Self-Cleaning and Captured Anyway: A Measured Primitive for Error in a Store an Agent Writes to Itself

Anonymous authors
Paper under double-blind review

Abstract

Agents increasingly write their own conclusions into a store they later retrieve from, closing a loop in which an error can be read back as evidence for itself. That contamination has been reported as a one-way process whose severity decays, and it is summarised by per-model averages. Against the store deployed systems actually use – one that only ever appends – the first description is wrong and the second has no referent. We take the loop to its infinite-tenure limit against an append-only store, prove that because writing never deletes the reachable state space has a hard upper edge at $(n{-}1)/n$, and measure the process with a single primitive carrying no fitted parameter: the copy function $\gamma(\varphi)$, the probability a model repeats what its store says. The outcome is then a choice between two edges rather than a decay: at $f_0{=}0.9$ the interval between the two modes holds 3.6% of 220 runs where a uniform spread would put 20.6%, and is strictly empty on the first 15. The primitive orders every model's capture rate to within one seed and survives the move to real facts, its sign against the critical value predicting the direction of drift on 353 of 360 real-fact runs, and scale does not rescue the store. Of four interventions with criteria frozen first, timing dominates fraction at matched budget, a consistency gate helps the false majority, and provenance marking moves nothing unless the mark correlates with truth. Two consequences travel past this apparatus: a falling exact-match score on such a store measures loss of determinism, not of information, at a measured mismeasurement of $+0.383$; and the resampling unit is the seed, at a design effect of 3.75 under which one of our own rank-correlation claims falls from Spearman $+0.98$ at three seeds to between $+0.31$ and $+0.80$ at forty-four. An append-only store's fate is therefore readable in advance from one parameter-free measurement, at a seed budget the field currently underspends.

1 Introduction

An agent that writes notes to itself and later reads them has closed a loop: its output becomes its input. This is the deployment-time analogue of training on generated data, with one difference that changes the mathematics: the store is append-only, so nothing is removed, only diluted. Let f be the false share of the store and n the number of records retrieved per step. The mean field is $\dot{f} = (\gamma_{\text{eff}}(f) - f)/n$, and because writing never deletes, the reachable state space has a hard upper edge at $(n{-}1)/n$, which for our apparatus is 0.9938. Under a bounded grounding budget R, $f_{\max}(R) = (n_{\text{false},0} + W)/(n_0 + W + R)$: the governing quantity is a budget rather than a fixed fraction of verified data, and the same budget spent early or late gives different outcomes – where our setting and the generational model-collapse literature come apart, since there real data is present every round and reusable without limit, so no scheduling question arises.

Contributions. (i) The outcome distribution is bimodal rather than a decaying mean, and one measured primitive with no fitted parameter, $\gamma(\varphi)$, says which way a model goes (§4.1); it transports – measured at the resolution it must predict at, calibrated on three topics it carries to forty-four, measured coarser it fails one-sidedly. (ii) A falling exact-match score on

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

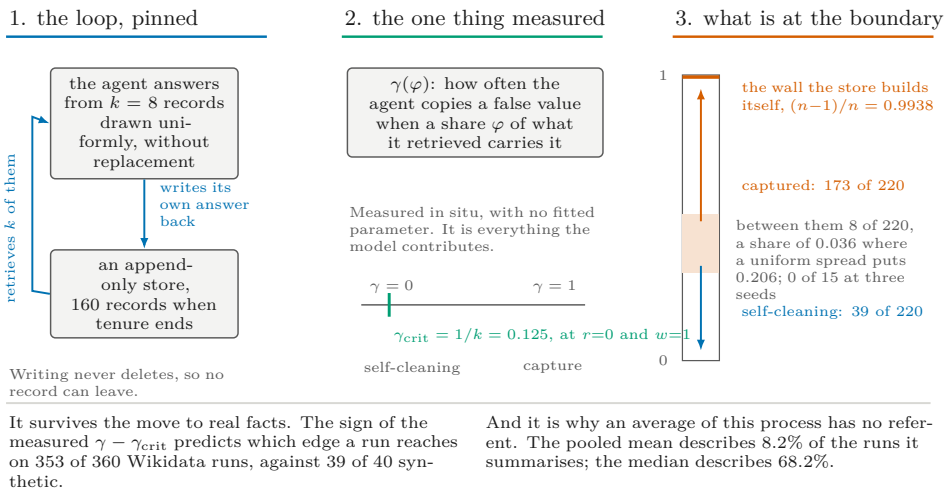

Figure 1. The apparatus, the primitive, and the boundary. An agent answers from a store it writes back into; because the store is append-only no record can leave it. Everything the model contributes travels through one measured function, the copy probability $\gamma(\varphi)$, and the sign of its distance from γ_{crit} says which edge a run reaches: not a decay towards a mean but a choice between the floor and a ceiling the store builds for itself.

a contaminated append-only store measures loss of determinism, not of information (§4.2). (iii) Four interventions with criteria frozen before their data (§4.3). (iv) Three axes of our own, taken to destruction against pre-registrations we wrote first. (v) A methodological result that reaches past this paper: the resampling unit is the seed, whose variance share reads 2.0% at three seeds and 68.8% at nine, a design effect of 3.75 (§4.6). (vi) That result applied to every row it had not reached (§4.6). (vii) The scope of claim 1, measured rather than argued: the empty interval is not a consequence of a two-valued answer space, and the regime it holds in has a boundary in how far a store's falsehoods scatter (§4.1).

What stands. Four explanations offered along the way fail their tests (the $1/n$ timescale, the variance bound, one of the two γ extrapolations, and the balance between the two modes), and what survives all of them:

(i) Bimodality, with the interval between the modes strictly empty on 15 runs and holding 3.6% of 220 against a uniform 20.6%, at four levels of f_0, and surviving at frontier scale against our registered prediction that it would not.

(ii) $\gamma(\varphi)$ is sufficient: one measured one-dimensional function, zero fitted parameters, orders every model's capture rate and transports across a seed split.

(iii) Timing dominates fraction at a matched budget (0.20/0.33/0.40 against 0.47, monotone in 5/5 models), the one result that transfers directly to a deployed schedule.

(iv) Score $= C \cdot s$: contamination costs determinism rather than information, so a score drop is not a capability drop.

(v) The retrieval rule the other four are measured under is no longer an assumption. The attached falsifier is fired: an embedding ranking moves capture by 0.500, so the results are scoped to uniform retrieval by name, and the resource that moves them is the variance of the evidence stream rather than its breadth (§4.4).

Every refuted item is an explanation offered for one of these, never the measurement.

What we do not claim. Not priority: that an agent's own outputs degrade its later behaviour is prior work (App. A). Not a location: no statistic here predicts where the boundary between the two outcomes sits. Not transfer to generational retraining, nor incidence in deployed stores – our worlds are synthetic because traceability requires it. No longer retrieval in general: §4.4 measures a 0.500 displacement under an embedding ranking, so every number here is a statement about uniform-without-replacement. Two concurrent manuscripts of the

What survives every attack in this paper		
1 bimodal, the middle sparse and not empty	0/15 in $[0.35, 0.50]$ at three seeds, 7/220 at 220; the graded share inside $(0.306, 0.512)$ is 0.036 against a uniform 0.206; survives $f_0 \in \{0.1, 0.3, 0.5\}$ and frontier scale (pooled 0.786)	§4.1
2 $\gamma(\varphi)$ is sufficient	zero fitted parameters; ordering exact at 44 seeds, $\rho = +1.00$, $p = 0.017$	§4.1
3 timing dominates fraction	0.20/0.33/0.40 vs 0.47, 5/5 models; ordering holds at 44 seeds, 220 runs per arm	§4.3
4 score $= C \cdot s$	pass@8 − pass@1 = +0.383 on stores nothing was removed from	§4.2
5 uniform retrieval is not innocuous	embedding ranking displaces capture by 0.500; cause is ranking determinism	§4.4

Explanations the data refuted, and what refuted each	
the $1/n$ timescale	Prop. 3: vacuous over 49 of 150 steps
the variance bound	1.07× against a required 2.83×
the mode balance	8/7 became 39/173 at 15× the sample
γ from three topics	12.1 seed units out of sample, one-sidedly
early copy rate orders capture	ρ +0.80 to +0.31 at 44 seeds, was +0.98 at three
the fill rule as the diagnosis	it fires for one model, the one arm A predicts exactly

Every refuted row is an explanation offered for a survivor, never a measurement. Two of the six fell to this paper's own resampling-unit result (§4.6) applied to rows it had not yet reached; the control those rows were measured against moves by 0.68 in capture rate between three seeds and forty-four, and no verdict changes when it is replaced.

Figure 2. The paper in one page: five results, and the six explanations the data refuted. Recomputed from results/ at draw time, so it cannot drift from the artifact it summarises. The five survivors are the whole of what we ask a reader to carry away; the six refuted explanations are kept as a ledger because a corrected record is part of the argument. Two of the six fell to this paper's own resampling-unit result (§4.6) applied to rows it had not yet reached.

same research programme exist and nothing below depends on either: one sends the item count to infinity against a fixed resolving power, one sends the instance count to infinity against a fixed per-paper testing budget. Every definition this paper borrows is restated here and no number is taken from either.

2 Related work

We ask one question of each neighbouring result: what must be true for its conclusion to hold, and does our apparatus put that assumption under strain? App. A tabulates every neighbour against what is added here; the four load-bearing distinctions are stated below.

Model collapse assumes the damage is summarised by a decaying statistic. Each result in that line is a trajectory of a summary statistic across generations, which needs the per-run outcome distribution to be unimodal enough for its mean to denote something. §4.1 measures the same loop at fixed tenure and finds otherwise, and a registered arm on the pooled mean itself finds it describes 8.2% of what it summarises against 74.5% for the best single centre at that width (App. B): this distribution summarises perfectly well; it is the mean that does not. Two structural differences keep the settings apart: retraining averages

over a corpus while an agent's read samples k of n records, so the variance driving our bimodality is averaged away there by construction; and real data is present every round and reusable without limit, the corner in which our schedule effect vanishes.

Agent-memory contamination assumes contamination is a quantity to be reduced. The shared move treats contamination as a level to drive down, which assumes the final level is a monotone function of the level along the way. §4.1 measures that it is not: the sign of the trajectory is fixed in the first few steps and the same f_0 produces both outcomes, so an intervention's value is set by when it lands rather than by how much it removes.

The closest result assumes the store is not append-only, which is exactly the assumption we pin. A steady-state occupancy of erroneous coordinates under a coordinate-transition model over a fixed, replaceable coordinate set is the same object as our f under the opposite structural assumption: replacement gives an ergodic chain with a stationary distribution, while append-only forbids replacement, puts a hard edge at $(n-1)/n$, and leaves a choice between edges. The two results do not compete: our ceiling proposition is vacuous the moment a record can be overwritten, and a stationary occupancy is undefined in our setting.

Memory poisoning assumes an adversary, and our store has none. Every false record after step 0 is written by the agent itself, in good faith, from what it read, so a detector trained to find injected content has no positive class here and the store still reaches 0.99 false. Contamination needs no adversary, and a defence keyed to one does not bound it.

3 The synthetic apparatus

One tracked topic, a generated entity with a generated numeric attribute; the false value is a one-digit corruption, plausible, never equal, unreachable by guessing. Store at t=0: n_0=10 records, $f_0 \in \{0.1, 0.9\}$. Each step retrieves k=8 records uniformly from the topic, asks the plain question a deployed assistant would be asked, and appends the answer; the prompt never states the truth or warns of error. Tenure is 150 steps, growing the store from 11 to 161 records; five models, three seeds, temperature 0. The world is generated because traceability requires it: a propagated error is attributable only if the truth is known by construction. Retrieval is uniform-without-replacement because that is what the dynamics below assume, and the choice carries a falsifier that §4.4 fires.

Table 1. The protocol. Every arm is this loop with one thing changed.

world	one tracked topic: a generated entity (Baldrix Holdings) with a generated numeric attribute (68.91M). The false value is a one-digit corruption: plausible, never equal, unreachable by guessing
store at t=0	n_0=10 records, $n_{\text{false}} \in \{1, 9\}$ of them asserting the corruption, giving $f_0 \in \{0.1, 0.9\}$
each step	retrieve k=8 records uniformly from the topic; ask the plain question a deployed assistant would be asked; append the answer
prompt	never states the truth, never warns of error, so γ measures ordinary behaviour rather than a response to a warning
tenure	150 steps, growing the store from 11 to 161 records: a 15× range in n, which is what makes the $1/n$ term visible within one run
models	five, three seeds, temperature 0

Definition 1 (The copy function). $\gamma(\varphi)$ is the probability that the agent asserts a false value when a fraction φ of the k retrieved records assert it. With f the false share of the store, $\varphi \sim \text{Bin}(k, f)/k$, so the per-step assertion probability is $\gamma_{\text{eff}}(f) = \mathbb{E}_\varphi[\gamma(\varphi)]$ and, with one write per step,

$$\dot{f} = (\gamma_{\text{eff}}(f) - f)/n. \tag{1}$$

γ is measured in situ, pooling over steps, from the same runs whose trajectories it predicts – never fitted to a trajectory.

The dynamics license a per-step statement, and this paper's empirical content is not carried by it. Three readings of the equation's $1/n$ denominator were tried and two failed: read as fixing how long the outcome stays open, refuted by Prop. 3, whose bound is vacuous over the interval in which the outcome is actually decided (t=49 of 150, against a measured decision point of t=2); read as fixing the spread of outcomes, refuted by an n_0 sweep at 1.07× against a required 2.83×. What survives is the per-step bound and the ceiling, both exact and both about what the store can hold. The empirical content – the existence of two outcomes, which model goes which way, the ordering – is carried entirely by measurements of $\gamma(\varphi)$ put back into the loop.

Proposition 1 (Append-only ceiling). Let no record ever be removed. Then the false share obeys

$$f(t) \le \big(n(t) - (n_0 - n_{\text{false}}(0))\big)/n(t) < 1 \tag{2}$$

for all finite t, with equality exactly when every write after t=0 is false. Under a grounding budget of R steps, $f_{\max}(R) = (n_{\text{false}}(0) + W)/(n_0 + W + R)$, W the number of ungrounded writes actually made.

Proposition 2 (Log-dilution). A grounded write at step t multiplies f by $n/(n{+}1)$, $n = n_0{+}t$; R grounded writes at steps $t_1 \dots t_R$ give total log-dilution

$$L = \sum_i \tfrac{1}{n_0+t_i} + O(\cdot), \quad L_{\text{front}} \approx \ln \tfrac{n_0+R}{n_0}, \quad L_{\text{unif}} \approx \tfrac{R}{T} \ln \tfrac{n_0+T}{n_0}, \quad L_{\text{back}} \approx \ln \tfrac{n_0+T}{n_0+T-R} \tag{3}$$

for the three schedules placing the R writes first, evenly, and last – strictly ordered $L_{\text{front}} > L_{\text{unif}} > L_{\text{back}}$ for every $0 < R < T$, with no free parameter. The ordering is the claim; the magnitudes are not, since a bounded observable must compress an unbounded one.

