects, and it is why `gera-shams.py` applies the 30-day window as well as the path and floor predicates. *(2026-10-04: the principle stands; the pool it was applied to did not. See §4.0.1c: `gera-shams.py` drew from the 115-chunk pool of a corpus that did not serve the trial; the served pool has 108. And eligibility is not enough: only 36 of the 89 non-designated items in the served pool can receive a bonus at all, so the run of §4.0.1c draws its shams from those 36.)*
- **The predicate has an independent control.** Our SQL with the age window returns 115 eligible chunks, and the replay independently *observes* pool: 115 at the same `t-ref`. Two routes, one number; without that agreement we would not know whether the pool we sample from is the pool the mechanism sees. *(2026-10-04: both routes read the same corpus, `corpus-preservado-20260908.db`, so their agreement could not detect that the corpus itself was the wrong one. On the corpus that served the trial the pool is 108 in 22/22 measured states; see §4.0.1c.)*

One limitation, declared. The corpus of the published anchor (`e20260826T060003Z.db`) no longer exists; it was pruned. ~~This runs on `corpus-preservado-20260908.db`, the deliberately preserved corpus of the trial, so it cannot reproduce the published anchor and does not try to: the comparison is internal, real against sham on **one** corpus, which is what specificity needs.~~

Correction (2026-10-04). The struck sentence is factually wrong on two counts. `corpus-preservado-20260908.db` is not the corpus that served the trial: it was preserved from the live database on 2026-09-08, after 60 `memory/lessons.md` chunks had been re-ingested on 2026-09-07, and production never served them (DEVIATIONS §10.10, §10.14; §4.0.1c). The corpus that served is the one the serving process held open from 2026-09-03, recovered from its file descriptor as `corpus-SERVING-REAL-e20260903-recuperado.db`. And “one corpus” does not rescue the comparison when, on that corpus, the real arm cannot move at any dose (§4.0.1c).

Artifact: `CONTROLES-JANELA-COMPLETA-2026-09-21.json`.

4.0.1c The specificity control: invalid as first configured, then run

What was launched, and what it measured. The design was 21 replays (`measurement/roda-sham.sh`): the real designation as baseline and 20 sham designations of 19 non-designated items each, at the same `w`. It was launched on 2026-09-22 at 01:40Z. The first 12 runs each hit our own 3 600 s time limit (`exit 124`) and wrote no output, and the loop was aborted at 13:45Z (project log `docs/HANDOFF.md` at the repository root, entry of 2026-09-22). `exit 124` is our timeout, not a verdict: **the launched sham measured nothing**.

The probes, and the reading we gave them. Two probes then replayed the real designation alone, at `w = 4` and `w = 100 000`: one state (`sonda2`), and the **132** states in which production recorded `ids_tratado`

$\neq$ `ids_controle (sonda3)`. Both returned `churn = 0` in every state at both doses (`out/NOGO-replay-sonda{2,3}-2026-09-23.json`). `sonda3`'s replayed *control* set matched production's in **0 of 132** states, and none of production's fresh-slot ids appeared in any of them. Our log of 2026-09-23 read this as a test without power (the real arm is 0, no sham can exceed it, so $p = \#\{\text{sham} \geq \text{real}\} / 21 = 21/21 = 1.0$ "by construction") and named, as the remaining explanation, a hypothesis it declared untested: that the serve-state the replay derives today loses the ties in `last_served` that production saw, leaving strata of one, where an intra-stratum bonus never decides. **That reading made a property of the configuration look like a property of the design.** The tie hypothesis is refuted, and $p = 1.0$ would have followed from the corpus and the cut, not from the replay or from the sham design.

The cause, measured on 2026-10-04. Two defects: the first alone yields exactly zero at any dose, the second alone nearly so (1/22 at $w = 100\ 000$):

1. **The wrong corpus (the main cause).** `sonda3` and `roda-sham.sh` both read `corpus-preservado-20260908.db`. It holds 60 `memory/lessons.md` chunks re-ingested on 2026-09-07 that production never had: its maximum id is 309 497, against 308 752 in the corpus that served, and the serving log contains no id above 308 752 before 2026-09-21T09:52:04Z. That the serving process stayed on the fd-pinned corpus through 2026-09-20 is an **inference** from that log and from DEVIATIONS §10.14, not a direct observation. In the coverage order (`last_served` IS NULL first, then `last_served` ASC, ties broken by `salience + boost; freshSlots = 2`), those 60 never-served chunks form the boundary stratum in **22/22** measured states, with **no** designated item in it; the first designated item sits at pool position 60–66. The bonus breaks ties only inside a stratum, so no w reaches a designated item there.
2. **The inclusive cut (an aggravating defect).** `--corte inclusivo (served_at <= T)` puts the brief's own serves into its own serve-state. Even on the served corpus it reproduces production's control in only 2/22 states and moves 1/22 at $w = 100\ 000$. The exact cut is `rowid (id < the brief's first row)`.

The control of the diagnosis. The same, unmodified `replay-oportunidade.mjs`, on the 22 states for which the trial's `brief_log` survives off the production host:

corpus	serve-state cut	control / churn / entering id = production	moved at $w = 2$	moved at $w = 100\ 000$
preserved	inclusive (<i>the sham's configuration</i>)	0 / 0 / 0 of 22	0/22	0/22
preserved	rowid	0 / 0 / 0 of 22	0/22	0/22
served	inclusive	2 / 1 / 0 of 22	0/22	1/22
served	rowid	22 / 22 / 22 of 22	17/22	22/22

With the served corpus and the exact cut the replay **reproduces production exactly** and **responds to dose**. The stratum structure of production's state (served corpus, `rowid`) also refutes the tie hypothesis directly: the pool is 108 chunks with none never-served; the boundary stratum has a median of 4 members (range 3–4) and holds 1–3 designated items in 22/22 states; production's entering chunk ties with its leaving chunk in 22/22; and the median number of singleton strata per state is 0. Ties were not lost. The inclusive cut thins them (enter/leave tie in 8/22), which is the only part of the hypothesis with any truth, and it is secondary.

The sham pool was wrong too. `gera-shams.py` draws shams from the 115-chunk pool of the preserved corpus (it asserts `len(cand) == 115`). Minus the 19 designated, that leaves 96 candidates, **60** of which are the never-served chunks that, on that corpus, form the only stratum where a bonus decides. A sham of 19 drawn from those 96 is expected to contain 11.9 of them, and the probability that a sham contains none is 1.5×10^{-10} (hypergeometric arithmetic on these counts). On that corpus, then, shams could move and the real designation could not: running the 21 replays would most likely have returned "`sham  $\geq$  real`", which reads as **failed specificity** and would have been the corpus's failure. This is a **prediction** from the measured strata and that arithmetic; no sham designation was ever replayed.

What a valid sham needs, and how each need was met. (i) The served corpus, `corpus-SERVING-REAL-e20260903-recuperado.db`; (ii) the `rowid` cut; (iii) shams redrawn from the served pool (108 chunks,

89 non-designated); (iv) a `brief_log` covering the trial through 2026-09-20. The first three are in hand. The fourth is not: the copy we can work on off-host ends at 2026-09-08 05:52:04, and the trial's `brief_log` after that exists only on the production host. Exporting it is **pending authorization**. At about one minute of wall clock per 22 states, cost is not the obstacle. ~~The valid sham is runnable; it has not been run.~~ (*rc3: all four were met by 2026-10-04.*) For (iv) we worked on an export of the trial's `brief_log`; where it overlaps the `brief_log` held in the served and in the preserved corpus, its rows are hash-equal to theirs (B-sham-v2/REPORT.md §1.2). (iii) needed a further correction, given with the result below. On the corpus: twelve daily readings (2026-09-10 to 09-21) of the sha256 of the file the serving process held open, taken through its file descriptor, return one value, equal to the served copy. A restart would have opened a newer epoch, and a file descriptor cannot return to a deleted file, so with the same file measured on 09-08 and the process start dated to 2026-09-03 17:23:30Z (DEVIATIONS §10.10), every state from then to 09-21 was served by that file (B-sham-v2/REPORT.md §1.1). This turns the **inference** stated in defect 1 above into daily measurements plus an argument for the gaps between them.

What was measured and what was not.

	states	status
fidelity and dose response, served corpus + rowid	22 of 132	measured — all timestamped between 2026-09-01T11:07Z and 2026-09-02T08:37Z, i.e. inside the single treatment epoch 09-01 (09:00 UTC boundary)
the same, for the states timestamped 2026-09-12 to 09-16	110 of 132	~~not measured~~ — needs (iv). For these states only two facts are measured: the probe's control mismatch (0/132) and that production never served the chunks the probe put in the fresh slots. That the dose would also respond there is an expectation measured (rc3) : control set, churn, entering id and boost count equal production in 110/110; at <code>w = 4</code> the dose moves 110/110, as in production (B-sham-v2/fidelity-110/)
every brief state of the four <code>w = 4</code> epochs (09-01, 09-12, 09-14, 09-15), real designation	2,646	measured (rc3) : control set equal to production in 2,646/2,646 (630/630 at 09-01, 672/672 in each of the others); churn and entering id at <code>w = 4</code> equal in 2,646/2,646; 132 states moved, production's 132
every reconstructible brief state of the trial window (18 epochs from 2026-09-03T17:23:39Z; 09-01 and 53 states excluded), real designation	11,812	measured (rc23) : control set equal to production in 11,812/11,812; churn and entering id at <code>w = 4</code> equal in the 2,016 states of 09-12, 09-14 and 09-15
a sham designation replayed	0 20	~~not run run (rc3)~~ : 20 of 20 validated, 2,646 states each, 0 errors; again over the window (rc23) : 20 of 20 validated, 11,812 states each, 0 errors
byte-identity of the serving <code>dist/</code> used with the build that served	—	not measured ; it was rebuilt from source (2026-09-29 for the 22-state test, again for the sham run, and again for the whole-window run, whose three pinned files came out byte-identical to the previous build) whose <code>src/api/brief.ts</code> matches the probe's recorded hash; the 22/22, 110/110, 2,646/2,646 and 11,812/11,812 exact reproductions are the functional check

The sham, run (2026-10-04). Twenty-one replays (`measurement/sprint-roda-sham-v2.sh`): the real

designation and $K = 20$ sham designations of 19 non-designated items each, on the served corpus with the `rowid` cut, at $w = 4$ and, as positive control, at $w = 100\,000$, over the same 2,646 brief states. The statistic is the number of states whose served brief differs from the undosed one (`mexeu`, by membership), and total churn.

statistic, $w = 4$	real	20 shams (min-max)	shams $\geq$ real	$p = (1 + \#\{\text{sham} \geq \text{real}\}) / 21$
states moved (<code>mexeu</code>)	132	81–122	0	0.0476
total churn	146	84–132	0	0.0476

At $w = 100\,000$ the real designation moves 155 states and the shams 81–122. The real run is identical in its dose detail to the earlier calibration run of the real designation, and it reproduces production in 2,646/2,646 states (table above). The replay is deterministic; three sham runs were slowed by CPU throttling on the host, which changed their wall time and not their output.

What it shows, and what it does not. No sham reached the real designation on either statistic, so the rank p-value is $1/21 = 0.0476$, the floor at $K = 20$: it says the real designation exceeded every sham, and at this K it cannot say more. It is evidence that the dose effect on **what was served** is specific to the designated items rather than the background churn that **these 20** severity- and bonus-matched designations produce at $w = 4$; it does not extend to every designation of that severity and bonus mass. Nor is the rank p-value a calibrated randomization p-value: the shams are matched on severity and bonus mass but not drawn by the real designation’s rule (one item per signature group), so real and sham are not exchangeable draws of one mechanism (third limit below). The specificity is relative: the matched shams also move 81–122 states, so most of the real designation’s churn is what a comparable designation produces, and the real designation exceeds the largest sham by 10 states. It says nothing about the trial’s outcomes; whether serving the designated items changed repeated failures is H1c (§4.1), a different question. Four declared limits:

- **The state set is restricted to the $w = 4$ epochs.** This is a deviation from the configured instrument (`roda-sham.sh`, which replayed the whole serving log; the configuration was ours, not pre-registered). The reasons: $w = 4$ is the registered “same w ”; the epochs were randomized, so selecting them does not depend on the outcome; the corpus that served is verified in every one of these states and production can be checked directly against them; and the configured whole-log set (19,567 states, about 8,000 of them from the shadow phase, whose corpus is not identifiable) would have taken about 60 h, and the hash-verified trial window (11,865 states) about 36 h as estimated then, against a projected 8 h for this set. A replay over the whole trial window was then run (`job-janela2`, relaunched 2026-10-05T09:39:53Z, runner start 2026-10-05T09:40:32Z, completed 2026-10-07T02:09:46Z): 11,812 reconstructible states of 18 epochs, which excludes epoch 09–01, whose corpus has no hash proof, and 53 states whose serve-state cut cannot be reconstructed ($11,865 - 53$). Its result follows these limits.
- **The shams are drawn from the 36 boostable items, not from all 89 non-designated.** The serving code gives a bonus only to ids in the designation, real or sham, that have a `p2_verdict` row of severity S1–S4 written at least one day before the epoch; 53 of the 89 non-designated pool items have none and receive a bonus of zero by code. Shams drawn from all 89 (generated, not run) carry 21–50% of the real designation’s bonus mass, which would bias the test toward “real > sham”. The 20 shams are drawn from the 36, stratified as the real designation is (10 S1, 9 S2), so each carries the same bonus mass, $0.301 \cdot w$. Drawn from 36, the shams overlap: 6–14 shared items per pair (mean 10.4), none shared with the real designation.
- **The match is on severity and bonus mass, not on signature group** (§4.0.1b, $K = 20$ bullet), so the rank p-value is not a calibrated randomization p-value.
- **p is at its floor.** With $K = 20$ no result could give a smaller p, and a single sham at or above the real designation would have moved it to $2/21 = 0.095$.

The whole-window sham, run (2026-10-05 to 2026-10-07). The same 21 designations, the same instrument (28 of its 29 pinned hashes equal to the run above; the 29th is the state list) and the same two databases, over every reconstructible brief state of the hash-verified trial window: 11,812 states of 18 epochs, from the first brief after the serving process opened the `e20260903` corpus (2026-09-03T17:23:39Z) to the end

of the trial. It excludes epoch 09-01, whose corpus has no hash proof, and 53 states (52 on 2026-09-21, one on 2026-09-06, all of one agent) whose brief cannot be located in the `brief_log` by agent, second and control ids, so that no `rowid` cut exists for them under any designation; the exclusion is a property of the data, fixed when the first attempt stopped on them, and does not depend on the outcome. The dose is replayed at $w = 4$ in every state, whatever arm production served there. The statistic, the p rule and the positive control are those above, fixed before the job ran.

statistic, $w = 4$, 11,812 states	real	20 shams (min-max)	shams $\geq$ real	$p = (1 + \#\{\text{sham} \geq \text{real}\}) / 21$
states moved (mexeu)	790	449-583	0	0.0476
total churn	868	484-612	0	0.0476

No sham equals the real designation on either statistic; the only ties are between shams. At $w = 100\,000$, the positive control, the real designation moves 894 states and the shams 462-591. The real run reproduces production's control set in 11,812/11,812 states and, in the epochs where production served $w = 4$ (09-12, 09-14, 09-15), its churn and entering id in 2,016/2,016. Those 2,016 states are also in the run above, and their records are identical in all 21 runs (4,032/4,032 per run, both doses), and the real run's records for the 200 states of the calibration sample are identical to that sample (400/400): the replay is deterministic, and this run is not independent of the first, it extends it to the other 15 epochs and leaves out 09-01. In 235 states of every run no bonus is emitted; all of them fall after the designation's expiry at 2026-09-20 22:51:23 (§3.0.1).

What it adds. At $w = 4$ the real designation exceeds every one of the 20 matched shams over the whole window as well, by 207 states over the largest (790 against 583), and the shams' 449-583 is the range produced by these 20 matched designations at $w = 4$ in these states. Outside 09-12, 09-14 and 09-15 production did not serve $w = 4$ (control epochs served no dose, the others $w = 2$ or $w = 7.5$), so there the replay is a counterfactual: it measures what $w = 4$ would have changed in those states, not what was served. It lifts the first limit above, the restriction to the $w = 4$ epochs, for that counterfactual question only, and as far as the hash-verified window allows: 09-01, the 53 states and the shadow phase are still out. The evidence about what was served remains the 2,646 states of the $w = 4$ epochs. The other three limits are unchanged (shams from the 36 boostable items; a match on severity and bonus mass, not on signature group; p at its floor), and it says nothing about outcomes.

Our error, stated. DEVIATIONS §10.14-§10.15 had measured this mechanism on 2026-09-09: on the recovered corpus `mexeu(2) = 20` and `mexeu(105) = 37`, matching production's `churn > 0 = 20`; on the realigned corpus `0 / 0`; cause, a never-served cohort holding the `freshSlots`. The sham was nevertheless configured on a corpus of the family §10.15 shows to be blind, and the HANDOFF hypothesis of 2026-09-23 reopened a question §10.15 had already answered.

What this does to the claims of this paper. Nothing in §4.1-4.7 rests on the sham. The positive control (11/11) establishes that the mechanism changed briefs in every treatment epoch; ~~until a valid sham is run, nothing here establishes that the change is specific to the designated items rather than background churn at the same w~~ the sham establishes, on the $w = 4$ epochs and again over the whole hash-verified trial window, that the change is larger for the designated items than for any of 20 matched sham designations at the same w ; on the $w = 4$ epochs that change is the one production served, and over the rest of the window it is the change $w = 4$ would have made. Neither control bears on whether the change reached outcomes. The null of §4.1 stands as reported; the two controls make it a null about a lever that changed 2.83-6.85% of briefs per treatment epoch (§4.0.1b) and that, in the $w = 4$ epochs, changed more briefs through the designated items than any of 20 matched sham designations would have.

Artifacts: `out/NOGO-replay-sonda{2,3}-2026-09-23.json`; `_sprint-2026-10-04/B-replay-fidelity.md`, its eight runs and stratum files in `_sprint-2026-10-04/B-replay-fidelity/` (summary `RE-SUM0.txt`); `measurement/sprint-replay-estratos.mjs`; `measurement/sprint-replay-fidelidade-resumo.py`. For the sham run: `_sprint-2026-10-04/B-sham-v2/REPORT.md` (§7 holds the result);

B-sham-v2/job-v2b/RESUMO.json (every number of the sham table and the positive control); B-sham-v2/job-v2b-runs.tgz (the 21 runs) with B-sham-v2/job-v2b/RUNS.sha256; B-sham-v2/fidelity-110/ (the 110 states); the generator, runner, test and summarizer `measurement/sprint-gera-shams-v2.py`, `sprint-roda-sham-v2.sh`, `sprint-testa-roda-sham-v2.sh`, `sprint-resume-sham-v2.py`. For the whole-window sham: `_sprint-2026-10-04/B-sham-v2/JANELA-LANCAMENTO.md` (launch record, exclusions and result); B-sham-v2/job-janela2/RESUMO.json (every number of its table and the positive control); B-sham-v2/job-janela2-runs.tgz (the 21 runs) with B-sham-v2/job-janela2/RUNS.sha256; B-sham-v2/job-janela2/DETERMINISMO.json, written by `measurement/sprint-determinismo-janela2.py` (the shared states).

4.0.2 The registered test against the bootstrap, and the Holm rule on H1a

outcome	re-randomization p , unadjusted (registered analysis; sensitivity)	rejects under the registered decision rule?	what the bootstrap said (BCa; percentile in the sensitivity analysis)
H1c (primary in this report)	0.4006 ; 0.1603	no (primary, $\alpha = 0.05$; in the Holm family, adjusted 1.0)	contains zero in both — agree
H1a	0.2599 ; 0.0855	no : decided in the Holm family, adjusted 1.0 ($m = 5$ and $m = 4$); 0.43 or 0.34 in the sensitivity analysis	contains zero in the registered analysis — agree ; excluded zero in the sensitivity analysis — disagree
H1 (the deposited primary)	0.3294 ; 0.0127	registered: no under either reading (alone at $\alpha = 0.05$; Holm-adjusted 1.0 under the switch). Sensitivity analysis: yes under the deposited reading only (Holm-adjusted 0.0762 with $m = 6$, 0.0635 with $m = 5$)	contains zero in the registered analysis, excluded zero in the sensitivity analysis — agree in both

H1a’s unadjusted p -value falls below 0.05 in the registered analysis and did not in the sensitivity analysis; the offsets of the partial epochs and the washout’s removal from the denominator move it (§4). It is still not a rejection. The registration puts H1a in the Holm family with H1b, H1c and the two components of H2, and under Holm the first threshold is 0.01 (0.0125 without H1b); 0.0168 exceeds both, so the step-down stops before H1a and nothing in the family rejects. Reading 0.0168 against 0.05 would apply a decision rule the registration does not contain. (*rc10: the struck paragraph described rc9’s analysis, in which H1a’s unadjusted p -value was 0.0168. With sessions counted at their start it is 0.2599, and nothing in the family comes near a threshold. In rc9’s analysis H1a was not rejected either: the registration decides it in the Holm family, where 0.0168 exceeds the first threshold, 0.01.*)

Different estimand, stated so the numbers are not compared naively. The permutation statistic is a difference of arm means over per-epoch outcomes residualized on study-day; the ITT of §4.1 is a ratio of weighted totals. They answer related questions, not the same one, and the observed statistic here (-0.0174 for H1c; -0.0323 in the sensitivity analysis) is not the -0.0140 of §4.1 (-0.0199). What transfers is the verdict, not the magnitude, and the registered scope is narrow by construction: this tests the sharp null of *zero total effect*, and rejection alone does not attribute magnitude.

What this settles, and it is not in our favour rhetorically. §4.2 argued that H1a “does not bear weight” from the sparse-session mechanism. The registered decision rule reaches the same verdict by a route that does not need that argument, and it is the inference we weight (§8.3). In the sensitivity analysis the disagreement between test and interval is what an under-covering interval would produce, and it is also what two different estimands and a multiplicity rule can produce; this does not demonstrate any of them. In the registered analysis there is no disagreement left to explain.

H1 does not reject in the registered analysis under either reading ($p = 0.3294$, tested alone; Holm-adjusted 1.0), nor on the registered sensitivity without the sessions that cross an epoch boundary ($p = 0.1957$; §5). It rejected under the deposited reading in the sensitivity analysis (0.0127) and in the analyses of rc8 and

rc9 (0.0133 and 0.0302), and under the switch’s Holm rule in none (§1). Those versions reported it as an unexplained rejection, conditional on the deposited reading, of a magnitude the dose’s reach did not account for. The earlier rejection disappears when sessions are attributed to their start epoch. That correction changes both exposure and outcome attribution; it establishes that the rejection was not robust to the registered rule, not that the dose had no effect (§4.3).

Artifacts: `out/ITT-REGISTRADO-v4-2026-10-05.json` (field `multiplicidade` holds the deposited and the switch readings, each with both Holm family sizes, for the registered analysis, the registered sensitivity without boundary-crossing sessions and the analyses of rc8 and rc9 rebuilt by switching later switches off); `RERANDOMIZACAO-2026-09-21.json` for the sensitivity analysis; seed prefix `p2-rerand-2026-09-21` declared for both.

4.1 H1c: the primary outcome, null and not detectable at 80% power

leg	treatment	control	difference	95% CI	
locked (the 19 epochs of the analysis set)	0.0663	0.0803	−0.0140	[−0.0464; +0.0110]	contains zero
pre-committed sensitivity —	0.0684	0.0896	−0.0212	[−0.0623; +0.0073]	contains zero
all partials removed					
post-hoc sensitivity —	—	—	−0.0182	[−0.0509; +0.0073]	contains zero
without 09–14					
registered	0.0694	0.0803	−0.0109	[−0.0425; +0.0138]	contains zero
sensitivity —					
without the sessions that cross an epoch boundary					
<i>rc9’s analysis</i>	0.0675	0.0796	−0.0121	[−0.0445; +0.0125]	contains zero
(sessions split by epoch) — locked					
<i>sensitivity analysis</i> (rc7: 20 epochs, three-family panel, no window cut, percentile)	0.0696	0.0896	−0.0199	[−0.0560; +0.0086]	contains zero
— locked					
<i>sensitivity analysis</i> —	0.0721	0.0994	−0.0273	[−0.0725; +0.0061]	contains zero
pre-committed					
sensitivity					
<i>sensitivity analysis</i> —	—	—	−0.0244	[−0.0619; +0.0051]	contains zero
without 09–14					

The first four rows are the registered analysis, with BCa intervals (`out/ITT-REGISTRADO-v4-2026-10-05.json` for the locked leg and the sensitivity without boundary-crossing sessions; the two other sensitivity legs run through the same estimator in `_sprint-2026-10-04/B-rc15/checks-rc15.json`, block A, which first reproduces the artifact’s locked leg exactly). H1c does not divide by session-hours, but counting each session at its start moves episodes between epochs and arms, so its rows changed in rc10; the fifth row is rc9’s locked leg, which rc8 also reported. The pre-committed leg removes the partials 09–01, 09–03 and 09–20 from the 19, leaving 10 treatment and 6 control epochs.

Which sensitivity is the registered one, and our error in the first draft. §9.1 of the analysis spec pre-commits “the standard sensitivity that removes *all* the partials as a block”: 09–01, 09–03, 09–20. The first draft of this paper reported instead a sensitivity that removes 09–14, which was chosen after seeing that

one epoch dominates exposure and was never pre-committed. Both are now reported, the registered one first and labelled as such. We state the direction: the pre-committed leg is *more* favourable to our reading than the post-hoc one we had picked, so the substitution gained us nothing. That makes it a process failure rather than a self-serving one, and still a failure. (*rc8: the direction held for the percentile intervals of the analysis now reported as a sensitivity. Under the registered BCa interval it reverses for H1a, where the pre-committed leg excludes zero and only the post-hoc leg contains it (§4.2); so for H1a the leg we had picked after seeing the data was the more favourable one.*) (*rc9: with the washout out of the denominator, H1a excludes zero on both legs, so for H1a neither leg is more favourable to our reading; §4.2.*) (*rc10: with sessions counted at their start, H1a contains zero on both legs; §4.2.*)

This is the stable outcome of the study: all three legs contain zero in the registered analysis, so does the registered sensitivity without boundary-crossing sessions, and all three legs contain zero in the sensitivity analysis.

4.1.1 Our two uncertainty statements rest on different constructions

§3 of the analysis spec pre-commits this sentence:

“Under the realized inclusion criterion and $ICC = 0.0985$, not even total elimination of repeated failures is detectable at 80% power.”

Control sits at $H1c = 0.0803$ in the registered analysis, so “total elimination” is an effect of -0.0803 , and the observed interval excludes it by 0.0339 (0.0351 in *rc9*’s analysis; 0.0896 , -0.0896 and 0.0335 in the sensitivity analysis, computed from the unrounded bound -0.056038 and control rate 0.089552 ; from the rounded values shown the distance is 0.0336). That alone is not a contradiction: a power statement is about the probability of rejection (80%), not about every realized interval, and a single realized interval can exclude an effect that the design detects with less than 80% probability.

The contradiction is in the standard error each estimate implies. A minimum detectable effect at 80% power and two-sided 5% implies a standard error of $MDE / (z_{0.975} + z_{0.8}) = MDE / 2.80$. The power artifact (~~out/H1C-POWER-REALIZADO-2026-09-10.json~~, with its own $p0 = 0.0782$) finds even $p1 = 0$ undetectable, so its MDE is at least 0.0782 and the standard error it implies is at least 0.0279 (0.0320 if total elimination is taken at the realized control rate, 0.0896). The observed interval spans 0.0646 , which under the same normal approximation implies a standard error of 0.0165 , 59% of the smaller bound. Both are ours; they are incompatible estimates of the same standard error.

Correction (rc7). The struck paragraph compared two numbers that do not measure the same thing. The planning power approximation and the realized bootstrap interval use different variance constructions and assumptions: `measurement/potencia-h1c.py` uses one variance term under the null and another under the alternative, so $MDE / 2.80$ does not identify a single standard error, and the width of a percentile interval divided by 3.92 is not the bootstrap’s standard deviation. Their difference does not establish a contradiction, and it does not quantify any under-coverage of the bootstrap. The power conclusion remains conditional on its planning inputs, and the bootstrap’s coverage has not been calibrated for this design. The bounds 0.0279 and 0.0320 and the 59% are withdrawn as inferential claims.

We name the known limits of each:

statement	known limit
MDE saturated at 100%	the $ICC = 0.0985$ comes from the pre-registration and was estimated for density per session-hour , not for proportion per opportunity. The spec flags this itself and computes the margin: the verdict flips on an 11.5% drop in ICC (5.6% under whole-epoch counting)

statement	known limit
95% CI of [-0.0464; +0.0110]	a BCa cluster bootstrap on 11 and 8 clusters (the sensitivity analysis: percentile, on 11 and 9) may under-cover; the few-cluster literature [cameron2008bootstrap; webb2014reworking] motivates caution but does not calibrate this ratio estimator. Separately, the resampling does not explicitly propagate the stratum-B sampling (§4.1.2); some of that variation enters through between-epoch differences, and the resulting coverage error, including its direction and magnitude, has not been calibrated

Both limits were known to us before this section existed. Our reading is that neither statement is a calibrated measure of this study’s uncertainty, and that the null in H1c (non-rejection by the registered re-randomization test, $p = 0.4006$) does not rest on the interval; the point estimate, a 17% relative reduction (22% in the sensitivity analysis), is not distinguishable from zero at this N. A reader should treat every interval in this paper as indicative of sign and order of magnitude, not as calibrated coverage.

The same limit disciplines §4.3: an interval “excluding zero” on H1 or H1a, as in the sensitivity analysis and in earlier versions, is a claim made with the same machinery, whose coverage has not been established; that is a further reason, independent of the denominator argument, not to read those as findings. The limit applies to every interval here, including the null one, and not only where it is convenient. The first draft attached it only to H1; an adversarial reviewer pointed out that a method cannot be optimistic selectively.

The under-powering is structural, not bad luck. The analysis spec counts partial epochs by fractional exposure (its §3) and forbids rounding a partial epoch to a whole one (its §7), and that counting is the power statement of this paper. Under it, and under the planning assumptions, no reduction from $p_0 = 0.0782$, including total elimination, reaches 80% power: the effective size is 84.8 over the 19 served epochs (`out/H1C-POWER-FRACIONARIA-2026-09-10.json`), which are the registered analysis set, against a critical 95.26. At the 20 epochs of the sensitivity analysis, with 09-02 added as a full ninth control epoch (10.9375 treatment, 8.23375 control), it is 91.49, and there the verdict flips on a 4.1% drop in ICC. The statement is about reductions only: `mde()` in `measurement/potencia-h1c.py` searches p_1 between 0 and p_0 , while the registered test is two-sided, and an increase can reach 80% power (at the effective size of 84.8, $p_1 = 0.25$ gives about 86%). The fractional counts take 09-01 as $630/672 = 0.9375$ of an epoch, its share of briefs served in active mode, while the spec’s clock exposure for it is $22.38 \text{ h} / 24 = 0.9325$ (its §1); 09-03 enters by its brief share ($441/672$) and 09-20 by clock ($13.86 \text{ h} / 24$). Taking 09-01 by clock moves the effective size from 84.80 to 84.78 and changes no verdict. Whole-epoch counting, recomputed with the unchanged `measurement/potencia-h1c.py`, is reported as a sensitivity. With the 19 clusters the power calculation counted (11 treatment, 8 control; `out/H1C-POWER-REALIZADO-2026-09-10.json`) it agrees: effective size 90.2, which the script reproduces from the artifact’s own inputs. Those 19 clusters are the registered analysis set, so under the registered analysis both countings give the same verdict. At the sensitivity analysis’s 11 treatment and 9 control the verdict flips under whole-epoch counting: the effective size is 96.41 against 95.26, so total elimination ($p_1 = 0$) becomes detectable at 80% (relative MDE 0.996). The spec computed that case itself (96.4) and set it aside because 09-02 served no brief, which is also why the registered analysis excludes 09-02 (§3). The spec’s statement that no defensible inclusion criterion inverts the verdict therefore holds at the registered analysis set under both countings, and fails only under whole-epoch counting at the sensitivity analysis’s allocation, where the design is borderline for total elimination and reaches 80% power only for reductions of about 99.6% or more. Inputs and outputs: `_sprint-2026-10-04/REVIEW-B-rc2-evidence/power-*.json`. Reporting the interval without that sentence would be the exact defect family this project spent eight weeks documenting; the spec makes the declaration mandatory in the abstract, not in a footnote, and that is where it is.