Proposition 3 (Per-step bound). Each step appends exactly one record, so $|f(t{+}1) - f(t)| \le 1/(n_0{+}t{+}1)$ and the cumulative bound $\sum_{s=t}^{T-1} 1/(n_0{+}s{+}1) = \log \frac{n_0+T}{n_0+t} + O(n_0^{-1})$ is logarithmic in the remaining tenure and constrains nothing until it falls below 1: at n_0=10, T=150 that is t=49, so for the first 49 steps the bound permits f to reach either endpoint from anywhere – the measured decision point is t=2, so the timescale is a measurement and not a consequence of the $1/n$ factor.

Proposition 4 (Resolution at s seeds). A two-valued f_{end} with choice probability p over s seeds has

$$\text{se}(\hat{p}) = \sqrt{p(1-p)/s} \approx 0.29 \text{ at } s{=}3; \tag{4}$$

pooled over m models sharing s seeds with intraclass correlation ρ, the effective sample size is $sm/(1 + (m{-}1)\rho)$. A level (pooled across models) carries the full design effect, measured at 3.56–3.66 at 44 seeds; a within-seed contrast cancels the seed effect, measured at 1.14–1.47 on the same runs – conflating the two is this paper's own §4.6 failure mode, corrected in §4.6.

4 Results

4.1 Bimodal outcomes, predicted by γ

At f_0=0.9, every run sorted: eight end below 0.31, seven above 0.51, the interval between them empty on the first 15 of 220 and holding 3.6% overall against a uniform 20.6%. Replacing the agent with a rule that asserts false with the measured $\gamma(j)$ and simulating the identical loop puts every model's capture rate within one seed of the observed count, nothing fitted: Spearman +0.82 at three seeds, with one adjacent inversion (simulated gap 0.02, unresolvable at three seeds). Two objections to the empty interval, both settled without an API call. A two-valued answer space does not by itself empty it: a state-independent reader on the identical urn puts 113 of 300 runs inside the dead band at p=0.3; what empties it is the copy probability rising with the number of false records retrieved. Whether a scattered store (many distinct falsehoods) still empties the interval is answered oppositely by two arms – on the synthetic world the highest-scattering model has 0 of 44 runs inside it, while on 36 real-fact (Wikidata) runs the ones whose store went multi-valued have 6.4× the interval occupancy of those that did not – and a registered arm built to decide between them (300 runs, three falsehood-distinguishability cells) instead falsified the mechanism common

to both readings: held out over 44,989 steps, γ as a function of $\varphi = j/k$ beats a scripted largest-single-value-block reader by 1491 in log-likelihood, so no scattering mechanism can act through γ, because γ does not see scattering. The two-arm disagreement stays split, narrowed rather than closed (App. S).

Claim 2 out of sample, at a 14× harder bar. The published γ curve, measured on three topics, misses real capture on the 44 topics it must predict, one-sidedly (worst 8.5σ); re-measured on the predicted topics with the identical simulator it lands within 1.32σ everywhere. $\gamma(\varphi)$ is therefore sufficient, but γ measured on three topics is not γ: a split-half re-partition shows fitting on unseen topics costs 15–16% of accuracy, not a factor of anything, and the one-sidedness disappears (App. T). The mean-field limit's own variance prediction fails. The drift half of the limit stands; four sweeps at $n_0 \in \{10, 20, 40, 80\}$ refute the variance half, an eightfold increase shrinking within-model spread by 1.07× against a required 2.83× – conditional independence between successive reads fails, and no mechanism for that correlation is derived anywhere in this paper (App. F).

4.2 What a falling score measures

Neither observed mode is an attractor of γ_{eff}: one is escape toward f=0, the other saturation against the structural ceiling 0.9938, where 4 of 7 captured runs finish within two records of it. In an append-only store information cannot be lost, only diluted: the truth is retrievable on any draw with probability $1 - f^k > 0$, at least 0.57 at f=0.9, k=8, so a falling exact-match score here is necessarily not a loss of what the store holds. Contamination is loss of determinism, not loss of information. Writing C for the fraction solvable at all and s for the share of draws committing to one version, an exact-match score reads $C \cdot s$; the control measures the resulting mismeasurement at +0.383 (single-draw accuracy against the pass@8 rate). This scope condition needs an append-only store: under eviction or pruning, truth records can leave, C is no longer conserved, and the decomposition is untested.

4.3 Four interventions

Each intervention has its criteria in prereg/, committed before its data existed, and is reported at whatever the criteria say; Table 2 lists what was at risk and what came back. Grounding schedule. Fifteen truth writes in 150 steps, three schedules (front/uniform/back), everything else identical: pooled capture rate 0.20/0.33/0.40 against an ungrounded control's 0.47, monotone in all five models; a further sweep at $R \in \{5, 45\}$ confirms the ordering at every budget, and at R=45 front-loading leaves capture at 0.07 against back-loading's 0.47 – forty-five grounding writes placed late buy exactly nothing. Log-dilution (Prop. 2) predicts the same order at $L = 0.947/0.332/0.099$ ($1 : 0.35 : 0.10$) against a measured $1 : 0.52 : 0.26$: compressed, as a bounded observable must compress an unbounded one; the magnitude bar itself failed through a calibration error, one model captured in every arm and another in none. Majority admission gate helps or is neutral at low f_0 and is harmful at high f_0: once the store is majority-false, the gate admits exactly the falsehood and rejects the correction, the record-level analogue of adversarial-majority failure among agents. Cost-matched state slot helps at low f_0; a tie at high f_0 bounds only the smallest member of the parent framework's state-channel class (§5). Source marker is harmless at low f_0 but lowers both capture and γ at high f_0: marking moves outcomes only when the mark correlates with truth, and in this loop it does not.

4.4 The retrieval rule's own falsifier

Seven arms on one shared batch test whether uniform retrieval is innocuous. It is not, and the direction is the opposite of what we predicted. We predicted ranking would make collapse more likely; it makes it less at low contamination ($0.533 \to 0.033$ capture at f_0=0.1) and polarises at high contamination ($0.900 \to 1.000$ at f_0=0.9): uniform retrieval hedges in both directions, a deterministic ranking drives the store to whichever end it started nearer. The mechanism is not the one the preregistration was built on: the intended variable, effective retrieved breadth k_{eff}, misses the match it was defined to catch (ranking at k=8 reproduces

Table 2. What each intervention put at risk, and what came back. "Attacks" names the established account it was built to contradict; where there is none, it tests a claim of ours.

intervention	frozen criterion	attacks	outcome
grounding schedule	capture monotone in schedule, $\geq$ 4/5 models; front-loading beats no grounding by $\geq 1/3$	the fraction-of-verified-data rule (Gerstgrasser et al., 2024)	ordering supported 5/5; magnitude failed
majority admission gate	helps or is neutral at low f_0; harmful or no better at high f_0; both $\geq 4/5$	write-time admission control as an unconditional safeguard (Zhang & Li, 2026)	supported 5/5 and 5/5
cost-matched state slot	helps at low f_0, $\geq 4/5$; a direction at high f_0, $\geq 4/5$ same way	our own two readings of what a self-fed slot does	low supported; high a tie
source marker	harmless at low f_0; lowers capture at high f_0; lowers γ; each $\geq 4/5$	our own registered D13′	harmless supported; the other two failed

uniform-at-k=1 exactly, $k_{\text{eff}} = 1.40$ against 1.00); what lines the two up is the variance of the evidence stream, 0.240 against 0.060 – ranking freezes the retrieved set across steps where uniform resamples it. This is structural rather than an artefact of one embedder: three rankers (a dense embedder, BM25, character-trigram cosine) all fail to separate the truth from its one-digit corruption (largest separation 0.0052), because the query is the question and the corruption is in the answer, so no query–document score has any signal about the corrupted field. Every threshold in this paper is therefore scoped to uniform retrieval by name.

4.5 The synthetic world on a real axis

Thirty-six (entity, capital) pairs from Wikidata, stratified by measured prior knowledge p_0 (an empty-store probe), test whether the published dynamics are an artefact of a zero-prior synthetic world. The threshold transfers. $\text{sign}(\hat{\gamma} - \gamma_{\text{crit}})$ predicts $\text{sign}(f_{\text{end}} - f_0)$ on 353 of 360 real-fact runs (Wilson 95% $[0.960, 0.991]$) against 39 of 40 synthetic runs in the same batch – the denominators differ ninefold, so this is agreement on real facts rather than a matched comparison, and the threshold, $\gamma_{\text{crit}} = 1/k$ at zero retirement and one write per step, contains no term for what the model knows and needs none. The empty interval is not an artefact of zero-prior topics either: it is emptier on real facts, mass 0.014 against the synthetic control's 0.125. The copy probability does move with the prior as predicted: $\gamma(1)$ falls from 0.711 in the bottom p_0 tercile to 0.411 in the top (bootstrap $[0.125, 0.464]$). One registered prediction fails, and it is the one that matters for external validity. The synthetic world is not the $p_0 \to 0$ endpoint of the real curve: its $\gamma(1)$ is 1.000 against 0.700 at the lowest-p_0 real cell, so it bounds the copy probability from above by a measured 0.300 rather than instantiating it – the idealisation overstates the risk by a measured amount while preserving the threshold prediction.

4.6 The resampling unit, and the falsifier on the rows it had not reached

The seed selects the topic and the topic dominates the variance. A variance share reading 2.0% at three seeds reads 68.8% at nine seeds, a design effect of 3.75: each seed carries 1.33 runs of information however many models are run. Measured at 44 seeds on the schedule axis and separated by kind: a pooled level carries 3.56–3.66×; a paired within-seed contrast carries only 1.14–1.47× – the correction the four interventions' criteria need is a factor of about 1.2, not of 3.75, since each compares treated against control on the same seed. Re-evaluating each of the eight frozen intervention criteria seed by seed reproduces every published value exactly; seven of eight hold on all three seeds, and one (the low-f_0 state-slot arm) misses its own bar on one of three seeds and is downgraded to underpowered. Two of this paper's own rank-correlation claims are directly exposed, and they move in opposite

directions. Recomputed on 44-seed capture rates rather than three: the claim that early copy rate orders capture rate falls from Spearman +0.92–+0.98 (three seeds, $p \leq 0.067$) to +0.31–+0.80 (no window $p < 0.13$) – diagnosed in the predictor, whose five-model spread on the early copy rate itself collapses from 0.333 (three topics) to 0.061 (forty-four); while the claim that γ-simulated capture orders observed capture, reported at three seeds as "+0.82 with one adjacent inversion," becomes exact at forty-four (ρ=+1.00, p=0.017) – the inversion was an artefact of the three-seed denominator. The transferable rule: ask first whether a criterion is a level or a contrast; a level over m models buys $sm/(1+(m-1)\rho)$ observations ($\rho \approx 0.65$ here, so models are nearly free); a contrast is not penalised at few seeds provided it is checked seed by seed; neither rescues a per-model ordering, whose own independent variable is an estimate.

5 Discussion

Relation to the parent framework. This paper takes three things from the framework it descends from and deliberately takes no fourth. The parent's subject is a pair of channels, one compressive and composable, one a scalable verbatim index, and its claim is about the class holding both; our loop is an index channel with no state channel, and §4.3's cost-matched slot adds the smallest state channel the apparatus admits. What we return is a bound, not a verdict: the tie E-A returns bounds only that smallest member, a self-fed slot, and says nothing about a richer state channel or one fed by the stream rather than by the system's own output. No constant is taken from the family: a concurrent training-time manuscript's own load-accounting transplant of the parent's constant was excluded by its data with the direction reversed, so nothing here is calibrated against anything there.

Limitations. (i) Append-only is load-bearing, and the remedy this paper's limit recommends (forgetting) violates it, so the two halves of the paper hold in adjacent regimes; a 2026 remedy compatible with append-only exists (each observation stored as its own weighted candidate under a Noisy-OR update rather than a single overwritten conclusion), and whether an accumulating false majority drives that weight the way it drives our f is not derived here (§6, item 1). (ii) The bimodality and its timescale have no mechanism: what fails is conditional independence between successive reads, and what would fix it is a correlation model neither derived nor guessed at. (iii) The budget transient's peak is derived, not measured: our own falsifier sat on the wrong side of it, and a sweep past the peak on 44 fresh seeds did not resolve at $+0.46\sigma$ – the bill for 3σ rose to about 1895 seeds when the first instalment was paid, not fell. (iv) The f_0 divide is bought but coarse: four levels on one model rule out f_0=0.9 as the adversarial extreme (capture peaks near f_0=0.5) but cannot locate a basin boundary. (v) One retriever is one point: a character-level or hybrid index is a different experiment, and w>1 writes per step are unmeasured. (vi) The capacity coupling is vacuous here: the information-theoretic bound gives $\alpha_{\min} = 0$ for four of five models, so the surviving statement is a design rule (measure $\alpha_{\min}$ before spending on provenance) rather than a binding constraint. (vii) The synthetic world is bounded, not defended: whether deployed stores instantiate these dynamics in large measure remains unestablished.

6 What remains owed

None of the five arms below can move this paper's two central results (the append-only ceiling is proved, not measured; the bimodality is measured on seeds reported in full); each bears on mechanism and on how far the account transports to stores that are append-only in the record and mutable in the read-out. (1) Substitute the Noisy-OR read-out for our retrieval read-out, holding the append-only store fixed: one code change, no new data, and it settles whether the mutable part has merely moved from the record set to a weight vector. (2) A correlation model for successive reads – the single missing account that plausibly explains both the variance failure and the early decision timescale. (3) The budget-transient peak, priced at about 1895 seeds and left there; the ordering this paper reports never needed it. (4) A finer f_0 grid on more models, and the retriever family at w>1. (5) A stream-fed (rather

than self-fed) state slot – the only arm in this family that could move the parent framework's own status rather than describe this instrument; not run, because the apparatus needs a write-side hook it does not have.

7 Conclusion

A store an agent writes and later reads is a closed loop, and because writing never deletes, its dynamics are constrained before any model is chosen: the reachable states have a hard upper edge, information is diluted rather than lost, and what holds the outcome down is a bounded grounding budget rather than a fixed fraction of verified data. Outcomes are bimodal; a copy function calibrated once and fitted nowhere orders the models within a seed; a falling score reports commitment rather than capability, at a mismeasurement of +0.383; and at matched budget, when grounding is spent dominates how much of it there is. The most consequential result reaches past this paper: the resampling unit is the seed, and a variance share reading 2.0% at three seeds reads 68.8% at nine, so an apparatus of this shape should buy seeds and not models. Two registered arms turned the same treatment on the apparatus itself: uniform retrieval is not innocuous (an embedding ranking moves capture by 0.500 and polarises), and the synthetic world, this paper's largest exposure, is bounded rather than defended – on real facts the threshold retains 353/360 agreement while the copy probability runs a measured 0.300 lower, so the idealisation overstates the risk it bounds and preserves the prediction it makes.

Reproducibility statement

Every registered criterion's freeze commit and timestamp, read from git log rather than transcribed, is in Appendix N; every graded row, generated from the same result files the criterion judges so the index cannot disagree with the rows it indexes, is in Appendix X (87 rows: 50 for the paper, 37 against). Every number in the main text is computed by an analyzer from files in results/ at build time rather than typed by hand, and every pre-registration file, runner, grader and result file is included in the anonymised supplementary material, with MANIFEST.md mapping each cited artefact to its file. Model calls are the only cost this paper incurs; every table is regenerable from the shipped result files with no further API spend.