4.1.2 What the bootstrap does not resample

There are three layers of randomness the estimate inherits: the assignment of epochs to arms, the hash-ordered sampling of 800 episodes from 5 556 in stratum B, and the panel’s verdicts themselves. The cluster bootstrap resamples observed epochs within each arm, holding arm membership, arm sizes, the adjudication

sample and its labels fixed. It does not redraw the constrained assignment; the separate re-randomization test does that. The component $E[\text{Var_sampling}(B) \mid \text{epochs}]$ may therefore be under-represented in every interval that uses stratum B (H1, H1a, H1c; not the correlation of §4.6), since the frozen sample enters only through the between-epoch spread. A sampled stratum-B failure carries weight 6.945; its contribution to estimator variance has not been quantified. The finite-population correction is ≈ 0.856 , not zero.

We also treat a systematic (hash-ordered) sample as if it were simple random. If the hash order correlates with time, session or agent, the bootstrap does not correct for it.

The third was measured in rc12 (§4, §7): re-adjudicating 50 stratum-B episodes with the September settings changed 4 panel outcomes, two of them `failure` $\leftrightarrow$ `not_failure` in opposite directions, and a flip of a sampled stratum-B episode moves any weighted count it enters (failures, and repeats where the episode is one) by 6.945. Repeat-adjudication uncertainty is not explicitly propagated; some realized variation may enter through between-epoch differences, and its contribution to interval coverage has not been quantified. Its direction on the contrast is not known from 50 episodes.

The bootstrap does not explicitly propagate adjudication-sampling or repeat-adjudication uncertainty. Some realized variation enters through between-epoch differences, but the resulting coverage error, including its direction and magnitude, has not been calibrated; nor is the direction of the hash-order effect known. Combined with §4.1.1, the intervals reported here may be optimistic, and the primary result is a non-rejection ($p = 0.4006$, §4.0.2) that wider intervals would not overturn.

4.2 H1a: contains zero once sessions are counted at their start, is not rejected, and bears no weight

Correction (2026-10-04). This section was titled “*H1a — inverts its conclusion on one epoch*” and its table showed only the post-hoc leg that removes 09–14. That broke the rule §4.1 states for itself (the pre-committed sensitivity is reported first), and it made the inversion look specific to one epoch when it is not. Under the pre-committed leg (§9.1 of the spec: the three partials removed, 09–14 kept), the H1a interval already contains zero while the point estimate barely moves. What turns on 09–14 is the magnitude; the exclusion of zero is a property of the locked leg alone.

Correction (rc8). The struck sentences describe the percentile intervals of the analysis now reported as a sensitivity. Under the registered analysis the BCa interval excludes zero on the locked leg and on the pre-committed leg, and contains zero only on the post-hoc leg that removes 09–14; the previous title, “excludes zero only on the locked leg”, does not hold for it. What does hold is that the exclusion of zero rests on 09–14’s exposure and that the registered decision rule does not reject H1a: its unadjusted re-randomization p-value is 0.0241, and under the Holm rule it is adjusted to 0.1205 (0.0964 without H1b), which does not reject (§4.0.2).

Correction (rc9). The rc8 correction, struck above, describes the registered analysis with the washout still in the session-hour denominator. With the washout removed, as the registration requires (§2), the BCa interval of H1a excludes zero on all three legs, the post-hoc leg without 09–14 included: -97.85 [-162.77 ; -26.43]. The exclusion of zero therefore no longer rests on 09–14. It does rest on long sessions: the registered sensitivity that drops the four sessions with episodes in more than one epoch, which we compute only as a diagnostic (§5), contains zero, -51.12 [-115.60 ; $+22.07$]. What has not changed is the decision: the registered rule does not reject H1a (unadjusted $p = 0.0168$, Holm-adjusted 0.0840, or 0.0672 without H1b; §4.0.2).

Correction (rc10). The rc9 correction, struck above, describes an analysis that still split every session across the epochs it touched. Counted in the epoch of its start, as the registration requires (§2), H1a contains zero on all three registered legs, and on the registered sensitivity without the sessions that cross an epoch boundary. Every interval of H1a that excluded zero, in every version, was computed with a few long sessions cut into pieces and spread over the epochs, mostly of the treatment arm; counting them at their start moves their exposure and the epoch and arm of their episodes, and no interval excludes zero after it. The decision has not changed in any version: the registered rule does not reject H1a (unadjusted $p = 0.2599$, Holm-adjusted 1.0; §4.0.2).

	locked (the 19 epochs)	pre-committed sensitivity — partials 09-01, 09-03, 09-20 removed	post-hoc sensitivity — without 09-14
session-hours, treatment / control	324.95 / 57.06	324.64 / 56.35	324.49 / 57.06
H1a (opportunities/h), registered (BCa)	−15.22 · CI [−373.70; +310.82] · contains zero	−10.51 · CI [−362.55; +315.50] · contains zero	−15.76 · CI [−375.88; +295.33] · contains zero
H1a, registered sensitivity without boundary-crossing sessions (BCa; hours 3.52 / 3.06)	−51.12 · CI [−115.53; +22.23] · contains zero	—	—
H1a, rc9’s analysis (sessions split by epoch, BCa; hours 10.75 / 3.25)	−233.20 · CI [−351.70; −79.81] · excludes zero	−219.26 · CI [−348.56; −54.93] · excludes zero	−97.85 · CI [−162.77; −26.28] · excludes zero
H1a, rc8’s analysis (washout in the denominator too, BCa)	−163.52 · CI [−281.49; −61.02] · excludes zero	−138.91 · CI [−254.18; −28.88] · excludes zero	−78.27 · CI [−146.19; +2.11] · contains zero
H1a, sensitivity analysis (percentile, 20 epochs)	−143.92 · CI [−209.80; −15.34] · excludes zero	−133.34 · CI [−203.62; +7.41] · contains zero	−63.04 · CI [−120.45; +1.76] · contains zero

Sources: the registered locked leg and the sensitivity without boundary-crossing sessions from `out/ITT-REGISTRADO-v4-2026-10-05.json`, the two other registered legs from `_sprint-2026-10-04/B-rc15/checks-rc15.json` (block A); the rc9 and rc8 rows, recomputed with the arm-stratified acceleration, from `_sprint-2026-10-04/B-rc15/checks-rc15.json` (block R, which first reproduces, under the acceleration those versions used, the intervals they reported: `out/ITT-REGISTRADO-v2-2026-10-05.json` and `_sprint-2026-10-04/B-rc9/checks-rc9.json` for rc9, `out/ITT-REGISTRADO-2026-10-05.json` and `_sprint-2026-10-04/B-rc8/checks-rc8.json` for rc8); the sensitivity analysis’s pre-committed column from `ITT-SENSIB-PRECOMPROMETIDA.json` (field `sensibilidad`), the same artifact that holds its pre-committed row of §4.1 (session-hours 12.16 / 4.08 there; the post-hoc leg’s 5.57 / 5.16 is in `ITT-2026-09-21.json`). The registered decision rule does not reject H1a (unadjusted $p = 0.2599$, Holm-adjusted 1.0), and neither did the sensitivity analysis’s test ($p = 0.0855$; §4.0.2). The registered BCa upper bound of H1a is the 45th largest of 10 000 replicates (§4, *Uncertainty*), so it rests on the tail of the bootstrap distribution; its Monte-Carlo error was not measured.

The denominator measures idleness. For each eligible session, `span_por_sessao` computes `max((max(ts) - min(ts))/3600, 1/60)` hours: the observed event span with a one-minute minimum, the distance between first and last event, not time worked. A session that acts, sleeps, and returns six hours later contributes six hours of exposure.

Counted at their start, three sessions carry almost all the exposure of the trial. One session (`d37a5964...`), started in 2026-09-08 (treatment, $w = 2$), has 60 episodes over 11 epochs and spans 240.42 h: 09-08 holds 240.75 h, 74% of all treatment exposure. Two more make 09-12 (treatment, 81.37 h, one session of 67 episodes spanning 81.01 h) and 09-11 (control, 54.31 h, one session of two episodes 54.00 h apart). The other sixteen epochs fall between 0.23 and 0.62 h. Split across the epochs it touched, as up to rc9, the same session put 6.33 h into 09-14 (three episodes, one at 13:52 and two at 20:12) and smaller pieces elsewhere, and 09-14 then held 63% of treatment exposure; the sessions that actually worked in that epoch produced 56, 23 and 21 episodes after the washout, within 1.1 to 8.0 minutes each (with the washout in, 74, 65 and 56 episodes in ~~9 to 10 minutes~~ 8 to 10 minutes each (*corrected 2026-10-04: measured spans 9.71, 8.86 and 8.03 min, recomputed with the pilot’s own per-session grouping for Figure B1*); the first two lie inside the washout). Either way the quantity is the same: the span of a session that was mostly idle. Attributed to its start, it inflates one treatment epoch instead of several, and H1 and H1a lose their contrast; split, it inflated the treatment arm’s denominator less and in more places, and their intervals excluded zero.

The artifact is not introduced by this analysis. The function is the pilot’s, so the pilot’s own `hours_per_epoch` (1.1144) carries the same property. It is a feature of the registered definition that only manifests when an epoch contains a sparse session.

We did not change the denominator. Replacing it with “effective work” would be altering a locked definition *after* seeing that it yields an uncomfortable result, and would invalidate $\hat{r}$, the ICC and the `N` that `sizing.py` derived from them.

But that cuts both ways, and we did not say so in the first draft. If the exposure measure counts idleness, then $\hat{\mathbf{r}}$, the ICC and the sample size derived from that same measure inherit the defect. The power calculation that tells us this study is under-powered was computed on a denominator we are now calling artifactual. We do not know the direction: a denominator inflated by sparse sessions could have made the pilot's event rate look lower than it is, which would have over-sized the study, or the reverse. What we can state is that the under-powering claim of §4.1 rests on a quantity this section undermines, and that resolving it requires recomputing the pilot, which is future work, not a footnote. It does not rescue the trial either way: the realized N is 19 clusters regardless of what N *should* have been. The sensitivity is reported alongside and the divergence is not adjudicated in favour of either leg; the rule that says so was pre-committed in §2 of the analysis spec. The line the reader must take away is ~~the sensitivity one: a conclusion that turns on one sparse session is not a conclusion that both sensitivity legs contain zero (the registered one without touching 09-14, the post-hoc one by removing it) and that the registered test does not reject. We draw no conclusion from an interval that excludes zero on the locked leg only, on a denominator dominated by one sparse session: that the registered decision rule does not reject H1a, and that the only leg on which its interval contains zero is the one that removes 09-14. We draw no conclusion from intervals that exclude zero on a denominator dominated by one sparse session, under a construction whose coverage we have not established. (Corrected in rc8: the struck sentence held for the sensitivity analysis, where the pre-committed leg contains zero; under the registered BCa interval it does not.)~~ that the registered decision rule does not reject H1a. Its interval excludes zero on every registered leg and contains zero on the diagnostic leg without the sessions that span more than one epoch. We draw no conclusion from intervals that exclude zero on a denominator that long sessions dominate, under a construction whose coverage we have not established. (Corrected in rc9: with the washout in the denominator, the post-hoc leg without 09-14 contained zero; with it out, it does not.) that the registered decision rule does not reject H1a, and that its interval contains zero on every registered leg. The intervals that excluded zero in earlier versions did so on a denominator that a few idle sessions dominate, cut in a way the registration does not use, under a construction whose coverage we have not established. (Corrected in rc10.)

Figure B1. H1a: the intervals that excluded zero came from long sessions split across epochs; counted at their start, none does. (a) Session-hours per epoch, by designated arm, as the registered estimator computes them: for each session, the span between its first and last event with a one-minute minimum (`span_por_sessao`, imported from `pilot_replay.py`), counted in the epoch where the session started, inside the registered exposure windows and without the sessions that start inside the two-hour washout, over the 19 epochs of the analysis set (09-02 excluded). 2026-09-08 (treatment, $w = 2$) has 240.75 h, 74% of all treatment exposure (324.95 h over 11 epochs, against 57.06 h over 8 control epochs), and almost all of it is one session with 60 episodes spanning 240.42 h; 09-12 (81.37 h) and 09-11 (54.31 h) are one session each; the other sixteen epochs fall between 0.23 and 0.62 h. (b) The H1a difference, treatment – control in opportunities per session-hour, with its 95% cluster-bootstrap interval (epoch as the unit, 10 000 replicates, seed 20260921). Registered analysis (BCa): the locked leg, -15.22 [-373.70 ; $+310.82$], the pre-committed sensitivity, -10.51 [-362.55 ; $+315.50$], and the post-hoc sensitivity without 09-14, -15.76 [-375.88 ; $+295.33$], all contain zero, and so does the registered sensitivity without the sessions that cross an epoch boundary, -51.12 [-115.53 ; $+22.23$]. The fifth row is rc9's analysis, which split those sessions across epochs, -233.20 [-351.70 ; -79.81]; the sixth the sensitivity analysis's locked leg (percentile, 20 epochs, no window cut), -143.92 [-209.80 ; -15.34]; both exclude zero. Filled marker: the interval excludes zero; hollow marker: it contains zero. These intervals may under-cover; their coverage has not been established for this design (§4.1.1, §4.1.2). The registered re-randomization test, a different estimand (§4.0.2), gives an unadjusted $p = 0.2599$; H1a is decided in the Holm family, where it is adjusted to 1.0, and it is not rejected. Sources: panel (a) from `episodios-ensaio-20260921.jsonl`, `p2-serving.ndjson` (the windows) and `ASSIGNMENT-SERVING.json`; the script aborts unless the per-arm sums reproduce `horas_sessao` of the registered leg in `out/ITT-REGISTRADO-v4-2026-10-05.json` and every epoch reproduces `_sprint-2026-10-04/B-rc15/checks-rc15.json`. Panel (b) from those two files; the script aborts unless `checks-rc15.json` was produced from the same `ITT-REGISTRADO-v4` file (sha256), agrees with it on the registered leg, and the artifact is the six-switch analysis with the arm-stratified acceleration. Generated by `measurement/sprint-figB-h1a-inversao-registrado-v4.py` (rc15). The rc10 figure, `figures/figB1-h1a-inversao-registrado-v3.svg`, which plots the one-sample acceleration, the rc9 figure, `figures/figB1-h1a-inversao-registrado-v2.svg`, and the rc8 figure, `figures/figB1-h1a-inversao-registrado.svg`, are kept unchanged. The figure of ear-

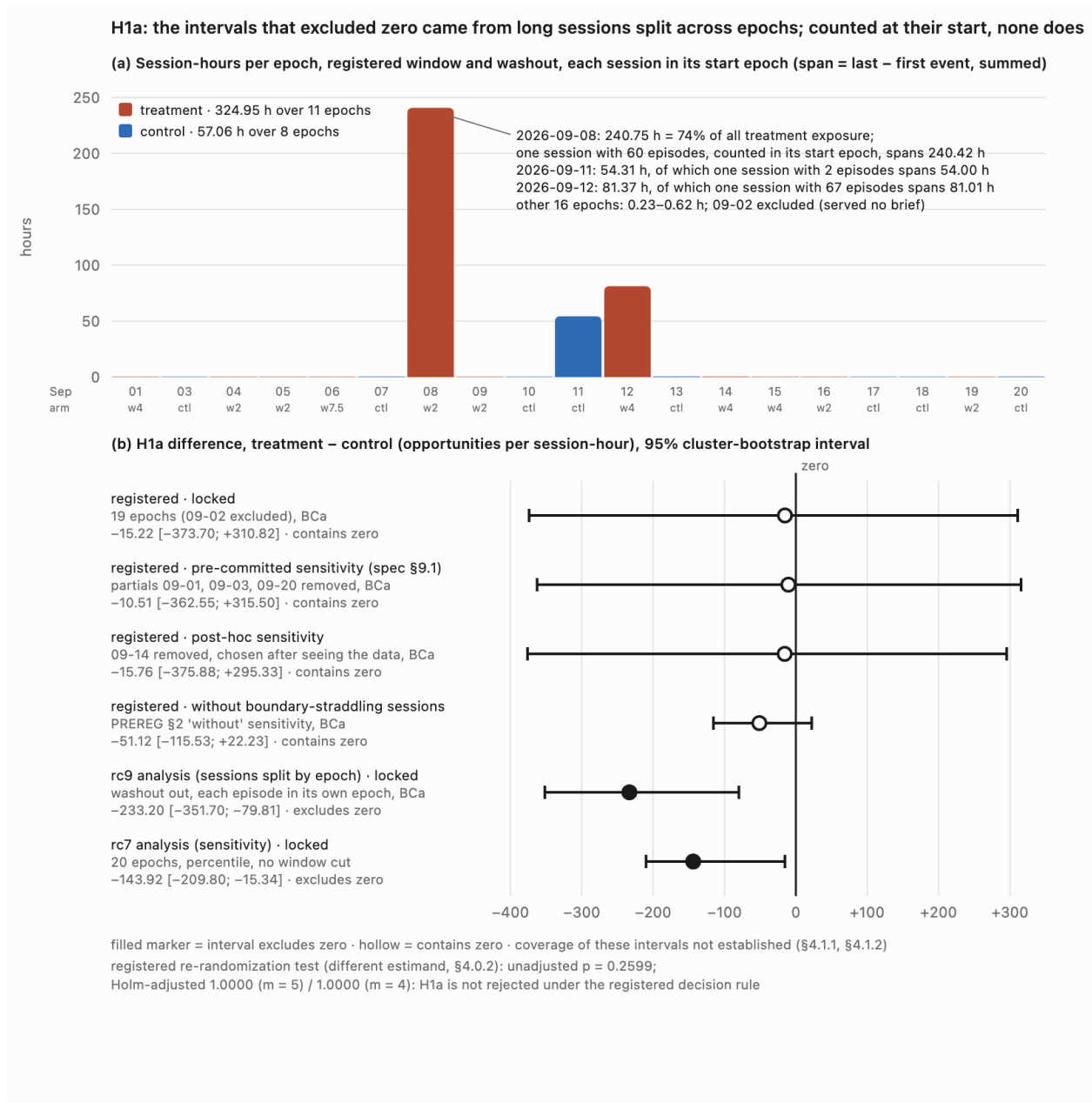


Figure 1: Figure B1

lier versions, `figures/figB1-h1a-inversao.svg` from `measurement/sprint-figB-h1a-inversao.py`, plots the sensitivity analysis and is kept unchanged as its record; its embedded headline (“*H1a turns on one epoch: the exposure denominator counts idle time*”) was never regenerated.

4.3 H1: the interval that excluded zero came from sessions split across epochs

H1 returns -1.27 , BCa CI $[-28.36; +23.15]$ per session-hour, containing zero on all three legs (pre-committed sensitivity -1.02 , CI $[-30.26; +24.02]$; post-hoc sensitivity without 09-14 -1.32 , CI $[-28.40; +18.79]$), and the registered test does not reject under either reading ($p = 0.3294$ alone; Holm-adjusted 1.0). On the registered sensitivity without the sessions that cross an epoch boundary it is -7.40 $[-16.32; +1.65]$, $p = 0.1957$ (§5).

H1 returns -19.85 , BCa CI $[-32.69; -9.89]$, excluding zero on all three legs (pre-committed sensitivity -21.30 , CI $[-34.94; -9.74]$; post-hoc sensitivity without 09-14 -11.63 , CI $[-20.31; -3.40]$), and the registered test rejects under the deposited reading ($p = 0.0302$; under the switch’s Holm rule it does not, §1). **Correction (rc10).** The struck sentence is rc9’s analysis, which split boundary-crossing sessions across epochs. The sensitivity analysis agrees with it on every leg: -14.62 $[-20.51; -4.79]$, -15.59 $[-22.43; -4.05]$ and -9.74 $[-16.21; -2.99]$, $p = 0.0127$, a rejection under the deposited reading. (2026-10-04: the pre-committed leg, from *ITT-SENSIB-PRECOMPROMETIDA.json*, was missing here for the same reason as in §4.2; adding it changes no conclusion of this section.) The rest of this section was written about that result, which earlier versions reported as an unexplained rejection under the deposited reading. We keep it, with the figures of rc9’s analysis, because it is the argument a reader of those versions saw, and because its conclusion, that H1 is not evidence of an effect of the dose, now has a plainer reason.

H1 is the product of the other two, so it is not an independent hypothesis:

$$H1 = \text{repeats} / \text{hours} = (\text{repeats} / \text{opportunities}) \times (\text{opportunities} / \text{hours}) = H1c \times H1a$$

Verified on the estimates, relative error $\leq 5.8 \times 10^{-6}$ in all six arm-by-leg cells of the sensitivity analysis (five distinct: 09-14 is a treatment epoch, so control is identical on the locked and post-hoc legs); the residue is the JSON rounding to six places, not a discrepancy. On the six cells of the registered analysis the relative error is at most 3.8×10^{-6} (`_sprint-2026-10-04/B-rc15/checks-rc15.json`, block A). This is an algebraic identity, so H1 carries no information that H1a and H1c do not already carry.

The split also falls exactly along the denominator. H1 and H1a both divide by session-hours; H1c does not: it is a ratio of two counts. In the analyses that excluded zero, the two quantities that did so were the two that divide by the measure §4.2 shows to be idleness, and the one quantity free of that denominator was null. Counting each session at its start, as registered, changes that denominator by a factor of about 30 in treatment and 18 in control, and removes both exclusions; it moves H1c, which has no denominator in hours, only from -0.0121 to -0.0140 .

What the argument did NOT do in rc9, and what replaced it. In rc9’s analysis H1 still excluded zero without 09-14, with hours per epoch nearly equal across arms (0.396 treatment against 0.406 control; 0.557 against 0.573 in the sensitivity analysis) and H1 at -11.63 (-9.74). We read the residual through volume: the control arm carries 137.7 opportunities and 10.96 repeats per epoch against 95.5 and 6.08 in treatment (132.3 and 11.85 against 93.4 and 6.08 in the sensitivity analysis), and offered two readings this design does not separate: the treatment reduces the volume of failure activity without moving the rate H1c measures, or the arms differ in baseline volume by chance (with 19 clusters and between-epoch volume spanning 73 to 234 episodes). Those readings still describe the volumes; with sessions at their start the registered analysis has 135.0 and 10.84 per control epoch against 109.5 and 7.26 per treatment epoch. What they no longer have to explain is an H1 interval that excludes zero.

The magnitude argument of earlier versions belongs to the same history. H1 was removed from primary on 2026-08-30, in a revision that was not deposited (§1), for requiring a 955% effect (DESIGN-REVISION-2026-08-30.md l.196, “*impossible by construction*”): the registered 30% MDE divided by the pre-trial calibration share of altered briefs, 3.14% (11 of 350 briefs at $w = 2$). In rc9’s analysis the observed reduction was 73.5% on the locked leg (7.149 against 27.003) and 43.1% without 09-14 (70.7% and 47.1% in the sensitivity analysis), 23.4 and 13.7 times total elimination at the calibration share (22.5 and 15.0), 16.1 and 9.4 times at the share

realized in the trial, 4.56% (15.5 and 10.3), and 10.7 and 6.3 times at the highest per-epoch share, 6.85% (10.3 and 6.9). These are proportional-dilution calculations, not causal bounds. In the registered analysis the point reduction is larger still, 83.8% (0.246 against 1.520), and its interval contains zero, so there is no rejection left to dilute.

We also owe a note on our own criterion. §4.2 discards H1a because the registered decision rule does not reject it, and H1 is discarded on the same ground: under neither reading does the registered test reject it. In earlier versions H1 was discarded on different grounds (the identity, the dilution) while its test rejected under the deposited reading; those grounds are stated above so a reader can weigh them without ours.

4.4 H1b: trivially 1.0, because two of our own locks collide

lock	date	text
Opportunity (§3)	2026-07-29	<i>“An executed action a ... for which the serving snapshot at session start contained ≥ 1 failure episode a_past with $\text{sig}(\mathbf{a_past}) = \text{sig}(\mathbf{a})$”</i>
H1b (§1)	2026-08-16	<i>“An opportunity yields a repeat attempt if the session emits at least one action whose signature equals that of a_past”</i>

These are different estimands. Under the July lock the opportunity is the action, and the action carries $\text{sig}(\mathbf{a_past})$ by definition $\Rightarrow$ H1b = 1.0 by construction, and the repeat-attempt set coincides with the opportunity set: the nesting remains, but H1b no longer provides the intended intermediate distinction. For H1b to have content, Opportunity would have to be a property of the *session*, which changes the denominator for the whole family, not just for H1b.

No later document resolves it: the analysis spec does not mention H1b once.

Correction of our own phrasing (2026-09-21, adversarial review). We first wrote that H1b is “unevaluable”. That conflates two things. Under the lock we kept, H1b is evaluable and equals 1.0: it is *trivial*, not unmeasurable, and no artifact computes it because none is needed. What is unanswerable is the question H1b was written to carry. Saying “unevaluable” made a definitional triviality sound like missing data, which is the more flattering of the two readings and the wrong one.

We kept the July lock (Opportunity is the action) because PREREG §420 records that $\hat{\mathbf{r}}$, $\hat{\mathbf{p}}_0$ and the ICC were all computed by replay under that model, and that the construction then locked was the one that “*keeps every locked number valid — the alternative constructions would have invalidated them*”. Adopting the session reading now would invalidate those three numbers and, by dependency, the `N_epochs` derived from them.

What is lost. H1b was the member of the family distinguishing “*the treatment makes the agent stop trying*” from “*makes it try and succeed*”; §1 of the pre-registration calls it “*the substantive one for this paper.*” That distinction is not reportable from this study. The loss is of mechanism, not of power: H1c was already declared under-powered at the realized N (§4.1.1), so nothing here worsens what was known about detectability.

The reason that goes in the record is a collision between two locks, unreconciled before the window closed, not “we did not measure it”, and not “it came out non-significant”.

4.5 Coverage by arm, and a pre-committed explanation it does not test

Coverage is post-randomization. PREREG §5 requires an ITT co-estimate over all post-washout epochs without coverage exclusion. That is the analysis reported here (§4.1–4.3); the registered coverage-filtered analysis was not computed (§5). Coverage is reported descriptively:

	treatment	control
briefs with ≥ 1 designated item	28.0% (2 068 / 7 392)	26.9% (1 385 / 5 145)

The treatment column counts all briefs of the treatment epochs, including the 42 that 09-01 served in shadow mode before the dose went active; over the 7 350 briefs served in active mode it is 2 057 / 7 350, still 28.0%.

Across 4 324 designated-serving occurrences, all 19 signatures were served, each at $\approx 5.3\%$: uniform.

~~This contradicts the projection the pre-commitment rests on.~~ DESIGN-REVISION-2026-08-30.md committed in advance that a null in H1c “*does not distinguish ‘the mechanism does not work’ from ‘the lessons it promotes are too generic’*”, on the grounds that 93.8% of coverage would concentrate in the bucket signature Bash|shell:outro. ~~The realized distribution is flat. The pre-committed alternative explanation is therefore weakened by the data, which is the opposite of what a pre-commitment usually does for the party that wrote it, and is why it is reported here rather than dropped.~~ **Correction (rc7).** The two numbers count different things. The 93.8% is 573 of the 611 covered action opportunities of the planning corpus, grouped by signature (out/CONCENTRATION-2026-08-30.json). The $\approx 5.3\%$ counts designated-chunk serving occurrences: `cobertura_e_m10.py` adds one per designated item in a served brief, without linking it to an action opportunity. Those occurrences were approximately uniform across the 19 items; because they are not the covered action opportunities the planning statistic counted by signature, the present measurement does not test that concentration projection, and it does not weaken the generic-lessons explanation. The pre-committed indistinction stands: the null in H1c does not distinguish a mechanism that does not work from lessons too generic to work.

Caveat. We record a near-miss: `boost_by_id` is not a coverage measure. It records boost *calculated* for every candidate, uniformly, and reading it as coverage would have produced a fabricated 139 650. Coverage comes from crossing the served id lists against the designation.

4.6 Exploratory arm–designated-item-presence correlation

These correlations measure designated-item presence among served briefs, not the session-level logging coverage registered for M10. PREREG §5 defines that coverage as the share of an epoch’s sessions with a `brief_log` row carrying a non-null `brief_id`, with the denominator taken from the session corpus. The registered M10 correlation and interval have not been computed, and the locked artifacts cannot compute them: `brief_log` is not among them, and the table has no session column (`serving-brief.ts` creates it with `chunk_id`, `scope`, `agent`, `served_at` and `brief_id`), so whether a session received a logged brief can only be decided by an agent-and-time matching rule that was never registered. The serving log we hold carries the same two keys and no session id. Its TOST is not evaluable at $K < 30$: the registered TOST at $|r| \leq 0.15$ requires $K \geq 30$ analyzed epochs; we have 19. §8 of the analysis spec declares it “**not evaluable at the realized K**”, which is *not* the same as “equivalence not established”. (*Renamed in rc9: earlier versions called the correlation below M10.*) What we computed is reported unconditionally:

leg	K	95% CI (epoch		
		r	bootstrap)	
primary — all	19	+0.1130	[−0.4773; +0.5088]	contains zero
without 09-20 (the 13.86 h partial)	18	−0.1871	[−0.5704; +0.3353]	contains zero
without 09-03 and 09-20	17	−0.1114	[−0.5114; +0.4518]	contains zero
without 09-14	18	+0.0837	[−0.5180; +0.4932]	contains zero

The 17-epoch leg keeps 09-01, the third partial. The pre-committed sensitivity, which removes all three partials as a block (§9.1 of the spec), was not computed for this correlation.

The sign flips. 09-20 is a control epoch with 14.6% coverage against $\approx 28\%$ everywhere else, low because the coverage denominator counts all 672 briefs of the epoch, of which 287 were served after the designation

expired and could not contain it (98 of the 385 before expiry, 25.5%): an artifact of the clock, not of the arm. That one epoch moves $\mathbf{r}$ from -0.19 to $+0.11$. All four legs contain zero with half-widths near 0.5: at $K = 19$ this correlation distinguishes nothing, and reporting only the primary leg would have sold a sign that belongs to a truncated epoch. Correlation with the dose rather than the binary arm: -0.0521 .

4.7 The single epoch at the top dose: chance, not truncation

Only one of the 20 designated epochs drew $\mathbf{w} = 7.5$, against 3.33 expected. The explanation that writes itself, *stopping at 20 of 234 biases against the top dose*, is mechanistic and plausible, and the measurement does not support it.

Measured over 2 000 seeds conditioned on the same realized dates, using the published assignment script:

leg	value	share of the -2.333 deficit
expected under the design ($20 \times 39/234$)	3.333	—
truncation — conditioning on the realized dates	+0.023	1.0% -1.0%
chance — from the conditional mean to the realized	-2.357	101.0%

(Corrected 2026-10-04: the truncation share was printed unsigned, so the two shares summed to 102%. With signs ($+0.0232$ and -2.3565 against a deficit of -2.3333 in *ITEM7-DOSE-TOP0-2026-09-21.json*), they are -1.0% and $+101.0\%$, which sum to 100%.)

Conditional mean 3.357, mode 3 (25.3%), $P(\mathbf{n} = 1) = 10.15\%$, $P(\mathbf{n} \leq 1) = 11.95\%$. The conditional mean for **control** is 9.991 against 9 realized, so truncation does not displace that either.

Two traceability notes (2026-10-04). (1) The **control** figure 9.991 is held by no artifact (it appears only in *DEVIATIONS-FOR-PAPER.md*) because *ITEM7-DOSE-TOP0-2026-09-21.json* stores only the $\mathbf{w} = 7.5$ distribution. An independent redraw of 20 000 designs with `assign_arms.assign` and a declared seed recipe (`_sprint-2026-10-04/figures/figB2-item7-crosscheck-20000.json`) gives 10.0124 ± 0.0146 for **control**, consistent with 9.991 within the Monte-Carlo error of a 2 000-seed run; it corroborates the number but does not make it traceable. (2) The seed recipe of the 2 000-seed run is not recorded in its artifact, unlike the re-randomization's `seed_prefix`; three guessed recipes did not reproduce its histogram, so the 20 000-seed run is a second measurement, not a reproduction. It gives $P(\mathbf{n} = 1) = 9.11\% \pm 0.20\%$ against 10.15% here, whose own Monte-Carlo s.e. at 2 000 seeds is about 0.7 pp: consistent, and a reminder that the 10.15% carries that much noise.

Figure B2. One epoch at the top dose is a tail draw, not a truncation effect. Distribution of $\mathbf{n}$, the number of the 20 realized epochs assigned $\mathbf{w} = 7.5$, over 2 000 designs re-drawn with the published `assign_arms.py` and conditioned on the same dates. The realized $\mathbf{n} = 1$ (highlighted) has probability 10.15%, and $P(\mathbf{n} \leq 1) = 11.95\%$. The mode is 3 (25.3%). The dashed line is the design expectation $20 \times 39/234 = 3.333$; the dotted line is the conditional mean, 3.357. The two lines are 0.023 apart, which is 0.7 Monte-Carlo standard errors (s.e. 0.035 over 2 000 seeds, computed from the histogram). Conditioning on the realized dates therefore moves the expectation by an amount whose sign this run does not resolve. The whole deficit of -2.333 is chance; the truncation shift is not distinguishable from zero (1% of the deficit, MC s.e. 1.5%). With $\mathbf{n} = 1$ at the top dose, there is no dose-response to read (analysis spec §7). Source: *ITEM7-DOSE-TOP0-2026-09-21.json*. The script aborts unless the histogram reproduces the artifact's summary fields, the design counts (39 of 234) come from `assign_arms.py`, and the realized count matches *ASSIGNMENT-SERVING.json*. Generated by `measurement/sprint-figB-sementes-dose-top0.py`.

Why truncation does not measurably bias here: stratification is (calendar half) $\times$ (weekday | weekend), and the first 20 days fall entirely inside `h1`, where the allocation is already balanced by construction. Conditioning moves the expectation by $+0.023$ — and in the direction opposite to the intuition, which is 0.7 Monte-Carlo standard errors from zero (s.e. 0.035 over 2 000 seeds); its sign is not resolved, and the independent 20 000-seed redraw gives -0.007 ± 0.011 . Truncation accounts for about 1% of the deficit in

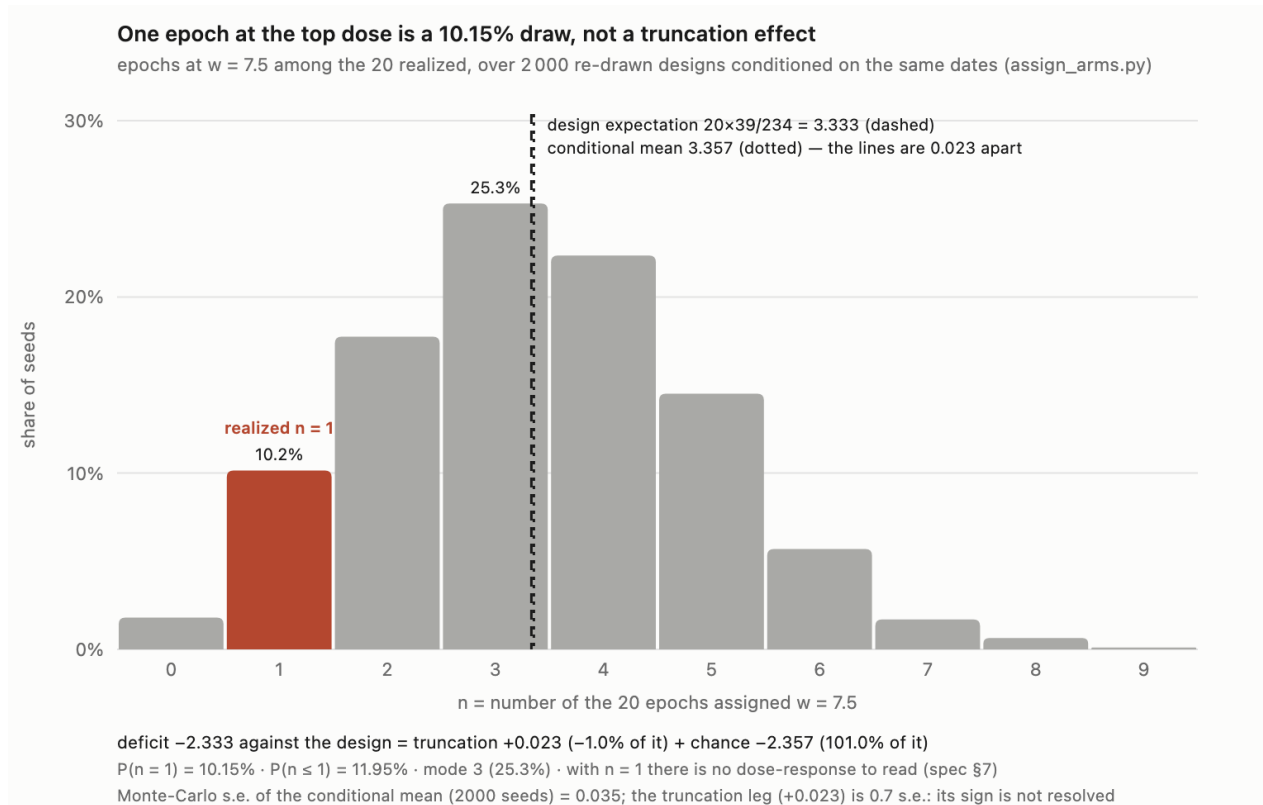


Figure 2: Figure B2

magnitude, in either direction. (*Corrected 2026-10-04: the first version read a direction into a difference inside the Monte-Carlo noise.*) This agrees with the spec’s earlier measurement ($23/300 = 7.67\%$, mode 3-4; the 95% interval of that proportion is $[4.7\%; 10.7\%]$ and contains 10.15%); what is new is the separation of the two causes, which the spec asserted without measuring.

The consequence for the paper is negative and is stated as such: with $n = 1$ at the top, there is no dose-response to read, and §7 of the analysis spec forbids reading one.

5. What is not evaluable, and why: declared, not omitted

registered rule	why not reportable	what stands in its place
H1b	evaluable and $= 1.0$ by construction, because the 2026-07-29 and 2026-08-16 locks collide; the mechanism question it carried is unanswerable (§4.4)	nothing; the mechanism question is not answered
M10 as registered: arm $\times$ session-level <code>brief_log</code> coverage, its interval, and the TOST at $ \text{tr} \leq 0.15$	not computed: needs a session $\leftrightarrow$ <code>brief_log</code> linkage the locked artifacts do not hold, and <code>brief_log</code> has no session column (§4.6); the TOST also requires $K \geq 30$, and $K = 19$	the exploratory designated-presence correlation of §4.6, a different quantity
the 95%-logging-coverage analysis set (epochs with <code>brief_log</code> coverage $\geq 95\%$, PREREG §5)	not computed, for the same reason	the ITT without coverage exclusion, which PREREG §5 makes mandatory, is what §4 reports

registered rule	why not reportable	what stands in its place
boundary-straddling sessions: assigned to the epoch of their start, flagged, with a with/without sensitivity; sessions longer than one epoch as their own stratum (PREREG §2)	~~not implemented as registered: the estimator assigns each episode to its own epoch. Four sessions in the window have episodes in more than one epoch; one of them makes most of 09-14's exposure (§4.2)~~ reportable, and reported from rc10 on: the registered analysis counts every session in the epoch of its start (§2, §4). Six sessions have episodes in more than one epoch; three start in the window, all three longer than one epoch (240.42, 81.01 and 54.00 h), and are listed in §4 as their own stratum; a fourth starts before the window	with (the registered analysis): H1 -1.27 [-28.36; +23.15], H1a -15.22 [-373.70; +310.82], H1c -0.0140 [-0.0464; +0.0110], p 0.3294, 0.2599, 0.4006. Without: H1 -7.40 [-16.32; +1.65], H1a -51.12 [-115.53; +22.23], H1c -0.0109 [-0.0425; +0.0138], p 0.1957, 0.1501, 0.4571. All contain zero and none rejects, under either multiplicity reading (out/ITT-REGISTRADO-v4-2026-10-05.json, leg sens_registrado_sem_sesoes_atravessadas)
dose-response reading rule (PREREG designation block; H3) H2: task regret, secondary confirmatory (two components, time and tokens, in the Holm family)	n = 1 at the top dose (§4.7)	~~per-dose effect with n on the same line, no gradient~~ not computed in this version the result below: neither component rejects
H3: retrieval metrics (nDCG@10, recall@10), exploratory	reportable, and reported from rc8 on: computed from the locked action archive with the locked instrument; three choices the registration does not fix are declared below not computable: nDCG@10 and recall@10 need graded relevance judgments per (query, item), and none were defined or collected for the served briefs, which are push context and answer no query; building a relevance label now would be a new estimator defined after unblinding	nothing
secondary model and co-estimates of PREREG §5 (the NB/binomial mixed model, the lag-1 and A→B co-estimates, the Appendix-B bounds)	not computed; they were outside the departures the registered re-analysis resolved (the six switches of §4)	nothing in this version
leave-one-agent-out	6 agents over 19 epochs	~~exploratory, no inference~~ not computed in this version
first-half / second-half contrast	window truncated inside the first half	not reported

These are listed before the discussion. And, except H1b (found after the close, §4.4), were listed in §8 of the analysis spec before the window closed, so that their absence cannot be read as omission. Four of them (the TOST, the dose-response rule, leave-one-agent-out and the half contrast) were listed in §8 of the analysis spec before the window closed, where the dose-response rule is labelled *H3*, which is not the registration's H3. H1b was found after the close (§4.4). H2 and H3 were listed nowhere before rc7: no version of this manuscript, of the analysis spec or of the deviation log mentions them, which is the omission this table exists to prevent, and we found it only because an adversarial reviewer did. The secondary model and co-estimates were not listed either, until rc8. Had a rule become unevaluable and simply been dropped, a reader could not tell it from a rule that was never there.

H2, computed (rc8). No regret pipeline had been run on trial data: `task_regret.py` was run once, in August, on the pilot corpus, to lock the winsorization points (p95: 7.45 s and 65 206 tokens, PREREG §4.2, locked 2026-08-15). The locked action archive pairs 5 951 of 5 951 `tool_use` / `tool_result` events onto the locked corpus by episode, with no orphan in either direction and no signature divergence; 37 signatures have a regret floor (at least five successful episodes) and 5 875 episodes carry regret. The registration does not fix three choices, which we declare, made without looking at the arms:

1. the “best known resolution” is the minimum over successful (**not is_error**) episodes of the same signature, with at least five of them, over the whole locked corpus (5 951 episodes, 2026-08-23 to 2026-09-21), both arms pooled;
2. the population is every post-washout action in the analysis epochs, not only the opportunities;
3. the effect is the difference of per-arm means pooled over episodes, with the same epoch bootstrap and BCa interval as H1; the sharp-null statistic is the per-epoch raw mean, trend-residualized like every permutation outcome of PREREG §5.

The winsorization applies to the estimator only; the test uses raw values, as locked.

component	treatment / control mean (winsorized)	difference, winsorized (95% BCa)	difference, raw (95% BCa)	re-randomization p (raw)
time (s)	1.795 / 1.478	+0.316 [+0.010; +0.689]	-1.18 [-5.51; +0.83]	0.8285
tokens	12 936 / 10 500	+2 437 [-1 066; +6 184]	+5 533 [+749; +12 308]	0.3372

Neither component rejects, alone or in the Holm family (adjusted 1.0); in the sensitivity analysis the raw p -values are 0.892 and 0.950. Two intervals exclude zero, both in the direction of a treatment that costs more: winsorized time (+0.32 s) and raw tokens (+5 533). Both exclude zero under the registered BCa construction only; the percentile interval contains zero for both (time [-0.011; +0.661], tokens [-390; +10 748]; field `ic95_percentil` of the same artifact). The winsorized-time lower bound, +0.009913 s, clears zero by less than 0.01 s, and its Monte-Carlo error was not measured. The registered test is the re-randomization test on the raw per-epoch means, and it does not reject either; the intervals carry the coverage caveats of §4.1.1. On time the winsorized and raw estimates have opposite signs (+0.32 s, treatment slower; -1.18 s): the raw tail is heavier in control. Source: `out/ITT-REGISTRADO-v4-2026-10-05.json`, field `H2`; the analysis covers 1 218 treatment and 1 089 control episodes in the 19 epochs. (*rc10: counting each session at its start changes the episode counts by +46 in treatment and -26 in control; in rc9's analysis, identical to rc8's, the table read time +0.284 [-0.030; +0.666], raw -1.13, $p = 0.831$; tokens +2 141 [-1 380; +6 230], raw +2 420 [-2 404; +7 760], $p = 0.684$, over 1 172 and 1 115 episodes.*) (*rc15: those are the intervals rc9 printed, with the one-sample acceleration; with the arm-stratified one the same analysis gives tokens [-1 375; +6 236], raw [-2 401; +7 770], and time unchanged at three decimals; out/ITT-REGISTRADO-v4-2026-10-05.json, field `H2.registrado_v2_equivalente`.*)

6. Instrument defects that changed a reported number

Consistent with Paper A's §6, defects that changed a number we had published are part of the contribution rather than an appendix.

1. **Ruler substitution (09-20):** delivery volume where the spec locks clock exposure; 16+3+1 published as 17+2+1, propagated to six places including a pushed commit (§3.1).
2. **Post-randomization conditioning on the arm:** mode of observed dose instead of designation; 0 divergences, but the premise was wrong (§3.1).
3. **"Zero inversions" is a tautology:** adding a vote to an odd panel under strict majority can only hold or tie, so zero reversals to the opposite strict majority is structurally guaranteed and a coin would achieve it. ~~Ties do change the final label under the tie rule; those are the 28 losses of item 6.~~ The 28 ties are the losses of item 6, and only 3 of them change the final label under the tie rule (failure to `not_failure`); the other 25 were already `not_failure` under the three-family majority (*corrected in rc4; out/C12-EMPATES-POR-BRACO-2026-10-05.json*). The metric with content is agreement with the three-family majority: 1 111 / 1 145 = 97.0%.
4. **A rate generalized from one batch:** "100% of abstentions are *xai*" was true in the first 100 episodes and false over 1 195 (*xai* 26 + 6 quota + 1 missing · *zhipu* 13 · *google* 11), and was never re-measured before being asserted.
5. **Claiming the data confirm a locked parameter:** τ is locked since July over five families; "confirming" it with outcome-period data is post-hoc. We also called marginal rates "agreement", and cited the S1/S2 boundary that τ absorbs while omitting S0/S1, which it does not and where the 11.9 pp marginal gap lives.
6. **A rationalized sign:** we explained a +7/-8 swing by claiming a retry had recovered what a fourth panelist rescued; the sets are disjoint, and the sign comes from the 28 losses, not the gains. Recomputed under the published rule: gains 20 (episodes the fourth family brings to three substantive verdicts), losses 28 (episodes it brings to an exact tie), balance -8.
7. **A panel composed after counting ties in the trial's own verdicts:** post-hoc by §9 of the spec, recorded as such. What stands: the fourth family enters only as a substitute where fewer than three substantive verdicts exist (the substitute-only rule adopted in DEVIATIONS §10.31; the deposited

registration mentions a fourth family only as a mitigation not adopted there, PREREG 1.695), never as a fourth vote, so it generates no ties; the four-vote set is published as declared sensitivity. That is the rule as stated, and the registered analysis implements it: the 36 episodes with fewer than three substantive verdicts from the three families receive the fourth family’s verdict, 20 of them reach the floor (13 as failures), and 16 stay **unknown** (§4). The analysis of earlier versions, now the sensitivity analysis, did not implement it: its verdict file holds only the three families, so those 36 stayed **unknown**.

Defects 1–7 were found by two mechanisms that catch disjoint classes: adversarial review by independent model families, and mechanical census. Six of the seven above came from the former on a body of work the latter had already passed.

7. Threats to validity

Power. Addressed throughout and not mitigated: the registered horizon outlived the intervention the trial ran (§3.0.1), the study is under-powered as a consequence, and no analysis choice repairs that. Under the spec’s fractional counting no reduction, including total elimination, reaches 80% power, and at the registered analysis set the same holds under whole-epoch counting; under whole-epoch counting at the sensitivity analysis’s allocation it is borderline (§4.1.1). The feasibility stop that §3-bis of the prospective estimand declared for Epoch 1 was met under its planning definition, not under the set promotable at the served dose, and was not executed (§3.0).

One-system, one-fleet. Everything is measured on a single production system. Nothing here establishes that the effect, or its absence, generalizes.

Reproducibility of the adjudication. Re-run in October on 50 stratum-B episodes with the September prompt, settings and model ids, the panel’s verdicts did not reproduce at the label level for two of the three families (47/50 and 46/50 identical labels; 50/50 for the third; §4, rc12); whether that is non-determinism under the same model id or a changed model, the record cannot say. That bounds how reproducible the adjudication is: the estimates here are those of the September run, the October re-adjudication did not reproduce all of the September labels. Repeat-adjudication uncertainty is not explicitly propagated; some realized variation may enter through between-epoch differences, and its contribution to interval coverage has not been quantified (§4.1.2).

Tie rule. The exact-tie rule resolves to **not_failure** by design. It cannot bind in the three-family panel, where three substantive verdicts are required and three cannot tie, nor in the registered substitution panel, where an adjudicated episode has exactly three (§4), but it binds in the four-vote sensitivity set. There it biases toward fewer failures; its effect on the treatment-control contrast depends on the arms the ties fall in, ~~which we have not counted, and a reader is entitled to weigh that~~ and counted in rc4: of the 28 ties, 10 fall in treatment epochs, 11 in control and 7 outside the window (10, 10, 7, and one control tie before 09–03’s exposure window under the registered window; rc8); under the rc10 analysis 4 fall in the analysed exposure of treatment epochs, 7 of control, 9 in the washout, 7 outside the dates and 1 before 09–03’s window (`_sprint-2026-10-04/B-rc15/checks-rc15.json`, block C). ~~it is nil for H1c~~ It changes no H1c label relative to the three-family majority (none of the 11 tied H1c opportunities was a failure there), but against the opposite resolution it holds 4 treatment and 7 control opportunities (HT weights 27.78 and 36.73) at **not_failure**; resolving the ties as **failure** instead moves the four-vote H1c point difference from -0.0140 to -0.0249 in the registered analysis (-0.0121 to -0.0211 under rc8’s registered window and epoch set), and from -0.0208 to -0.0266 under the sensitivity analysis’s, toward a larger reduction under treatment (§4; `out/C12-EMPATES-REGISTRADO-2026-10-05.json` and `out/C12-EMPATES-COMO-FAILURE-2026-10-05.json`, point only).

Post-randomization denominator. H1c divides by opportunities, which are executed actions and can themselves change under treatment (opportunities per epoch 109.5 treatment, 135.0 control; 100.3 and 132.3 in the sensitivity analysis). Its contrast is a policy-level ratio over realized opportunities, not a per-opportunity propensity for a fixed population.

Scope of the specificity control. The sham replay (§4.0.1c) was run twice with the same 20 shams: over

the 2,646 brief states of the four $w = 4$ epochs, and over the 11,812 reconstructible states of the hash-verified trial window, which still leaves out epoch 09–01 (no hash proof of its corpus), 53 states whose serve-state cut cannot be reconstructed, and the shadow phase. The two runs share 2,016 states with identical records, so the second extends the first rather than replicating it. In both, the shams are drawn from the 36 non-designated items that can receive a bonus and matched on severity and bonus mass, not on signature group; and $p = 1/21$ is the floor at $K = 20$, a rank against these 20 shams rather than a calibrated randomization p-value. It establishes that the real designation outranked the 20 tested shams on what was served in the $w = 4$ epochs, and on what $w = 4$ would have changed over the window, where the second run forces that dose in every state, control epochs included; it does not establish specificity against every matched designation, nor an effect on outcomes.

Denominator. §4.2. The exposure measure counts idleness, and long sessions dominate it; from rc9 the washout is out of it (§2), and from rc10 every session is counted in the epoch of its start, as the registration requires, which puts 74% of treatment exposure into one epoch and almost all of it into three sessions (§4.2, §5).

Our own bookkeeping. §1.1. The claim that the H1b estimand decision preceded the estimates, and that the primary switch preceded the assignment seed, rests on our own timestamps (written times, commit times and a file time); the §10.33 decision followed a preliminary estimate (§1.1).

A stopping rule that may have been met and was not executed. §3.0. Under its planning definition, a rule declared before the seed would have stopped the trial after Epoch 1 for design impossibility; under the set promotable at the served dose it would not. The stopping rule did not truncate the observed trial. Its nonexecution does not by itself establish that the continuation process introduced no bias. A reader who holds us to the undeposited prospective estimand should read §4 as produced after a declared stop; a reader who holds us to the deposit should not, since the deposit contains no such stop. The deposited safety abort, which the deposit does contain, has no record of having run at the epoch boundaries (§3.0).

An assignment rule replaced after the seed. §1. The arms served are those of the registered script, fixed before the round; the rule the seed declaration named would have given different arms to 11 of the 20 realized epochs, and the replacement was made after the round was emitted and retracted in writing the same evening, before Epoch 1.

8. Related work

Four bodies of work bear on this trial, each read for what it gives us and for what this trial does not do that its authors do. In short: the agent-memory literature has benchmarks, cost characterisations and (dated 1 October 2026) one randomised design, but no pre-registered randomised trial on live traffic that we could find; the methods this trial needs come from online experimentation, which has no memory-specific content; and pre-registration is argued for in ML and IR but rarely practised. Paper A’s §8 covers the exposure and surface literature and is not repeated here.

8.1 Agent-memory evaluation, and what the surveys ask for

Benchmarks compare systems on fixed tasks. MemoryArena [the2026memoryarena], EvoMemory [wei2025evomemory], LifelongAgentBench [zheng2025lifelongagentbench], MemoryAgentBench [hu2025memoryagentbench], LoCoMo [maharana2024locomo], LongMemEval [wu2025longmemeval] and InterruptBench [zou2026interruptbench] stress different competencies but share one structure: a curated task suite, run once per system, systems compared to each other. Their contrast is between *systems on fixed tasks*; ours is between *two policies of one system on traffic we did not choose*. Evo-Memory both fixes task order as a fairness device and measures that order matters: ExprAG’s average success is 0.57 under Easy→Hard and 0.69 under Hard→Easy (its Table 2), over two *sorted* orders, not sampled ones.

What the canonical survey asks for. The TMLR survey [huang2026survey] closes its evaluation discussion (§9.6) by asking for “closed-loop, longitudinal, and execution-grounded evaluation paradigms”,

in environments “where experience accumulation has real consequences”, “enabling comparison between memory-augmented agents and memory-free baselines under identical conditions”, with replayable state and provenance. This trial is a neighbour of that direction, not an instance of it: §9.6 contrasts memory against *no* memory; we contrast two policies of an always-on memory. We share the paradigm: a live system, a closed loop, per-brief provenance (`ids_tratado` / `ids_controle`) and a replayable serving state. §4.0.1c shows what “replayable” costs: the replay reproduces production only on the corpus that actually served and with an exact serve-state cut. On the version pinned in Paper A (sha256 497e9549..., 4 Aug 2026) we count zero occurrences of `random*` and of `pre-regist*`; the single A/B appears as a cost that user simulators spare (“reducing costs associated with live user studies and online A/B testing”). We read that as the absence of a convention, not a claim that no one has run such experiments. Three other surveys [zhang2024memorysurvey; hu2025memoryage; du2026memoryautonomous] return zero for `randomi[sz]ed`, A/B and `pre-?regist` (lower bounds: hyphenation not stitched).

Telemetry is the other axis. Omri et al. [omri2026agentmemory] attribute tokens, latency, utilisation and energy to memory construction, retrieval and generation across ten systems, with no randomised contrast. We measure whether a lever changed what agents *did*. We report action-level time and token regret (§5), but no per-phase resource attribution or energy measurement; a trial carrying their per-phase attribution would answer whether a lever that changes a brief a few percent of the time pays for itself.

8.2 Interventional studies of memory

A thin literature *intervenes* on memory. Three questions separate it from this trial: randomised exposure, advance registration, live traffic.

Work	What is varied	Setting	Randomised exposure	Pre-registered	Live traffic
Xiong et al. [xiong2025memorymanagement]	memory addition / deletion; record quality	four agents, controlled experiments	not determined	not determined	no
Feng et al. [feng2026memorytransformation]	memory structure × content, 2×2 factorial	LiveCodeBench → MATH, 3 seeds, 360+ runs	not determined	“six pre-registered validation gates”	no
Sun et al. [sun2026experiencesegmentation]	none / random / prompt / retrieved experience in prompts	production moderation workload; results on an offline benchmark	random- <i>content</i> control, not randomised units	not stated	in production; results offline
Tablan et al. [tablan2026learningonstaticdynamic]	no-memory vs. full back-learned stores vs. static RAG	τ -bench banking, simulated	not determined	not stated	no
Srivastava [srivastava2026cmi]	no / with / perturbed memory, per candidate	Causal-LoCoMo (87 examples)	not described in what was read	not stated	no
Behnam & Wang [behnam2026cmp]	which memories fill k reserved context slots	LongMemEval, LoCoMo, multi-hop QA; Mem0 replay	yes , known propensities	not stated	no
This trial	promotion dose w on 19 designated items	production fleet, 24 h epochs	yes , public-beacon seed	yes (OSF yf7d2); primary switched before any data, in an undeposited revision (§1); a declared feasibility stop, met under its planning definition, not executed, and the seed declaration’s assignment rule replaced after the round (§1, §3.0)	yes

“Not determined” means *we did not read far enough to say*; “not stated” means *not found in what was read*.

Causal Memory Policy. Behnam and Wang [behnam2026cmp] (arXiv, 1 Oct 2026, not peer-reviewed) show that, for a fixed query and context in which a memory has zero probability of being retrieved, randomizing its presence in the store cannot identify its utility conditional on retrieval: it is “not identified by any design that holds $M-m$ fixed while randomizing only whether m is included” (their Theorem 1; a *retrieval-level positivity violation*). They restore identification by randomising which memories occupy k reserved context slots, with design-fixed propensities and a self-normalised inverse-propensity estimator. Identification fails for 54% of required memories on LongMemEval and 67% on LoCoMo, and the failure persists in Mem0 replayed over benchmark histories.

Three points of contact. First, an instance of their failure is in our data. The 19 designated items leave eligibility together at 2026-09-20 22:51:23, after which no dose reaches them, because the boost is addressed by id and they are no longer candidates; treatment and control become the same intervention for 214 of 234 registered epochs (§3.0.1). CMP is the formal name and the first treatment of that failure we have found. We did not have the framing at registration (2026-08-18) and claim no priority. Whether the 20 *realised* epochs also contain a partial failure is a question for Paper A’s exposure measurements and §4 of this paper; this section does not assert it. Second, the estimands differ. CMP identifies an individual memory’s utility per interaction; we estimate an intention-to-treat effect of one lever with the fleet-wide epoch as unit, do not randomise within a brief, and have no propensity-weighted estimator. Third, the evidence differs in kind. Theirs is benchmark and replay evidence with a measured per-query cost; ours is live, pre-registered, with an interval whose coverage we have not established (§4.1.1). CMP-style slot randomisation *on* a live fleet *under* a registration is, as far as we found, empty; it is what a successor trial would need.

Randomised evaluations of deployed AI tools exist [peng2023copilot; becker2025metr; bean2026llmmedical], but randomise *people or tasks*, which presumes no interference between units. A memory store shared by a fleet violates that presumption.

8.3 Online controlled experiments

Randomise, then check the randomisation. The online-experimentation literature [kohavi2020trustworthy; kohavi2009controlled] is the standard reference for randomised experiments on live systems (cited for that role, not for a specific prescription); Fabijan et al. [fabijan2019srm] describe sample ratio mismatch as an indicator of data-quality problems. Our analogues are the registered assignment script with its committed hash and the positive (11/11) and dual negative (8/8) controls of §4.0.1b. The sham replay, the placebo-like specificity control, ~~was not executed: the configuration launched was invalid, and a valid one awaits data that exist only on the production host~~ was invalid as first launched and was run in a corrected configuration on 2026-10-04: the real designation ranks above all 20 shams on the $w = 4$ epochs, and again over the whole hash-verified trial window in a second run, 2026-10-05 to 2026-10-07, which forces $w = 4$ in every state (§4.0.1c). We used no covariate variance reduction [deng2013cuped] and did not interleave [radlinski2008clickthrough; chapelle2012interleaved; hofmann2016online], for which Hofmann et al. report one to two orders of magnitude more sensitivity than absolute metrics; we have no per-item credit signal to interleave on.

Why the unit is the epoch. When units share state, per-unit randomisation is biased; this is the problem around which marketplace [blake2014marketplace; johari2022twosided] and network [saveski2017network; aronow2017interference] experimentation are built. The switchback remedy randomises the whole system over time [bojinov2023switchback; hu2022switchback; basse2023minimax] and comes with randomisation-based inference [bojinov2019timeseries]. Ours is a fleet-level switchback with 24 h epochs and a two-hour washout. Unlike that literature, we fixed the washout *ex ante* instead of estimating the carryover order; we have 19 periods in the analysis (20 realized); and a memory store accumulates, so carryover through *state* may outlast any fixed washout. We claim the relation, not that these theorems apply to our state process.

Few clusters. Our percentile cluster bootstrap runs on 9–11 clusters per arm, the regime in which ordinary cluster bootstraps over-reject [cameron2008bootstrap], perform “poorly with fewer than eleven clusters” [webb2014reworking], and for which a six-point wild bootstrap is recommended [cameron2015practitioner].

The analysis has 19 clusters, split 11/8 (20, split 11/9, in the sensitivity analysis). The few-cluster literature motivates caution: ordinary cluster bootstraps can over-reject with few clusters [cameron2008bootstrap], and Webb’s six-point wild bootstrap is proposed for small total cluster counts [webb2014reworking; cameron2015practitioner]. That literature concerns cluster bootstraps for regression inference; it does not directly calibrate this BCa (or percentile) ratio estimator or establish the direction of its coverage error. We used none of these. The registered re-randomisation test (§4.0.2) is the finite-sample alternative (a Fisher randomisation test [imbens2015causal; bojinov2019timeseries]) and the inference to weight where the two disagree, as on H1a in the sensitivity analysis.