Ethics statement

Every corpus is synthetic (a generated entity and a one-digit numeric corruption) or drawn from public Wikidata facts by an automated, disclosed procedure; no human data was collected and no individual is identified or evaluated. The models tested are openly released or accessed through their public APIs under their standard terms; we introduce no new capability, only a measurement of how an append-only self-retrieval loop interacts with pre-existing model behaviour. We are not aware of a Code of Ethics concern raised by this work.

AI use statement

We report our use of generative AI tools against the categories of the ICLR 2027 AI Policy for Authors. Our use was confined to preparing the document.

Tasks in the required-disclosure list: none. Generative AI was not used to generate synthetic data sets, to develop the theoretical model or conceptual framework, to formulate the mathematical claims, to supply critical ingredients for the proofs or to write them, to propose or refine the hypotheses, to design the methodology or the experiments or to give feedback on them, to set or choose any pre-registered threshold, to implement methods, to clean or reformat any dataset, or to interpret the results. The synthetic worlds are emitted by fixed-seed builders, the loop runs of §3 were configured and launched by the authors, and every number reported here is computed by an analyzer the authors wrote, from the result files those runs produced; builders, analyzers, job scripts, result files and verdicts are all in

the supplementary material. Translation and qualitative or thematic data analysis are not applicable to this work.

Tasks in the recommended-disclosure list. Generative AI was used on the manuscript itself. It condensed a longer version of this work into the present conference format, which involved drafting and rewriting substantial parts of the exposition; it edited for readability, formatted tables and references, and assembled and anonymised the supplementary archive. It was not used to discover the research topic, identify the gap, select the literature, or decide what the paper claims: the claims, the numbers they quote and the qualifications attached to them are the authors', carried over from the longer version.

Verification. Every quantitative statement in this version was checked against the graded row it cites, and every registered outcome against the freeze commit tabulated in Appendix N. The one transformation AI-written code applied to a released artefact — rewriting absolute paths and host names in the supplementary archive for anonymisation — was scoped to provenance strings and verified not to have altered any result file, score, or reported number. We have reviewed all AI-assisted work and take responsibility for the final content of this paper, including its text, claims and artefacts.

References

Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G. Baraniuk. Self-consuming generative models go MAD. arXiv preprint arXiv:2307.01850, 2023.

Anonymous. The capability convergence hypothesis: Capability from access structure, not scale. arXiv preprint, 2026a. author list and repository withheld for double-blind review; a concurrent manuscript of the same research programme, arXiv identifier withheld.

Anonymous. Position: Preregistration Protects the Criterion, Not the Comparison: One Programme Audited Against Its Own Laws Seven Times. concurrent manuscript; author list and repository withheld for double-blind review, 2026b.

Anonymous. The Infinite-Rerun Limit: Computing what a criterion can decide, before it is frozen: a resolution, a rule, and the flip rate a band buys. concurrent manuscript; author list and repository withheld for double-blind review, 2026c.

Elias Bareinboim and Judea Pearl. External validity: From do-calculus to transportability across populations. Statistical Science, 29(4):579–595, 2014.

Apratim Dey and David Donoho. Universality of the $\pi^2/6$ pathway in avoiding model collapse. arXiv preprint arXiv:2410.22812, 2024.

Xin Ding, Xinrui Wang, Yifan Yang, Hao Wu, Shiqi Jiang, Qianxi Zhang, Liang Mi, Hanxin Zhu, Kun Li, Yunxin Liu, Zhibo Chen, and Ting Cao. MemCompiler: Compile, don't inject — state-conditioned memory for embodied agents. arXiv preprint arXiv:2605.07594, 2026.

Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Henry Sleight, John Hughes, Tomasz Korbak, Rajashree Agrawal, Dhruv Pai, Andrey Gromov, Daniel A. Roberts, Diyi Yang, David L. Donoho, and Sanmi Koyejo. Is model collapse inevitable? breaking the curse of recursion by accumulating real and synthetic data. arXiv preprint arXiv:2404.01413, 2024.

Hyeonjeong Ha, Jeonghwan Kim, Cheng Qian, Jiayu Liu, William M. Campbell, Yue Wu, Yuji Zhang, Kathleen McKeown, Dilek Hakkani-Tür, and Heng Ji. MemGuard: Preventing memory contamination in long-term memory-augmented large language models. arXiv preprint arXiv:2605.28009, 2026. Separates memories by functional role at storage and retrieval, so the reference class is typed rather than filtered on agreement.

Chuan He, Zebin Chen, Zhengyi Yang, Shaobo Qiao, Mingchen Ju, Jiate Liu, Dong Wen, and Guanfeng Liu. Minority sentinel: When to overturn majority voting in multi-agent LLM debates. arXiv preprint arXiv:2606.29270, 2026.

Junfeng Liao, Qizhou Wang, Jianing Zhu, Bo Du, Rui Yan, and Xiuying Chen. Belief memory: Agent memory under partial observability. arXiv preprint arXiv:2605.05583, 2026. Identifies self-reinforcing error in agent memory that stores each observation as a single deterministic conclusion, and replaces it with multiple candidate conclusions held as separate entries, each carrying a probability updated by a Noisy-OR rule as observations arrive, all candidates surfacing together at retrieval. Nothing is removed, so the design is compatible with an append-only store.

Ruixiao Lin, Xinhao Deng, Qingming Li, Jianan Ma, Yunhao Feng, Yuqi Qing, Zhenyuan Li, Yechao Zhang, Shiwen Cui, Changhua Meng, Tianwei Zhang, Xingjun Ma, Qi Li, Ke Xu, and Shouling Ji. Safety in self-evolving LLM agent systems: Threats, amplification, and case studies. arXiv preprint arXiv:2606.23075, 2026.

Jiayuan Liu, Shiyi Du, Weihua Du, Mingyu Guo, and Vincent Conitzer. The consensus trap: Rescuing multi-agent LLMs from adversarial majorities via token-level collaboration. arXiv preprint arXiv:2604.17139, 2026.

Judea Pearl and Elias Bareinboim. Transportability of causal and statistical relations: A formal approach. In Proc. 25th AAAI Conference on Artificial Intelligence, 2011.

Yian Wang, Agam Goyal, Yuen Chen, and Hari Sundaram. State contamination in memory-augmented LLM agents. arXiv preprint arXiv:2605.16746, 2026.

Guilin Zhang, Wei Jiang, Xiejiashan Wang, Aisha Behr, Kai Zhao, Jeffrey Friedman, Xu Chu, and Amine Anoun. Adaptive memory admission control for LLM agents. arXiv preprint arXiv:2603.04549, 2026. Admission on five retention factors – future utility, factual confidence, semantic novelty, temporal recency, content-type prior – on LoCoMo.

Yan Zhang and Shibo Li. ConsistencyGate: Preventing memory contamination in LLM agents via self-consistency admission control. arXiv preprint arXiv:2607.22962, 2026.

Shiding Zhu, Yudi Qi, Yajie Wang, Jiaze Li, Chao Song, Yaorui Shi, Yibo Miao, Hanqi Gao, and Kai Zhang. Escaping the self-confirmation trap: An execute-distill-verify paradigm for agentic experience learning. arXiv preprint arXiv:2606.24428, 2026.

Supplementary Material

This appendix follows the references and is supplementary. The main text is self-contained: every claim it makes is stated with the measurement that decides it, and the sections below hold the extended related work, the withdrawn/refuted axes in full, the proofs, the registration record, and the public scorecard of all 87 graded rows. Reviewers are not required to read it.

A Extended Related Work

We ask one question of each neighbouring result: what must be true for its conclusion to hold, and does our apparatus put that assumption under strain? A result that survives is one we depend on. A result that does not is not thereby wrong; its scope can now be stated more narrowly than its authors could, because we ran a limit they had no reason to run.

Feedback-loop results outside language modelling assume the loop is closed through the world. A deployed predictor that shapes the data it is later trained on is the subject of performative prediction, of runaway feedback in predictive policing, of degenerate feedback in recommenders and of model-driven amplification of dataset bias. Those loops close through an environment whose response is exogenous: the analysis turns on how the world reacts to the deployed model. Our loop closes inside one process, with no environment and no second agent, so none of that machinery applies and what is left is the reading behaviour of a single model against records it wrote itself. The lesson we keep is theirs: the fixed point of the loop, not the quality of any single prediction, is the object to analyse.

Table 3. Nearest prior work. Every row is a phenomenon this paper does not claim to have discovered.

Prior work	What it establishes	What is added here
Self-Confirmation Trap (Zhu et al., 2026)	the closed loop named: own trajectories written to memory and reused	the outcome is bimodal rather than cumulative, and one primitive predicts which branch
State contamination (Wang et al., 2026)	unsafe content survives summarisation below detector thresholds	a different failure axis; no dynamics, no copy probability there
Admission control (Zhang & Li, 2026)	write-time gating on self-consistency reduces contamination	the same gate has a regime where it erases the model's own cleaning
Majority aggregation traps (Liu et al., 2026; He et al., 2026)	one-shot majority over samples locks in correlated error	the write side over tenure, where the gate decouples composition from reader quality
Accumulate vs. replace (Gerstgrasser et al., 2024; Dey & Donoho, 2024)	accumulating real with synthetic data avoids collapse	a budgeted grounding event, where schedule dominates at fixed fraction
Exact-match and pass@k critiques	strict matching understates capability; pass^k separates capability from reliability	an intervention that conserves C while moving s
Threat amplification in self-evolving systems (Lin et al., 2026; Alemohammad et al., 2023)	a component's failure is amplified by the loop that reuses its output	the amplification as one number with a fixed point, whose attractor is Prop. 1 rather than a divergence
State-conditioned memory (Ding et al., 2026)	compiling a policy against state beats injecting retrieved text	the ceiling binds what the store contains, so it binds whatever the compiler reads

Knowledge conflict fixes the contradiction and varies the model; we let the loop generate it. Whether a model follows a context that contradicts what it has memorised is measured directly by the knowledge-conflict line, and which side wins is known to depend on how well the fact is represented in pretraining. Our $\gamma(\varphi)$ is that measurement made in situ: the conflicting evidence is produced by the agent's own earlier writes rather than authored by an experimenter, its strength is the retrieved fraction φ rather than a binary condition, and it is measured on the same runs whose trajectories it must predict. §4.5 puts the parametric axis those papers vary, our p_0, on the same plot as the loop, and finds $\gamma(1)$ falling from 0.711 to 0.411 across prior terciles while the drift threshold does not move.

Provenance assumes the mark carries information about truth. Tiered and watermarked memories and the tracing survey make each record's origin recoverable so a reader can discount derived records. §4.3 registers and measures that prediction: marking moves outcomes only when the mark correlates with truth, and in our loop it does not, the agent's own writes being as often right as its reads permit.

Consistency and majority reads assume the majority is more often right. Admission on repeated sampling, which inherits the majority rule from self-consistency decoding, majority voting failing under corrupted majorities, and the minority correct in roughly one disagreement in four all read the majority as evidence. §4.3 measures a majority admission gate inside our loop and finds it erases model differences by making every model worse alike: once the store is majority-false, the gate admits exactly the falsehood and rejects the correction. Two deployed admission controls are richer than ours and are the right comparison: one admits on five retention factors (future utility, factual confidence, semantic novelty, temporal recency and a content-type prior), so agreement with the incumbent store is one signal of five rather than the whole rule; another does what our own mechanism predicts would help, typing memories by functional role and keeping the types separate at storage and at retrieval, changing the gate's reference class instead of filtering within it, which is where we locate the harm. We do not claim our result transfers to either, and a harmful regime in the weakest gate is not conservative evidence about stronger versions, an inference that needs monotonicity in gate strength we have not tested.

The benchmarks measure retrieval; the object here is the loop. Memory benchmarks and the precision-aware evaluation line evaluate what an agent can retrieve from a memory it is given, and a separate line addresses the adjacent problem that retrieval corpora are increasingly model-generated. All hold the store fixed and vary the reader; we hold the reader fixed and let the store be written by it, making its composition endogenous rather than a condition of the experiment. §4.2 shows a system can score well on the first design and fail the second.

Guide to this appendix

The appendix is long because the paper grades seventy-two rows against nineteen preregistrations and 4,602 runs, and a graded row a reader cannot check is not graded. It is not one block: every section below is a named topic, nothing nested more than one level. Three groups, and a reader checking the paper rather than reading it wants the third.

(i) Arms that carry a claim in the body and are reported in full here (§H onward): the 220-run re-test of the empty interval, claim 2b out of sample, the frontier and f_0 scope arms the abstract quotes, the budget sweep past the peak, the accidental replicate, and the resampling-unit result that is contribution (v). These are here for length, not for weight.

(ii) Derivations, and the accounts the data refuted: the two limits of the dilution law, the variance bound and the sweep refuting it, the proofs, and the three axes we killed.

(iii) The record, generated at build time and not transcribed (§N onward): every criterion with the commit that froze it, the invariants checked on all 7,244 runs, per-model detail behind every pooled number, the four analyses this round added, and what the whole record cost. Every table in this group is written from results/ by make_appendix.py, so

an appendix that disagrees with the repository is a build failure rather than a reading someone has to catch.

Where to start if you are checking rather than reading. §R is the round-8 knife applied to the rows it had not reached, §T is the four estimators for γ, and §N is the freeze record every criterion ID resolves against. Those three are the ones that would change a verdict if they were wrong.

B E-mean: a single centre

The arm behind §2's statement that the defect is in the mean rather than in summarising, graded in Table 4.

Table 4. E-mean, 220 runs, criteria frozen before the data. "Share described" is the fraction of runs within ±0.10 of the centre, a half-width fixed in the prereg at about four seed units. The last row bounds the whole column. The whole tabular is generated from results/MEAN.json.

	centre	value	share described	verdict
P-T1-M1	pooled mean	0.792	$18/220 = 0.082$	supported (bar ≤ 0.40)
P-T1-M2	median	0.981	$150/220 = 0.682$	reported; narrows M1
	best single centre	0.894	$164/220 = 0.745$	upper bound on any one number

C Three axes and their falsifiers

Each of the following was a post-hoc reading of data in hand, turned into a pre-registration with fresh seeds and then refuted by it; §1 says why they are listed and does not need repeating here. What this appendix adds is each one in full.
(a) Damage scales with unaided cleaning ability. Three interventions appeared to hurt the cleanest model most, at rank correlation +1.00 over five models. Registered on seven fresh seeds with the generating data excluded, the ordering survives (0.264 vs 0.249) and the magnitude does not, the frozen bar being 0.30 against a measured gap of 0.015. The diagnosis is in the generating data: the "cleanest model" had a 3-seed baseline of 0.150 that becomes 0.729 on seven seeds, so the independent variable was noise. This is the failure mode of a superseded model of ours in which fixed points computed from 12-sample cells moved to zero when recomputed from 188.
(b) A marker's effect is a single function of label–truth correlation. Three points (anti-correlated, scrambled, correlated) fell monotonically in five of five models. Two new arms at partial correlation, whose positions the hypothesis predicts, break it: strict five-point monotonicity holds in 0/5, every model rising at the scrambled point above the partially anti-correlated one. Refuted; the surviving statement is the weaker P-L2.
(c) A sign reversal unifying two timescales. We proposed that one-shot aggregation failure and the store-level ratchet are the same mechanism at τ=1 and $\tau \to \infty$, distinguished by whether a better model is safer or more exposed. That rests on (a) and dies with it.