Under-power and counterfactual logging. Across twenty-five large field experiments, “the median confidence interval on return on investment is over 100 percentage points wide” [lewis2015unfavorable]. Ours is a *weaker* case: they were under-powered by the economics of the outcome, we by a registered horizon of 234 epochs that was never re-read against the intervention frozen after registration, whose 19 items left a 30-day eligibility window together after 20 epochs (§3.0.1). Observational comparison is not the remedy: observational methods often fail to recover the effects measured in the same advertising experiments [gordon2019comparison]. Our replay analyses belong to the family of Li et al. [li2011unbiased], and the closest analogue of our sham is ghost-ads logging of the exposures the treatment *would* have produced [johnson2017ghost]: in both, the counterfactual is only as good as the state it is computed from, which is exactly where our first sham configuration failed, and what the corrected one had to reproduce before its result could be read (§4.0.1c).

8.4 Pre-registration, and what a null is worth

Pre-registration in ML and NLP has been piloted, not adopted: the NeurIPS 2020 and 2021 workshops [bertinetto2020prereg; albanie2022prereg] ran registered-report-style review (the 2021 preface reports 22 proposals, 10 accepted and 3 results papers); it has been argued for in NLP [vanmilttenburg2021prereg], weighed against its costs [sogaard2023twosided] and adapted to predictive modelling [hofman2023prereg]. Vaccaro [vaccaro2026prereg] lists memory settings among the researcher degrees of freedom of experiments with AI agents (agents as subjects, so the fit is by analogy), and Wilder and Zhou [wilder2025evaluation] propose that ML venues require a declaration of preregistration for field experiments. In IR, a Dagstuhl report [bauer2023dagstuhl] recommends results-blind review. The only memory-specific pre-registered design we found is Feng et al.’s, on benchmarks. This trial is a *registration*, not a registered report (its protocol was never peer-reviewed before data), and its deviations are in Appendix A. The cross-field arguments [nosek2018prereg; chambers2022registeredreports; munaf02017manifesto] are the standing justification. Registration was associated with a change in reported findings: 17 of 30 large NHLBI trials before 2000 reported significant benefit, against 2 of 25 after mandatory prospective registration of outcomes [kaplan2015nullnhlbi]. This observational comparison does not identify registration as the cause.

Power language needs care. The critiques of low power [ioannidis2005false; button2013power; gelman2014beyond] and of post-experiment power calculations [hoenig2001abuse] apply here; Hoenig and Heisey also call power calculations “valuable in planning an experiment”. Our saturated minimum detectable effect (under the spec’s fractional counting and, at the registered 19 clusters, under whole-epoch counting too; whole-epoch counting at the sensitivity analysis’s 20 clusters flips it, §4.1.1) is prospective and outcome-independent, which is the planning form, but its inputs are the problem: the ICC was estimated for a different quantity (§4.1.1). The choice of significance test is known to change error rates on IR test collections [urbano2019significance]; we know of no corresponding analysis for live retrieval experiments.

8.5 Position

Not new: that memory evaluation scores systems rather than behavioural effect; randomising a retrieval-side lever with known propensities; the switchback design and its inference; pre-registering ML experiments. **What this trial is:** a pre-registered, beacon-seeded, fleet-level switchback on live agent traffic, with per-brief provenance, instrument controls, and a null reported with its power failure and the collisions between its own locks. **What it is not:** it does not identify per-memory utility, randomise within a brief, estimate carryover, or attribute resources to memory phases (it reports action-level time and token regret, §5, and no energy);

it compares two policies, not memory against none; it is one system on one fleet; and its specificity control covers what was served in the $w = 4$ epochs and, as a counterfactual at $w = 4$, the rest of the hash-verified trial window (without epoch 09–01), not outcomes (§4.0.1c). **What it adds, narrowly:** a live-traffic instance of the positivity failure CMP formalises (§3.0.1), and two collisions whose parts each looked complete alone: an internal contradiction in the registered estimand (§4.4), and a mismatch between the registered horizon and the subsequently frozen intervention (§3.0.1). **What would falsify this positioning:** a pre-registered randomised memory trial on live traffic published before ours. A bounded search (web and arXiv, 2026-10-03/04, one agent, not a systematic review) found none, nor an IR registered-report track.

Reading depth. Full or partial text was read for the TMLR survey (§9.6, with string counts on the pinned PDF), Omri et al., Evo-Memory, Behnam & Wang, Feng et al., the NeurIPS 2021 preface, Sogaard et al., Vaccaro and the Dagstuhl report. MemoryArena v2 was only string-searched. No book was opened, and several items are known at abstract or metadata level and are cited for their topic, not for a finding. Per-reference record: `_sprint-2026-10-04/B-related-work.verification-log.md`; long form of this section: `_sprint-2026-10-04/B-related-work.md`.

9. Discussion

In short, the registered horizon outlived the intervention the trial ran: the registration sized the trial for 234 epochs, and the designation frozen after it was eligible for 20 (§3.0.1); the under-powering follows from that. Under the analysis spec’s fractional counting, which governs the power statement, and under the planning assumptions, no reduction, including total elimination, reaches 80% power, and this was projection-robust under the spec’s cuts ten days before the window closed, and written down rather than discovered afterwards. At the registered analysis set of 19 epochs, whole-epoch counting gives the same verdict; only at the sensitivity analysis’s 11 treatment and 9 control does it flip (effective size 96.41 against 95.26), where total elimination becomes borderline detectable and 80% power holds only for reductions of about 99.6% or more (§4.1.1). The primary we report, H1c, replaced the deposited primary H1 before any data, in a revision that was not deposited; under H1, tested alone as the deposited reading has it, the registered test does not reject in the registered analysis ($p = 0.3294$), and under the switch’s Holm rule it does not either. It rejected under the deposited reading in versions up to rc9, whose analyses split long sessions across epochs, and §4.3 states why that rejection went with the denominator: it disappears once sessions are counted as registered, which shows it was not robust to the registered rule and does not test the dose (§1). Three further commitments made before the seed were not kept, and we report all three here for the first time: the feasibility stop we found from the artifacts while preparing rc8 (met in Epoch 1 under its planning definition, not under the set promotable at the served dose, and not executed; §3.0); the replaced assignment rule, retracted in the seed declaration on 2026-08-30 and not carried into the manuscript (§1); and the deposited census of the live study (PREREG §3), which a sample of 800 of the 5 556 non-`is_error` episodes replaced, and which a post-unblinding attempt could not restore because the panel’s labels did not reproduce at the declared 0.99 (§4, §7). What remains, in place of an effect estimate, is three observations about instruments, each of which cost us a published error to find:

A pre-registration can contradict itself, and the contradiction can survive to the outcome.

One contradiction was internal to the registration: two locks eighteen days apart defined one estimand incompatibly (§4.4), and neither review caught it. A second incompatibility arose between its horizon and the subsequently frozen intervention: the sample size was computed for 234 epochs, and the designation fixed after v1.12 had a 30-day age limit, leaving 20 eligible trial epochs, 8.5% of the registered 234 (§3.0.1). The second is the more instructive, because each number is correct in isolation: `sizing.py` did its arithmetic correctly, and the 30-day window is a documented default. The two were never read against each other, and no step in the process had the job of crossing them.

An interval that excludes zero can be the artifact of an unregistered denominator, and the registered non-rejection the reportable result. The only rejection under the deposited decision rules is on the hypothesis demoted, in the undeposited revision, as requiring a 955% effect at the calibration share of altered briefs, and under the switch’s Holm rule it is not a rejection. H1a’s unadjusted p-value (0.0168)

is below 0.05 and is not a rejection: the registration decides it in a Holm family, where it does not reject. Reading the rejection and ignoring the null would have inverted the paper; under the deposited registration that rejection is on the primary, which is why §1 reports both. In every version up to rc9 the only rejection under the deposited decision rules was on the hypothesis demoted, in the undeposited revision, as requiring a 95% effect at the calibration share of altered briefs; its intervals excluded zero on every leg in every version, and H1a's excluded zero on every registered leg in rc9 (on the locked leg only in the sensitivity analysis, and not on rc8's post-hoc leg; §4.2). All of that was computed with long, mostly idle sessions split across the epochs they touched; counted at their start, as registered, no interval of the H1 family excludes zero and nothing rejects (§4.3). The correction moves exposure and the attribution of outcomes together, so it shows that those results were not robust to the registered rule, not that the dose had no effect. Reading the rejection and ignoring the null would have inverted the paper; three revisions of it carried the rejection as unexplained before the registration's own attribution rule removed it.

~~A pre-committed alternative explanation can be refuted by the data it was written to protect. The concentration projection that made a null ambiguous did not hold; coverage was flat. That weakens a defence we had reserved for ourselves, which is the only circumstance in which a pre-commitment demonstrably did its job.~~

A planning statistic and a realized statistic can share a name and not a quantity. We first read the flat distribution of designated-chunk serving occurrences as refuting the pre-committed concentration projection, which counted covered action opportunities by signature. The two count different things: the projection is untested, and the pre-committed indistinction between a mechanism that does not work and lessons too generic to work stands (§4.5). (*Corrected in rc7.*)

We publish an under-powered null in full because the two alternatives (not publishing, or publishing a claim the design cannot support) are the behaviours that make under-powered trials worth less than nothing to the people who read them.

References

Citations in §8 are pandoc-style ([@key]) and resolve in `_sprint-2026-10-04/B-related-work.bib`, which carries the full author lists, URLs and the per-entry notes on what is not established (venue, publication status). The list below is generated from that file for the keys cited here; where the two differ, the `.bib` file is authoritative. Reading depth per reference is in `_sprint-2026-10-04/B-related-work.verification-log.md`. Artifacts of this study are cited by file name in the text and listed in Appendix B.

- Albanie, S. et al. (eds.) (2022). Proceedings of the NeurIPS 2021 Workshop on Pre-registration in Machine Learning. <https://proceedings.mlr.press/v181/>. [@albanie2022prereg]
- Aronow, P. M. and Samii, C. (2017). Estimating average causal effects under general interference, with application to a social network experiment. *The Annals of Applied Statistics* 11(4). doi:10.1214/16-AOAS1005. [@aronow2017interference]
- Basse, G., Ding, Y. and Toulis, P. (2023). Minimax designs for causal effects in temporal experiments with treatment habituation. *Biometrika* 110: 155–168. doi:10.1093/biomet/asac024. [@basse2023minimax]
- Bauer, C. et al. (2023). Report from Dagstuhl Seminar 23031: Frontiers of Information Access Experimentation for Research and Education. *Dagstuhl Reports* 13(1): 68–154. doi:10.4230/DagRep.13.1.68. [@bauer2023dagstuhl]
- Bean, A. M. et al. (2026). Reliability of LLMs as medical assistants for the general public: a randomized preregistered study. *Nature Medicine* 32(2): 609–615. doi:10.1038/s41591-025-04074-y. [@bean2026llmmedical]
- Becker, J. et al. (2025). Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity. arXiv preprint. arXiv:2507.09089. [@becker2025metr]
- Behnam, A. and Wang, B. (2026). Causal Memory Policy: Making Memory Utility Identifiable by Intervening on Retrieval. arXiv preprint. arXiv:2610.02070. [@behnam2026cmp]

- Bertinetto, L. et al. (2020). The pre-registration experiment: an alternative publication model for machine learning research. *NeurIPS 2020 Workshop*. <https://neurips.cc/virtual/2020/workshop/16158>. [[@bertinetto2020prereg](#)]
- Blake, T. and Coey, D. (2014). Why marketplace experimentation is harder than it seems: the role of test-control interference. *Proceedings of the Fifteenth ACM Conference on Economics and Computation*: 567–582. doi:10.1145/2600057.2602837. [[@blake2014marketplace](#)]
- Bojinov, I. and Shephard, N. (2019). Time Series Experiments and Causal Estimands: Exact Randomization Tests and Trading. *Journal of the American Statistical Association* 114(528): 1665–1682. doi:10.1080/01621459.2018.1527225. [[@bojinov2019timeseries](#)]
- Bojinov, I., Simchi-Levi, D. and Zhao, J. (2023). Design and Analysis of Switchback Experiments. *Management Science* 69(7): 3759–3777. doi:10.1287/mnsc.2022.4583. [[@bojinov2023switchback](#)]
- Button, K. S. et al. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* 14(5): 365–376. doi:10.1038/nrn3475. [[@button2013power](#)]
- Cameron, A. C., Gelbach, J. B. and Miller, D. L. (2008). Bootstrap-Based Improvements for Inference with Clustered Errors. *The Review of Economics and Statistics* 90(3): 414–427. doi:10.1162/rest.90.3.414. [[@cameron2008bootstrap](#)]
- Cameron, A. C. and Miller, D. L. (2015). A Practitioner’s Guide to Cluster-Robust Inference. *Journal of Human Resources* 50(2): 317–372. doi:10.3368/jhr.50.2.317. [[@cameron2015practitioner](#)]
- Chambers, C. D. and Tzavella, L. (2022). The past, present and future of Registered Reports. *Nature Human Behaviour* 6(1): 29–42. doi:10.1038/s41562-021-01193-7. [[@chambers2022registeredreports](#)]
- Chapelle, O. et al. (2012). Large-scale validation and analysis of interleaved search evaluation. *ACM Transactions on Information Systems* 30(1): 1–41. doi:10.1145/2094072.2094078. [[@chapelle2012interleaved](#)]
- Deng, A. et al. (2013). Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*: 123–132. doi:10.1145/2433396.2433413. [[@deng2013cuped](#)]
- Du, P. (2026). Memory for Autonomous LLM Agents: Mechanisms, Evaluation, and Emerging Frontiers. arXiv preprint. arXiv:2603.07670. [[@du2026memoryautonomous](#)]
- Fabijan, A. et al. (2019). Diagnosing Sample Ratio Mismatch in Online Controlled Experiments. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*: 2156–2164. doi:10.1145/3292500.3330722. [[@fabijan2019srm](#)]
- Feng, Z., Yao, M. and Lewis, D. S. (2026). Memory Transplants for LLM Agents: Disentangling Architecture and Content Transfer under a Code-to-Math Shift. *ICLR 2026 Workshop on Memory for LLM-Based Agentic Systems (MemAgents)*. <https://openreview.net/forum?id=AIJsJlqfsp>. [[@feng2026memorytransplants](#)]
- Gelman, A. and Carlin, J. (2014). Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspectives on Psychological Science* 9(6): 641–651. doi:10.1177/1745691614551642. [[@gelman2014beyond](#)]
- Gordon, B. R. et al. (2019). A Comparison of Approaches to Advertising Measurement: Evidence from Big Field Experiments at Facebook. *Marketing Science* 38(2): 193–225. doi:10.1287/mksc.2018.1135. [[@gordon2019comparison](#)]
- He, Z. et al. (2026). MemoryArena: Benchmarking Agent Memory in Interdependent Multi-Session Agentic Tasks. arXiv preprint. arXiv:2602.16313. [[@he2026memoryarena](#)]
- Hoenig, J. M. and Heisey, D. M. (2001). The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis. *The American Statistician* 55(1): 19–24. doi:10.1198/000313001300339897. [[@hoenig2001abuse](#)]
- Hofman, J. M. et al. (2023). Pre-registration for Predictive Modeling. arXiv preprint. arXiv:2311.18807. [[@hofman2023prereg](#)]
- Hofmann, K., Li, L. and Radlinski, F. (2016). Online Evaluation for Information Retrieval. *Foundations and Trends in Information Retrieval* 10(1): 1–117. doi:10.1561/15000000051. [[@hofmann2016online](#)]
- Hu, Y. and Wager, S. (2022). Switchback Experiments under Geometric Mixing. arXiv preprint. arXiv:2209.00197. [[@hu2022switchback](#)]
- Hu, Y. et al. (2025). Memory in the Age of AI Agents. arXiv preprint. arXiv:2512.13564.

- [@hu2025memoryage]
- Hu, Y., Wang, Y. and McAuley, J. (2025). Evaluating Memory in LLM Agents via Incremental Multi-Turn Interactions. arXiv preprint. arXiv:2507.05257. [@hu2025memoryagentbench]
 - Huang, W. et al. (2026). A Survey of Agent Memory in the Second Half: Towards Self-Evolving and Long-Horizon Agents. *Transactions on Machine Learning Research*. arXiv:2602.06052. [@huang2026survey]
 - Imbens, G. W. and Rubin, D. B. (2015). Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. *Cambridge University Press*. doi:10.1017/CBO9781139025751. [@imbens2015causal]
 - Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine* 2(8): e124. doi:10.1371/journal.pmed.0020124. [@ioannidis2005false]
 - Johari, R. et al. (2022). Experimental Design in Two-Sided Platforms: An Analysis of Bias. *Management Science* 68(10): 7069–7089. doi:10.1287/mnsc.2021.4247. [@johari2022twosided]
 - Johnson, G. A., Lewis, R. A. and Nubbemeyer, E. I. (2017). Ghost Ads: Improving the Economics of Measuring Online Ad Effectiveness. *Journal of Marketing Research* 54(6): 867–884. doi:10.1509/jmr.15.0297. [@johnson2017ghost]
 - Kaplan, R. M. and Irvin, V. L. (2015). Likelihood of Null Effects of Large NHLBI Clinical Trials Has Increased over Time. *PLOS ONE* 10(8): e0132382. doi:10.1371/journal.pone.0132382. [@kaplan2015nullnhlbi]
 - Kohavi, R. et al. (2009). Controlled experiments on the web: survey and practical guide. *Data Mining and Knowledge Discovery* 18(1): 140–181. doi:10.1007/s10618-008-0114-1. [@kohavi2009controlled]
 - Kohavi, R., Tang, D. and Xu, Y. (2020). Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing. *Cambridge University Press*. doi:10.1017/9781108653985. [@kohavi2020trustworthy]
 - Lewis, R. A. and Rao, J. M. (2015). The Unfavorable Economics of Measuring the Returns to Advertising. *The Quarterly Journal of Economics* 130(4): 1941–1973. doi:10.1093/qje/qjv023. [@lewis2015unfavorable]
 - Li, L. et al. (2011). Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*: 297–306. doi:10.1145/1935826.1935878. [@li2011unbiased]
 - Maharana, A. et al. (2024). Evaluating Very Long-Term Conversational Memory of LLM Agents. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*: 13851–13870. doi:10.18653/v1/2024.acl-long.747. [@maharana2024locomo]
 - Munafò, M. R. et al. (2017). A manifesto for reproducible science. *Nature Human Behaviour* 1(1): 0021. doi:10.1038/s41562-016-0021. [@munaf02017manifesto]
 - Nosek, B. A. et al. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences* 115(11): 2600–2606. doi:10.1073/pnas.1708274114. [@nosek2018prereg]
 - Omri, Y. et al. (2026). Agent Memory: Characterization and System Implications of Stateful Long-Horizon Workloads. arXiv preprint. arXiv:2606.06448. [@omri2026agentmemory]
 - Peng, S. et al. (2023). The Impact of AI on Developer Productivity: Evidence from GitHub Copilot. arXiv preprint. arXiv:2302.06590. [@peng2023copilot]
 - Radlinski, F., Kurup, M. and Joachims, T. (2008). How does clickthrough data reflect retrieval quality? *Proceedings of the 17th ACM Conference on Information and Knowledge Management*: 43–52. doi:10.1145/1458082.1458092. [@radlinski2008clickthrough]
 - Saveski, M. et al. (2017). Detecting Network Effects: Randomizing Over Randomized Experiments. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*: 1027–1035. doi:10.1145/3097983.3098192. [@saveski2017network]
 - Søgaaard, A., Hershovich, D. and de Lhoneux, M. (2023). A Two-Sided Discussion of Preregistration of NLP Research. *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. doi:10.18653/v1/2023.eacl-main.6. [@sogaard2023twosided]
 - Srivastava, S. S. (2026). Causal Intervention-Based Memory Selection for Long-Horizon LLM Agents. arXiv preprint. arXiv:2605.17641. [@srivastava2026cmi]
 - Sun, L., Zhang, H. and Zhang, X. (2026). External Experience Serving in Production LLM Systems: A Deployment-Oriented Study of Quality-Cost Trade-offs. arXiv preprint. arXiv:2606.11806.

- [@sun2026experienceserving]
- Tablan, V., Taylor, S. and Bernhem, K. (2026). Learning on the Job: Continual Learning from Deployment Feedback for Frozen-Weights Agents. arXiv preprint. arXiv:2607.22157. [@tablan2026learningonthejob]
 - Urbano, J., Lima, H. and Hanjalic, A. (2019). Statistical Significance Testing in Information Retrieval: An Empirical Analysis of Type I, Type II and Type III Errors. *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*: 505–514. doi:10.1145/3331184.3331259. [@urbano2019significance]
 - Vaccaro, M. (2026). Preregistration for Experiments with AI Agents. arXiv preprint. arXiv:2606.11217. [@vaccaro2026prereg]
 - van Miltenburg, E., van der Lee, C. and Krahmer, E. (2021). Preregistering NLP research. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. doi:10.18653/v1/2021.naacl-main.51. [@vanmiltenburg2021prereg]
 - Webb, M. D. (2014). Reworking Wild Bootstrap Based Inference for Clustered Errors. *Queen’s University, Department of Economics*, Queen’s Economics Department Working Paper 1315. http://qed.econ.queensu.ca/working_papers/papers/qed_wp_1315.pdf. [@webb2014reworking]
 - Wei, T. et al. (2025). Evo-Memory: Benchmarking LLM Agent Test-time Learning with Self-Evolving Memory. arXiv preprint. arXiv:2511.20857. [@wei2025evomemory]
 - Wilder, B. and Zhou, A. (2025). Fostering the Ecosystem of AI for Social Impact Requires Expanding and Strengthening Evaluation Standards. arXiv preprint. arXiv:2510.18238. [@wilder2025evaluation]
 - Wu, D. et al. (2025). LongMemEval: Benchmarking Chat Assistants on Long-Term Interactive Memory. arXiv preprint. arXiv:2410.10813. [@wu2025longmemeval]
 - Xiong, Z. et al. (2025). How Memory Management Impacts LLM Agents: An Empirical Study of Experience-Following Behavior. arXiv preprint. arXiv:2505.16067. [@xiong2025memorymanagement]
 - Zhang, Z. et al. (2024). A Survey on the Memory Mechanism of Large Language Model based Agents. arXiv preprint. arXiv:2404.13501. [@zhang2024memorysurvey]
 - Zheng, J. et al. (2025). LifelongAgentBench: Evaluating LLM Agents as Lifelong Learners. arXiv preprint. arXiv:2505.11942. [@zheng2025lifelongagentbench]
 - Zou, H. P. et al. (2026). When Users Change Their Mind: Evaluating Interruptible Agents in Long-Horizon Web Navigation. arXiv preprint. arXiv:2604.00892. [@zou2026interruptbench]

Appendix A: relation to the pre-registration

Deviations are logged in `DEVIATIONS-FOR-PAPER.md`, which is append-only, and the substantive ones for this paper are §10.14 and §10.22 (the closure decision of §3.0) and §10.29 through §10.34; the primary switch, the omission of H2 and H3 until rc7, the washout kept in the denominator until rc8, the unexecuted stopping rule, the safety abort without an execution record, the registered M10 not computed, the boundary-straddling attribution not applied until rc10, the sample that replaced the registered census of the live study, and the replaced assignment rule below are not in it (the assignment-rule retraction is in `ASSIGN-SEED-2026-08-30.md`), and are declared here and in §1, §2, §3.0, §4, §4.6 and §5. The list is not summarized here because a summary of a deviation log is a second copy that will diverge from the first.

~~Three Seven Nine Twelve Thirteen Twelve Eleven Twelve~~ (the struck item below is no longer a deviation) that a reader cannot reconstruct from the estimates alone:

- the primary outcome: the deposited registration (v1.12) names H1, with H1a–H1c as a Holm-corrected co-primary family; H1c became the primary on 2026-08-30 (`DESIGN-REVISION-2026-08-30.md` §3-ter; `PROSPECTIVE-ESTIMAND-2026-08-30.md` §3-bis), before the assignment seed was drawn and before any randomized epoch, in documents dated by our commit log and not deposited. For the deposited primary, H1, the registered test does not reject under either reading ($p = 0.3294$ tested alone; Holm-adjusted 1.0 under the switch); under the deposited Holm correction no member of the family rejects, H1a included (unadjusted 0.2599, adjusted 1.0). Under the deposited reading H1 rejected in the sensitivity analysis and in rc8’s and rc9’s analyses (§1);

- nothing after v1.12 was deposited: a v1.13 amendment prepared on 2026-08-27 was held (`deposit/PLAN-v1.13.md`); the start of the trial was authorized by `PROSPECTIVE-ESTIMAND-2026-08-30.md`, and the analysis was fixed by the analysis specification of 2026-09-10, both undeposited and dated by our commit log only;
- the registered H2 (task regret) and H3 (retrieval metrics) were not reported, and until rc7 their absence was not declared anywhere; from rc8, H2 is reported with three declared choices the registration does not fix and does not reject, and H3 is declared not computable (§5);
- the analysis of versions up to rc7 departed from the registered one in four places: percentile bootstrap instead of BCa, 09-02 included against the spec’s §2, no cut at the designation’s expiry inside 09-20 and no exposure offsets, and no fourth-family substitution in the panel. From rc8 the registered analysis is reported and that one is a sensitivity (§4). Under the deposited reading, H1 rejects in the earlier sensitivity analysis but not in the registered analysis (v4); under the switched-primary reading, neither analysis rejects H1. Two parts of the registered analysis could not be implemented as written and are interpretations: the jackknife runs over the 19 realized epochs, not the 234 registered (§4), and the spec’s “offset” for partial epochs is implemented as an exposure window (§3.0.1). Up to rc8 the session-hour denominator also kept the two-hour washout that the outcomes excluded, a defect `estimador_itt.py` shares and that rc8’s byte-for-byte control reproduced; rc9 removes it, which moves H1 and H1a and no verdict under either multiplicity reading (§2, §4). Up to rc9 the analysis also counted every episode in the epoch of its own timestamp, where the registration counts a session in the epoch of its start; rc10 applies the registered rule, which takes H1’s deposited-reading rejection away ($p\ 0.0302 \rightarrow 0.3294$) and every interval of H1 and H1a that excluded zero (§2, §4.3). Up to rc14 the BCa acceleration was the one-sample formula over the pooled jackknife values, which does not match the arm-stratified bootstrap; rc15 computes it arm by arm, which moves most BCa bounds slightly and changes no zero-inclusion conclusion and no p-value (§4, *Uncertainty*);
- a stopping rule declared before the seed (`PROSPECTIVE-ESTIMAND-2026-08-30.md` §3-bis: close the interventional arm with a null if Epoch-1 coverage falls below 36.7%): the reconstructed Epoch-1 coverage was below the threshold under the planning definition (32.4%, signatures promotable at $w = 2$) and above it under the set promotable at $w = 4$, the dose Epoch 1 was served at (37.4%); we found no contemporaneous record of evaluating the rule, and the arm continued (§3.0). Relative to the deposit, which forbids optional stopping, continuing is consistent;
- the horizon: data collection ended after 20 randomized epochs, before either deposited horizon condition (234 randomized epochs, or the calendar cap) was reached. Expiry of the fixed designation was not a deposited stopping condition. The dose was switched off on 2026-09-21 at 09:43:05Z, after the designation expired at 2026-09-20 22:51:23Z (§3.0, §3.0.1);
- the deposited mechanical safety abort (arm-blind, evaluated by script at every epoch boundary) has no record of having been evaluated at the 20 boundaries: it was deployed dormant, its one recorded run (2026-08-21) refused to evaluate under the 14-day gate, and the crontab records hold no entry for it. Ex post, no adjudicated episode of the window reaches a panel-majority S3 or S4 (§3.0);
- PREREG §3 required census adjudication of the live study. The reported analysis instead uses a census of the `is_error` stratum and a sample of 800 of the 5 556 remaining corpus episodes, weighted by 6.945. This departs from the deposited sampling plan; the six-switch reanalysis does not restore the census. Its bootstrap conditions on the realized adjudication sample (§4, §4.1.2). Restoring the census was attempted on 2026-10-05, after unblinding, under gate rules written before the first call: the model-identity gate passed, and the test-retest gate failed (label agreement 47/50 and 46/50 for two of the three families, against a declared 0.99), so the census was not run and the sample stays a deviation (§4, `_sprint-2026-10-04/B-censo/GATES.md`);
- the assignment rule named in the seed declaration (`ASSIGN-SEED-2026-08-30.md`, a global sort) is not the registered `assign_arms.py`; it was replaced by the registered script 24 min 38 s after the drand round was emitted, and the two assign different arms to 11 of the 20 realized epochs (§1). The retraction is in `ASSIGN-SEED-2026-08-30.md`, filed the same evening, before Epoch 1, and not in the deviation log; it overrode the declaration’s own failure clause, under which a divergence between its verification hash and `ASSIGNMENT.json` meant that the study does not start;
- the registered M10 (arm $\times$ session-level `brief_log` coverage, with its interval) and the 95%-coverage analysis set were not computed: the session $\leftrightarrow$ `brief_log` linkage is not in the locked artifacts, and

the table has no session column; the correlation earlier versions called M10 measures designated-item presence among served briefs (§4.6, §5);

- ~~boundary-straddling sessions are not assigned to the epoch of their start as PREREG §2 requires; the registered ‘without’ leg is computed only as a diagnostic, and on it H1 and H1a contain zero (§5); (rc10: implemented; see the item on the registered analysis above, and §5)~~
- the analysis stratum migrated from S2 to $\geq$ S1 by panel agreement (κ 0.87–0.93 at $\geq$ S1 against 0.31–0.53 for the S1/S2 split), so the S1/S2 division became an instrument finding rather than an analysis boundary;
- the designation was recorded as an open defect in v1.12 and was closed on 2026-08-26, after v1.12 was published, by the fixed 19-item designation, which was not deposited (§1, §3.0.1).

The dose band {2.0, 4.0, 7.5} was retained when the sizing formula was corrected. The later calibration replay changed 11/15/17 of 350 states; this does not contradict that commitment. Up to rc15 this list carried the band as a twelfth deviation, reading “*what does not move, and could not*” (PREREG, the design-effect correction of 2026-08-17) as a promise that the dose could not change briefs; the phrase names the parameters that correction left unchanged. rc16 removed the item, and the paragraph on the two items where the registration promises less, which rested on it. **DEVIATIONS-FOR-PAPER.md** still carries the withdrawn reading in its front table (the band row, marked $\Downarrow$): it is a dated, append-only log, and it is not edited.

Appendix B: artifacts

artifact	holds
out/ITT-REGISTRADO-v4-2026-10-05.json · measurement/estimador_itt_registrado.py (six switches; arm-stratified BCa acceleration) · _sprint-2026-10-04/B-registered/RESULTADO-v4.md	the registered analysis (reported from rc15): the v3 analysis with the BCa acceleration computed arm by arm (multi-sample), every leg, H2, the Holm families and the controls of the row below, plus a control that, with the v3 acceleration carried beside each interval, gives back every BCa interval, acceleration and bias correction of out/ITT-REGISTRADO-v3-2026-10-05.json (field controle_v3_registrado). Script sha256 b3095740..., frozen at _sprint-2026-10-04/B-registered/estimador_itt_registrado-v4-b3095740.py; artifact sha256 3ed7637e.... Run with --aceleracao uma_amostra the script emits the v3 file in every field except gerado_em and the script sha256 (checked in rc15). The acceleration is tested in _sprint-2026-10-04/B-registered/teste_aceleracao_v4.py · teste-aceleracao-v4.json (a hand-computed two-arm case in exact fractions, and SciPy’s multi-sample acceleration on the registered and the ‘without’ legs and the four H2 estimators). (rc16: the test also compares the unrounded accelerations, to 1e-12, as its docstring already said it did; largest difference 1.4×10^{-16} ; script sha256 d4d960db..., output 041bd4a9.... The rc15 bytes, which RESULTADO-v4.md records, are kept as _sprint-2026-10-04/B-registered/teste_aceleracao_v4-rc15-bb9c3197.py and teste-aceleracao-v4-rc15-ae9e8fe9.json.) (rc17: RESULTADO-v4.md now ends with a dated note naming the rc16 test bytes; nothing above the note was edited, so every hash it records still resolves.) In the ballast manifest from rc23 (working list 17)

artifact	holds
out/ITT-REGISTRADO-v3-2026-10-05.json · _sprint-2026-10-04/B-registered/estimador_itt_registrado-v3-0aa9202b.py (six switches) · _sprint-2026-10-04/B-registered/RESULTADO-v3.md	the registered analysis as reported in rc10 to rc14, with the one-sample BCa acceleration (superseded in rc15 by the row above, which reproduces it): every leg, the one-switch-at-a-time deltas including the washout and start-epoch switches, the Holm families under the deposited and the switch readings ($m = 5 / 4$ and $m = 6 / 5$) for the registered analysis, the registered sensitivity without boundary-crossing sessions and the rc8- and rc9-equivalent legs, H2, the list of sessions with episodes in more than one epoch, provenance with the sha256 of the script, of its six imported modules and of every input; the controls that rebuild ITT-2026-09-21.json and RERANDOMIZACAO-2026-09-21.json byte for byte and rc8's and rc9's registered legs field for field. Script sha256 0aa9202b... from rc11, frozen at _sprint-2026-10-04/B-registered/estimador_itt_registrado-v3-0aa9202b.py; rc11 corrected four stale comments and docstrings, and the regenerated artifact (41a0a0ee...) equals rc10's in every field except gerado_em and the script sha256. The rc10 bytes of both (b5135f04..., 36f6421a...), which checks-rc10.json and the rc10 Figure B1 pin, are kept in _sprint-2026-10-04/B-registered/(estimador_itt_registrado-v3-rc10-b5135f04.py, ITT-REGISTRADO-v3-rc10-36f6421a.json). rc13 corrected one sentence of RESULTADO-v3.md, which said that no within-stratum contrast of the own stratum is estimable, and rc15 corrected it again. No standalone inferential interval or test is reported for this stratum. The descriptive per-hour contrasts are reported in §4; H1c cannot be contrasted because control has no eligible opportunities. No number changed, and the corrected file is covered by the rc23 ballast manifest through the B-registered/ directory hash. In the ballast manifest from rc23 (working list 17)
out/ITT-REGISTRADO-v2-2026-10-05.json · _sprint-2026-10-04/B-registered/estimador_itt_registrado-v2-ac9d2f05.py · _sprint-2026-10-04/B-registered/RESULTADO-v2.md	the registered analysis of rc9 (five switches; sessions split by epoch), now a sensitivity, and the frozen copy of the script that produced it (sha256 ac9d2f05..., the value its provenance records). In the ballast manifest from rc23 (working list 17)
out/ITT-REGISTRADO-2026-10-05.json · _sprint-2026-10-04/B-registered/estimador_itt_registrado-v1-c28b064f.py · _sprint-2026-10-04/B-registered/RESULTADO.md	the registered analysis of rc8 (four switches; washout in the denominator), and the frozen copy of the script that produced it (sha256 c28b064f..., the value its provenance records). In the ballast manifest from rc23 (working list 17)
ITT-2026-09-21.json	the estimates of the sensitivity analysis (the analysis of versions up to rc7), both legs, bootstrap parameters
ITEM7-DOSE-TOPO-2026-09-21.json	the full 2 000-seed distribution of §4.7
ASSIGNMENT.json · ASSIGNMENT-SERVING.json	designation and served arms
DESIGNATION-2026-08-26.json	the 19 designated items
estimador_itt.py · assign_arms.py · pilot_replay.py	the instruments
p2-serving.ndjson · episodios-ensaio-20260921.jsonl	serving log and episodes
ensaio-20260921-PRIMARIO-3fam.jsonl	3 592 rows: 3 585 verdicts (1 195 per family) and 7 superseded xAI non-verdicts (6 quota, 1 missing), three families
ensaio-20260921-SENSIB-deepseek.jsonl	1 195 verdicts, fourth family (sensitivity)
COBERTURA-M10-2026-09-21.json · cobertura_e_m10.py	coverage by arm, per-signature share, and the designated-presence correlation of §4.6 with its four legs (the file names keep the label M10; it is not the registered M10)
RERANDOMIZACAO-2026-09-21.json · rerandomizacao.py	the registered sharp-null test, 10 000 redesigns (§4.0.1a, §4.0.2)
CONTROLES-JANELA-COMPLETA-2026-09-21.json · controles_instrumento.py	instrument controls over all 19 served epochs, 09-02 having served none (§4.0.1b)
out/CONTROLES-2026-09-10.json	the same controls at the trial's midpoint — 6/6 and 3/3, superseded by the row above
out/H1C-POWER-REALIZADO-2026-09-10.json	the MDE saturation of §4.1, computed before the window closed
out/H1C-POWER-FRACIONARIA-2026-09-10.json · out/H1C-POWER-SEM-0903-2026-09-10.json	the two other inclusion cuts of the power margin (§4.1.1)
out/expiracao-designados-2026-09-09.json	the designation's expiry measurement of §3.0.1 — created_at and the 30-day window
MANIFESTO-LASTRO-P2.json	SHA-256 hashes of the 115 artifacts currently covered, for loss detection: from rc23, every artifact in this table (excepted: the manifest itself, B-censo/raw/ (see its row), and the two records of the final Codex read of rc23, written after the manifest, see their row)

artifact	holds
out/C12-EMPATES-REGISTRADO-2026-10-05.json	the four-vote ties under the registered window and epoch set, the four-vote H1c point under the paper's rule and with ties as failure , and the count of ties in the substitution panel (0); written by checks-rc8.py block C, which aborts unless it reproduces the rc4 counts and points under the sensitivity analysis's window; the rc4 files are unchanged (added in rc8; in the ballast manifest from rc23)
measurement/sprint-figB-h1a-inversao-registrado.py · _sprint-2026-10-04/figures/figB1-h1a-inversao-registrado.{svg,png,run.json}	Figure B1 under the registered analysis of rc8; aborts unless panel (a) reproduces the registered session-hours and panel (b) the registered artifact and checks-rc8.json (added in rc8; in the ballast manifest from rc23)
_sprint-2026-10-04/B-rc10/checks-rc10.py · checks-rc10.json	the registered analysis of rc10 on the two sensitivity legs, per-epoch session-hours and volumes, the dilution ratios, the H1 identity and the BCa quantile ranks (block A); the washout's removals (block W); the sessions with episodes in more than one epoch, their start epoch, arm, span and episodes (block S); the four-vote ties under the rc10 analysis (block C). It aborts unless it first reproduces the atual, registered, 'without' and rc9-equivalent legs of out/ITT-REGISTRADO-v3-2026-10-05.json exactly and rc9's registered leg (block G) (added in rc10; in the ballast manifest from rc23)
_sprint-2026-10-04/B-rc9/checks-rc9.py · checks-rc9.json	the registered analysis of rc9 on the two sensitivity legs, per-epoch session-hours and volumes, the dilution ratios, the H1 identity and the BCa quantile ranks (block A); the washout's removals and 09-14's sessions (block W); the sessions with episodes in more than one epoch (block S). It aborts unless it first reproduces the atual, registered and rc8-equivalent legs of out/ITT-REGISTRADO-v2-2026-10-05.json exactly and rc8's registered leg (block G) (added in rc9; in the ballast manifest from rc23)
_sprint-2026-10-04/B-registered/f4_promoviveis_w4.py · f4-promoviveis-w4.json	the signatures promotable at each w , re-derived from out/dose-350-v3.json and DESIGNATION-2026-08-26.json ; the Epoch-1 coverage readings at w = 2 and w = 4 ; aborts unless the w = 2 set equals the planning artifact's seven and the w = 2 reading reproduces checks-rc8.json block D (added in rc9; in the ballast manifest from rc23)
_sprint-2026-10-04/B-registered/f3_abort_ex_post.py · f3-abort-ex-post.json	the deposited safety abort evaluated ex post over the adjudicated episodes of the window, both panels: panel-majority S3 / S4 counts per epoch and the two clauses (added in rc9; not the registered evaluation; in the ballast manifest from rc23)
measurement/sprint-figB-h1a-inversao-registrado-v3.py · _sprint-2026-10-04/figures/figB1-h1a-inversao-registrado-v3.{svg,png,run.json}	Figure B1 under the registered analysis of rc10; aborts unless panel (a), each session in its start epoch, reproduces the registered session-hours and every epoch of checks-rc10.json , panel (b) agrees with the v3 artifact, and the artifact is the six-switch analysis (added in rc10; in the ballast manifest from rc23)
_sprint-2026-10-04/B-rc15/checks-rc15.py · checks-rc15.json	the checks of checks-rc10.json recomputed with the arm-stratified acceleration (blocks A, W, S and C; only the BCa intervals, adjusted quantiles and ranks differ); block V, which aborts unless the v3 acceleration carried beside each interval gives back checks-rc10.json exactly; block R, rc8's and rc9's analyses on the three legs with the arm-stratified acceleration, which aborts unless the v3 acceleration gives back the intervals of checks-rc8.json and checks-rc9.json . It aborts unless it first reproduces the atual, registered, 'without' and rc9-equivalent legs of out/ITT-REGISTRADO-v4-2026-10-05.json exactly (block G) (added in rc15; in the ballast manifest from rc23)
measurement/sprint-figB-h1a-inversao-registrado-v4.py · _sprint-2026-10-04/figures/figB1-h1a-inversao-registrado-v4.{svg,png,run.json}	Figure B1 under the registered analysis of rc15; panel (a) unchanged from rc10; aborts unless the artifact is v4 (arm-stratified acceleration on every registered interval, v1 and v3 controls identical), checks-rc15.json pins its sha256 and agrees with it on the registered leg, and every epoch of panel (a) reproduces checks-rc15.json (added in rc15; in the ballast manifest from rc23)

artifact	holds
<code>measurement/sprint-figB-h1a-inversao-registrado-v2.py ·</code> <code>_sprint-2026-10-04/figures/figB1-h1a-inversao-registrado-v2.{svg,png,run.json}</code>	Figure B1 under the registered analysis of rc9; aborts unless panel (a) reproduces the registered session-hours without the washout and every epoch of <code>checks-rc9.json</code> , panel (b) agrees with the v2 artifact, and the artifact is the five-switch analysis (added in rc9; in the ballast manifest from rc23)
<code>_sprint-2026-10-04/B-censo/GATES.md · _sprint-2026-10-04/B-censo/GATE-RULES-PREDECLARED.md · _sprint-2026-10-04/B-censo/gate2-ids.txt · _sprint-2026-10-04/B-censo/census-ids.txt · _sprint-2026-10-04/B-censo/run_census.sh · _sprint-2026-10-04/B-censo/SHA256SUMS</code>	the census restoration attempted after unblinding (§4, §7, App. A): the gate rules, written before the first call; the gate verdict (gate 1 passed, gate 2 failed), with per-provider and panel-outcome agreement and the cost; the 50 gate-2 ids (sha256 6377c80b...) and the 4 756 census ids (46c5ad60...); the census wrapper, prepared and not run; and the sha256 of every file of the directory, including the comparison (<code>gate2-compare.json</code> , <code>compare_gate2.py</code>) and the raw panel input and output in <code>B-censo/raw/</code> , which holds episode excerpts and panelist reasons and is kept out of the public repository (added in rc12; in the ballast manifest from rc23, <code>B-censo/raw/</code> excepted). rc13 replaced the absolute personal paths in <code>run_census.sh</code> , <code>gate_harness.py</code> , <code>census_tools.py</code> and <code>compare_gate2.py</code> with paths derived from each script's location and regenerated <code>SHA256SUMS</code> and <code>census-pins.sha256</code> ; nothing else in those files changed, and the sha256 that <code>GATES.md</code> records for <code>gate_harness.py</code> and <code>census_tools.py</code> are those of the bytes that ran the gates, which differ from the current files only in that line
<code>_sprint-2026-10-04/B-rc18/zenodo-22110203-record.json ·</code> <code>_sprint-2026-10-04/B-rc18/zenodo-22110203-files.json</code>	a public snapshot of Zenodo v1.12 (record 22110203), fetched 2026-10-05 for rc18: the record's creation time (2026-08-26T12:01:06Z) and the file list with the md5 and size of every deposited file, <code>DECISION-designacao-2026-08-25.md</code> among them (§1, §3.0.1) (added in rc19; in the ballast manifest from rc23)
<code>_sprint-2026-10-04/B-rc19/zenodo-21978476-record.json ·</code> <code>_sprint-2026-10-04/B-rc19/zenodo-21978476-files.json</code>	a public snapshot of Zenodo v1.11 (record 21978476), fetched 2026-10-05 for rc19: the record's creation time (2026-08-17T18:32:45Z) and the file list, with <code>assign_arms.py</code> at md5 3fa3f710..., the bytes v1.12 carries (§1) (added in rc19; in the ballast manifest from rc23)

B.1 Where each number in the text comes from

Added 2026-09-21 after an adversarial reviewer observed that the header promised universal traceability and several numbers had no artifact named beside them. That review was right, and understated: two of them, the coverage figures and the whole of the correlation then called M10 (§4.6), had no artifact at all; an ad-hoc script computed them, and they were never saved. `cobertura_e_m10.py` now produces them, and reproduces the ad-hoc values exactly.

number	§	artifact
registered analysis: H1, H1a, H1c, every BCa interval and re-randomization p-value of the locked leg; session-hours 324.95 / 57.06, opportunities and repeats per arm; Holm-adjusted p-values under the deposited reading (all 1.0) and under the switch (all 1.0, H1 included; with H1c tested alone, 1.0); the sensitivity analysis's 0.0762 / 0.0635; H2; panel counts 1 179 / 16 / 20 / 13; the expiry cut (33 episodes, 22.835 weighted opportunities, 2 repeats, 0.5902 → 0.4820 h on the registered leg, fields <code>pernas.registrado_sem_janela</code> and <code>pernas.registrado</code>, <code>por_epoch.2026-09-20.horas</code>; the 0.6235 → 0.5154 h of the block <code>corte_pos_expiracao</code> keeps the washout in the denominator and is a diagnostic); the one-switch deltas of §4, the washout's and the start epoch's included; the registered sensitivity without boundary-crossing sessions and its Holm families; the rc9-equivalent leg (H1 0.0302, H1a 0.0168, Holm 0.0840 / 0.0672 and 0.151 / 0.1208 / 0.0906); BCa adjusted quantiles 99.51% / 99.56%; (<i>rc11</i>) H2's percentile intervals (time [−0.011; +0.661], tokens [−390; +10 748], field <code>ic95_percentil</code>); the own stratum's descriptive differences +0.01867 repeated failures/hour and +0.43543 opportunities/hour (registered leg minus the 'without' leg, per arm: 6.00 / 321.43 and 139.96 / 321.43 in treatment, 0 / 54.00 in control) the registered analysis of rc8 (washout in the denominator): H1 −14.12, <i>p</i> = 0.0133; H1a −163.52, <i>p</i> = 0.0241, Holm 0.1205 / 0.0964; session-hours 12.56 / 4.33; the rc8 rows of §4.2 (points and session-hours as rc8 reported them; intervals recomputed in rc15) registered analysis (rc10) on the pre-committed and post-hoc legs; per-epoch session-hours (240.75 h = 74% at 09-08, 81.37 h at 09-12, 54.31 h at 09-11, the other sixteen 0.23-0.62 h); reduction 83.8% (0.246 against 1.520); per-epoch volumes 109.5 / 135.0 and 7.26 / 10.84; H1c 17% and 0.0339; identity error 3.8×10^{-6}; the 50th and 45th largest replicates as BCa upper bounds; the washout's 972 episodes (603 / 369); six sessions in more than one epoch, three starting in the window, their spans and episodes; the four-vote ties 4 / 7 / 9 / 7 / 1 and −0.0140 → −0.0249; (<i>rc11</i>) the latest epoch a boundary-crossing session touches (09-18) and the epochs of the session that starts on 2026-08-27 (09-02...09-07 in the window)	Abstract, 1, 3.0.1, 4-4.3, 5, 8 4, 4.2 Abstract, 3.0.1, 4-4.3, 5, 7	<code>out/ITT-REGISTRADO-v4-2026-10-05.json</code> (rc15, script sha256 b3095740...; its <code>gerado_em</code> timestamp, in UTC, falls on 2026-10-06, the day after the date in the file name; every number but the BCa intervals, accelerations and adjusted quantiles equal to <code>out/ITT-REGISTRADO-v3-2026-10-05.json</code>, added in rc10 and regenerated in rc11 with identical numbers, script sha256 0aa9202b...); the rc9 p-values and points also in <code>out/ITT-REGISTRADO-v2-2026-10-05.json</code> <code>out/ITT-REGISTRADO-2026-10-05.json</code> and <code>_sprint-2026-10-04/B-rc8/checks-rc8.json</code>, block A (added in rc8); from rc15 the §4.2 rows of rc8's and rc9's analyses carry the arm-stratified acceleration, <code>_sprint-2026-10-04/B-rc15/checks-rc15.json</code>, block R <code>_sprint-2026-10-04/B-rc15/checks-rc15.json</code>, blocks A, W, S and C (rc15; the same numbers, but the BCa intervals and ranks, in <code>_sprint-2026-10-04/B-rc10/checks-rc10.json</code>, added in rc10)

number	§	artifact
registered analysis (rc9) on the pre-committed and post-hoc legs; per-epoch session-hours (0.25–0.94 h, 6.79 h = 63%); reductions 73.5% / 43.1% and ratios 23.4 / 13.7, 16.1 / 9.4, 10.7 / 6.3; per-epoch volumes 103.5 / 137.7, 95.5 / 137.7 and 6.08 / 10.96; hours per epoch 0.396 / 0.406; H1c 15% and 0.0351; identity error 6.0×10^{-6} ; the 6th-smallest replicate as BCa lower bound; the washout's 1 007 episodes (638 / 369); four sessions in more than one epoch	Abstract, 4–4.3, 5, 7	_sprint-2026-10-04/B-rc9/checks-rc9.json, blocks A, W and S (added in rc9)
Epoch-1 coverage 45 / 139 = 32.4% (Wilson [25.2%; 40.5%]); 34 / 101, 30 / 75, 54.5% weighted; 92.1% / 89.4% for the 19 designated groups	Abstract, 3.0, App. A	_sprint-2026-10-04/B-rc8/checks-rc8.json, block D, from episodios-ensaio-20260921.jsonl and out/CONCENTRATION-2026-08-30.json (added in rc8)
the ten signatures promotable at $w = 4$; Epoch-1 coverage at $w = 4$ 52 / 139 = 37.4% (Wilson [29.8%; 45.7%]), 40.6% active, 48.0% after the washout, 65.1% weighted; planning corpus 677 / 1 526 = 44.4%	Abstract, 3.0, 7, App. A	_sprint-2026-10-04/B-registered/f4-promoviveis-w4.json, from out/dose-350-v3.json, DESIGNATION-2026-08-26.json and out/CONCENTRATION-2026-08-30.json (added in rc9)
safety abort ex post: 696 adjudicated episodes of the window, 0 at panel-majority S3 or S4, 10 with a panelist at S3 or above, 1 at S4	3.0, App. A	_sprint-2026-10-04/B-registered/f3-abort-ex-post.json (added in rc9); the deployment and its one run in DEPLOY-C2-C3-2026-08-21.md and SHADOW-ARMED-2026-08-21.md
declared vs registered assignment rule: 130 of 234 and 11 of 20 epochs with a different arm; 10 / 10 against 9 / 11; 2426d13d... reproduced	Abstract, 1	_sprint-2026-10-04/B-rc8/checks-rc8.json, block E; commit times 21:33:46Z and 21:56:42Z from the commit log (added in rc8)
sensitivity analysis: H1, H1a, H1c and every interval	4.1–4.3	ITT-2026-09-21.json — except the pre-committed legs:
sensitivity analysis: pre-committed sensitivity of H1c, H1a and H1; session-hours 12.16 / 4.08	4.1–4.3	ITT-SENSIB-PRECOMPROMETIDA.json, field sensibilidade (added 2026-10-04)
sensitivity analysis: session-hours per arm, opportunities, repeats	4.2	idem
7.13 h at 09–14 with the washout in, 6.79 h without; session d37a5964..., 3 episodes, span 6.33 h, started 2026-09-08; 74/65/56 episodes with the washout in, 56/23/21 without	4.2	episodios-ensaio-20260921.jsonl; _sprint-2026-10-04/B-rc9/checks-rc9.json block W and figures/figB1-h1a-inversao-registrado-v2.run.json (rc9)
coverage 27.98% / 26.92%; 4 324 occurrences; 19/19 signatures at $\approx 5.3\%$	4.5	COBERTURA-M10-2026-09-21.json
the fabricated 139 650 that <code>boost_by_id</code> would yield	4.5	idem, field nota_boost
r and all four legs of the designated-presence correlation; -0.0521 against dose	4.6	idem, field M10 (the label of the file, not the registered M10)
truncation/chance split, 2 000 seeds	4.7	ITEM7-DOSE-TOP0-2026-09-21.json
Monte-Carlo s.e. 0.035 of the conditional mean; signed shares -1.0% / $+101.0\%$	4.7	computed from the histogram in ITEM7-DOSE-TOP0-2026-09-21.json; recorded in _sprint-2026-10-04/figures/figB2-sementes-dose-topo.run.json
control conditional mean 9.991	4.7	no artifact — DEVIATIONS-FOR-PAPER.md only; corroborated (10.0124 ± 0.0146) by _sprint-2026-10-04/figures/figB2-item7-crosscheck-20000.json
spans 9.71 / 8.86 / 8.03 min of the three busiest 09–14 sessions	4.2	episodios-ensaio-20260921.jsonl; recorded in _sprint-2026-10-04/figures/figB1-h1a-inversao.run.json

number	§	artifact
every number of §4.0.1c (0/132; 22/22; 17/22; strata; pool 108 / 115)	4.0.1c	<code>_sprint-2026-10-04/B-replay-fidelity/RESUMO.txt</code> and the runs beside it; 11.9 and 1.5×10^{-10} are hypergeometric arithmetic on those counts (<code>B-replay-fidelity.md</code> §4)
~~implied standard errors 0.0279 / 0.0320 (power) and 0.0165 (bootstrap)~~ (struck text only; withdrawn in rc7 as inferential claims)	4.1.1	arithmetic on out/H1C-POWER-REALIZADO-2026-09-10.json (p0, detectavel_no_limite_p1_igual_zero: false) and the H1c interval of ITT-2026-09-21.json; $z_{0.975} = 1.960$, $z_{0.8} = 0.842$ (added 2026-10-04)
power at the sensitivity analysis's 11T/9C (effective size 96.41 against 95.26, relative MDE 0.996, i.e. 80% power only for reductions of about 99.6% or more) and fractional with 09-02 (91.49, margin 4.1%); at the registered 11T/8C, the reproduction of the artifact (90.2)	4.1.1	<code>_sprint-2026-10-04/REVIEW-B-rc2-evidence/power-*.json</code> , from the unchanged <code>measurement/potencia-h1c.py</code> (added in rc3)
sensitivity analysis: H1 reductions 70.7% / 47.1%; 22.5 / 15.0 and 10.3 / 6.9 times total elimination; H1c relative difference -22%	Abstract, 4.0.2, 4.1.1, 4.3	<code>_sprint-2026-10-04/REVIEW-B-rc2-evidence/checks.py</code> on ITT-2026-09-21.json; 3.14% (the pre-trial calibration share at $w = 2$, 11 of 350) from out/H1C-POWER-REALIZADO-2026-09-10.json (added in rc3)
realized altered-brief share 335 / 7 350 = 4.56% in active mode (the artifact's 337 / 7 392 includes 2 shadow changes); 09-01 22 / 630 = 3.49%; ratios 15.5 / 10.3 at the realized share in the sensitivity analysis; coverage 2 057 / 7 350 in active mode	Abstract, 4.0.1b, 4.0.2, 4.3, 4.5	<code>_sprint-2026-10-04/B-rc7/checks-rc7.json</code> (blocks A, B), from <code>p2-serving.ndjson</code> , after reproducing the artifact's 337 / 7 392 and 2 068 / 7 392 (added in rc7)
<code>p1 = 0.25</code> reaches about 86% (84.8) and 89% (91.49, the sensitivity analysis's set); 09-01 by clock gives 84.78 / 91.47	4.1.1	<code>_sprint-2026-10-04/B-rc7/checks-rc7.json</code> (block C), with <code>poder_z</code> copied unchanged from <code>measurement/potencia-h1c.py</code> (added in rc7)
adjudication sample, declared as a deviation from the registered census (rc11): 395 <code>is_error</code> episodes (census) and 800 of the 5 556 others (sample), Horvitz-Thompson weight $6.945 = 5\,556 / 800$	4, App. A	ITT-2026-09-21.json (<code>n_estrato_b_amostrado</code> , <code>n_resto_no_corpus</code> , <code>peso_estrato_b</code>); <code>episodios-ensaio-20260921.jsonl</code> (5 951 episodes, 395 with <code>is_error</code>) and <code>estrato-b-ids-20260921.txt</code> (800 ids), counted in rc11; the lock in <code>PREREG-DRAFT.md</code> §3
census restoration (rc12): 4 756 census ids; rules file created and last modified 23:13:58Z, first call 23:14:49Z, header time 2026-10-05T23:15Z; gate 2 on 50 ids, label agreement 50/50, 47/50, 46/50 (Wilson [0.93; 1.00], [0.84; 0.98], [0.81; 0.97]); panel outcome 46/50 (Wilson [0.81; 0.97]); criteria 0.99 and 0.90; the 99/100 test-retest; 154 calls, about US\$1.18	4, 7, App. A	<code>_sprint-2026-10-04/B-censo/GATES.md</code> , <code>GATE-RULES-PREDECLARED.md</code> (and its file-system time), <code>gate2-compare.json</code> , <code>census-ids.txt</code> ; the call times and the usage totals in <code>B-censo/raw/</code> (<code>gate1-calls.jsonl</code> , <code>cost-gate1.json</code> , <code>cost-gate2.json</code> ; out of the repository); <code>STABILITY-TEST.md</code> §7 (added in rc12)
gate-2 rules, declared noise property (rc13; 4 independence stated in rc15): if the 150 agreement indicators were independent Bernoulli draws with agreement probability 0.99, the point criterion over 150 pairs would fail with probability $1 - 0.99^{150} \approx 0.779$; the 99/100 measurement's Wilson [0.9455; 0.9982]; <code>GATES.md</code> dates gate 1 from 23:14:52Z, the second call	4	<code>_sprint-2026-10-04/B-censo/GATE-RULES-PREDECLARED.md</code> , <code>GATES.md</code> ; the call times in <code>B-censo/raw/gate1-calls.jsonl</code> (out of the repository) (added in rc13)
1 195 submitted, 1 159 with three substantive verdicts, 36 <code>unknown</code> (three families; the registered substitution counts are in the first row)	4, 6	<code>_sprint-2026-10-04/B-rc7/checks-rc7.json</code> (block D), through <code>carregar_verdicts</code> (added in rc7)

number	§	artifact
93.8% = 573 / 611 covered action opportunities	4.5	out/CONCENTRATION-2026-08-30.json; _sprint-2026-10-04/B-rc7/checks-rc7.json (block E) (added in rc7)
ITT-PRELIMINAR.json: file time 16:16:44 local (19:16:44Z), 2 000 replicates, H1a -143.92	1.1	the file itself and _sprint-2026-10-04/B-rc7/checks-rc7.json (block F) (added in rc7)
1 056 s (designation round 31657512); 21:18:35Z and 21:32:04Z, 13 min 29 s (assignment round 31774052); commits of 16:51:56Z and 20:51:20Z	1	DESIGNATION-2026-08-26.json (declaracao, emissao_de_R); ASSIGN-SEED-2026-08-30.md; the commit log (added in rc7)
v1.12 published 12:01Z; md5 35abeb68... and 7 036 bytes of the deposited DECISION-designacao-2026-08-25.md, the blob of commit d42f950; sig_primary dropped from the key at 19:40Z; decision 2 h 46 min after publication	1, 3.0.1	_sprint-2026-10-04/B-rc18/zenodo-22110203-record.json (created) and _sprint-2026-10-04/B-rc18/zenodo-22110203-files.json (the file's checksum and size); DESIGNATION-SEED-2026-08-26.md 1.31 (14:47Z) and 1.96-97 (19:40Z) (added in rc19)
assign_arms.py last changed 2026-08-17; Zenodo v1.11 (record 21978476) published 2026-08-17T18:32Z with that file at md5 3fa3f710..., the bytes v1.12 carries; thirteen days before round 31774052 was emitted	1	_sprint-2026-10-04/B-rc19/zenodo-21978476-record.json (created) and _sprint-2026-10-04/B-rc19/zenodo-21978476-files.json; _sprint-2026-10-04/B-rc18/zenodo-22110203-files.json; commit 3199ec1 (added in rc19)
opportunities per epoch 109.5 / 135.0 (registered; 1 203.98 / 11 and 1 079.91 / 8); 103.5 / 137.7 in rc9's analysis (1 138.47 / 11 and 1 101.75 / 8) and 100.3 / 132.3 (sensitivity analysis)	7	out/ITT-REGISTRADO-v3-2026-10-05.json; out/ITT-REGISTRADO-v2-2026-10-05.json (unchanged from rc8); ITT-2026-09-21.json (1 103.75 / 11 and 1 191.14 / 9) (added in rc3; registered in rc8; rc10)
09-20: 385 briefs before expiry, 98 with a designated item (25.5%); 287 after, 0	4.6	_sprint-2026-10-04/REVIEW-B-rc2-evidence/checks.py on p2-serving.ndjson and DESIGNATION-2026-08-26.json (added in rc3)
sham: 132 / 146 against 81-122 / 84-132; p = 0.0476; 155 at w = 100 000; 2,646/2,646; 110/110	4.0.1b, 4.0.1c	_sprint-2026-10-04/B-sham-v2/job-v2b/RESUMO.json; 110/110 and the 36 / 89 counts (53 = 89 - 36) in B-sham-v2/RESUMO.txt; cost and state counts (19,567; 11,865) in B-sham-v2/REPORT.md §3 (added in rc3)
whole-window sham: 790 / 868 against 449-583 / 484-612, 0 shams $\geq$ real, p = 0.0476; 894 against 462-591 at w = 100 000; 11,812/11,812 and 2,016/2,016; 235 states without a bonus, all after the expiry; 207 = 790 - 583; 11,812 = 11,865 - 53 states of 18 epochs, 09-01 excluded	Abstract, 4.0.1b, 4.0.1c, 7	_sprint-2026-10-04/B-sham-v2/job-janela2/RESUMO.json (rc23); the 4,032/4,032 shared records per run and the 400/400 of the calibration sample in B-sham-v2/job-janela2/DETERMINISMO.json; the state lists, the exclusions, the hashes and the completion in B-sham-v2/JANELA-LANCAMENTO.md (added in rc4; result in rc23)
four-vote ties: 28 (10 treatment, 11 control, 7 outside the window); 4 / 7 of them H1c opportunities, 0 failures among those; 3 change a label	4, 6, 7	out/C12-EMPATES-POR-BRACO-2026-10-05.json, from ensaio-20260921-*.jsonl, episodios-ensaio-20260921.jsonl and ASSIGNMENT-SERVING.json (added in rc4)
four-vote ties under the registered window and set: 10 treatment, 10 control, 7 outside the dates, 1 before 09-03's exposure window; 4 / 7 tied opportunities; four-vote H1c -0.0121 $\rightarrow$ -0.0211 with ties as failure; 0 ties in the substitution panel	4, 7	out/C12-EMPATES-REGISTRADO-2026-10-05.json (added in rc8)

number	§	artifact
Figure B1 (rc10): 324.95 / 57.06 h; 240.75 h = 74% at 09-08, one session of 60 episodes spanning 240.42 h; 09-12 81.37 h (67 episodes, 81.01 h) and 09-11 54.31 h (2 episodes, 54.00 h); the other sixteen epochs 0.23–0.62 h	4.2	_sprint-2026-10-04/figures/figB1-h1a-inversao-registrado-v3.run.json (added in rc10) and, with the panel (b) intervals of rc15, _sprint-2026-10-04/figures/figB1-h1a-inversao-registrado-v4.run.json
Figure B1 (rc9): 10.75 / 3.25 h; 6.79 h = 63%; other eighteen epochs 0.25–0.94 h; busiest sessions 56 / 23 / 21 episodes within 1.1–8.0 min	4.2	_sprint-2026-10-04/figures/figB1-h1a-inversao-registrado-v2.run.json (added in rc9); rc8's 12.56 / 4.33 h and 7.13 h = 57% in figB1-h1a-inversao-registrado.run.json
tied opportunities' HT weights 27.78 / 36.73; four-vote H1c point difference under the sensitivity analysis's window −0.0208 under the paper's rule, −0.0266 with ties resolved as failure (rerun −0.026571, first order −0.026566; the sixth decimal differs because the rerun uses a per-arm proportion stored rounded, 0.09189)	4, 7	out/C12-EMPATES-COMO-FAILURE-2026-10-05.json (por_braco.*.empates_oportunidade_peso; contrafactual_empate_como_failure), the same script with --empates-como-failure-out (added in rc4, after the independent check)
fractional effective size 84.8 over the 19 served epochs; the spec's own 96.4 for 9 control epochs	4.1.1	out/H1C-POWER-FRACIONARIA-2026-09-10.json; SPEC-ANALISE-2026-09-10.md §3 (added in rc4)
agreement 1 111/1 145; abstentions 26+6+1 / 13 / 11; gains 20, losses 28	6	ensaio-20260921-*.jsonl, recomputed in DEVIATIONS-FOR-PAPER.md §10.31
20 designated, 19 served, 16+3+1	3	ASSIGNMENT-SERVING.json + p2-serving.ndjson

Several of these are outside the repository and several are large. `scripts/manifesto-lastro-p2.py` hashed the first 20 (154 MiB) on 2026-09-21; `scripts/estende-lastro-p2.py` appended 30 more on 2026-10-05 and 65 more on 2026-10-07, each time without regenerating it, for 115 artifacts and 176 MiB (sha256 d310e4e9...; the earlier versions kept). The 65 of rc23 are the whole-window sham and every artifact added from rc7 to rc23 that this paper cites (ITT-PRELIMINAR.json, the registered analyses v1 to v4 with their scripts and B-registered/, checks-rc7 to checks-rc15, the four registered versions of Figure B1 with their scripts, out/C12-EMPATES-REGISTRADO-2026-10-05.json, the census-attempt files of B-censo/ without raw/, STABILITY-TEST.md, the Zenodo snapshots in B-rc18/ and B-rc19/, and receipts/). `scripts/backup-lastro-p2.sh` copies the 79 that live outside the repository, with the hash recomputed **at the destination**; the other 36 are versioned in the repository and were recomputed from `origin/main`, **36/36**. The inputs of the registered analysis are the ones the manifest hashed on 2026-09-21, and its provenance records their sha256. B-censo/raw/, which holds episode excerpts and panelist reasons, is not in the manifest, nor are the two records of the final Codex read of rc23 (Appendix B). **Done: both legs now verified at the destination** (2026-09-22 01:48Z, again 2026-10-05 13:03Z after item 10, and 2026-10-07 12:02Z after item 17, receipt RECIBO-ITEM15-20261007T120404Z.txt): local **12/12**, then **16/16**, then **79/79**, off-machine **12/12**, then **16/16**, then **79/79** on a host that is not the one that served the trial: it has no `/root/.openclaw`, its epoch pointer is frozen at 2026-08-23, and it already holds Paper 1's ballast. Identified by **capability**, **never by address**. A manifest proves the bytes are the bytes; it does not prove a copy exists, which is why the copy is verified separately and carries a dated receipt.