D E-retriever: three costs

§4.4 states the ranker-family check and its result in full: no ranker of the three separates a truth record from its one-digit corruption, so the mechanism is structural rather than the embedder's. It is not repeated here. What this appendix adds is the arm's cost, which the body gives in one sentence.

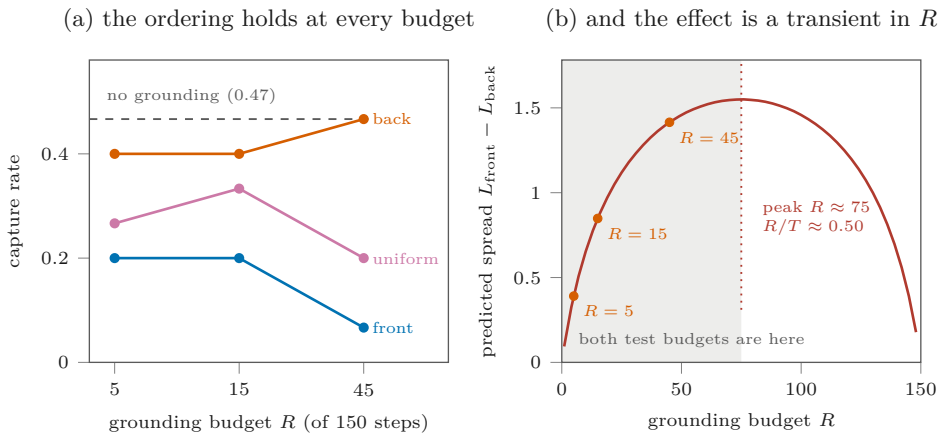

Figure 3. The ordering survives every budget; the effect's size is a transient we did not test. (a) Capture rate by schedule at $R \in \{5, 15, 45\}$ against the ungrounded control; at R=45 back-loading has returned to the control exactly. (b) The closed form (3): the predicted spread vanishes at R=0 and at R=T, where all three schedules are one arm, so it peaks near R=75. Both budgets we registered a falsifier on sit on the rising side, which is why that falsifier failed upward.

Three costs of this arm, two of them ours and one paid in dollars. (i) P-T1-R3's ladder has almost no dynamic range and is not monotone: k_{eff} runs only 1.00 to 1.95 across $k = 1$ to 8, with capture 0.000, 0.400, 0.233, 0.533 along it. The amendment predicted this compression before the data and reduced R3's weight in advance; an interpolation across a non-monotone four-point ladder cannot carry much either way. (ii) P-T1-R4 is supported but does not mean what its name says. The deduplicated arm displaces capture by 0.367 against ranking's 0.500, so deduplication does reduce the displacement; but the same amendment recorded that with only two distinct on-topic values, one-record-per-value fills two slots and pads the other six from off-topic, measured here at 59.8%. That arm tests dedup-plus-padding, and its k_{eff} of 3.22 is distractor variety rather than evidence. (iii) The run halted on its own spend guard at \$5.06 against a hard stop of \$5.00, having cost \$5.13 against a preregistered estimate of \$2.06. Thirty-three runs were dropped, all in one model's high-contamination cells, so the floor clause excludes qwen3-8b entirely and two of three models are judged. Every number above is on $n = 2$ models $\times$ 10 seeds.

One post-hoc observation, because it points at this paper's own headline. If ranking is a variance sink, its signature is a thinner middle rather than a shifted mean. Using §4.1's own cuts, unrefitted, the share of runs inside the empty interval $(0.306, 0.512)$ is 0.050 under the control and 0.000 under both ranking arms. The direction agrees, but two runs against zero out of forty is an observation and not evidence. It is recorded because it is the cheapest thing to register next.

E Two limits of the dilution law

This appendix carries items (iii) and (iv) of §4.3 in full. Both are derived from (3) and neither is measured in this paper.

(iii) The budget dependence is a transient, derived but not measured. spread(R) = $L_{\text{front}} - L_{\text{back}}$ vanishes at both ends: at R=0 nothing is grounded, and at R=T every step is, so the three schedules are one arm (Table 5, Figure 3). A quantity positive in between and zero at both ends has an interior maximum, and (3) puts it near $R/T \approx 0.5$: spread runs $0.39, 0.85, 1.42, 1.55, 1.22, 0$ at $R = 5, 15, 45, 75, 120, 150$.

(iv) The transient is in the fraction; in the tenure the effect is unbounded, and the two limits do not commute. Item (iii) is stated at fixed T and read alone invites the conclusion that schedule stops mattering once the budget is large. The apparatus's own limit says the opposite. Write $\alpha = R/T$ and let $T \to \infty$ with α fixed. Then

$$L_{\text{back}} \longrightarrow \ln \frac{1}{1-\alpha} \quad \text{(a constant)}, \qquad L_{\text{front}} \approx \ln \frac{n_0 + \alpha T}{n_0} \longrightarrow \infty \quad \text{(like } \ln T\text{)}. \tag{5}$$

L_{back} carries a hard ceiling no tenure can lift, 0.105 at α=0.1 and 0.693 at α=0.5, reached to three decimals by T=15000, while L_{front} has no ceiling. Three consequences follow from (3), none needing a new run:

Table 5. The three limits of (3) at n_0=10. The peak's location is scale-free and its height is not, which is why (iii) and (iv) are both true. The last row is the only place the three schedules coincide.

limit taken	what happens to the spread	at T=150/1500/15000	regime
R fixed, $T \to \infty$	back $\to$ 0, front $\to$ const.	—	finite budget, long tenure
α fixed, $T \to \infty$	diverges like $\ln T$	1.55/3.69/5.98 at α=$\frac{1}{2}$	ours
peak over R at fixed T	at $R^*/T = 0.50$ always	height $\approx \ln \frac{n_0+T/2}{n_0} - \ln 2$	—
$\alpha = 1$ exactly	identically 0, every T	0/0/0	free re-grounding

First, the peak of (iii) sits at $R^*/T = 0.50$ at $T = 150$, 1500 and 15000 alike: its location is a fraction and does not move with tenure. Second, its height does move, growing as $\ln T$ (1.55, 3.69, 5.98). A transient in α and an unbounded quantity in T are therefore not in tension; they are one surface read along two axes. Third, and this is the sharp one, the spread diverges for every $\alpha < 1$ and is identically zero at $\alpha = 1$, so

$$\lim_{\alpha \to 1} \lim_{T \to \infty} \text{spread}(\alpha, T) = \infty \qquad \text{while} \qquad \lim_{T \to \infty} \lim_{\alpha \to 1} \text{spread}(\alpha, T) = 0.$$

The order of the limits decides the answer. Free re-grounding is not a regime this one approaches smoothly as the budget grows; it is a single point at which the schedule axis collapses, and one step off it the axis reopens without bound.

A frozen criterion of ours embeds an assumption this derivation exposes. We registered "the schedule effect must shrink as the budget grows" as the falsifier for the boundary claim and picked $R \in \{5, 45\}$ to test it. The derivation, written after the criterion was frozen and before any of that data was read, puts the peak near $R/T \approx 0.5$, so both R=15 and R=45 lie on the rising side and the mechanism predicts the spread to grow. The criterion was set on the wrong side of a transient, and item (iv) says why the mistake was easy to make: "larger budget" is ambiguous between R at fixed T, where the effect is a transient, and α at growing T, where it is unbounded. We wrote the falsifier for the first reading and the apparatus runs in the second. We report it as frozen, do not rewrite it, and do not treat a failure in the predicted direction as a confirmation: confirming a transient needs a sweep past $R/T > 0.5$, which we have not run. Claim (iii) is therefore derived and untested, and prereg/E_BUDGET_PREREG.md records when each half was written.

F The variance bound under sweep

The decomposition behind §4.1's refutation of the $1/\sqrt{n_0}$ bound.

The natural derivation, and the bound it predicts. One write moves f by $(X - f)/(n{+}1)$ with $X \sim \text{Bernoulli}(\gamma_{\text{eff}}(f))$, so treating successive writes as conditionally independent,

$$\underbrace{\textstyle\sum_{n_0}^{N} \frac{\gamma_{\text{eff}}(1-\gamma_{\text{eff}})}{n^2} = O(1/n_0)}_{\text{total variance the process can ever accumulate}}, \qquad \underbrace{\textstyle\sum_{n_0}^{N} \frac{\gamma_{\text{eff}}(f)-f}{n} \approx (\gamma_{\text{eff}}-f) \ln \frac{N}{n_0}}_{\text{total drift}}. \tag{6}$$

The variance sum converges: unlike the drift, which grows logarithmically without bound, the noise a run can accumulate over an unbounded tenure is finite and set by how small the store was at the start. That predicts a bound: the spread of the limiting f should fall as $1/\sqrt{n_0}$, so the two outcomes should merge as the initial store grows (Figure 4).

The bound is wrong. Four sweeps at $n_0 \in \{10, 20, 40, 80\}$, three models and three seeds each, run for $3n_0$ steps so all four cover the same stretch of n/n_0, decompose the spread of f_{end} into within- and between-model parts:

n_0	10	20	40	80
within-model sd	0.318	0.282	0.291	0.298
$\times\sqrt{n_0}$	1.01	1.26	1.84	2.67
between-model sd	0.084	0.083	0.038	0.072

An eightfold increase in n_0 should shrink the within-model spread by $\sqrt{8} = 2.83$. It shrinks by 1.07. The bound fails flatly rather than marginally, so what breaks is the conditional-independence step in (6): successive reads of the same store by the same model are not independent draws from γ_{eff}, and whatever correlates them does not weaken as the store grows. The between-model part is flat, as a property of γ rather than of n_0 should be, which is what makes the within-model failure interpretable rather than a decomposition artefact.

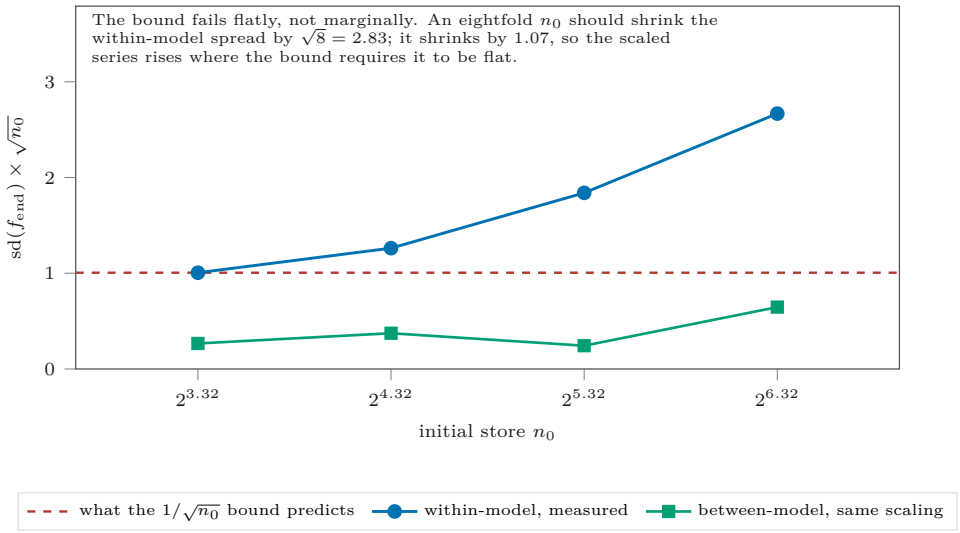

Figure 4. The bound as a flat line, and the measurement that is not one. If the spread of f_{end} fell as $1/\sqrt{n_0}$, multiplying it by $\sqrt{n_0}$ would give the same value at every store size, which is the dashed line drawn through the first point. The measured within-model series rises instead: an eightfold increase in n_0 shrinks the spread by 1.07 where the bound requires 2.83. The between-model series is drawn on the same scaling and for the same reason: flat in the raw units, as a property of γ rather than of n_0 should be, so the within-model failure is a failure of the conditional-independence step in (6) and not an artefact of the decomposition. Recomputed at build time.

G Proofs

Proof of Prop. 1. Append-only means the record set is non-decreasing under inclusion, so the $n_0 - n_{\text{false}}(0)$ true records at t=0 are true records at every t, whence $n(t) - n_{\text{false}}(t) \geq n_0 - n_{\text{false}}(0)$, which rearranges to the middle term of (2). That term is < 1 whenever the store starts with at least one truth; the apparatus has exactly one at $f_0 = 0.9$, which is why the ceiling is tight rather than decorative. Equality holds iff every one of the $n(t) - n_0$ writes after t=0 was false. For the budgeted form, a run making W ungrounded and R grounded writes has $n = n_0 + W + R$ and, grounded writes being true by construction,

$n_{\text{false}} \le n_{\text{false}}(0) + W$. Note what the bound does not say: it constrains the endpoint and says nothing about the path, which is why §4.1 needs a separate argument for the shape of the outcome distribution. □

Proof of Prop. 2. A grounded write is true and appends one record, so $f \mapsto f \cdot n/(n{+}1)$. Composing the R writes and taking logarithms, $L = \sum_i \log(1{+}(n_0{+}t_i)^{-1}) = \sum_i (n_0{+}t_i)^{-1} + O(\sum_i (n_0+t_i)^{-2})$. The three closed forms follow from $\sum_{t=a}^{b-1}(n_0+t)^{-1} = \log\frac{n_0+b}{n_0+a} + O(n_0^{-1})$ applied to $\{0,\dots,R{-}1\}$, to an arithmetic grid of spacing T/R, and to $\{T{-}R,\dots,T{-}1\}$.

The ordering needs no approximation and no parameter. The summand $(n_0{+}t)^{-1}$ is strictly decreasing, and the three schedules assign R writes to the R smallest indices, an interior grid, and the R largest. A sum of a strictly decreasing function over the R smallest indices strictly exceeds the sum over any other R-subset, which strictly exceeds the sum over the R largest; for $0 < R < T$ the uniform grid is neither extreme. Hence $L_{\text{front}} > L_{\text{unif}} > L_{\text{back}}$ for every n_0, T and R in range. □

Proof of Prop. 4. The first half is the binomial standard error of a mean of s independent Bernoulli(p) draws. For the second, index a run by (seed j, model k) with $\mathrm{Var}(X_{jk}) = \sigma^2$, $\mathrm{Corr}(X_{jk}, X_{jk'}) = \rho$ for $k \neq k'$ (runs sharing a seed share the topic and the planted falsehood, which generates ρ) and independence across seeds. Then $\mathrm{Var}\big(\frac{1}{sm}\sum_{j,k} X_{jk}\big) = \frac{\sigma^2}{sm}[1 + (m - 1)\rho]$, the variance of $sm/[1 + (m-1)\rho]$ independent observations. The bracket is the design effect; it equals 1 only at $\rho = 0$ and grows linearly in m, so adding models to a fixed seed set buys information at a rate that falls to zero. □

Proof of Prop. 3. A write changes n_{false} by 0 or 1 and n by exactly 1, so $|f(t{+}1) - f(t)| = |n_{\text{false}}(t{+}1) - f(t)(n(t)+1)|/(n(t)+1) \le 1/(n(t)+1)$ with $n(t) = n_0 + t$. Summing and comparing with $\int_t^T (n_0 + s)^{-1}\, ds$ gives the logarithm. The threshold is the least t with $\sum_{s=t}^{T-1}(n_0 + s + 1)^{-1} < 1$, computed directly. □

H The falsifier on 220 further runs

One registered claim is graded failed on exactly the ground that "the generating estimate was three seeds of a two-valued variable." The split above is three seeds of a two-valued variable. Same n, same kind of outcome, opposite treatment, and §M, which is also ours, measured that the seed carries 1.33 runs of information and that its variance share is 2.0% at three seeds against 68.8% at nine. Three seeds is precisely the sample in which the seed effect is invisible, so the fifteen runs above are worth about four independent observations. We therefore froze prereg/E_BIMODAL2_PREREG.md, committed its grader before the data existed, and ran the identical arm on forty-four seeds that had never been used: 220 runs, no grounding, every other setting copied field by field and asserted at load time (Table 6).

Table 6. E-bimodal2, seeds 9–52, 220 runs, against the fifteen at the same cuts. The cuts are the paper's own: 0.306 is the largest escaped run above and 0.512 the smallest captured one, so the middle column is the interval this section calls empty.

	≤ 0.306	in $(0.306, 0.512)$	≥ 0.512	share in the middle
3 seeds (n=15)	8	0	7	0.000
44 seeds (n=220)	39	8	173	**0.036**
uniform baseline for an interval of width 0.206				0.206
P-T1-B1′, frozen before the data				bar ≤ 0.10

P-T1-B1′ and P-T1-B2 are both supported, so claim 1 keeps its headline at n=220. The interval holds 3.6% of runs where a single mode spread over the same range would put 20.6%; the two cluster means are 0.161 and 0.953, a gap of 0.791 against a frozen bar of 0.60; and the middle band's density is 0.15 of the clusters' against a bar of 0.25. P-T1-B5 confirms no single model carries the verdict, the per-model counts inside the interval are $2, 2, 2, 2, 0$.

Clustering over the forty-four seeds gives a standard error of 0.0131 on the middle share against 0.0126 unclustered, a design effect of 1.08; that is far below the 3.75 §M measured, because this indicator is nearly constant within a seed while f_{end} is not.

> Those figures moved when the cuts stopped being typed, and the move is in our favour, so we state it here rather than in a footnote. The grader's cuts were the constants 0.306 and 0.512, described in its own comment as max(escaped) and min(captured) over the fifteen runs of §4.1, and they are those two values rounded down, so the run at 0.30625 that defines the lower edge fell strictly above it and counted as middle. The two rows of Table 6 were then counted against different thresholds: the forty-four-seed row against the constants (37/10/173, share 0.045) and the three-seed row against the definition (8/0/7). On the definition throughout, the forty-four-seed row is 39/8/173 and the share is **0.036**. P-T1-B1′ is supported under both, and the corrected number is the more favourable one, which is the direction that obliges us to print the arithmetic rather than the conclusion.

What did not reproduce is the split. Eight clean against seven captured became thirty-nine against one hundred and seventy-three. The emptiness of the interval is a property of the dynamics and it survived a fifteen-fold sample; the balance between the two modes was a property of three topics. Anything in this paper that reads the near-even split as evidence about how often contamination locks in is unsupported; what the data supports is that at f_0=0.9 these five models lock in about four times in five.

One amendment ran against us and one specification error was ours. P-T1-B1 as originally frozen put the clusters at $f_{\text{end}} \leq 0.10$ and ≥ 0.90; those are not this paper's numbers, and applied to the fifteen runs above they split them 3/6/6 rather than 8/0/7. The criterion tested a claim stronger than any we had made. We found this from the old data before reading a single f_{end} from the new, amended the prereg on the record, and graded the replacement; the original is marked mis-specified, not failed, under the same rule that protects §4.3's apparatus failures. Second, the prereg's variance clause compared this experiment's seed share to §M's 68.8%, but those are different arms, grounded at R=45 there, ungrounded here. The clean same-arm check was already on disk: 2.0% at three seeds, 68.8% at nine, **67**.**6**% at fifty-three. 68.8% was not a small-sample artifact; it is stable on its own arm, and the 57.9% measured here is a second condition in which the seed again dominates the model (6.0%). We record the specification error rather than deleting it.

Two scope questions, both registered and both answered against us, are set out in Appendix H: whether the effect is an artefact of the number of models, and whether it survives the one apparatus choice we could not vary. Neither changes a verdict.

I Claim 2b given an actually out-of-sample test

Arm B is in sample, and calling it out of sample was our own inertia. Arm B estimates γ on seeds 9–52 and predicts seeds 9–52, so its landing is consistent both with $\gamma(\varphi)$ being a sufficient primitive and with a nine-cell curve fitted to 44 seeds reproducing those 44 seeds, which would be worth nothing. Closing this costs only a re-partition of data already on disk: gamma_splithalf.py holds out half the seeds, fits γ on the other half, and compares against the same held-out rates, so the only variable is where γ came from (Table 7). Two partitions are run, contiguous and parity, since reporting one would let the partition do the arguing, and the reading was written before the numbers: within 1.5× of in-sample means γ transports, beyond 3× means arm B was measuring its own fit.

γ transports. Fitting it on unseen topics costs 15–16% of accuracy rather than a factor of anything, and the diagnostic that condemned arm A is gone: arm A's errors were one-sided at signed mean $z = -5.15$ with all five models under-predicted, while out of sample the signed mean z is +0.32 and +0.35. In absolute terms the out-of-sample error is 0.068–0.081 in capture rate against arm A's 0.28. Claim 2b therefore stands on a test that could have killed it. Two things are not flattering. The worst single cell degrades from 1.5σ to 2.7σ and is granite-4.1-8b both times, though its four out-of-sample z values (+2.41, −0.50, −0.84,

Table 7. Claim 2b given an actually out-of-sample test. Each row holds out 22 seeds and predicts them twice, differing only in the fitting set; errors are in seed units of the held-out half, where one unit is 0.045. Not pre-registered and not graded. Arm A is carried down from the out-of-sample capture comparison of the main text (§4.1) for scale, in its own units.

	mean $\lvert\text{sim} - \text{obs}\rvert$			worst $\lvert z\rvert$	
split	in-sample	out-of-sample	ratio	in	out
contiguous	1.3	1.5	**1.15**×	1.44	2.41
parity	1.5	1.8	**1.16**×	1.55	2.71
signed mean z out of sample: +0.32 and +0.35					
arm A, for scale: 12.1 seed units of 44, worst z 8.50, signed mean z −5.15, under-predicts 5/5					

+2.71) swing in sign, so this is variance 22 seeds cannot suppress rather than a missing variable. The whole test is post hoc: it was run because writing this paragraph exposed the gap, which is a worse reason to have found it than a prereg.

The failure has a name, and naming it says what would have prevented it. Arm A is a transport failure in the sense of Pearl & Bareinboim (2011); Bareinboim & Pearl (2014): a quantity estimated in one population is used to predict in another, and licensing that use requires a selection diagram marking the respects in which the two populations differ. Here the two populations are two topic samples, the respect in which they differ is the topic, and no diagram was written, the published curve was carried across the difference silently, on the strength of γ being called a primitive. What arm B shows is that the machinery is right and the transport was not licensed: re-estimating within the target population repairs the prediction entirely. The general form of the lesson is that a constant is not a primitive because it is written as a function; it is a primitive when the function's argument list is complete, and φ was not the whole argument list. A companion paper (Anonymous, 2026b) counts how many such licences a growing programme owes and finds the count quadratic in the number of populations.

The one-sidedness is ours, and it runs opposite to the bias we declared in advance. A γ cell with no observations behind it is filled from the nearest measured cell below it, and baseline_urn.py calls that "the conservative direction for a monotone-increasing γ." On three topics, 17 of 45 cells have fewer than thirty reads behind them, against 3 of 45 on forty-four topics; qwen3-8b's curve had zero observations at j=0..4 and eight at j=5, so most of it was the fill rule rather than a measurement. That rule is conservative in the curve and not in the outcome: this loop is positive feedback, so a γ biased low is a capture rate biased low, and the bias cannot cancel. Worse for us, the simplification we did declare in advance points the other way, the scripted reader asserts only the seeded false value, so it "should if anything capture more easily than the model it imitates." We named one bias and shipped its opposite, and the reason we could not see it is that at three seeds the bar was thirty times wider than the error.

Decided early, but not for the reason we gave. $\text{sign}(f_t - f_0)$ predicts capture in 12 of 15 runs at t=2, and no better at t=30. The natural reading is (1)'s restoring force carrying $1/n$, the decision being taken while n is smallest. Prop. 3 refutes it: the $1/n$ bound permits f to reach either endpoint for the first 49 of 150 steps, so it cannot explain a decision at t=2 (Figure 5). The early decision is measured and not derived, and the mechanism that concentrates it is open.

We attacked it, from logs already on disk, and the attempt failed. The diagnosis frozen in prereg/E_TIMESCALE_PREREG.md was that we had been looking in the wrong state variable: (1) governs f, whose timescale is $O(n)$, but what decides a write is $\varphi_t = j_t/k$, the false share of the k retrieved records, and φ lives on $\{0, \ldots, k\}$ with k=8 constant in t. The

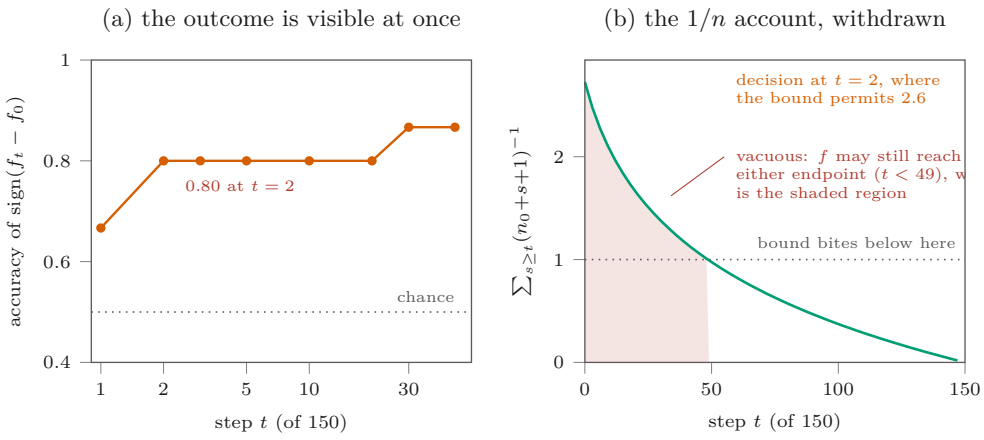

Figure 5. Decided early, and the $1/n$ account of why fails. (a) The sign of the first move predicts the final outcome from step 2 and does not improve thereafter; this is the measurement and it stands. (b) The account that fails. Plotting $1/n$ invites shading "the window in which the outcome can still move". Prop. 3 shows that window is not there: the quantity that actually constrains f is the cumulative bound $\sum_{s \geq t}(n_0+s+1)^{-1}$, plotted here, and it forbids nothing until it falls below 1 at t=49. At the measured decision point t=2 it permits 2.55, more than the whole range of f. The shaded region is where the $1/n$ explanation has no content.

hypothesis was that the outcome is set by which absorbing boundary φ reaches first, in $O(1)$ steps. On 220 runs, zero API cost:

- P-T1-D1 supported: median first passage to $\varphi \in \{0, 1\}$ is 2 steps, and the preregistration predicted this would pass cheaply at f_0=0.9, so it establishes little.
- P-T1-D2 not supported, at 219/219 = 1.000: every run leaves the boundary after touching it. φ crosses; it does not absorb, and the framing is dead rather than dented.
- P-T1-D3 not supported: Spearman$(\tau_{\text{abs}}, T_{\text{commit}}) = +0.108$ against a bar of 0.50 on the 57 runs that ever contradict their own outcome. We found a faster process and did not show it is the mechanism.

The preregistration said in advance what to do if D3 failed, and we do it: §5(ii′) stands word for word, with this recorded as a route tried and not gone through. A mechanism that the data on disk cannot separate from its alternative is not a mechanism we have measured.

What did come out, and it makes the refuted account worse rather than better. P-T1-D4 asked whether the realised step is far below Prop. 3's bound at the steps where the decision is observed. It is: median $|f_{t+1} - f_t| = 0.0055$ against a median bound of 0.0625 over $t \leq 10$, a factor of 11.4. The reason is exact and post hoc (a write moves f by $(1-f)/(n+1)$ if it is false and by $-f/(n+1)$ if it is true, an identity with residual 10^{-16} over 2420 logged steps) so the bound $1/(n+1)$ is attained only at $f \in \{0, 1\}$ and discards a factor of $\min(f, 1-f)$. Prop. 3's bound is therefore not merely vacuous before t=49; it is loose by an order of magnitude exactly where the decision is taken. That is a statement about the argument we withdrew, not about the dynamics, and it supplies no replacement.

J Two scope questions, both registered

Everything above is five open-weight models of 8–27B at a single initial composition $f_0 = 0.9$. Two objections follow, and neither is answerable by argument. Is this a small-model artifact?, to which our own data seemed to say yes: over the five models Spearman(parameters, capture) $= -0.671$ ($p = 0.215$), the 8B mean is 0.841 against 0.705

for 14–27B, and the largest model captures least. Is $f_0 = 0.9$ doing the work?, nine of ten starting records false is an adversarial initial condition, and a store that starts mostly false ending mostly false is not obviously a finding. Both were preregistered (prereg/E_FRONTIER_PREREG.md, prereg/E_F0_PREREG.md), both frozen before their runners existed, and in the first case we wrote our own prediction into the record and it was wrong.

Table 8. E-frontier: the phenomenon at frontier scale, against our own registered prediction that it would vanish. The bimodal2 apparatus field for field (asserted by the grader, not assumed) on seeds 9–28, a prefix of the 44 the open models used, so runs pair seed for seed. The prereg states, before the run, that we expected pooled capture below 0.5; the reasoning was the scale trend above. P-T1-F1's bar was 0.50 and no answers were lost on either model.

model	runs	calls	lost (%)	capture	
anthropic/claude-sonnet-4.5	20	3000	0.00	**1.000**	captured on every seed
openai/gpt-4o	20	3000	0.00	0.700	
pooled	40	6000	0.00	**0.850**	P-T1-F1 bar 0.50: supported

Two further frozen criteria were graded on this arm and the paper had not printed either. P-T1-F3 asks whether bimodality survives at frontier scale, separately from the rate: the share of runs landing in the interval §4.1 calls empty must be ≤ 0.10. It comes out $3/40 = 0.075$, supported, and the split is informative, 0/20 for claude-sonnet-4.5 at mean $f_{\text{end}} = 0.923$ and 3/20 for gpt-4o at 0.671. Two peaks survive whatever the rate does, so f_0 and model class select which peak is occupied rather than whether peaks exist, and that is the structural half of the claim rather than its direction.

P-T1-F4 is the harder one and it splits. Re-measuring γ on each frontier model's own runs and simulating asks whether the primitive is sufficient for a model family it was never fitted on, a stronger extrapolation than §4.1's cross-topic split-half. The bar was 2.0 seed units, one seed unit being 0.050 at twenty seeds. claude-sonnet-4.5 is exact (1.000 observed, 1.000 simulated, 0.0 seed units); gpt-4o is outside it, 0.700 observed against 0.901 simulated, 4.0 seed units, and the sign is over-prediction where §4.1's arm A under-predicted every model (Table 8). The two failures of γ we have measured run in opposite directions, which is what a sufficient primitive with an under-powered estimator looks like and is not what a missing variable looks like, but we have two models and cannot separate those readings here.