Working list: struck in the commit that closes it

1. ~~**Ballast:** manifest + two verified copies~~ → **Done, 2026-09-22:** 20 artifacts hashed, 154 MiB; the 12 that live outside the repository copied, **both legs 12/12 recomputed at the destination**; the other 8 versioned in the repository. Three defects of ours found and fixed along the way: a directory hash reimplemented in shell that diverged by one trailing newline, an `rsync` that ran *before* the check and so restored the corrupted byte it was meant to detect, and an `ssh` inside a `while read` that ate the loop's stdin and verified **1 of 12** while reporting "0 divergem". **Caveat (2026-10-04):** the manifest does not

- cover ITT-SENSIB-PRECOMPROMETIDA.json, the only source of the pre-committed rows of §4.1–4.3, so a published number rests on an artifact the ballast does not protect. Open as item 10; closed 2026-10-05.
2. **Run the registered re-randomization** → **Done, 2026-09-21** (§4.0.1a, §4.0.2): 10 000 redesigns, 9 941 distinct patterns, control of 300/300 reproducing the spec. It **disagrees with the bootstrap on H1a**, and H1 is now reported as an unexplained rejection.
 3. **Report the pre-committed instrument controls** → **Done, two of three, 2026-09-21** (§4.0.1b): positive **11/11**, negative dual **8/8**, both re-measured over the full window. **Critical. Still open: the sham replay**, which needs the serving code re-executed; our attempt to do it by counting over the log was invalid and is recorded as such. **Correction (2026-10-04): the sham was not executed** (§4.0.1c). The run launched on 2026-09-22 produced no output, and its configuration was invalid: wrong corpus, inclusive cut, shams drawn from the 115-pool instead of the 108-pool. On the served corpus with the exact cut, the replay reproduces production 22/22 and responds to dose. ~~Still open as item 9.~~ → **Done, rc3 (2026-10-04)**: the corrected sham ran (item 9); all three pre-committed controls are now reported.
 4. **Measure the 30-day freshness question** → **Done, 2026-09-21** (§3.0.1): a dosed $N = 234$ was infeasible: the designation expires 2026-09-20 22:51:23, giving 20 eligible epochs of 234. §9 (*numbered §8 before v2*) rewritten around it. Measured 2026-09-09; incorporated into the manuscript 2026-09-21. Only dose eligibility, not continued data collection, was limited to 20 epochs.
 5. **Adversarial review of this manuscript** → **Done, 2026-09-21**: five families launched, **four delivered with exit: 0 receipts** (DeepSeek, Grok, GLM on this paper; Kimi on Paper A). Their findings produced §1.1, §3.0, §3.0.1, §4.0.1, §4.0.2, §4.1.1, §4.1.2, Appendix B.1 and the corrections in §3.1, §4.1–4.5. The fifth (Codex) returned a review whose citations do not resolve against the file and is recorded as invalid in REVISAO-ADVERSARIAL-2026-09-21.md; a review that cites sections the document does not have is not evidence about the document.
 6. **Figures**: the H1a sensitivity (the single-epoch inversion) and the §4.7 seed distribution. ~~Both derive from locked artifacts.~~ → **Done, 2026-10-04**: Figure B1 (§4.2) and Figure B2 (§4.7), each generated by a script that aborts if its values diverge from the locked artifacts. Making them surfaced four corrections to the text (§4.2 pre-committed leg; “8 to 10 minutes”; the signed truncation share; the unresolved sign of +0.023); see the changelog. **Caveat**. The headline embedded in the B1 SVG still reads “*H1a turns on one epoch*”; regenerate it with the caption title of §4.2 (item 11). (*rc27, 2026-10-07: the caveat is closed by item 11’s rc8 note; the registered Figure B1 carries a headline that matches its caption, and the old SVG is kept as a record.*)
 7. **Related work** — Paper A’s §8 covers the surface literature, not trials of memory interventions. → **Done, 2026-10-04**: §8, condensed (~2 000 words) from `_sprint-2026-10-04/B-related-work.md` (~4 100 words), with references resolving in `B-related-work.bib`. Before submission: re-check Behnam & Wang (arXiv 2610.02070, 1 Oct 2026, not peer-reviewed) for a revised version, and re-read MemoryArena v2 in full. (*rc27, 2026-10-07: neither check was done; both concern submission to a venue, not this deposit, and §8.5 states the reading depth, MemoryArena v2 only string-searched.*)
 8. **Deposit as a new version of the registration, declaring the deviations and the result in one record**. Blocked by 9 (valid sham) and 10 (ballast gap) and 15 (whole-window sham). Also blocked by 17 (rc7 ballast) and 18 (registered re-analysis), added in rc7; the record must declare the primary switch of §1. (*rc8: 18 is done; 17 now also covers the rc8 artifacts. The record must also declare the unexecuted stopping rule of §3.0 and the replaced assignment rule of §1.*) (*rc19: 17 is still open and still blocks this deposit; it now also covers the rc15–rc18 artifacts and the Zenodo snapshots in B-rc18/ and B-rc19/.*) (*rc23: 15 and 17 are done, so none of the items named here blocks the deposit any longer; the deposit itself is not done.*) → **Done (rc27)**: deposited as version 2.0 of the registration record, version DOI 10.5281/zenodo.23223385. (*rc28: the note no longer dates the deposit; publication is done by hand, after this text is frozen.*)
 9. **Valid sham**: needs the trial `brief_log` 2026-09-08..09-20 from the production host (authorization pending). Everything else for it is in hand: served corpus, `rowid` cut, shams redrawn from the 108-pool (89 non-designated). `gera-shams.py` / `roda-sham.sh` must first be fixed, since both hard-code the preserved corpus and `--corte` *inclusivo* (§4.0.1e). → **Done, 2026-10-04** (§4.0.1c): new generator, runner, test and summarizer (`sprint-*-sham-v2*`; the old scripts untouched); 21 runs on the served corpus with the `rowid` cut, 2,646 states of the $w = 4$ epochs, shams drawn from the 36 boostable items;

- real 132 against 81–122, $p = 1/21$. The larger set (the 11,865 states of the trial window, about 36 h 11,812 reconstructible states of the trial window) is configured and not run is running as item 15 was run as item 15 (rc23): real 790 against 449–583, $p = 1/21$.
10. **Ballast gap:** add `ITT-SENSIB-PRECOMPROMETIDA.json` and the artifacts marked † in Appendix B to `MANIFESTO-LASTRO-P2.json` and to both verified copies. → **Done, 2026-10-05:** 30 entries appended without regenerating the manifest: the 24 † rows, the five `out/` rows of 2026-09-09/10 that had never been hashed, and `B-sham-v2/RESUMO.txt`. The existing entries are byte-identical. It now hashes **50** artifacts (165 473 805 bytes = 158 MiB; sha256 172c5382..., previous versions kept). The 4 new ones that live outside the repository (`ITT-SENSIB-PRECOMPROMETIDA.json` and the three C12 files) are in both copies, **16/16 recomputed at the destination on each leg**, after a pre-check on both legs with nothing copied (receipts `RECIBO-ITEM10-20261005T125933Z.txt`, `RECIBO-ITEM10B-20261005T130316Z.txt`). The other 26 are versioned, **26/26 recomputed from origin/main**. The whole-window sham (item 15) and `JANELA-LANCAMENTO.md` are not yet in it; they enter together when the job completes. (*rc23: both entered on 2026-10-07, with the rc7–rc23 artifacts; item 17.*)
 11. **Traceability of §4.7 and Figure B1:** save the control distribution (the 9.991) and the seed recipe of the 2 000-seed run in an artifact, or cite the 20 000-seed crosscheck instead; regenerate the B1 SVG headline. (*rc8: Figure B1 is now `figB1-h1a-inversao-registrado.svg`, generated with a headline that matches its caption; the old SVG is kept unchanged as the record of the sensitivity analysis. The §4.7 part stays open.*) → **Closed, rc27 (2026-10-07), as a declared gap:** the control distribution and the seed recipe were not saved, so 9.991 stays the one number held by no artifact; §4.7 and B.1 declare it and cite the 20 000-seed crosscheck as a corroboration, not a reproduction.
 12. **Second adversarial review of this manuscript** → **Done, 2026-10-04:** five families launched, two valid (DeepSeek, Gemini); GLM, Grok and Kimi did not receive the file. 28 findings, 18 confirmed and applied in rc2, 2 rejected (`_sprint-2026-10-04/REVIEW-B-2026-10-04.md`).
 13. **Two computations rc2 declares as not done:** the H1e MDE at the ITT’s 20 clusters (9 control epochs; §4.1.1), and M10 under the pre-committed leg that removes all three partials (§4.6). (*rc3: the first is done; §4.1.1 reports it.*) (*rc9: the correlation of §4.6 is not the registered M10, which is not computable from the locked artifacts; §5.*) → **Closed, rc27 (2026-10-07), the second not computed:** §4.6 declares the all-partial leg not computed for its correlation, and §5 declares the registered M10 not computable.
 14. **Third adversarial review of this manuscript (rc2)** → **Done, 2026-10-04:** one valid voice of four (Codex); 24 findings, 22 confirmed and applied in rc3, 2 rejected (`_sprint-2026-10-04/REVIEW-B-rc2-2026-10-04.md`). rc3 itself has not been reviewed. (*rc27, 2026-10-07: superseded; rc4, built on rc3, had the regression review of item 16.*)
 15. **Whole-window sham:** the replay over the trial window, 11,812 reconstructible states of 18 epochs (without epoch 09–01, whose corpus has no hash proof, and without 53 states whose serve-state cut cannot be reconstructed), job `job-janela2`, relaunched 2026-10-05T09:39:53Z, expected to finish about 2026-10-07T03:00Z. Add its result to §4.0.1c, §7 and B.1 before deposit. → **Done, rc23 (2026-10-07):** `job-janela2` (launcher 2026-10-05T09:39:53Z; runner start 2026-10-05T09:40:32Z in `RECIBO.txt`, added in rc24) `CONCLUIDO` 2026-10-07T02:09:46Z, 21/21 runs validated, instrument and databases re-hashed on the host after completion; real 790 changed states against 449–583 (churn 868 against 484–612), 0 shams at or above the real designation, $p = 1/21$; positive control 894 against 462–591; the 2,016 states shared with the first run identical in all 21 runs. Reported in §4.0.1b, §4.0.1c, §7, §8.3, §8.5, the Abstract and B.1 (`_sprint-2026-10-04/APPLY-B-rc23.md`).
 16. **Regression review of rc4** → **Done, 2026-10-05:** one voice (Codex), 5 findings, 5 confirmed and applied in rc5 (`_sprint-2026-10-04/APPLY-B-rc5.md`). rc5 itself has not been reviewed. (*rc27, 2026-10-07: superseded; rc6, built on rc5, had the two full reads of item 19.*)
 17. **Ballast for the rc7 and rc8 evidence:** add `ITT-PRELIMINAR.json` (the preliminary run that dates the §10.33 decision, §1.1) and `_sprint-2026-10-04/B-rc7/checks-rc7.json` with its script to `MANIFESTO-LASTRO-P2.json` and to both verified copies; (*rc8*) also `out/ITT-REGISTRADO-2026-10-05.json` with `measurement/estimador_itt_registrado.py` and `B-registered/RESULTADO.md`, `B-rc8/checks-rc8.json` with its script, `out/C12-EMPATES-REGISTRADO-2026-10-05.json`, and the registered Figure B1 with its script; (*rc9*) also `out/ITT-REGISTRADO-v2-2026-10-05.json` with the five-switch script, the frozen rc8 script, `B-registered/RESULTADO-v2.md`, `B-rc9/checks-rc9.json`

- with its script, the two B-registered/f* analyses with their scripts, and the rc9 Figure B1 with its script; (rc10) also out/ITT-REGISTRADO-v3-2026-10-05.json with the six-switch script, the frozen rc9 script, B-registered/RESULTADO-v3.md, B-rc10/checks-rc10.json with its script, and the rc10 Figure B1 with its script; (rc11) also the frozen v3 script B-registered/estimador_itt_registrado-v3-0aa9202b.py and the rc10 bytes kept beside it (estimador_itt_registrado-v3-rc10-b5135f04.py, ITT-REGISTRADO-v3-rc10-36f6421a.json); (rc15) also out/ITT-REGISTRADO-v4-2026-10-05.json with the frozen v4 script (B-registered/estimador_itt_registrado-v4-b3095740.py), B-registered/RESULTADO-v4.md, B-registered/teste_aceleracao_v4.py with teste-aceleracao-v4.json, B-rc15/checks-rc15.json with its script, and the rc15 Figure B1 with its script; (rc16) also the revised acceleration test (B-registered/teste_aceleracao_v4.py with teste-aceleracao-v4.json) and the rc15 bytes kept beside it. (rc19) None of the rc15-rc18 artifacts listed here is in the manifest yet (checked 2026-10-05: no B-rc15/ to B-rc18/ path among the 50 artifacts of MANIFESTO-LASTRO-P2.json), and neither are the Zenodo snapshots B-rc18/zenodo-22110203-{record,files}.json and B-rc19/zenodo-21978476-{record,files}.json, which §1, §3.0.1 and B.1 now cite; until they are, this item blocks the deposit (item 8). → **Done, rc23 (2026-10-07)**: 65 entries appended under the label item15-janela2 without regenerating the manifest, the 50 existing ones byte-identical: every artifact listed here, the whole-window sham of item 15 with JANELA-LANCAMENTO.md, STABILITY-TEST.md and receipts/ (B-censo/raw/ left out: it holds episode text). The manifest now hashes **115** artifacts (184 509 309 bytes = 176 MiB; sha256 d310e4e9..., previous versions kept). The 79 that live outside the repository are in both copies, **79/79 recomputed at the destination on each leg**, after a pre-check with nothing copied (16/16 on each leg); the 36 versioned ones, **36/36** recomputed from origin/main (receipt RECIBO-ITEM15-20261007T120404Z.txt, with recibos/backup-20261007T120120Z.txt and ...T120248Z.txt).
18. **Registered re-analysis**: replace every REANALISE block of rc7 with the analysis as registered (panel substitution, the expiry cut with the spec's offsets, BCa with the leave-one-epoch-out jackknife, 19 epochs without 09-02) and report the analysis of this version as a sensitivity; state why H2 and H3 are not reported (§5). → **Done, rc8 (2026-10-05)**: all 108 blocks replaced (85 values, 23 sentences), the rc7 analysis reported as the sensitivity, H2 reported (§5), H3 declared not computable (_sprint-2026-10-04/APPLY-B-rc8.md).
 19. **Review of rc6** → **Done, 2026-10-05**: two full reads (Codex, Fable), both NO-GO; findings verified against the artifacts and applied in rc7, except those that depend on the registered re-analysis, which are marked REANALISE (item 18; _sprint-2026-10-04/APPLY-B-rc7.md). rc7 itself has not been reviewed. (rc27, 2026-10-07: superseded; rc8, built on rc7, had the two full reads of item 20.)
 20. **Review of rc8**: the registered analysis, the two findings of §3.0 and §1 (the unexecuted stopping rule and the replaced assignment rule) and the rc8 sweep have not been reviewed by any voice. → **Done, 2026-10-05**: two full reads (Codex, Fable), both NO-GO; 12 findings, each verified against the code, the artifacts and the registration and applied in rc9 (_sprint-2026-10-04/REVIEW-B-rc8-2026-10-05.md, _sprint-2026-10-04/APPLY-B-rc9.md).
 21. **Review of rc9**: the washout switch, the multiplicity under both readings, the $w = 4$ reading of the stopping rule, the safety-abort record and the sweep have not been reviewed by any voice. → **Closed, rc27 (2026-10-07), superseded**: rc9 had no read of its own; rc10, built on it, had the two full reads of item 23, and the later candidates the reads of item 24.
 22. **Boundary-straddling sessions**: implement the registered attribution of a session to the epoch of its start (PREREG §2), which only the 'without' leg approximates here, and decide where its result is reported (§5). It moves H1 and H1a, whose intervals contain zero on the 'without' leg. → **Done, rc10 (2026-10-05)**: the author decided that the analysis reported is the registered one, so PREREG §2's attribution is the sixth switch of measurement/estimador_itt_registrado.py and is on in the registered analysis (out/ITT-REGISTRADO-v3-2026-10-05.json); the 'without' leg is its registered sensitivity. Reported in §2, §4, §4.2, §4.3 and §5; the three sessions that start in the window and last longer than one epoch are listed as their own stratum in §4.
 23. **Review of rc10**: the sixth switch, its effect on H1, H1a and H2, and the rewritten §4.2 and §4.3 have not been reviewed by any voice. → **Done, 2026-10-05**: two full reads (Fable, Codex), both NO-GO; 16 findings, one raised by both, each verified against the code, the artifacts and the registration and applied in rc11 (_sprint-2026-10-04/REVIEW-B-rc10-2026-10-05.md, _sprint-2026-10-04/APPLY-

B-rc11.md).

24. **Review of rc11:** the census deviation, the stratum contrasts and the rc11 sweep have not been reviewed by any voice. (rc12: *nor has the census attempt of §4, §7 and Appendix A.*) (rc13: rc12, which contains all of the above, had one full read by Fable (REVIEW-B-rc12-2026-10-05.md, NO-GO: two MEDIUM and five LOW, all verified and applied in rc13); the Codex read was not run; rc13 itself has not been reviewed.) (rc15: rc14, which contains rc13, had two full reads (REVIEW-B-rc14-2026-10-05.md): Fable GO, Codex NO-GO with three MEDIUM and three LOW, and Fable two LOW; all verified and applied in rc15. rc15 itself, including the arm-stratified acceleration, has not been reviewed.) (rc16: rc15 had two full reads (REVIEW-B-rc15-2026-10-05.md): Fable GO with four LOW and an optional note, Codex NO-GO with two MEDIUM and two LOW; all verified and applied in rc16. rc16 itself has not been reviewed.) (rc17: rc16 had two full reads: Codex NO-GO with one MEDIUM and two LOW, Fable with three LOW; all verified and applied in rc17. rc17 itself has not been reviewed.) (rc18: rc17 had one full read, Fable NO-GO with three MEDIUM and four LOW and no HIGH; all verified, and applied or adapted in rc18 (APPLY-B-rc18.md). rc18 itself has not been reviewed.) (rc19: rc18 had one full read, Fable GO with six LOW; all verified and applied in rc19 (APPLY-B-rc19.md). rc19 itself has not been reviewed.) (rc20: rc19 had one full read, Codex NO-GO with one MEDIUM and two LOW; all verified and applied in rc20 (APPLY-B-rc20.md). rc20 itself has not been reviewed.) (rc21: rc20 had one full read, Fable NO-GO with two MEDIUM and six LOW and no HIGH; all verified, and applied or adapted in rc21 (APPLY-B-rc21.md). rc21 itself has not been reviewed.) (rc22: rc21 had one full read, Codex GO with three LOW (adversary-receipt-codex-2026-10-06T133927-60502.txt); all verified and applied in rc22 (APPLY-B-rc22.md). rc22 is the text frozen until the whole-window sham result is integrated.) (rc23: rc23 integrates the whole-window sham result and closes items 15 and 17 (APPLY-B-rc23.md); it has not been reviewed.) (rc24: rc23 had two full reads, both GO with no MEDIUM or HIGH: Fable (an agent run, four LOW) and Codex (adversary-receipt-codex-2026-10-07T093115-73165.txt, REVIEW-B-rc23-codex-2026-10-07.md, five LOW); all verified and applied in rc24 (APPLY-B-rc24.md). rc24 changes wording only; it has not been reviewed.) (rc25: Fable's diff check of rc23 → rc24 (an agent run, no file) gave GO with three LOW in Appendix B and B.1; all verified and applied in rc25 (APPLY-B-rc25.md). rc25 changes wording only; it has not been reviewed.) → **Closed, rc27 (2026-10-07):** rc24 had Fable's diff check (rc23 → rc24), GO. rc25 changes wording only and rc26 changes only the status header (and adds its changelog block); neither had a read of its own, and each passed a parity check against its predecessor (B-rc25/parity-rc25.py, B-rc26/parity-rc26.py). rc27 changes only this working list and the changelog (B-rc27/parity-rc27.py). The deposited package gets a final read by Grok, Codex and Fable before publication. (rc28, 2026-10-07: what happened instead. The package built from rc27 had two final reads: Grok (adversary-receipt-grok-2026-10-07T181540-7884.txt, in receipts/), NO-GO with six MEDIUM and five LOW, and Fable (an agent run, no receipt), NO-GO with one HIGH, four MEDIUM and two LOW. Every finding was verified against the text, the artifacts and the public Zenodo API, and applied in rc28 or in the deposit description (APPLY-B-rc28.md, B-rc28/parity-rc28.py). The Codex read did not run: the call exited 1 after 4 s, the account being out of credits (adversary-receipt-codex-2026-10-07T181828-14394.txt, in receipts/). rc28 itself has not been read; a confirmation read of it is planned before publication.)
25. **The registered census of the live study** (§4, Appendix A): restoring it needs the 4 756 stratum-B episodes not in the sample adjudicated by the panel, a paid run; the author's decision is pending, and until then the sample is declared as a deviation. → **Closed, rc12 (2026-10-05): attempted, gates failed, not run.** The author authorized the gates and, only if they passed, the census. Gate 1 (model identity) passed and gate 2 (test-retest, 50 stratum-B episodes) failed for google and zhipu, so the census was not started and the sample stays a declared deviation (§4, §7, Appendix A; _sprint-2026-10-04/B-censo/GATES.md). The census wrapper is prepared and refuses to run without an explicit override of the failed gates.

Caveat. Items 2 and 3 were **registered analyses missing from the first draft**, not enhancements; the sham (item 9) was run in rc3. A manuscript that omits its own pre-registered inference test and its own instrument controls is not ready, and the fact that three adversarial reviewers had to tell us is itself the §6 pattern repeating on this document.

Changelog: v2 (2026-10-04)

Prepared from MANUSCRIPT-B.md as of 2026-10-03, in `_sprint-2026-10-04/B-v2.md`; the original is untouched. Release candidate rc2 is `_sprint-2026-10-04/B-v2-rc2.md`, prepared from the form pass B-v3.md (a file name only; the manuscript version is v2). Every content change below traces to a sprint report with its artifacts; old text that was wrong is kept struck through beside its correction, as this manuscript already does.

Sham / replay fidelity (source: `_sprint-2026-10-04/B-replay-fidelity.md` and `B-replay-fidelity/`):

1. §4.0.1b table, specificity row: “running since 2026-09-22 01:40Z” struck → “not executed; the configuration launched was invalid; a valid one is runnable and not yet run”.
2. §4.0.1b: the “It is now running” paragraph struck, with a correction note; the bullets on the 30-day pool and on the “two routes, one number” control annotated: both routes read the same, wrong corpus, so their agreement could not catch it (served pool 108, not 115).
3. §4.0.1b limitation: the sentence calling `corpus-preservado-20260908.db` “the deliberately preserved corpus of the trial” struck and corrected: it is not the corpus that served (60 `memory/lessons.md` chunks re-ingested 2026-09-07 that production never had; served corpus = fd-recovered `e20260903`).
4. New §4.0.1c. The “no power by construction / $p = 1.0$ ” reading of 2026-09-23 is presented as superseded: the tie hypothesis is **refuted**. The zero came from (1) the wrong corpus and (2) the inclusive cut. With the served corpus and the `rowid` cut, the replay reproduces production 22/22 and responds to dose: 17/22 states at $w = 2$, 22/22 at $w = 100\ 000$. Also declared: `gera-shams.py` drew from the 115-pool rather than the 108-pool (a prediction of spurious “failed specificity”, not a measurement). A valid sham is runnable and **not run**: it needs the trial `brief_log` after 2026-09-08, which exists only on the production host. Measured: 22 states, all in epoch 09-01. Not measured: 110 of 132.
5. Working list item 3 updated; new item 9 (valid sham, authorization pending).

Figures (source: `phase2-inputs/B-figures.json`, `_sprint-2026-10-04/figures/`):

6. Figure B1 inserted in §4.2 and Figure B2 in §4.7, with captions, as relative links `figures/*.svg` from this file’s location.
7. §4.2: the **pre-committed** sensitivity leg (`ITT-SENSIB-PRECOMPROMETIDA.json`: $H1a -133.34$, CI $[-203.62; +7.41]$, contains zero; session-hours 12.16 / 4.08) is added and reported before the post-hoc leg, per §4.1’s own rule. The section title “inverts its conclusion on one epoch” is replaced by “excludes zero only on the locked leg”, and the abstract sentence is corrected to match. The B1 caption title is written to agree; the headline embedded in the SVG does not yet (item 11).
8. §4.3: $H1$ ’s pre-committed leg added (-15.59 , CI $[-22.43; -4.05]$, excludes zero). Same artifact, same omission; no conclusion changes.
9. §4.2: “9 to 10 minutes” → “8 to 10 minutes” (measured 9.71 / 8.86 / 8.03 min).
10. §4.7: truncation share “1.0%” → “-1.0%” (the unsigned shares summed to 102%); “+0.023 ... in the direction opposite to the intuition” → the sign is not resolved (0.7 MC s.e.; 20 000-seed redraw gives -0.007 ± 0.011); notes that the `control` 9.991 has no artifact, that the 2 000-seed recipe was not recorded, and that $P(n = 1) = 10.15\%$ carries ≈ 0.7 pp of MC noise.
11. Appendix B: rows for `ITT-SENSIB-PRECOMPROMETIDA.json`, the replay probes, the replay-fidelity runs, the figure scripts and the 20 000-seed crosscheck, marked † as not covered by `MANIFESTO-LASTRO-P2.json`. B.1: rows for the pre-committed legs, the §4.7 MC numbers, the 9.991 (declared as having no artifact), the session spans and §4.0.1c. Working list: items 10 and 11 added; item 1 annotated.

Related work (source: `_sprint-2026-10-04/B-related-work.md`, `.bib`, verification log):

12. New §8 “Related work” (~2 000 words), condensed from the ~4 100-word draft. It keeps the positioning against Behnam & Wang (arXiv 2610.02070) and the TMLR survey §9.6. Section references inside the draft were re-pointed to the manuscript’s actual sections: “§1” for the under-covering interval and the

ICC became §4.1.1. Sentences saying the sham was “still running” were replaced by the §4.0.1c status. The draft’s relative dates (“two days old”) became absolute ones.

13. New “References” section, generated from `B-related-work.bib` for the 61 keys cited; citations are pandoc-style `[@key]`. Four `.bib` entries are not cited in the condensed text (`frauen2026causalmethods`, `kohavi2012puzzling`, `garcin2014offline`, `rossetti2016contrasting`).
14. Renumbering: Discussion §8 → §9. Cross-references fixed: §1.1 (“One phrase in §9 overstates”) and working-list item 4. References to “§8 of the analysis spec” and to “Paper A’s §8” are to other documents and are unchanged.

Form only:

15. The duplicated heading “### 3.1 Two classification errors of ours, and what they cost” was removed; one copy remains. Status header updated.

rc2: adversarial review of 2026-10-04 and assembly fixes (source: `_sprint-2026-10-04/REVIEW-B-2026-10-04.md`, two valid voices of five, 18 confirmed findings C1–C18; the phase-2 assembly report):

16. §3.0.1 (C1): “closed 51 minutes before the designation expired” replaced. The window closes **at** expiry (2026-09-20 22:51:23Z); the dose was switched off later, 09:43:05Z on 2026-09-21 (`desliga-dose.ndjson`, archive stamp 20260921T094305Z). No artifact held the 51 minutes.
17. §4.1.1 (C2): “19 analyzable clusters” named as the power calculation’s population (11 treatment, 8 control, `out/H1c-POWER-REALIZADO-2026-09-10.json`); the ITT has 20 and the MDE was not recomputed at 9 control epochs (working list 13).
18. §4 and §7 (C3): “a 2-2 tie” → an exact tie ($n/2$ failures, even n). Checked in `pilot_replay.py` (`carregar_verdicts`: fewer than three substantive verdicts are not adjudicated): the rule cannot bind in the three-family primary panel and binds in the four-vote sensitivity set; the §7 threat now says so.
19. Abstract (C4): H1a excludes zero on the locked leg only; H1 excludes zero on all three legs and is an unexplained rejection, set aside for the impossible 955% effect, with the denominator explaining part of it but not the residual (§4.3).
20. Appendix B caveat and working list 1 (C5): “all 14” and “17 artifacts” → the manifest hashes **20** (161 608 733 bytes = 154 MiB); 12 copied and verified 12/12 on both legs, 8 versioned in the repository (`MANIFEST0-LASTRO-P2.json sha256 556ef20c...`, receipt `recibos/backup-20260922T014836Z.txt`).
21. §4.1.1 (C6): the “an interval must contain both 0 and -0.0896 , hence $\text{span} \geq 0.0896$ ” derivation and “they cannot both be right” are replaced by a comparison of implied standard errors: ≥ 0.0279 from the saturated MDE (0.0320 at $p_0 = 0.0896$) against 0.0165 from the bootstrap interval (59%). The “72%” is withdrawn with the derivation.
22. §4.4 heading and §5 table (C7): “unevaluable” → “trivially 1.0”; §5 column “why unevaluable” → “why not reportable”, H1b row rewritten.
23. Status header (C8): the universal traceability promise now declares the 9.991 of §4.7.
24. §4.6 (C9): M10 row “without both partials” → “without 09-03 and 09-20”; the all-partial leg is declared not computed for M10 (working list 13).
25. Appendix B (C10): “3 592 verdicts” → 3 592 rows = 3 585 verdicts + 7 superseded xAI non-verdicts.
26. §4.0.1b and Appendix B (C11): “all 20 epochs” → “all 19 served epochs” (11 positive + 8 negative dual, `CONTROLES-JANELA-COMPLETA-2026-09-21.json`).
27. §4.3 (C12): “ $\approx 5 \times 10^{-6}$ in all four arm-by-leg cells” → “ $\leq 5.8 \times 10^{-6}$ in all six (five distinct)”, recomputed on both ITT artifacts.
28. Abstract (C13): H1c “on both the locked and the sensitivity leg” → “on all three legs”.
29. §9 and §1.1 (C14): “we knew this ten days before” → “this was projection-robust ten days before”; the §1.1 caveat now refers to the earlier wording.
30. §4 structure (C15, assembly report): bold “(a)” and “(b)” became headings §4.0.1a and §4.0.1b; §4.0.2 moved after §4.0.1c; the bootstrap-layers section moved after §4.1.1 and renumbered §4.4.1 → §4.1.2. Order is now 4.0.1 (a, b, c), 4.0.2, 4.1, 4.1.1, 4.1.2, 4.2–4.7. Cross-references to §4.4.1 updated (§4.1.1 table, Figure B1 caption, working list 5); every other §4 reference already pointed at an unchanged number.
31. Working list 8 (C16): “Blocked by 1, 2 and 3” → “Blocked by 9 and 10”. Working list 5 (C17): “parecer” → “review”.

32. §1.1 (C18): the two pieces of evidence speak to our practice, not to the ordering of the 2026-09-21 decisions, which rests on commit timestamps alone.
33. Assembly report, terms: “designated chunks” → “designated items” throughout (15 places, including “non-designated”), matching Paper A; “lands on the floor or ceiling of its exact share” → “receives its exact share rounded down or up”; “our own 3 600 s ceiling” → “our own 3 600 s time limit”. The technical “exposure ceiling” (§1) and the eligibility “floor predicates” (§4.0.1b) are unchanged.
34. Version label: v2 throughout, with rc2 named in the status header; working list items 12 (struck) and 13 (open) added; B.1 row for the §4.1.1 standard errors added.

rc3: adversarial review of rc2 and the sham result (sources: `_sprint-2026-10-04/REVIEW-B-rc2-2026-10-04.md`, one valid voice of four, 22 confirmed findings C1–C22 with numbers recomputed in `REVIEW-B-rc2-evidence/`; `_sprint-2026-10-04/B-sham-v2/REPORT.md` §7 and `B-sham-v2/job-v2b/RESUMO.json`; per-finding record in `_sprint-2026-10-04/APPLY-B-rc3.md`). The title is unchanged.

35. Abstract, §1 table, §4.1 heading, §4.4, §9 (C1): “could not have found otherwise”, “undetectable by construction” and “no possible result” → not detectable at 80% power at the N the calculation counted.
36. Abstract, §1.1, §4.1.1, §8.4, §9 (C2): the saturation is stated for the 19 clusters the calculation counted (11 treatment, 8 control). §4.1.1 reports the recomputation at the ITT’s 20: whole-epoch 11T/9C detectable at 80% (effective size 96.41 against 95.26), fractional with 09-02 not (91.49, margin 4.1%). Working list 13, first half, closed.
37. Abstract (C3): the contradiction is stated through the implied standard errors (≥ 0.0279 against 0.0165), not through the interval excluding total elimination.
38. Abstract, §4.0.2, §4.3 (C4): “the 955% effect it would imply” → the observed reduction (70.7%; 47.1% without 09-14) would require 22.5 times total elimination in the 3.14% of altered briefs (at least 6.9 times at 6.85%). 955% kept where it is the 2026-08-30 demotion reason (abstract, §4.3, §9).
39. §4.0.2 (C5): the registered test no longer “shows” the H1a bootstrap interval was the artifact; the disagreement is consistent with under-coverage and does not demonstrate it.
40. Abstract, §4.1.1, §4.1.2, §9 (C6): the null is the non-rejection by the registered test ($p = 0.1603$), not the point estimate (a 22% relative reduction).
41. §4.1.1 table, §4.1.2, abstract (C7): under-coverage cited as expected from the literature, not measured; the frozen stratum-B sample “may be under-represented” in the intervals that use it (not M10); the direction of the systematic-sample error is not known.
42. Abstract, §3.0 (C8): the analysis specification is dated by our commit log and was not deposited with the registration (deposited version v1.12).
43. §3.0.1 (C9): “could have an effect” → “the dose could reach a designated item”.
44. §7 (C10): new threat, post-randomization denominator (opportunities per epoch 100.3 treatment, 132.3 control).
45. §6 item 3 (C11): zero reversals to the opposite strict majority; ties change the label, and those are the 28 losses of item 6.
46. §4 and §7 (C12): the tie rule biases toward fewer failures; its effect on the contrast is not counted. §7 paragraph retitled “Tie rule”.
47. §2 and §6 item 5 (C13): “12 pp disagreement” → an 11.9 pp difference in marginal failure rates, a lower bound on paired disagreement.
48. §4.6 (C14): 09-20 coverage is low because 287 of its 672 briefs were served after expiry (0 covered); before expiry 98 of 385 (25.5%). B.1 row added.
49. §4.7 and Figure B2 caption (C15): “cannot bias” → “does not measurably bias”; the truncation shift is not distinguishable from zero (1% of the deficit, MC s.e. 1.5%).
50. §4.0.1b (C16): $1/21$ is a rank p-value, a randomization p-value only to the extent the shams are drawn by the real designation’s mechanism (one per signature group), which neither sham generator does.
51. §5 (C17): the per-dose and leave-one-agent-out cells → “not computed in this version”.
52. §5 (C18): H1b excepted from “listed in §8 of the analysis spec before the window closed”.
53. §4 (C19): condition (i) defined in place (PREREG §4.1; the review’s fix said §3, where the Opportunity lock lives; the repeated-failure definition is in §4.1).
54. §1.1 (C20): “every estimate” excepts the 9.991 of §4.7 and is conditioned on the deposit.
55. §4.3 (C21): the H1a criterion restated as “excludes zero only on the locked leg, on a denominator

dominated by one sparse session”.

- 56. §3.0 heading, §4.0.1b heading, working-list caveat (C22): headings no longer say the feasibility question is unanswered or that a control is not runnable.
- 57. Sham integration: STATUS block; abstract (one sentence on the controls); §4.0.1b table row, heading and the eligible-pool bullet; §4.0.1c heading, the “what a valid sham needs” paragraph, the measured/not-measured table (110/110 and 2,646/2,646 now measured), a new result table with its four declared limits, the claims paragraph and the artifact list; §7 (new “scope of the specificity control”); §8.3 (two places); §8.5; Appendix B and B.1 rows; working list 3, 8, 9 and the caveat. Result: real 132 states moved and churn 146 against 81–122 and 84–132 for 20 shams, $p = 1/21 = 0.0476$ on both; positive control 155 at $w = 100\ 000$; real run equal to calibration and to production in 2,646/2,646 states.
- 58. Working list: item 14 added (this review, struck). Status header: rc3.

rc4: author decisions of 2026-10-05 and the tie census (sources: the author’s decisions on the title, the power statement, the sham scope and C12; B-sham-v2/JANELA-LANCAMENTO.md; out/C12-EMPATES-POR-BRACO-2026-10-05.json; per-change record in _sprint-2026-10-04/APPLY-B-rc4.md).

- 59. Title: “*Under-powered by construction: intention-to-treat results from a pre-registered interventional trial of agent-memory dosing*” → “*A registration that outlived its intervention: a pre-registered randomized trial of memory dosing in a production agent fleet*”. The thesis sentences follow it: abstract (two places), §1.1 (“the core claim”), §7 (power), §9 (opening). The under-powering is still reported, as a consequence of the registration outliving the intervention.
- 60. Abstract, §1 table, §1.1, §4.1.1, §8.4, §9 (decision on C2): the spec’s fractional counting governs the power statement (effective size 84.8 over the 19 served epochs, 91.49 at the ITT’s 20 with 09–02 as a full control epoch, against 95.26: not detectable). Whole-epoch counting is a sensitivity: it agrees at 11 treatment and 8 control (90.2) and flips at 11 treatment and 9 control (96.41 > 95.26). §4.1.1 adds that §7 of the spec forbids rounding partial epochs and that the spec computed the 9-control case itself (96.4) and set it aside.
- 61. §1.1 caveat and §9: “projection-robust” holds only under the spec’s cuts; it fails at 11 treatment and 9 control under whole-epoch counting.
- 62. §1.1, first list: the designation’s expiry (§3.0.1) added as outcome-independent evidence, with its artifact.
- 63. STATUS block and §4.0.1c (first declared limit): a sham replay over the whole trial window (11,812 reconstructible states of 18 epochs; without 09–01, whose corpus has no hash proof, and without 53 states whose serve-state cut cannot be reconstructed) is running; its result will be added before deposit. Working list: item 9 updated, new item 15, item 8 blocked by 10 and 15. The rc3 result ($w = 4$, $p = 1/21$) is unchanged.
- 64. §4 and §7 (C12): the exact ties of the four-vote set counted per arm: 28 in all, 10 in treatment epochs, 11 in control, 7 outside the window; 4 and 7 are H1c opportunities, none of them a failure under the three-family majority, so the tie rule removes no repeated failure from either arm and leaves condition (i) unchanged.
- 65. §6 item 3 (found while counting C12): “Ties do change the final label ... those are the 28 losses of item 6” struck. Only 3 of the 28 ties change the label; 25 were already **not_failure** under the three-family majority. §6 item 6 now defines gains (20: brought to three substantive verdicts) and losses (28: brought to an exact tie).
- 66. Appendix B: † row for out/C12-EMPATES-POR-BRACO-2026-10-05.json and its script. B.1: rows for the tie counts, the fractional effective size (84.8) and the spec’s 96.4, and the whole-window sham launch.
- 67. §4 and §7 (independent check of rc4, D-B1): “counted in rc4 it is nil for H1c” struck. The tie rule changes no H1c label relative to the three-family majority, but against the opposite resolution it holds 4 treatment and 7 control tied opportunities (HT weights 27.78 and 36.73) at **not_failure**. Measured with the C12 script and the imported rule (out/C12-EMPATES-COMO-FAILURE-2026-10-05.json): with every tie resolved as **failure** the four-vote H1c point difference goes from -0.0208 to -0.0266 ; opportunities do not change, and the first-order value is the same equal to it only at four decimals (corrected in rc5, item 73). Appendix B † row and B.1 row added.

68. STATUS block, §4.0.1c, working list items 9 and 15, B.1 (D-B2): the whole-window sham is `job-janela2`, 11,812 states (11,865 – 53), relaunched 2026-10-05T09:39:53Z, expected to finish about 2026-10-07T03:00Z. Item 9’s “11,865 states ... about 36 h” struck; in §4.0.1c the 11,865 / 36 h stays as the estimate that motivated the $w = 4$ set.

rc5: regression review of rc4 (source: one Codex review of rc4, five findings, all confirmed against the code and artifacts; per-change record in `_sprint-2026-10-04/APPLY-B-rc5.md`).

69. Abstract (two places), §4.0.2, §4.3 (H1): “would require more than 100% in the 3.14% of briefs the dose alters” and “a reduction above 100% is impossible, so an interval excluding zero there cannot be read as the treatment working” replaced. The 22.5 / 15.0 and 10.3 / 6.9 ratios are now stated as proportional-dilution calculations, not causal bounds: the altered share of briefs need not equal the share of failures they carry, and an effect can travel through activity volume (§4.3, reading 1) or state (§8.3). H1 remains an unexplained rejection.
70. §4 (combined estimator): condition (i) is now described as implemented. The estimator requires a same-signature failure ($\text{severity} \geq \tau$) whose epoch begins at least one epoch length earlier; it does not check membership in the serving snapshot that PREREG §4.1 names, and `pilot_replay.py` declares the reconstruction an approximation.
71. Abstract, §4.0.1c, §7 (sham): the specificity claim is limited to the 20 tested shams (“any designation of the same severity and bonus mass” struck in substance), and the rank p-value is stated not to be a calibrated randomization p-value, because the shams are matched on severity and bonus mass but not drawn by the designation’s one-per-signature-group rule. The numbers and the term “rank p-value” are unchanged.
72. §4.1.1, §9, B.1 (power): “under-powered for any smaller effect” and “anything less stays out of reach” replaced: under whole-epoch counting at 11 treatment and 9 control, 80% power holds only for reductions of about 99.6% or more (relative MDE 0.996).
73. §4, B.1 (C12): the rerun with ties as **failure** gives -0.026571 and the first-order calculation -0.026566 ; both are now stated, and item 67 is corrected.
74. Working list: item 16 added (this review, struck). Status header: rc5.

rc6: ballast gap closed, and a writing pass (sources: `_sprint-2026-10-04/LASTRO-B-item10.md` and the `avoid-ai-writing` pass; per-change record in `_sprint-2026-10-04/APPLY-B-rc6.md`).

75. Appendix B: † caveat dropped; the manifest now covers every row of the table, 50 artifacts (working list 10; `LASTRO-B-item10.md`).
76. Working list 10 struck as done, item 8 no longer blocked by it, item 1’s caveat closed; the Appendix B caveat now counts 50 artifacts, 158 MiB, 16 copied and verified at the destination and 34 versioned.
77. The passages that depend on the whole-window sham (`job-janela2`) are wrapped in `<!-- SHAM-JANELA -->` comments with their wording unchanged; the list is in `APPLY-B-rc6.md`.
78. Writing pass over the rest of the text (em dashes, bold, contrast pivots, closing aphorisms). No number, unit, reference, identifier, path, quote or claim changed. Status header: rc6.

rc7: review of rc6 (sources: `_sprint-2026-10-04/REVIEW-B-rc6-2026-10-05.md`, a Codex and a Fable full read, both NO-GO, with the author’s two decisions of 2026-10-05; numbers added here are recomputed in `_sprint-2026-10-04/B-rc7/checks-rc7.json`; per-finding record, with what was verified and the list of REANALISE blocks, in `_sprint-2026-10-04/APPLY-B-rc7.md`). Every number and sentence that the registered re-analysis will replace is wrapped in a REANALISE comment and still carries the rc6 value (working list 18).

79. Abstract, §1 (table and a new paragraph), §4.0.2 table, §4.3, §8.2 table, §9, Appendix A (Fable HIGH-1; author decision): the deposited primary is H1, with H1a–H1c as a Holm-corrected co-primary family; H1c is reported as primary under the undeposited revision of 2026-08-30 (`DESIGN-REVISION-2026-08-30.md` §3-ter, commit 16:51:56Z; `PROSPECTIVE-ESTIMAND-2026-08-30.md` §3-bis, commit 20:51:20Z), before the assignment seed. The result under the deposited primary is stated beside it (H1 rejects), and the Holm statement is added.
80. §5, §1, Appendix A (Fable HIGH-2): rows for H2 (task regret) and H3 (retrieval metrics), with their reason pending the re-analysis; the row “dose-response / H3” renamed to the registration’s dose-response

- reading rule; the closing paragraph no longer says every row was listed in the spec before the close.
81. §4.5 (heading and text), §1.1, §9 (Codex 1): the flat $\approx 5.3\%$ counts designated-chunk serving occurrences, the 93.8% counted covered action opportunities (573 / 611); the measurement does not test the projection, and the pre-committed indistinction stands (now also in the abstract, as the pre-commitment required). The third observation of §9 is replaced.
 82. Abstract, §4.1.1 (heading, text and table), B.1 (Codex 4): the comparison of implied standard errors is struck; the two statements use different variance constructions, and their difference is not a contradiction.
 83. Abstract, §4.0.1b, §4.0.2, §4.3, §4.5, B.1 (Codex 5, Fable MEDIUM-7 and LOW-10): 3.14% is named the pre-trial calibration share (11 of 350 at $w = 2$); the realized share is 335 of 7 350 briefs served in active mode (4.56%), the denominator chosen because the artifact's 337 of 7 392 includes two shadow-mode counterfactual changes on 09-01; ratios at the realized share 15.5 and 10.3; coverage over active briefs 2 057 / 7 350.
 84. Abstract, §1 table, §4.1.1, §7, §9 (Codex 6): the power statement covers reductions only; an increase can reach 80%.
 85. Abstract, §4.1.1, Figure B1 caption, §4.1.2, §4.3, §8.3 (Codex 7): intervals *may* under-cover; the few-cluster literature is cited for caution, not as calibration.
 86. §1, §1.1 (Codex 8, Fable MEDIUM-3): the 1 056 s belong to designation round 31657512; the assignment round 31774052 has its own precedence (21:18:35Z against 21:32:04Z).
 87. §8.2 (Codex 9): CMP's Theorem 1 is stated for a memory with zero retrieval probability.
 88. §8.4 (Codex 10): the NHLBI comparison is observational.
 89. Abstract (Codex 11): the "evidence of nothing" sentence replaced.
 90. §1.1, §7, working list 17 (Fable MEDIUM-6): ITT-PRELIMINAR.json (file time 16:16:44 local) precedes the §10.33 decision (16:20), so the claim that no number existed at either decision is withdrawn for §10.33.
 91. §4, B.1 (Fable LOW-8): the sixth-decimal difference between rerun and first-order value is the rounding of a stored per-arm proportion (0.09189), not feedback of condition (i); item 73 is left as written and corrected by this item.
 92. §9 (Fable LOW-9 and LOW-11): "the only significance" → the only rejection by the registered test; "three weeks apart" → eighteen days apart.
 93. §4.1.1 (Fable LOW-12): 09-01 enters the fractional counts by brief share (0.9375), not by clock (0.9325); the difference moves 84.80 to 84.78.
 94. §3, §3.0.1, §4, §6 item 7, Appendix A (Codex 2 and 3, Fable MEDIUM-4 and MEDIUM-5): declared, not yet resolved: the panel substitution was not applied (1 159 of 1 195 adjudicated, 36 unknown), the window is selected by epoch date with no cut at expiry, the bootstrap is percentile instead of BCa, and 09-02 is included against the spec. All are marked REANALISE.
 95. Sweep of the classes the review named, outside the SHAM-JANELA blocks: §4.0.2 (the disagreement can also come from the estimands), §4.7 ("false" → not supported by the measurement), §8.5 (wording of the contradictions).
 96. Working list: item 8 also blocked by 17 and 18; items 17 and 18 added; item 19 (this review, struck). Status header: rc7.

rc8: the registered analysis reported, and two findings of ours (sources: out/ITT-REGISTRADO-2026-10-05.json and _sprint-2026-10-04/B-registered/RESULTADO.md; _sprint-2026-10-04/B-rc8/checks-rc8.json; the author's decision of 2026-10-05 that the registered analysis is the reported one and the analysis of rc7 a sensitivity; per-block record in _sprint-2026-10-04/APPLY-B-rc8.md). No REANALISE marker remains.

97. Abstract, §1, §3, §3.0.1, §4-§4.3, §6 item 7, §7, §8.3, §8.4, §9, Appendix A, B.1: every REANALISE block of rc7 (85 values, 23 sentences) replaced by the registered analysis, with the rc7 value kept as the sensitivity where the sentence reports both. H1 rejects in both (0.0133; 0.0127); H1c does not reject in either (-0.0121 [-0.0445 ; $+0.0125$], 0.434; 0.1603); H1a's unadjusted p is 0.0241 and it is not rejected under the registered Holm rule (adjusted 0.1205 with $m = 5$, 0.0964 with $m = 4$).
98. §4: what moves each number (09-02's exclusion moves H1c most; the offsets move H1a; the expiry cut is negligible; the panel goes from 1 159 to 1 179 adjudicated; BCa changes only intervals), and the

estimator’s byte-for-byte control.

99. §4.0.2 (heading and table) and §4.2 (heading, table, take-away and Figure B1): under the registered BCa interval H1a excludes zero on the locked and pre-committed legs and contains zero only without 09-14; “only on the locked leg” is struck as a statement about the sensitivity analysis. §4.1 notes that the direction of the pre-committed against the post-hoc leg reverses for H1a.
100. §4.1 and §4.2 tables: the registered analysis on the pre-committed and post-hoc legs (`checks-rc8.json` block A), with the sensitivity analysis’s three legs beside it.
101. §4.1.1, abstract, §1.1, §7, §8.4, §9: at the registered 19 epochs both countings give the same power verdict; the whole-epoch flip belongs to the sensitivity analysis’s 20.
102. §5: H2 computed and reported (three declared choices; neither component rejects); H3 declared not computable; a row for the secondary model and co-estimates of PREREG §5, not computed.
103. §4, §7, B.1: the four-vote ties recounted under the registered window and set (`out/C12-EMPATES-REGISTRADO-2026-10-05.json`); the tie rule cannot bind in the substitution panel (0 ties).
104. §3.0 (struck sentence and a new correction), abstract, §1.1, §7, §9, Appendix A, §8.2: the stopping rule of PROSPECTIVE-ESTIMAND-2026-08-30.md §3-bis was met in Epoch 1 (32.4%, 45 of 139, against 36.7%) and not executed (`checks-rc8.json` block D).
105. §1 (new paragraph), abstract, §1.1, §7, §9, Appendix A, §8.2: the assignment rule declared in ASSIGN-SEED-2026-08-30.md is not `assign_arms.py`; it was replaced 24 min 38 s after the round was emitted, and the two differ on 11 of the 20 realized epochs (`checks-rc8.json` block E).
106. Figure B1 regenerated from the registered artifact (`measurement/sprint-figB-h1a-inversao-registrado.py`, new output files; the earlier figure is kept as the record of the sensitivity analysis).
107. Appendix B and B.1: the registered artifact, `checks-rc7.*`, `checks-rc8.*`, the registered C12 file and the registered figure listed, each marked as not yet in the ballast manifest. Working list: 8 and 17 extended, 11 annotated, 18 struck, 20 added. Status header: rc8.

rc9: review of rc8 applied (sources: `_sprint-2026-10-04/REVIEW-B-rc8-2026-10-05.md`; `out/ITT-REGISTRADO-v2-2026-10-05.json` and `_sprint-2026-10-04/B-registered/RESULTADO-v2.md`; `_sprint-2026-10-04/B-rc9/checks-rc9.json`; per-finding record in `_sprint-2026-10-04/APPLY-B-rc9.md`). The SHAM-JANELA blocks are unchanged.

108. §2, §4, §4.0.2, §4.2, §4.3, abstract, Appendix A, B.1 (Codex C1): the washout left the outcomes but stayed in the session-hour denominator. A fifth switch removes it before both; H1 -19.85 [-32.69 ; -9.89] ($p = 0.0302$), H1a -233.20 [-352.07 ; -80.31] ($p = 0.0168$, Holm 0.0840 / 0.0672), H1c unchanged. Under the new denominator H1a excludes zero on all three legs, so §4.2’s heading, table, take-away and Figure B1 change.
109. Abstract, §1, §1.1, §4.0.2, §4.3, §9, Appendix A (Fable F2): H1 rejects only under the deposited reading; under the switch it is a secondary in the Holm family and does not reject (adjusted 0.151 with $m = 6$, 0.1208 with $m = 5$).
110. §4.6 (renamed), §3, §4.1.2, §5, Appendix A, Appendix B, B.1 (Codex C2): the correlation called M10 measures designated-item presence among served briefs; the registered M10 and the 95%-coverage set are not computable from the locked artifacts (no session $\leftrightarrow$ `brief_log` linkage), and are declared.
111. Abstract and §6 item 7 (Codex C3): the substitution rule is attributed to DEVIATIONS §10.31, not to the deposited registration.
112. Abstract, §1, §9, Appendix A (Codex C4, Fable F1, F5, F7): the assignment-rule retraction of 2026-08-30 (before Epoch 1) is reported with its date, the declaration’s own failure clause, and the discrepancy between “hours after” and 24 min 38 s.
113. Appendix A and §7 (Codex C5): the stopping disclosure is limited to what the record shows; nonexecution does not establish absence of bias.
114. §3.0, §7, Appendix A (Fable F3): the deposited horizon and safety abort stated; no record that the abort ran; an ex-post count over the adjudicated episodes.
115. Abstract, §3.0, §1.1, §7, §8.2, §9, Appendix A (Fable F4): the promotable set at $w = 4$ re-derived; Epoch-1 coverage there is 37.4%, above the threshold.
116. §4.1.1 (Fable F6): the 0.0335 is from unrounded values; 0.0336 from the rounded ones.
117. §5, §4.2, §4.3, Appendix A, working list 22 (found while verifying C1): boundary-straddling sessions are not attributed as PREREG §2 registers; the ‘without’ leg, computed as a diagnostic, contains zero

for H1 and H1a.

118. Appendix B, B.1, working list 17, 20 (struck), 21 and 22: rc9 artifacts listed, each not yet in the ballast manifest. Status header: rc9.

rc10: sessions in the epoch of their start (sources: the author’s decision of 2026-10-05 that the analysis reported is the registered one; `out/ITT-REGISTRADO-v3-2026-10-05.json` and `_sprint-2026-10-04/B-registered/RESULTADO-v3.md`; `_sprint-2026-10-04/B-rc10/checks-rc10.json`; per-item record in `_sprint-2026-10-04/APPLY-B-rc10.md`). The SHAM-JANELA blocks are unchanged.

119. §2, §4, §4.0.2, §4.1, §4.2, §4.3, abstract, §1, Appendix A, B.1 (working list 22): a sixth switch attributes every session to the epoch of its start, as PREREG §2 registers. H1 $-1.27 [-28.15; +23.50]$ ($p = 0.3294$), H1a $-15.22 [-372.20; +314.65]$ ($p = 0.2599$), H1c $-0.0140 [-0.0464; +0.0110]$ ($p = 0.4006$); every Holm-adjusted p-value 1.0 under both readings. H1 no longer rejects under the deposited reading, and no interval of the H1 family excludes zero on any registered leg. The rc9 text of §4.2, §4.3, §4.0.2 and the abstract is struck where it stated the old result, and kept where it records the argument earlier versions made.
120. §5 (straddling row), §4.2 and §4: the registered ‘without’ sensitivity (H1 -7.40 , H1a -51.12 , H1c -0.0109 ; p 0.1957, 0.1501, 0.4571) and the three sessions longer than one epoch, reported as their own stratum.
121. §5 (H2): the H2 table under the rc10 analysis; two intervals exclude zero (winsorized time, raw tokens), neither component rejects; the rc9 values kept in a note.
122. §4 (ties), §7, §8: the four-vote ties under the rc10 analysis (4 / 7 in the analysed exposure, $-0.0140 \rightarrow -0.0249$ as failure); opportunities per epoch 109.5 / 135.0; the lesson paragraph of §8 rewritten.
123. Figure B1 regenerated (`sprint-figB-h1a-inversao-registrado-v3.py`); the rc9 and rc8 figures kept unchanged.
124. Appendix B, B.1, working list 17, 22 (struck) and 23: rc10 artifacts listed, each not yet in the ballast manifest. Status header: rc10.

rc11: review of rc10 applied (sources: `_sprint-2026-10-04/REVIEW-B-rc10-2026-10-05.md`; `out/ITT-REGISTRADO-v3-2026-10-05.json`, regenerated with identical numbers; `_sprint-2026-10-04/B-rc10/checks-rc10.json`; `PREREG-DRAFT.md` §3 and §5; per-item record in `_sprint-2026-10-04/APPLY-B-rc11.md`). The SHAM-JANELA blocks are unchanged.

125. §4 and Appendix A (Codex 1): the census of the live study that PREREG §3 locked was replaced by a census of `is_error` and a sample of 800 of the 5 556 other episodes, weighted by 6.945; declared as a deviation. Working list 25 added.
126. §4.0.2, §9, abstract (Codex 2, Fable MEDIUM-1): the switch-6 correction moves exposure and outcome attribution together, so it shows the earlier rejection was not robust to the registered rule, not that the dose had no effect; §9’s lesson heading names the unregistered denominator.
127. §4 (Codex 3, Fable LOW-4): the own stratum’s descriptive differences, $+0.01867$ repeated failures/hour and $+0.43543$ opportunities/hour, with no interval or test; H1c not contrastable; the session that starts on 2026-08-27 takes its window episodes with it.
128. §4.2 and the Figure B1 caption (Codex 4): `span_por_sessao` has a one-minute minimum.
129. §7 Denominator and Appendix A (Fable HIGH-1, Codex 5): the status sentences that predated rc10 rewritten; one verdict differs between the registered and the sensitivity analysis, under the deposited reading.
130. §4.5 (Codex 6): the no-exclusion ITT is PREREG §5’s mandatory co-estimate, not its primary; the coverage-filtered primary was not computed.
131. §8.1 and §8.5 (Codex 7): H2 reports action-level time and token regret; what is not measured is per-phase attribution and energy.
132. Appendix B (Codex 8): the ballast manifest covers 50 artifacts, not every artifact in the table.
133. §3.0 and Appendix A (Codex 9): the safety abort has no execution record, rather than being unexecuted.
134. Abstract and §5 (Fable MEDIUM-2): H2’s two intervals exclude zero under BCa only; the percentile intervals contain zero.
135. §3.0.1 (Fable LOW-5): the expiry cut is applied to episodes attributed to 09–20, and no earlier session reaches past expiry. §7 (Fable LOW-6): the ties under the rc10 analysis. §5: “the four departures” of

the co-estimates row is now the six switches.