The phenomenon survives, and the model that survives it worst is the newest one. claude-sonnet-4.5 is captured on 20 of 20 seeds, which is the same rate as qwen3-8b at 8B and higher than every 14–27B model. The -0.671 rank correlation was therefore not a scale law; it was five points with three of them at one size. We report this as the correction of a reading of ours rather than as a prediction confirmed, because our registered prediction was the opposite: we expected capture below 0.5 and wrote down why. Parameter counts for both frontier models are undisclosed, so they are excluded from any correlation rather than replaced with a proxy, substituting price or context length would report a different known quantity as the measurement.

What forty runs at two models buys, computed rather than asserted. This is the paper's most-quoted sentence and its thinnest arm, so the accounting belongs beside it. The resampling unit is still the seed, so the relevant factor is $1 + (m-1)\rho$ with m=2, not the 3.75 measured at m=5: on this arm's own twenty shared seeds $\rho = 0.254$ for f_{end}, giving a design effect of 1.25 and 31.9 effective observations of 40. For the headline the arithmetic is simpler still, because claude-sonnet-4.5 captured on every seed: 20/20 puts a one-sided 95% lower bound of 0.861 on its capture probability, against a bar registered at 0.50. The sentence survives its own power accounting. Three weaknesses are not fixed by that. Parameter counts for both models are undisclosed, so neither can enter a correlation. Served models at temperature 0 are not reproducible (Table 13's two realisations are the only place we can price that), so a replication may differ by more than seed noise. P-T1-F4 also splits, on two models, which is the smallest sample on which a split can be reported at all. Owed and registered: four or more frontier models at 44 seeds, and one end-to-end replication inside a deployed memory framework reporting only whether the interval is empty. Neither

is run, and the second is the single measurement that would most change what this paper is evidence about.

Table 9. E-f0: the empty interval is empty at every initial composition, and $f_0 = 0.9$ is not the worst one. gemma-3-27b at the bimodal2 apparatus, seeds 9–28, one model rather than five because §M's own finding is that this apparatus should buy seeds and not models. "Void" is the interval $(0.306, 0.512)$ that §4.1 calls empty. P-T1-F0-1's bar was a share ≤ 0.20, frozen before the runner existed. The whole tabular is generated, and it reads the band from the same artifact Table 6 does, so the two cannot disagree about where the empty interval is, which they did while both were typed: this table's f_0=0.9 row said 3/44 in the void, the count against the rounded cuts, where the cuts' own definition gives 2/44.

f_0	n	capture	mean	median	in void	
0.1	20	0.500	0.459	0.441	**0/20**	and the mean lands inside the void
0.3	20	0.700	0.676	0.956	0/20	
0.5	20	**0.750**	0.723	0.969	0/20	highest capture of the four
0.9	44	0.682	0.724	0.994	2/44	the arm §H reports

Three readings, and the middle one answers the objection outright.

(i) The void is empty at every f_0 we bought, 0/20 three times over against a frozen bar of 0.20. Two peaks are therefore a property of the dynamics; f_0 selects which peak a run lands on, not whether peaks exist. This is the structural half of claim 1 and it does not depend on any capture rate.

(ii) Capture is not monotone in f_0, and $f_0 = 0.9$ is not the adversarial extreme. It reads 0.500/0.700/0.750/0.682: the highest capture in this sweep is at $f_0 = 0.5$, where the store starts balanced. Hence "you started from ninety percent falsehood" does not describe what drives the outcome, a balanced store is worse than a mostly-false one, which (1) anticipates, since a store already at the append-only ceiling has little room left to move while a balanced one sits where the per-step variance is largest. We did not freeze a criterion on the shape and do not add one now.

(iii) At $f_0 = 0.1$ the pooled mean falls inside a strictly empty interval. The mean is 0.459 and no run lies in $(0.306, 0.512)$; the runs split 10–10 between escape and capture. §2's objection to summarising this distribution by its mean was measured on the $f_0 = 0.9$ arm at 8.2% of runs described (P-T1-M1), and $f_0 = 0.1$ is the sharper case: there the summary statistic denotes a region containing no observation at all. That is the strongest form of the claim and we did not preregister it, so it is reported and not graded.

A loader bug found in E-f0's first output, and what it did and did not affect. run_sched_gate.py writes every run twice, one shard, then again into a merged file, so the grader's first glob counted each run twice and reported 40 runs where 20 seeds were bought. Shares were unaffected, because duplicating a set preserves its proportions, so P-T1-F0-1's verdict is the same either way. What was wrong was the n, and a wrong n misstates precision, which is worse than a wrong share because a reader cannot see it. The loader now deduplicates on the full identifying configuration.

K The sweep past the peak

The sweep past the peak graded untested. Three further budgets, $R \in \{75, 120, 148\}$ of 150 steps, all else identical, five models and three seeds each. The design discriminates rather than confirms: (3) puts the spread down 22% by R=120, while §4.1's early-decision result keeps back-loading harmless until its first grounding write enters the $t \leq 2$ window, at step $T-R$. The accounts agree at R=148 and disagree at R=120, and the criteria were frozen on that disagreement (prereg/E_TRANSIENT_PREREG.md).

P-T1-T1 and P-T1-T2 are untested for apparatus failure. At R=120 and R=148 all three arms give capture 0.00, so by the floor clause registered with E-budget those cells carry no

Table 10. The sweep past the peak. Capture is the frozen observable and it dies: a grounding write is a truth record, so the ceiling itself carries the budget, $f_{\max}(R) = (n_{\text{false}} + Tw)/(n_0 + Tw + R) = 159/(160+R)$, and by R=148 a run must reach 97% of its own ceiling to be counted as captured. The last column divides that ceiling out, per run: a step whose answer did not parse writes nothing and is not logged (123 of 168,660 across the whole programme), so each run's ceiling uses its own write count rather than the nominal T. Using the nominal value instead shifts s_{norm} by at most 0.0012. s is back minus front.

R	R/T	back's first write at step	$f_{\max}(R)$	capture (frozen) front / back	s	$\overline{f_{\text{end}}}/f_{\max}$ (post hoc) front / back	s_{norm}
5	0.03	145	0.964	0.20 / 0.40	0.20	0.275 / 0.468	0.193
15	0.10	135	0.909	0.20 / 0.40	0.20	0.275 / 0.489	0.214
45	0.30	105	0.776	0.07 / 0.47	0.40	0.211 / 0.507	0.295
75	0.50	75	0.677	0.07 / 0.33	0.27	0.190 / 0.396	0.205
120	0.80	30	0.568	0.00 / 0.00	no info	0.213 / 0.331	0.117
148	0.99	2	0.516	0.00 / 0.00	no info	0.216 / 0.219	0.003

information. Prop. 1 says why they are empty rather than unlucky: at R=148 the append-only ceiling has fallen to $f_{\max} = 0.516$, only 0.016 above the capture threshold, so a run must land in a window one sixtieth as wide as the state space to be scored captured. Had those cells been used, P-T1-T1 would have read $s(148)$=0.00 < $s(75)-0.15$ and been marked supported on the strength of no run anywhere reaching the threshold. The observable ran out, not the mechanism, and the defect is ours: $f_{\max}(R)$ follows from (2) and was derivable before the sweep was launched.

Read on a scale-free observable the transient appears, and the discriminator still fails to discriminate. Dividing out $f_{\max}(R)$ gives the last column of Table 10: a rise to 0.295 at R/T=0.30, then 0.205, 0.117, and 0.003 at R=148, where the three arms sit within 0.009 of one another. Changing the observable after the frozen one floors is how a failed prediction is normally rescued, so we grade nothing on it, and the frozen thresholds applied to it land inside the undecided band, -0.088 against ≤ -0.10 for dilution and ≥ -0.05 for early decision. Neither account wins, and by §M neither available interval is usable. The peak's location is claimed in neither direction anywhere in this paper.

The scale-free retest, on six fresh seeds, splits. Two scale-free criteria were frozen and six unused seeds bought (prereg/E_TRANSIENT_PREREG.md amendment 1). P-T1-S1 uses f_{end} over the run's own ceiling; P-T1-S2 uses capture at $\frac{1}{2} f_{\max}(R)$, and exists so that the conclusion cannot depend on which scale-free observable we happened to pick. Both required spread(120) $\leq$ spread(45) $- 0.15$.

Table 11. E-transient2, seeds 3–8 only; seeds 0–2 generated the hypothesis and are excluded. 30 runs per arm. Standard errors are clustered over the six seeds, these arms share their seed grid by construction, so the contrast is within-seed and clustering reduces the interval rather than widening it (§M).

observable	spread(45)	spread(120)	difference	bar	verdict
$f_{\text{end}}/f_{\max}(R)$	0.153	0.093	-0.060	≤ -0.15	failed
capture at $\frac{1}{2} f_{\max}(R)$	0.167	0.000	-0.167	≤ -0.15	supported

P-T1-S4 is a split, and the prereg requires reporting it rather than choosing. The collapse's existence depends on which of two equally defensible readings is used. Two disclosures make it worse: the supported side clears its bar by 0.017, half of one seed unit, and the R=120 cell carrying that support has its arms at 0.700/0.733/0.700, a span of exactly one seed unit, which by the standard registered for the capture-rate observable is no information at all (Table 11). We do not use this to overturn the verdict, since amendment 1 carried no floor clause and adding one after the data is moving the goalposts.

We then bought the resolution we had declared unbuyable, and it did not resolve. P-T1-S3 stated in advance that separating the peak needs "about 220 runs per cell, about 44 seeds, and we do not buy it." That records a budget rather than a limit of nature. We bought it: $R \in \{45, 75, 120\}$, three schedules, five models, seeds 9–52, 1980 runs at 150 steps, criteria frozen in prereg/E_TRANSIENT3_PREREG.md with the grader committed before the data existed. Seeds 0–8 generated the hypothesis and are excluded, dropping 315 runs already on disk. Every cell is complete and balanced.

Table 12. E-transient3, seeds 9–52, 220 runs per arm. s_{norm} is $\overline{f_{\text{end}}}$(back) − $\overline{f_{\text{end}}}$(front) with each run divided by its own ceiling. Standard errors are clustered over the 44 seeds. The ordering front < uniform < back holds at all three budgets; it is the decline that was under test. The whole tabular is generated from results/TRANSIENT3.json, an artifact that did not exist until this table did, because the grader printed these means on every run and persisted none of them.

R	front	uniform	back	s_{norm}
45	0.587	0.666	0.773	0.186
75	0.578	0.631	0.743	0.165
120	0.564	0.591	0.677	0.113

At 44 seeds the discriminator is undecided rather than resolved. The contrast $s_{\text{norm}}(120) - s_{\text{norm}}(75)$ is -0.052 at 1.14σ, against -0.088 at three seeds, which was outside the band; the shape criterion reads -0.074 against a bar of -0.15 at 1.62σ and fails on magnitude and on significance alike. What survives is the ordering, front < uniform < back at every budget, which is what the schedule axis rests on and never needed the transient to assert (Table 12). The bill went up rather than down when we paid the first instalment: the effect fell from 0.115 at three seeds to 0.021 at forty-four while the standard error shrank 21% more slowly than $1/\sqrt{n}$, so 3σ now needs about 1895 seeds, or 9475 runs per cell. Claim (iii)'s statement about the peak's location is marked not resolvable at ≤1980 runs and left there.

The boundary this draws. In generational retraining, real data is a fixed corpus present in every round and re-usable without limit, so "when to ground" is not a question that arises; the established results there concern accumulation and proportion (Gerstgrasser et al., 2024; Dey & Donoho, 2024). Their setting is the free-regrounding limit of ours, $R \to T$, where (3) makes the three schedules identical and only the fraction is left to vary. By item (iv) that limit is a single point rather than a neighbourhood: the schedule spread is zero at α=1 and diverges like $\ln T$ at every $\alpha < 1$, so no amount of increasing their re-use budget turns their setting into ours and no amount of decreasing ours turns it into theirs. Ours is the budgeted case, where the same fraction delivered at a different time is worth an order of magnitude less. The containment runs along the budget, not along the store size, accumulation itself is theirs (Gerstgrasser et al., 2024).

L An accidental replicate

This sweep was run twice. The second run was launched by a command intended only to count configurations, and it overwrote the first run's raw data; we record the process failure rather than omit it. What it bought is a measurement this design otherwise lacks, since every other figure here prices seed variation alone.

The point estimates move by one run in fifteen on two of three arms, while the ordering, its 5/5 model-level agreement and all three frozen verdicts are identical. The schedule result is robust in the form it is claimed, an ordering, and the pooled rates should be read as ± one run rather than as two-decimal quantities. Every other number here carries the same unpriced replication variance, and this table is the only place we can bound it.

Table 13. The same protocol, same seeds, same budget, run twice. Only the second run's raw data survives; the first survives as its logged pooled figures. Note that the agent runs at temperature 0, so the two realisations differ only because the served model is not reproducible at that setting.

pooled capture rate	front	uniform	back	frozen criteria
first realisation	0.20	0.27	0.33	P-S1 5/5, P-S2 ✓, P-S3 ×
second realisation (on disk)	0.20	0.33	0.40	P-S1 5/5, P-S2 ✓, P-S3 ×
movement	0	+1/15	+1/15	identical

A write-side gate erases the differences between models. Admitting the agent's answer only when it agrees with the majority of retrieved records is the most conservative deterministic form of consistency-gated admission, and weaker than any deployed one: Zhang et al. (2026) admit on five factors of which agreement is one, and Ha et al. (2026) type memories by role rather than filtering on agreement at all. A harmful regime in the weakest gate is not "conservative evidence about stronger versions": the inference needs harm to be monotone in gate strength, we have not tested a second gate, and Ha et al. (2026)'s mechanism is a reason to expect non-monotonicity, since typing the store changes the reference class our own diagnosis blames. What the arm supports is a statement about this gate. The registered prediction was a sign flip, helpful at low contamination and harmful at high, and it holds in five of five models at each end. The numbers show more than that. On seven fresh seeds the ungated baselines span 0.729, 0.745 and 0.932; gated, all three sit at 0.9936, 0.9937 and 0.9935, separating only in the fourth decimal.

> The gate's reference class is the store the agent itself wrote. It therefore fixes the store's composition upstream of the reading behaviour, and a model whose cleaning lived in that behaviour has nothing left to clean with. The observable consequence is not that the gate is harmful on average but that it removes the differences between models: a 0.2 spread in unaided outcome becomes a 0.001 spread.

That majority aggregation can lock in error is established for one-shot aggregation over samples, where the mechanism is error correlation and a better model is therefore safer (Liu et al., 2026; He et al., 2026). Two things differ here: the filter sits on the write side rather than the read side, and it acts over a tenure rather than at one decision, which is what decouples composition from reading. A third proposed difference, that the model-quality dependence reverses sign between the two timescales, was killed; §1 reports its death.

A cost-matched state slot: a tie, reported as one. The parent framework's central object is a pair of channels, one compressive and composable, and it flags the systems-level version of its claim as untested (Anonymous, 2026a). The smallest such channel this apparatus admits is one line carrying the agent's own previous answer, with one retrieved record removed to pay for it, eight lines per decision in both arms, so the comparison is cost-matched at the line level. We registered two competing readings: the slot supplies invalidation throughput and helps at both contamination levels, or the slot is a self-reinforcing write channel and helps only at low contamination. They agree at f_0=0.1 and disagree at f_0=0.9, which is the discriminating cell.