136. **measurement/estimador_itt_registrado.py** (Fable LOW-3): four stale comments and docstrings corrected, no logic change; re-run, every field identical except **gerado_em** and the script sha256 (0aa9202b...); frozen copy and the rc10 bytes in **B-registered/**. Appendix B, B.1, working list 17, 23 (struck), 24 and 25. Status header: rc11.

rc12: the census restoration attempted after unblinding (sources: **_sprint-2026-10-04/B-censo/GATES.md**, **_sprint-2026-10-04/B-censo/GATE-RULES-PREDECLARED.md**, **_sprint-2026-10-04/B-censo/gate2-compare.json**; per-item record in **_sprint-2026-10-04/APPLY-B-rc12.md**). The SHAM-JANELA blocks are unchanged.

137. §4: the attempt to restore the PREREG §3 census, after unblinding, under gate rules written before the first call; gate 1 (model identity) passed, with the alias caveat; gate 2 (test-retest on 50 stratum-B episodes) failed for two of the three families; the panel outcome agreed on 46/50; the census was not run, and the sample stays a declared deviation; cost about US\$1.18.
138. §7: the label-level drift between the September and October runs bounds how reproducible the adjudication is.
139. Appendix A: the census item records the attempt and its outcome.
140. Appendix B, B.1, the ballast caveat: the census-attempt files listed, not yet in the ballast manifest. Working list 24 extended, 25 closed. Status header: rc12.

rc13: review of rc12 applied (source: **_sprint-2026-10-04/REVIEW-B-rc12-2026-10-05.md**, Fable; each finding verified against the manuscript and the census record before it was applied; per-item record in **_sprint-2026-10-04/APPLY-B-rc13.md**). The SHAM-JANELA blocks are unchanged.

141. Abstract and §9 (Fable MEDIUM-1): three commitments made before the seed were not kept, not two; the third, the census PREREG §3 locked, is the one in the deposit. The abstract now also reports the census attempt and its failed test-retest gate.
142. §4.1.2, and the abstract's uncertainty paragraph (Fable MEDIUM-2): three layers of randomness, not two; the panel's run-to-run variation measured in rc12 is in no interval.
143. §4 (Fable LOW-1): the header time of the rules file is rounded, falls after the first call read literally, and is not evidence of the order; **GATES.md** dates gate 1 from the second call.
144. §4 (Fable LOW-2): the rules' declared noise property of the 0.99 criterion (about 78% failure on noise alone at 150 pairs), and why this failure is not that case. (*rc15: the 78% holds only if the 150 agreement indicators are independent, and "this failure is not that case" is withdrawn: the Wilson intervals are evidence against 0.99, not proof of a changed model or of lower reliability than in September; item 152.*)
145. §7 (Fable LOW-3): "drifted" replaced by "did not reproduce"; the record cannot separate non-determinism from a changed model.
146. **B-registered/RESULTADO-v3.md** and Appendix B (Fable LOW-4): its sentence on the own stratum corrected (no inferential contrast; descriptive differences in §4).
147. Appendix A (Fable LOW-5): twelve deviations, not thirteen; the struck item is no longer one.
148. **B-censo/** scripts: absolute personal paths replaced by paths derived from each script's location; **SHA256SUMS** and **census-pins.sha256** regenerated; Appendix B, B.1, working list 24. Status header: rc13.

rc14: writing pass (source: **_sprint-2026-10-04/APPLY-B-rc14.md**; the **avoid-ai-writing** skill, mode edit, voice technical, over the prose changed since rc6, the Abstract and §9; parity in **_sprint-2026-10-04/B-rc14/parity-rc14.py**). No number, interval, quotation, table, struck span or changelog line changed, and no claim moves. The SHAM-JANELA blocks are unchanged. This block was added in rc15 (Fable F2).

149. Abstract, §3, §4.0.2, §4.1.1, §4.1.2, §4.5, §9 and B.1: eleven wording edits (E1–E11 of **APPLY-B-rc14.md**): misplaced parentheticals, restatements and run-in bold labels removed; every qualifier kept. Status header: rc14 (recorded in rc15).

rc15: review of rc14 applied (source: **_sprint-2026-10-04/REVIEW-B-rc14-2026-10-05.md**, Fable GO and Codex NO-GO; each finding verified against the code, the artifacts and the census record before it was applied; per-item record in **_sprint-2026-10-04/APPLY-B-rc15.md**). The SHAM-JANELA blocks are unchanged.

150. `measurement/estimador_itt_registrado.py` (Codex C1): the BCa acceleration is computed arm by arm (multi-sample: arm-specific jackknife influences, centred within arm and scaled by $(n_g - 1)/n_g$), as the stratified bootstrap requires; checked against a hand-computed case and SciPy’s multi-sample acceleration. `out/ITT-REGISTRADO-v4-2026-10-05.json`: most BCa bounds move slightly; no interval changes whether it contains zero, and no p-value or point estimate changes. The v1, v2 and v3 controls pass under the earlier acceleration, which a switch restores. Frozen script `b3095740...; B-registered/RESULTADO-v4.md`.
151. Abstract, §4, §4.1, §4.2, §4.3, §5, Figure B1, Appendix A, Appendix B and B.1: every reported BCa interval switched to v4, the rc8 and rc9 rows of §4.2 included (`B-rc15/checks-rc15.json`, block R); §4 *Uncertainty* states the construction and the correction; the upper bounds of H1 and H1a are the 50th and the 45th largest replicates, and the claims that they “carry visible Monte-Carlo noise” and that one is “the least stable number in the table” are replaced by what was measured (the rank) and what was not (the Monte-Carlo error). Figure B1 regenerated from v4 (`measurement/sprint-figB-h1a-inversao-registrado-v4.py`).
152. §4, census attempt (Codex C2): the 0.779 failure probability of the all-agree criterion is stated under independence; the Wilson intervals are evidence against 0.99, not proof of a changed model or of lower reliability; “instrument change” withdrawn. §7: the October re-adjudication did not reproduce all of the September labels. Item 144 annotated; B.1.
153. §4.1.2, §4.1.1 and the abstract (Codex C3): the bootstrap resamples observed epochs within each arm and does not redraw the assignment; it does not explicitly propagate adjudication-sampling or repeat-adjudication uncertainty, and the direction of the resulting coverage error has not been established.
154. Appendix B and `B-registered/RESULTADO-v3.md` (Codex C4): no standalone inferential interval or test is reported for the own stratum, and H1c cannot be contrasted. §4 (Codex C5): the header time is inconsistent with the file-system timestamp. The estimator’s H1b docstring and output (Codex C6): trivially 1.0 under the retained action-level definition, and the mechanism question is not identifiable.
155. §9 and the abstract (Fable F1): H1a’s intervals excluded zero on every registered leg only in rc9, and on the locked leg in every version that split sessions. Status header and changelog record rc14 and rc15 (Fable F2). Working list 17 and 24, Appendix B’s ballast caveat. Status header: rc15.

rc16: review of rc15 applied (source: `_sprint-2026-10-04/REVIEW-B-rc15-2026-10-05.md`, Fable GO and Codex NO-GO; each finding verified against the registration, the code and the artifacts before it was applied; per-item record in `_sprint-2026-10-04/APPLY-B-rc16.md`). The SHAM-JANELA blocks are unchanged.

156. Appendix A (Codex K1): the dose band is removed from the deviation list (twelve $\rightarrow$ eleven), and the paragraph on the two items where the registration promises less is withdrawn. The registration retained the band when the sizing formula was corrected and did not say that the dose could not change briefly; the replay fact stays, restated. `DEVIATIONS-FOR-PAPER.md`, a dated log left unchanged, carries the same misreading.
157. §4.1.2, §7 and the status header (Codex K2, Fable L2): a sampled stratum-B failure carries weight 6.945, and its contribution to estimator variance has not been quantified; repeat-adjudication uncertainty is not explicitly propagated, some realized variation may enter through between-epoch differences, and its contribution to interval coverage has not been quantified.
158. §4.4 (Codex K3): the repeat-attempt set coincides with the opportunity set; the nesting remains, and H1b no longer provides the intended intermediate distinction.
159. `B-registered/teste_aceleracao_v4.py` (Codex K4): the unrounded comparison its docstring promised ($1e-12$) is added; rerun, 10 of 10 real cases pass, largest difference 1.4×10^{-16} . The rc15 bytes are kept beside it. Appendix B.
160. §4 *Uncertainty* (Fable L1, L4): the zero-inclusion claim cites `checks-rc15.json` blocks V and R and `RESULTADO-v4.md`, and the construction states the $(n_g - 1)/n_g$ scaling. B.1 (Fable L3, and the review’s note on `gerado_em`): the intervals of the rc8 rows were recomputed in rc15; the v4 artifact’s timestamp falls on 2026-10-06 in UTC. §5 (Fable, optional): the winsorized-time lower bound clears zero by less than 0.01 s, and its Monte-Carlo error was not measured. Working list 17 and 24. Status header: rc16.

rc17: review of rc16 applied (source: the Codex read of rc16, receipt `adversary-receipt-codex-2026-10-05T221948-70096.txt`, NO-GO with one MEDIUM and two LOW, and the Fable read, three

LOW; each finding verified against the deposited v1.12, the amendment, the designation record and the prospective estimand before it was applied; per-item record in `_sprint-2026-10-04/APPLY-B-rc17.md`). The SHAM-JANELA blocks are unchanged. The title was changed by the author after the review was applied (`APPLY-B-rc17.md`, author decision of 2026-10-05): “A registration that outlived its intervention” → “A registered horizon that outlived its intervention”, line 1 only, because the intervention that expired was not the registration’s own (item 161); no deposit metadata for this paper exists yet, and the one written at deposit must carry this title (item 164).

- 161. Abstract, §1, §1.1, §3.0.1, §7, §8.5 and §9 (Codex M1): the fixed 19-item designation was adopted on 2026-08-26, after v1.12 was published, and was not deposited; the deposited rule designates at brief composition and is recomputed at every brief (`AMENDMENT-v1.12.md` §5.2-bis), and the amendment records it as an open defect. The horizon/expiry mismatch is therefore stated as one between the registered horizon and the subsequently frozen intervention, present before the trial began, and no longer as a property of the registration or as a contradiction between two of its locks; the H1b collision of §4.4 stays the one contradiction internal to the registration. Appendix A’s designation item says the same.
- 162. Abstract (two places) and §9 (Codex LOW): “every earlier version” → “versions up to rc9”. B.1 (Codex LOW): the v4 numbers that differ from v3 include the accelerations.
- 163. Appendix A (Fable LOW): `DEVIATIONS-FOR-PAPER.md` still carries the withdrawn band reading in its front table, and it is a dated, append-only log left unedited. §1 (Fable LOW): the nesting sentence is in the H1b lock of 2026-08-16 in §1 of the pre-registration (`PREREG` 1.308). `B-registered/RESULTADO-v4.md` (Fable LOW): a dated note naming the rc16 test bytes is appended, nothing above it edited; Appendix B. Working list 24. Status header: rc17.

rc18: review of rc17 applied (source: the Fable read of rc17, NO-GO with three MEDIUM and four LOW and no HIGH; each finding verified against its primary source before it was applied, and two adapted where the source did not hold; per-item record in `_sprint-2026-10-04/APPLY-B-rc18.md`). The SHAM-JANELA blocks are unchanged.

- 164. rc17 changelog block (Fable M-1): it said the title was not changed, but the author changed it after the review was applied; the block now records the change, and the deposit metadata, when written, must carry the new title.
- 165. §1 and §3.0.1 (Fable M-3): the deposited bytes of `DECISION-designacao-2026-08-25.md` (md5 35abeb68... in the Zenodo v1.12 file list, equal to the blob of commit d42f950 of 2026-08-25) are the open version: options A–D, a recommendation of option B, marked as awaiting the author’s decision. Both sections now say that the deposit carries the per-brief rule retracted as an open defect and that recommendation, and that the decision, the seed, the key layout that was used and the 19 items were not deposited; §3.0.1 adds that the deposited option B keys on `sig_primary`, dropped at 19:40Z. “Only the horizon is in the registration” → “only the horizon is locked in the registration”. Correction, found while verifying: v1.12 was published at 12:01Z (the Zenodo record’s creation time, 2026-08-26T12:01:06Z), not 14:01Z; `deposit/PLAN-v1.13.md` (1.13) and `AMENDMENT-DRAFT-band-collapse-2026-08-26.md` (1.5) give the time in UTC+2 labelled Z, and neither is edited. The decision therefore came 2 h 46 min after publication, not 46 min as the review proposed.
- 166. §8.3 (Fable M-2): “a registration whose sample size ignored a 30-day eligibility window” → a registered horizon of 234 epochs never re-read against the intervention frozen after registration, whose 19 items left the 30-day window together after 20 epochs. No other sentence of the body carries that framing.
- 167. §1 (Fable L1, L2): the designation “went live” at 20:28Z, with its source named; the 1 056 s is counted from the declaration’s commit time (20:07:24Z), and the 1 053 s of the declaration file from its own 20:07:27Z. §3.0.1 heading (Fable L3): “infeasible by construction” → “infeasible once the designation was frozen”. §3.0.1 caveat (Fable L4): `PREREG` 1.513–519 and 1.550–558 are cited. Working list 24. Status header: rc18.

rc19: review of rc18 applied (source: the Fable read of rc18, GO with six LOW; each finding verified against its primary source before it was applied, and one (L4) stated more exactly than proposed; per-item record in `_sprint-2026-10-04/APPLY-B-rc19.md`). The SHAM-JANELA blocks are unchanged.

- 168. §1 and §3.0.1 (Fable L1): “the per-brief rule retracted as an open defect” → “the per-brief rule declared

- an open defect”. **AMENDMENT-v1.12.md** §5 declares the designation not validly frozen, and its §5.1 retracts the threshold model behind **CUT_FRESH**, not the per-brief rule. Item 165 keeps rc18’s wording as the record of rc18.
169. §1 (Fable L2): the deposited bytes of **DECISION-designacao-2026-08-25.md** are named (md5 35abeb68..., 7 036 bytes, the blob of commit d42f950), and the repository file is said to have been extended after publication to record the decision. B.1: a row for 12:01Z, the md5, 19:40Z and 2 h 46 min. Appendix B: a row for the **B-rc18/** snapshot, and the snapshots added to the artifacts not yet in the ballast manifest.
 170. §3.0.1 (Fable L3): “both numbers” are named, the horizon and the 20-epoch life of the fixed set.
 171. §1 (Fable L4, verified): “fixed and deposited two weeks before the round” → fixed on 2026-08-17 (its last commit, 3199ec1) and deposited in Zenodo v1.11 (record 21978476, published 2026-08-17T18:32Z, the day before the OSF registration), in the bytes v1.12 carries (md5 3fa3f710...), thirteen days before round 31774052 was emitted (2026-08-30T21:32:04Z); “deposited with v1.12” → “deposited with v1.11 and v1.12”. The proposed “registered on OSF on 2026-08-18, twelve days before the round” was not used: the deposit that dates the file is v1.11’s of 2026-08-17, and the OSF registration attaches the document, not the script. No other “two weeks” in the body. B.1 and Appendix B: the v1.11 snapshot in **B-rc19/**.
 172. rc18 changelog item 165 (Fable L5): **AMENDMENT-DRAFT-band-collapse-2026-08-26.md** l.5 carries the same mislabelled 14:01Z and is named beside **deposit/PLAN-v1.13.md**; neither file is edited.
 173. Working list 8, 17 and 24 (Fable L6): none of the rc15–rc18 artifacts, and neither Zenodo snapshot, is in the ballast manifest, and item 17 blocks the deposit. Status header: rc19.
- rc20: review of rc19 applied** (source: the Codex read of rc19, NO-GO with one MEDIUM and two LOW, receipt **adversary-receipt-codex-2026-10-06T095645-63516.txt**, **exit: 0**; each finding verified against the manuscript, PREREG 1.795 and 1.797–801, DEVIATIONS §10.14, §10.22 and §10.29 and **out/ITT-REGISTRADO-v4-2026-10-05.json** before it was applied; per-item record in **_sprint-2026-10-04/APPLY-B-rc20.md**). The SHAM-JANELA blocks are unchanged.
174. Abstract, §3.0.1 and Appendix A (Codex M1): “The trial did not stop early: 234 was infeasible from the start” → the trial ended before its registered 234-epoch horizon, after the fixed designation exhausted its eligibility, which could support dose exposure in only 20 trial epochs, a limitation fixed before Epoch 1. §3.0.1: “The trial did not stop early” → it ended before the deposited data-collection horizon, which was reached by neither 234 epochs nor the calendar cap, and expiry is not a deposited stopping condition; the dose was switched off only after the designation had expired. §3.0.1 heading: “N = 234 was infeasible once the designation was frozen, and the trial ran exactly as long as it could” → “The fixed designation could support dose exposure in only 20 of 234 registered epochs”. §3.0.1: “the expiry stops the trial” → “the expiry ends the intervention”. Appendix A: a twelfth item, the horizon not reached; the closure itself is in the deviation log (§10.14, §10.22, §10.29), so the list of items absent from it is unchanged.
 175. §3.0.1 and B.1 (Codex L1): 09–20’s session-hours under the expiry cut are quoted on the registered leg, 0.5902 → 0.4820 (legs **registrado_sem_janela** and **registrado**); the 0.6235 → 0.5154 of the block **corte_pos_expiracao** keeps the washout in the denominator and is named in B.1 as a diagnostic. The 33 episodes, 22.835 weighted opportunities and 2 repeats are the same on both legs.
 176. §9 (Codex L2): “had a 30-day life, 8.5% of the registered design” → “had a 30-day age limit, leaving 20 eligible trial epochs, 8.5% of the registered 234”.
 177. §1: the double parenthesis after “(option B)” merged into one. Working list 24. Status header: rc20.
- rc21: review of rc20 applied** (source: the Fable read of rc20, NO-GO with two MEDIUM and six LOW and no HIGH; each finding verified against the manuscript and DEVIATIONS §10.13, §10.14, §10.22, §10.28 and §10.29 before it was applied, and three adapted where the proposed text said more than the source; per-item record in **_sprint-2026-10-04/APPLY-B-rc21.md**). The SHAM-JANELA blocks are unchanged.
178. §3.0.1 (Fable M-1): “in a session that recorded it and left the design decision open” → “on the day the decision was taken not to widen the window and so to end the trial at the 09–20 epoch” (§10.13 records the morning decision to widen, §10.14 its reversal the same afternoon, §10.22 the instruction). Not “in the session that took the decision”: the log does not establish that the measurement and the

decision were one session.

- 179. Abstract (Fable M-2): “Three commitments made before the seed were not kept” → “Three further commitments ..., besides the primary outcome and the horizon above, ...”; the Abstract names the primary switch and the horizon before this sentence.
- 180. §3.0 (Fable L1, L2): the two DEVIATIONS sections are named for what each holds (§10.14, the decision not to widen the window; §10.22, the instruction of 14:38 local time). “Not a data-dependent stop” → “not an outcome-dependent stop”, resting on measurements, not on outcomes, among them the expiry and the positive-control reading of 09-08 (20 of 672 briefs changed). “Among them”: §10.14 also weighs the cost of widening the window, and cites a replay on the realigned corpus that §10.15 later withdrew as a ground.
- 181. §1.1 (Fable L3): the closure-decision date is added to what requires trusting us, with what it rests on (the deviation log’s written times; the dry-run receipt of 2026-09-09 17:42:31Z quoted in §10.28).
- 182. §3.0.1 (Fable L4): “the dose could reach a designated item” → “could have reached”; 09-20 was a control epoch.
- 183. rc20 changelog item 174 (Fable L5): “which neither 234 epochs nor the calendar cap reached” → “which was reached by neither 234 epochs nor the calendar cap”.
- 184. §9 (Fable L6): “the feasibility stop we found ... while preparing this version” → “while preparing rc8” (rc8 changelog item 104). Working list 24. Status header: rc21.

rc22: final read of rc21 applied (source: Codex’s final read of rc21, GO with three LOW, receipt `adversary-receipt-codex-2026-10-06T133927-60502.txt`, `exit: 0`; each finding verified against the artifacts and the deviation log before it was applied; per-item record in `_sprint-2026-10-04/APPLY-B-rc22.md`). rc22 is the text frozen until the whole-window sham result is integrated. The SHAM-JANELA blocks are unchanged.

- 185. §4 Adjudication (Codex LOW 1): the unknown share of opportunities on the registered leg 1.02% → 1.00% (`out/ITT-REGISTRADO-v4-2026-10-05.json`, `pernas.registrado.unknown_ponderado_sobre_oportunidades = 0.009998`; 1.02% was the v2 value, 0.010193). The 1.34% of the sensitivity analysis (leg `atual`, 0.013412) holds.
- 186. Working list 4 (Codex LOW 2): the measurement dated 2026-09-09 and its incorporation into the manuscript 2026-09-21; only dose eligibility, not continued data collection, was limited to 20 epochs.
- 187. Appendix A (Codex LOW 3): §10.14 and §10.22, the closure decision of §3.0, named beside §10.29 through §10.34. Working list 24. Status header: rc22.

rc23: whole-window sham result integrated; ballast extended (source: the job `job-janela2`, CON-CLUIDO 2026-10-07T02:09:46Z, 21/21 runs validated, summarized by the unchanged `sprint-resume-sham-v2.py` under the statistic and p rule fixed before it ran; per-item record in `_sprint-2026-10-04/APPLY-B-rc23.md`). The ten SHAM-JANELA blocks are replaced by their result text and the markers removed. No number of the H1 family moves.

- 188. §4.0.1c: “The whole-window sham, run” and “What it adds” (real 790 / 868 against 449–583 / 484–612, 0 shams ≥ real, $p = 1/21$; positive control 894 against 462–591; fidelity 11,812/11,812 and 2,016/2,016; 4,032/4,032 shared records per run; 235 states without a bonus, all after the expiry); the first limit now points to it; two rows of “What was measured”; the `dist/` row; “What this does to the claims”; the artifacts (`B-sham-v2/job-janela2/RESUMO.json`, `DETERMINISMO.json`, `JANELA-LANCAMENTO.md`).
- 189. §4.0.1b table, specificity row: the window result beside the $w = 4$ one.
- 190. Abstract: two sham replays, the second containing the states of the first.
- 191. §7 “Scope of the specificity control”, §8.3 and §8.5: the window, the 2,016 shared states, what still lies outside.
- 192. §1.1: `ITT-PRELIMINAR.json` entered the ballast manifest only in rc23.
- 193. Appendix B: rows for the whole-window sham and for `ITT-PRELIMINAR.json`, `STABILITY-TEST.md` and `receipts/`; nineteen rows of rc7–rc19 “in the ballast manifest from rc23”; the manifest row; the ballast paragraph (115 artifacts, 79/79 on each leg, 36/36 from `origin/main`). B.1: the whole-window sham row.
- 194. Working list 8, 9, 10, 15 (done), 17 (done), 24. Status header: rc23.

rc24: final reads of rc23 applied (source: two full reads of rc23, both GO with no MEDIUM or HIGH:

Fable, an agent run with four LOW, and Codex, receipt `adversary-receipt-codex-2026-10-07T093115-73165.txt`, `exit: 0`, verdict `REVIEW-B-rc23-codex-2026-10-07.md`, five LOW; each finding verified against the artifacts before it was applied; per-item record in `_sprint-2026-10-04/APPLY-B-rc24.md`). rc24 changes wording only: no number of the analysis or of the sham results moves.

195. §4.0.1c (Fable LOW 1): the calibration-sample comparison is the real run's records against that sample (400/400), not all 21 runs (`B-sham-v2/job-janela2/DETERMINISMO.json`, `REAL_vs_amostra200`; there is no sham-against-sample field).
196. Status header and working list 24 (Fable LOW 2): both GO reads of rc23 recorded.
197. §4.0.1c first limit and working list 15 (Fable LOW 3): the runner start 2026-10-05T09:40:32Z (`RECIBO.txt`) beside the launcher time 2026-10-05T09:39:53Z.
198. Status header (Fable LOW 4): 790 "of the 11,812 reconstructible brief states", not "of the trial window".
199. §4.0.1c limits (Codex LOW 1): the rank p-value is "not a calibrated randomization p-value".
200. §4.0.1c "What it adds" (Codex LOW 2): 449–583 is "the range produced by these 20 matched designations".
201. §8.3 (Codex LOW 3): the whole-window run is a second run, 2026-10-05 to 2026-10-07, not the replay of 2026-10-04.
202. Appendix B, v3 row (Codex LOW 4): the corrected `RESULTADO-v3.md` is covered through the `B-registered/` directory hash of the rc23 manifest (recomputed: 2555baf8..., 19 files).
203. Appendix B, manifest row (Codex LOW 5): the manifest itself and the two records of the final Codex read of rc23 named among the exceptions; a row for those two records, which were written after the manifest. Working list 24. Status header: rc24.

rc25: diff check of rc24 applied (source: Fable's diff check of rc23 → rc24, an agent run that left no file, GO with three LOW, all in Appendix B and B.1; each verified before it was applied; per-item record in `_sprint-2026-10-04/APPLY-B-rc25.md`). rc25 changes wording only: no number of the analysis or of the sham results moves.

204. Appendix B, row of the rc23 Codex read (Fable LOW 1): the receipt's stamp 2026-10-07T093115 is local time, UTC–3 (`scripts/adversary-run.sh` calls `date` without `-u`), so the read ran at 12:31Z, after the 12:02Z extension of the manifest.
205. B.1, ballast paragraph (Fable LOW 2): the two records of the final Codex read of rc23 are named beside `B-censo/raw/` as not in the manifest.
206. Appendix B, manifest row (Fable LOW 3): "`B-censo/raw/` (see its row)" in parentheses, so the exception list reads as three items. Working list 24. Status header: rc25.

rc26: header for the deposit (source: the author's decisions for the deposit of this version, 2026-10-07). rc26 changes the status header and adds this block; nothing else in the text changes, and no number of the analysis or of the sham results moves.

207. Status header: rewritten for the deposit. It states that this text is version 2.0 of the pre-registration record, with the concept DOI as recorded in the v1.12 snapshot (`_sprint-2026-10-04/B-rc18/zenodo-22110203-record.json`); that the text was reviewed in full by two model families, both GO (Fable on rc23 and on the rc23 → rc24 diff; Codex on rc23, receipt `adversary-receipt-codex-2026-10-07T093115-73165.txt`); and the DOI of this version, as a single placeholder token that the deposit build replaces with the DOI reserved for this version. The rc3–rc25 history the header carried is in this changelog.

rc27: working list closed for the deposit (source: the author's decision that the deposited text must not read as pending, 2026-10-07). rc27 changes the working list and adds this block; the status header and everything else in the text are rc26 byte for byte, and no number of the analysis or of the sham results moves. Each item is closed the way the list closes items: the title struck, its history kept, and a dated rc27 note appended.

208. Working list 8 (deposit): closed as deposited as version 2.0 of the registration record, with the version DOI written as the same placeholder token as the status header, which the deposit build replaces in both places.

209. Working list 24 (reviews from rc11 on): closed. rc24 had Fable’s diff check, GO; rc25 (wording) and rc26 (header) had no read of their own and each passed a parity check; rc27 changes only the working list and the changelog; the deposited package gets a final read by Grok, Codex and Fable before publication. That read is stated as planned, not as done.
210. Working list, other lines that read as open: 21 (review of rc9) closed as superseded by the reads of rc10 (item 23); 11 (§4.7 traceability) closed as the declared gap of §4.7 and B.1; 13 (M10, all-partials leg) closed as not computed, declared in §4.6 and §5; notes on 6 (B1 headline, closed by 11), 7 (two literature checks not done, for submission, not this deposit) and on the “itself has not been reviewed” lines of 14, 16 and 19 (each superseded by the read of the next candidate).
- rc28: final reads of the rc27 package applied** (source: two final reads of the package built from rc27, Grok (receipt `adversary-receipt-grok-2026-10-07T181540-7884.txt`) and Fable (an agent run), both NO-GO, 2026-10-07; the Codex read did not run). Every finding was verified before it was applied (`APPLY-B-rc28.md`). No number of the analysis or of the sham results moves; `B-rc28/parity-rc28.py` reverts the declared replacements and gets rc27 byte for byte.
211. Title (Grok M4; the author’s decision): “A registered horizon that outlived its intervention” → “A registered horizon that outlived the intervention the trial ran”, line 1, the wording §1.1, §7 and §9 already use. The deposit title follows line 1.
212. Status header (Fable H1): “Versions 1.0 to 1.12 of the record are the registration and its amendments” was wrong. The public API lists three published versions before this one: 1.9 (record 21964094, 2026-08-17), 1.11 (record 21978476, 2026-08-17), a correction whose own description says 1.10 was written and never deposited, and 1.12 (record 22110203, 2026-08-26), an amendment; the v2.0 draft reads `versions.index` 4, which is those three plus one (`_sprint-2026-10-04/B-rc28/zenodo-21964093-versions.json`).
213. Status header (Fable M2): it now says that rc27 changes only the working list and the changelog (items 208–210), that the rc27 package had two final reads, both NO-GO, applied here, that a confirmation read of rc28 is planned, and “rc3 to rc28”.
214. Abstract (Grok M5): “The trial retained the registered 234-epoch horizon while adopting a fixed designation that was eligible for only 20 epochs” → the deposited registration was not amended and still said 234 epochs; the trial adopted an undeposited designation eligible for 20 epochs and closed when it expired (§3.0, §3.0.1, Appendix A).
215. Abstract (Grok M3, M1): the whole-window replay forces $w = 4$ in every state, control epochs included (§4.0.1c: “The dose is replayed at $w = 4$ in every state, whatever arm production served there”), so its 790 against 449–583 is a counterfactual outside the $w = 4$ epochs. The claim that the dose acted specifically on what was served now rests on the 2,646 states of the $w = 4$ epochs (132 against 81–122), where the replay reproduces production; “contains the states of the first” → shares 2,016 states with it (the $w = 4$ epochs except 09–01; 2,646 – 630).
216. §4.0.1b (specificity row), §4.0.1c (“What it adds”, “What this does to the claims”) (Grok M3): the same distinction; “more through the designated items than through matched others” is now limited to the $w = 4$ epochs.
217. §7 (scope of the specificity control), §8.3 and §8.5 (Grok M3): the same distinction.
218. §1.1, caveat (Grok LOW): “On 2026-09-10 nine of the epochs were still projection” → the spec marked nine as projection (09–11...09–19); 09–20, entered as partial by its clock hours, had not begun either, and 09–10 was still open (`SPEC-ANALISE-2026-09-10.md`, table of §1 and its note).
219. Working list 8 (Fable LOW): “Done, rc27 (2026-10-07)” → “Done (rc27)”, with an rc28 note; the deposit is published by hand after the text is frozen, so the note does not date it.
220. Working list 24 (Fable M3): the rc27 sentence “The deposited package gets a final read by Grok, Codex and Fable before publication” is struck and an rc28 note says what happened: Grok and Fable read the package, both NO-GO, findings applied here; Codex did not run (exit 1, account out of credits); rc28 itself has not been read.
221. Deposit description (outside this text; Fable H1, M1, LOW; Grok M1, M2, M3, M6 and four LOW): the published versions as in 212; “declares every deviation” → Appendix A lists the twelve a reader cannot reconstruct from the estimates and the packaged append-only log holds the rest; six of them listed, said to be six; the expiry instant and the 20th epoch; 09–02 served no brief; the served arms came from

- `assign_arms.py`; the sensitivity analysis's H1 rejection under the deposited reading ($p = 0.0127$); H2's intervals exclude zero under BCa only, and the raw time difference has the opposite sign; the power statement's planning assumptions; the sham paragraph as in 215; the grouped amendment amends this Zenodo record, and OSF yf7d2 is unchanged (its `date_modified` is 2026-08-18T07:56:44Z).
222. `receipts/`: the Grok receipt of the final read and the Codex receipt of the call that did not run, copied with the host name and the home directory path redacted.