At f_0=0.1 the slot helps or is neutral in four of five models (P-A1). At f_0=0.9 it gives no verdict: zero models better, one worse, four tied (P-A2). The per-model spread is what the tie hides, one model moves +0.84 in capture and another −0.30, so this is a genuine failure to separate rather than a small effect. What follows is a boundary rather than a result: a self-fed slot is not the state channel of the parent framework, whose channel is fed by the stream rather than by the system's own output, and testing that distinction needs a stream-fed slot this apparatus does not yet have.

A source marker is not a directional lever. Our registered prediction, that marking each record's origin lowers contamination at high f_0, fails at 2/5 (P-P2); its mechanism claim, that the marker lowers γ, fails at 3/5 (P-P3). It helps sharply in some models (γ from 0.799

to 0.220) and hurts in others (0.309 to 0.580). What does replicate is a null we did not predict: with the label scrambled, identical tokens, identical wording, no information, the outcome sits at 0.856 against a two-end midpoint of 0.664, nearer the lying end than the middle, in five of five models.

An uninformative provenance marker is not a neutral cost. It occupies the same budget as content and demands a per-record judgement it cannot support, which is the behavioural reading of the capacity coupling: source bits and content bits are drawn from one budget.

M Unresolvability from the boundary structure

The three deaths below share one cause, and §4.2 is what makes it calculable rather than a lesson learned. If the outcome is a choice between two boundaries, escape to $f{=}0$ or saturation at $(n{-}1)/n$ (then f_{end} is nearly a two-valued random variable, and a per-model mean over s seeds is an estimate of the choice probability p) which is Prop. 4 (Table 14).

Both halves matter and the second is the one that caught us. (4) is a per-model statement; the design-effect factor $1+(m{-}1)\rho$ applies to everything pooled, which is most of this paper. The decomposition below measures ρ and gets a design effect of 3.75 at nine seeds, so 45 runs carry the information of 12; measured again at 44 seeds on two further arms it is 3.08 and 3.56–3.66, and pair by pair it runs 1.73–4.95 (§4.6).

The resampling unit, and why every error bar in this paper was the wrong size. (4) is a per-model statement and it survives: with one run per seed the marginal variance of a Bernoulli outcome is $\bar{p}(1{-}\bar{p})$ whatever the topic-to-topic spread in p. What does not survive is any quantity pooled across models, and that is most of them. The seed is not a noise knob (it seeds the world and the corruption draw, so it selects the topic and the false value) and on the $R{=}45$ front arm the variance decomposes as

	between-seed	between-model	residual
at 3 seeds	2.0%	94.4%	3.6%
at 9 seeds	68.8%	14.2%	17.0%

The dominant variance component is invisible in the sample that was used to decide it was small. At three seeds the topic contributes 2% and the model looks like the whole story; six more seeds make it 69%. Four of those nine seeds are topics on which all five models end contaminated (front-arm f_{end} 0.64–0.71) and five are topics on which all five end clean (0.14–0.20), the two groups separated by 5.0σ.

Two consequences, and they point in opposite directions, so we state them per quantity rather than issuing one correction. With $m{=}5$ models per seed and an intra-seed correlation of 0.69, the design effect is $1+(m{-}1)\rho = 3.75$: forty-five runs carry the information of twelve, so each seed is worth 1.33 runs no matter how many models we run on it. For a pooled level the seed effect adds, and treating runs as independent understates the standard error by 1.86× (0.056 against 0.105). For a within-seed contrast, which is what every schedule result is, since the arms share the grid by construction, the seed effect cancels, and treating runs as independent overstates it by 1.11× (0.037 against 0.034). The same mistake made our levels look too precise and our contrasts look too noisy.

No verdict in this paper changes. Every frozen criterion was set on a point estimate against a bar, not on a σ: the schedule ordering is 5/5 models monotone per seed, the budget falsifier failed on 0.20 against 0.40, the variance bound failed on 1.07× against a required 2.83×. What changes is the design advice, and it is the opposite of what we did: this apparatus should have bought seeds and not models. The five-model grid was chosen to price cross-model generality and it does that, but it contributes almost nothing to the precision of anything pooled. All bootstrapped intervals below use a cluster bootstrap over seeds and report the cluster count, because with three clusters a clustered interval is itself

untrustworthy, measured here, clustered and unclustered standard errors differ by up to 2.6× at three seeds and agree to within 2% at nine.

This also settles the failure recorded in §4.1. A diffusion would accumulate variance as $\sum 1/n^2 = O(1/n_0)$ and its spread would fall as $1/\sqrt{n_0}$; a boundary choice has variance $p(1-p)$ times the ceiling squared, in which n_0 does not appear at all. Measured, the within-model spread is flat in n_0 to 1.07× over an eightfold range. The variance bound failed and the ceiling result holds for the same reason, and the two findings are one: this process does not diffuse to a stopping point, it picks a side.

Table 14. What it takes to resolve a cross-model difference in the choice probability, at 80% power and 95% confidence, against what we ran.

difference in p to be resolved	0.4	0.3	0.2
seeds required per model	24	44	98
seeds we ran	3, then 7 in the registered re-test		

The axis that died was unresolvable by construction, and (4) says so without any data. A cross-model correlation whose independent variable is a 3-seed estimate of a boundary-choice probability is a correlation against an estimate with a standard error of 0.29 on a $[0, 1]$ scale. That is why the rank correlation of +1.00 evaporated: the model we called "cleanest" had $\hat{p}$ from three seeds, and seven seeds moved its mean from 0.150 to 0.729. Nothing was measured; the ordering was a draw.

The transferable form of this is a design rule for anyone reading a per-model number off a self-writing loop: first ask whether the outcome is interior or a boundary choice. If it is a boundary choice, per-model means need tens of seeds before they can carry a comparison, and no amount of care in the downstream arithmetic substitutes, which is the same lesson this project learned once before, when fixed points computed exactly from 12-sample cells moved to zero at 188 samples. The failure mode is not sloppy arithmetic on noisy inputs; it is exact arithmetic on noisy inputs, which is harder to see.

N Every criterion this paper grades, with its freeze date

Each row is a preregistration file: the criteria it freezes, the commit that introduced it, and that commit's UTC timestamp. The dates are read from git log --reverse at build time rather than transcribed. Amendments are counted separately and never overwrite the original, so a criterion that was mis-specified stays in the record alongside its replacement; §H discusses the one case where that mattered.

O The invariants, checked on every run in the repository

verify_invariants.py asserts each of the following over every run on disk, and refuses the build if one fails. They are not statistics; they are properties the apparatus must have if it implemented the protocol at all, and several of them exist because one build did not. The count in the right-hand column is the number of individual assertions evaluated, 7,929,204 in total, all of which hold.

P E-bimodal2 per model

P-T1-B5 puts every verdict on the five models pooled and forbids replacing a pooled verdict with a per-model one. The per-model figures are printed here because forbidding their use as evidence is not a reason to withhold them: a reader who suspects one model carries the result should be able to check. All 220 runs, seeds 9–52.

Table 15. The preregistration record, generated from git by make_appendix.py.

experiment	frozen at	criteria	amend.	words
E_A	a46631d9, 2026-08-06 07:15	P-A1, P-A2, P-A3	0	161
E_BIMODAL2	2072fee0, 2026-08-07 01:51	P-T1-B1, P-T1-B2, P-T1-B3, P-T1-B4, P-T1-B5, P-T1-B1', P-T1-B1"	3	513
E_BUDGET	a46631d9, 2026-08-06 07:15	P-B1, P-B2, P-B3	0	314
E_F0	45d6f80b, 2026-08-07 16:28	P-T1-F0	1	330
E_FRONTIER	a4095483, 2026-08-07 16:19	P-T1-F1, P-T1-F2, P-T1-F3, P-T1-F4, P-T1-F5	0	315
E_GATE	a46631d9, 2026-08-06 07:15	P-G1, P-G2	0	79
E_LABEL	a46631d9, 2026-08-06 07:15	P-L1, P-L2, P-L3	0	117
E_MEAN	5ddb9f4e, 2026-08-07 06:35	P-T1-M1, P-T1-M2, P-T1-M3, P-T1-M4	0	236
E_MULTIVALUE	60dddef7, 2026-08-18 07:42	P-T1-V1, P-T1-V2, P-T1-V3, P-T1-V4, P-T1-V5	2	812
E_PROV	a46631d9, 2026-08-06 07:15	P-P1, P-P2, P-P3	0	224
E_RATCHET	a46631d9, 2026-08-06 07:15	P-R1, P-R2, P-R3	0	140
E_REAL	0ad31a69, 2026-08-09 03:01	P-T1-N0, P-T1-N1, P-T1-N2, P-T1-N3, P-T1-N4, P-T1-N5, P-T1-N6	1	3140
E_RETRIEVER	87f0c5d7, 2026-08-08 14:28	P-T1-R0, P-T1-R1, P-T1-R2, P-T1-R3, P-T1-R4, P-T1-R5	1	2771
E_SCHED	a46631d9, 2026-08-06 07:15	P-S1, P-S2, P-S3	0	116
E_TEMP	d76ccf6a, 2026-08-09 01:35	P-T1-X0, P-T1-X1, P-T1-X2, P-T1-X3, P-T1-X4	0	1415
E_TIMESCALE	b8c38446, 2026-08-08 11:08	P-T1-D1, P-T1-D2, P-T1-D3, P-T1-D4	0	322
E_TRANSIENT3	fd89cc33, 2026-08-06 07:56	P-T1-U1, P-T1-U2, P-T1-U3, P-T1-U4, P-T1-U5	0	312
E_TRANSIENT	cbf4a12c, 2026-08-06 05:19	P-T1-T1, P-T1-T2, P-T1-T3, P-T1-T4, P-T1-S1, P-T1-S2, P-T1-S3, P-T1-S4	1	550
E_VERIFY	b5aab439, 2026-08-08 11:20	—	0	283

Q The two γ curves, with the denominators behind each cell

§4.1 turns on the difference between γ measured on three topics and γ measured on forty-four, and the mechanism is visible only with the support printed, the number of reads behind each cell. A cell with no observations is filled from the nearest measured cell below it, which baseline_urn.py calls the conservative direction for a monotone-increasing γ; it is conservative in the curve and not in the outcome, because the loop is positive feedback. Rows marked A are the published curve (seeds 0–2); rows marked B re-measure on seeds 9–52. The bold cells are those with fewer than thirty reads behind them.

Cells with fewer than thirty reads: 17 of 45 on three topics, 3 of 45 on forty-four. That ratio is the whole of §4.1's mechanism.

R The resampling unit on the remaining rows

§M establishes that the seed is the resampling unit and that a three-seed sample cannot see it. That finding killed one axis and downgraded one row, and was not carried to the four intervention arms of §4.3 or to the two per-model orderings that carry claims 2 and 4. The second omission is the worse of the two. Everything below is post hoc by construction, since it re-reads frozen criteria on samples that did not exist when they were frozen; no failed row becomes measured here.

The design effect of 3.75 is a property of levels, not of contrasts. The four intervention arms compare a treated run against a control run on the same seed, so their criteria are paired contrasts, and a contrast cancels whatever the seed does to both arms. That defence cannot be checked at three seeds – the content of §M is precisely that three seeds read the seed variance as near zero when it is not – so it is checked where a 44-seed sample exists, on the schedule axis, and only then transported.

Table 16. Machine-checked invariants, run at build time.

	invariant	assertions
✓	answer labels partition	1,114,218
✓	api_calls >= logged steps	7,544
✓	bimodal2 config == converge's n_false0=9 arm	220
✓	bimodal2 uses only fresh seeds (>=9)	220
✓	budgets matched within an R	6
✓	converge n_false0=9 arm is internally uniform	15
✓	converge's arm and bimodal2 share no seed	1
✓	dropped steps are absent from the log, not zero-filled	7,544
✓	f0 == n_false0/n0	7,544
✓	f_any*n is an integer	1,114,218
✓	f_end == last logged f_any	7,544
✓	f_seed <= f_any	1,114,218
✓	false count <= planted + own writes (eq:ceiling)	1,114,218
✓	gamma denominators sum to logged steps	7,544
✓	gamma numerator <= denominator	46,949
✓	gamma_crit == 1/k + r/w	7,544
✓	gamma_hat == the j=1 cell	4,790
✓	gamma_n == the j=1 denominator	4,790
✓	ground_steps distinct and in range	7,544
✓	logged steps <= configured steps	7,544
✓	n_false_retrieved <= k	1,114,218
✓	observed_growth == (f_end > f0)	7,544
✓	one configuration, one run within a collection	7,544
✓	predicted_growth == (gamma_hat > gamma_crit)	4,790
✓	shard id uniquely determines cfg	1
✓	store never shrinks	1,106,674
✓	store size accountable per step	1,114,218
	total	7,929,204

Table 17. E-bimodal2 by model. Middle counts runs in the interval $(0.306, 0.512)$ that §4.1 calls empty; the pooled bar frozen in the prereg is 10% against a uniform baseline of 20.6%.

model	n	≤ 0.306	middle	≥ 0.512	middle share	mean f_{end}
qwen3-8b	44	0	0	44	0.000	0.939
ministral-8b	44	6	2	36	0.045	0.808
phi-4	44	10	2	32	0.045	0.757
granite-4.1-8b	44	10	3	31	0.068	0.735
gemma-3-27b	44	11	3	30	0.068	0.724
pooled	220	37	10	173	0.045	0.792

The correction the intervention criteria need is a factor of 1.14–1.47 and not of 3.56–3.66, and for a contrast the naive standard error is if anything conservative (clustered/naive 0.93–1.15). What three seeds do cost is stated next, and it is a different thing.

Each frozen criterion, re-evaluated one seed at a time. Every criterion in these arms is a count over five models of a comparison between two three-seed means. Evaluated one seed at a time it becomes three counts, and a criterion that clears its bar on one seed and misses on another was decided by the seed rather than by the intervention.

One criterion of eight is seed-fragile: E-A low (P-A1). It is graded underpowered in the scorecard rather than measured. Two entries carry a different defect: E-gate's treated arm has a cross-model spread of 0.0003 against a control spread of 0.3000, so its 5/5 is one collapse observed five times and not five independent confirmations – which is §4.3's own finding about that gate, stated as a limit on its own evidence. No verdict changes when the control is replaced, though the control moves a long way:

Table 18. $\gamma(j)$ per model and the support of each cell, j false records retrieved of k=8. Generated by make_appendix.py from the same functions gamma_outsample.py uses, so the two cannot disagree.

model	arm	j=0	j=1	j=2	j=3	j=4	j=5	j=6	j=7	j=8
gemma-3-27b-it	A	0.00	0.00	0.03	0.02	0.03	0.00	0.00	0.76	1.00
	n	92	102	76	44	32	31	25	29	19
	B	0.00	0.03	0.05	0.03	0.03	0.04	0.10	0.93	1.00
	n	351	469	392	244	183	142	126	827	3866
granite-4.1-8b	A	0.00	0.00	0.11	0.14	0.18	0.30	0.47	0.84	1.00
	n	42	56	65	58	33	20	17	49	110
	B	0.00	0.01	0.13	0.23	0.27	0.40	0.59	0.86	1.00
	n	109	270	411	422	322	363	390	1099	3214
ministral-8b-2512	A	0.00	0.00	0.00	0.10	0.03	0.35	0.57	0.82	1.00
	n	23	43	37	31	32	49	30	68	137
	B	0.00	0.03	0.04	0.10	0.21	0.33	0.71	0.86	1.00
	n	138	196	178	184	189	243	458	1427	3587
phi-4	A	0.00	0.00	0.01	0.11	0.24	0.22	0.33	0.84	1.00
	n	52	75	72	47	25	18	6	25	130
	B	0.00	0.05	0.06	0.11	0.21	0.31	0.52	0.91	1.00
	n	224	371	376	276	221	203	202	983	3744
qwen3-8b	A	0.00	0.00	0.00	0.00	0.00	1.00	0.84	0.94	1.00
	n	0	0	0	0	0	8	25	122	295
	B	0.00	0.00	0.31	0.42	0.54	0.68	0.88	0.90	1.00
	n	0	2	16	38	84	199	590	1758	3913

Table 19. The same runs, the same 44 seeds, two quantities. Levels reproduce §M's design effect of 3.75 measured on nine seeds; paired contrasts do not carry it. The right-hand column is what a three-seed subsample of the identical data reports, and for levels it is wrong by a factor of four – which is why the contrast column may not be read off three seeds either, and is read off forty-four.

quantity, at 44 seeds	ρ	design effect	clustered/naive se	ρ at 3 seeds
level R=45 front	0.665	3.66	1.85	0.084
level R=75 front	0.662	3.65	1.83	0.141
level R=120 front	0.639	3.56	1.80	0.170
contrast R=45 front-back	0.118	1.47	1.15	0.027
contrast R=75 front-back	0.066	1.26	0.99	0.081
contrast R=120 front-back	0.035	1.14	0.93	0.284

The largest move is 0.682 in capture rate. Note also what no amount of care in the analysis can buy back: with three seeds as the unit, the smallest two-sided p a sign test can attain is 0.25, whatever the data say. At seven seeds it is 0.0156.

One ρ for five models is an assumption. Prop. 4 writes the design effect as $1 + (m-1)\rho$ with a single ρ. Measured pair by pair at 44 seeds, ρ is not one number:

The design effect implied by the extreme pairs spans 3.22, which is the size of the effect itself. Every correction's direction is unchanged and the mean ρ is +0.53, so 3.75 survives as a pooled figure; it does not survive as a per-pair one, and §M now quotes the range.

The two per-model orderings, at forty-four seeds. Claims 2 and 4 are rank correlations over five models whose dependent variable is a capture rate. At three seeds that rate has a standard error of 0.29 on a $[0, 1]$ scale by (4), which is the construction §M's claimbox blames for a spurious +1.00 elsewhere in this paper. Both are recomputed here on the 44-seed capture rates. They move in opposite directions.

Claim 4 does not survive: at 44 seeds the correlation runs +0.31 to +0.80 with no window reaching $p < 0.13$, against +0.92 to +0.98 at $p \leq 0.067$ as published. The reason is visible in the last column rather than in the correlation: the predictor's cross-model spread falls from 0.333 to 0.061, so there is nothing left to order with. Claim 2 moves the other way. Its published statistic is $\rho = +0.821$ with one adjacent inversion, on a simulation from three-topic γ against three observed seeds ($p = 0.133$); against the 44-seed capture rates the same construction gives $\rho = +0.900$ from the published curve and $\rho = +1.000$ ($p = 0.017$)

Table 20. The eight frozen criteria of the four 3-seed arms. Column three is the published value and is reproduced exactly by this script, which is the check that the re-evaluation is of the same criterion. Column four evaluates the identical rule on each seed alone. Column five replaces the three-seed control by bimodal2's 44-seed control, available for the high-f_0 cells at no cost.

arm	f_0	criterion	published	per seed	vs 44-seed control	reading
E-gate	low	P-G1 low	5/5	5/5/5	—	not a count of five
E-gate	high	P-G1 high	5/5	5/5/5	5/5	not a count of five
E-A	low	P-A1	4/5	5/3/5	—	seed-fragile
E-A	high	P-A2 worse	1/5	1/1/1	0/5	stable
E-prov	low	P-P1	4/5	5/5/4	—	stable
E-prov	high	P-P2	2/5	1/0/1	3/5	stable
E-label	high	P-L2 (row 11)	5/5	5/5/5	—	stable
E-label	high	P-L1 (row 13)	0/5	0/0/0	—	stable

Table 21. The ungated, unmarked, no-slot control at f_0=0.9, as the four intervention arms measured it and as bimodal2 measures the identical configuration. Nothing was re-run; the second column is 44 seeds that were already on disk.

model	capture, 3 seeds	capture, 44 seeds	$\bar{f}$, 3	$\bar{f}$, 44
gemma-3-27b-it	0.00	0.682	0.150	0.724
granite-4.1-8b	0.33	0.705	0.444	0.735
phi-4	0.33	0.727	0.408	0.757
ministral-8b-2512	0.67	0.818	0.537	0.808
qwen3-8b	1.00	1.000	0.971	0.939

from the curve measured on the topics it predicts. The inversion was an artefact of three observed seeds, and claim 2's ordering is exact – which no reader could have known from the number we published, because we published the weaker one.

S The answer space and the empty interval

The apparatus plants one false value and corrupts one digit, so the standing objection is that the outcome variable is effectively two-valued and that a two-valued outcome is bimodal for reasons that need no language model. The objection has an arithmetic half and a mechanism half, and they have different answers. Everything here is scripted or read off runs already on disk; no API calls, nothing graded.

First, the premise. It is not uniformly true. Over the 220 runs of E-bimodal2 the median run writes no novel value at all, but the mean is 9.78 and one run writes 128 of 150. Writing q for the share of non-truth writes that took a value the store had not seen, the apparatus realises $q = 0.083$ pooled and $q = 0.331$ in its highest-scattering model. That model has 0 of 44 runs in the dead band.

The same split on real facts, and it points the other way. E-real plants another real capital rather than a one-digit corruption (realfacts.py), so its falsehoods are semantically distinct from the truth; its models produce novel errors at a pooled q of 0.91; and, unlike E-bimodal2, its runs log the number of distinct values in each retrieved window, so the store's multi-valuedness is measured rather than inferred from a count of novel writes. The split below is post hoc and confounded with model identity, which is why it is also given within model.

The two arms disagree by a factor of 6.4 and they differ in exactly one thing the scripted largest-block reader is sensitive to: a synthetic novel value is another one-digit variant of the same numeral, a real one is a different city. The operative variable is plausibly how distinguishable a store's falsehoods are rather than how many there are, a hypothesis neither arm was built to test, which is why prereg/E_MULTIVALUE_PREREG.md was amended to vary distinguishability at fixed count before it is run, and why row 1k is graded split rather than resolved either way.

Table 22. Intra-seed correlation over the ten model pairs, and the design effect each pair implies. The pooled figure this paper quotes is a mean over these.

arm	pairs	ρ range	mean ρ	design effect range
bimodal2 f0=0.9, 44 seeds	10	+0.29 to +0.85	+0.53	2.14–4.39
transient3 front R=45, 44 seeds	10	+0.24 to +0.95	+0.65	1.94–4.79
transient3 front R=120, 44 seeds	10	+0.18 to +0.99	+0.62	1.73–4.95

Table 23. Claim 4's ordering, at three seeds and at 44. p is exact over all 120 permutations, not asymptotic. The last two columns are the cross-model spread of the predictor: at three topics the early copy rate separates the five models by 0.33, at forty-four by 0.06.

window	ρ at 3 seeds	p	ρ at 44	p	predictor spread, 3 → 44
$W = 3$	+0.921	0.067	+0.500	0.450	0.333 → 0.061
$W = 5$	+0.921	0.067	+0.308	0.667	0.400 → 0.068
$W = 8$	+0.975	0.033	+0.564	0.400	0.458 → 0.065
$W = 10$	+0.975	0.033	+0.800	0.133	0.500 → 0.080

The arithmetic half: a two-valued space does not empty the interval. $f_{\text{end}} = (n_{\text{false},0} +$ false writes$)/(n_0 + T)$, so a reader whose copy probability does not depend on the store lands where that constant puts it, in the interior. The interval is empty only if the copy probability rises with the number of false records retrieved, which is a statement about the reader and is what $\gamma(j)$ measures. Scripted, on the same two-valued space:

The mechanism half, where the objection has force. Distinct falsehoods do not reinforce one another: a reader shown five different wrong values has no majority to copy. baseline_urn.py's docstring asserted the direction of that effect and never tested it. Below, the scripted reader is given the scattering knob q and two decision rules – copy with probability $\gamma(j)$ where j counts non-truth records, or where j is the largest single-value block among them. Only the second can feel scattering at all, and it is the realistic one.

Read together with the first table this is a bound rather than a dismissal. The apparatus reaches $q = 0.33$ in one model and the interval is empty there; the scripted sweep puts the boundary between $q = 0.25$ and $q = 0.5$ under the rule that can feel scattering. Claim 1 is therefore a statement about stores whose falsehoods concentrate, the measured regime includes q up to 0.33, and a language-model arm with three or more candidate answers is owed and registered: it is the one experiment that would move claim 1's scope rather than its confidence.

T Four estimators for γ

The fill rule is the estimator, and it has to be bounded rather than described. Monotonicity gives a bracket for nothing: if γ is non-decreasing in j, filling an unobserved cell from the nearest measured cell below is a pointwise lower bound on the curve and filling from above is a pointwise upper bound, so simulating both brackets the capture rate a fully measured curve would have produced. Isotonic regression on the observed cells, weighted by their support, is a third estimator that uses every read rather than the nearest one; dropping cells under thirty reads is a fourth. Cell-level uncertainty is a cluster bootstrap over seeds, because the reads in a cell are not independent draws but a handful of runs contributing many reads each.

No estimator repairs arm A, and the reason is that the fill rule barely fires. It invents cells for qwen3-8b alone, which is the one model arm A predicts exactly; the four models that under-predict have every cell measured, so their bracket has zero width and isotonic regression moves the answer by 0.1 of a seed unit. Dropping thin cells and filling from below makes arm A worse (22.9 seed units, under-predicting 5/5); dropping them and filling from above, which replaces measured data with an upper bound, still leaves 8.0 seed units. The one-sidedness lives in the measured cells: arm A is below arm B by +0.029 on average over

Table 24. The realised answer space of E-bimodal2, per model, and the outcome distribution conditional on it. q is the share of non-truth writes taking a fresh value. The split is observational and confounded with model identity, which is why it is also given per model.

model	novel writes, median	max	q	capture	in dead band
gemma-3-27b-it	2	10	0.02	0.682	1/44
granite-4.1-8b	0	0	0.00	0.705	2/44
ministral-8b-2512	0	0	0.00	0.818	2/44
phi-4	0	0	0.00	0.727	2/44
qwen3-8b	43	128	0.33	1.000	0/44
novel writes 0	$n = 161$			0.795	6/161
novel writes 1-4	$n = 27$			0.667	0/27
novel writes 5-19	$n = 6$			0.167	1/6
novel writes >=20	$n = 26$			1.000	0/26

Table 25. E-real's runs split by whether the store went multi-valued. The band is the registered void at f_0=0.1 and the dead band at f_0=0.9. The multi-valued subset has the fuller band, which is the direction the scripted sweep predicts and the opposite of what E-bimodal2's highest-q model shows.

subset	n	in band	share	capture	$\bar{f}$	distinct retrieved
f0=0.1 (the registered level, P-T1-N6), band $(0.306, 0.512)$						
two-valued (novel = 0)	338	4	0.012	0.027	0.031	1.21
multi-valued (novel $\geq$ 1)	22	1	0.045	0.455	0.389	2.13
multi-valued (novel $\geq$ 5)	19	1	0.053	0.526	0.445	2.27
within granite-4.1-8b	two-valued 1/163, multi-valued 1/17					
within ministral-8b-2512	two-valued 3/175, multi-valued 0/5					
f0=0.9, band $(0.35, 0.5)$						
two-valued (novel = 0)	311	6	0.019	0.103	0.160	1.66
multi-valued (novel $\geq$ 1)	49	5	0.102	0.592	0.544	2.39
multi-valued (novel $\geq$ 5)	41	5	0.122	0.707	0.636	2.51
within granite-4.1-8b	two-valued 1/141, multi-valued 4/39					
within ministral-8b-2512	two-valued 5/170, multi-valued 1/10					

well-supported cells (18/28 of them positive) and +0.044 over thin ones (8/12). The fill rule is not the diagnosis of arm A's one-sidedness: three topics are not forty-four, and that is the whole of it.

The bootstrap column carries a second correction, in the other direction. §4.1's z scores divide by the binomial error of the observed rate and treat the simulated rate as exact. It is not exact: it is a function of a γ estimated from three seeds, and the widest band here is 0.966 wide. With that uncertainty put back, arm A's observed rate falls inside its own band for 3/5 models, so arm A's failure is established for two models of five rather than for the five the z scores imply. Arm B's is inside for 5/5, which is what a landed prediction looks like. Both corrections were found by building the bounded estimator.

U Three rankers on the same stores

§4.4 attributes a displacement of 0.500 to ranking determinism and measures it with one embedder. The two quantities that attribution rests on (whether a score can separate a truth record from its one-digit corruption, and how many distinct values survive into the top-k) are properties of the ranker and of the stored text, so they can be computed offline for any ranker, on the exact form retriever.record_text produces, over stores of the shape the loop builds. Query: What is the Q3 revenue of Palgrove Dynamics?; truth 28.35M; seeded falsehood 28.36M.

Every ranker's retrieved value set is identical on bm25 12/12; char3 12/12; minilm 12/12 store compositions as the store grows. The largest truth-versus-falsehood separation over

Table 26. A state-independent reader, copy probability p, on the identical urn. Its dead band is not empty, so bimodality is not a consequence of the answer space having two values.

p	capture	in dead band	$\bar{f}_{\text{end}}$	min	max
0.1	0.000	0/400	0.150	0.094	0.225
0.3	0.000	153/400	0.337	0.231	0.431
0.5	0.723	111/400	0.528	0.438	0.625
0.7	1.000	0/400	0.711	0.588	0.806
0.9	1.000	0/400	0.901	0.831	0.956

Table 27. Scattering sweep on the measured γ of the model whose realised answer space is richest. Under a largest-block rule the empty interval is a statement about a regime, and the regime has a boundary between q=0.25 and q=0.5.

q	rule	capture	in dead band	$\bar{f}_{\text{end}}$
0.0	count non-truths	1.000	0/400	0.951
0.25	count non-truths	1.000	0/400	0.957
0.5	count non-truths	1.000	0/400	0.948
0.75	count non-truths	1.000	0/400	0.950
1.0	count non-truths	1.000	0/400	0.953
0.0	largest block	1.000	0/400	0.952
0.25	largest block	0.963	14/400	0.697
0.5	largest block	0.260	241/400	0.448
0.75	largest block	0.000	67/400	0.288
1.0	largest block	0.000	0/400	0.196

all three is 0.0052, and for bm25, char3 it is exactly zero. That is structural rather than a weakness of any of them: the query is the question and the corruption is a digit inside the answer, so no query–document score carries information about the corrupted field. MiniLM's k_{eff} runs 1–4 against 1–1 for BM25 and 1–1 for character trigrams, so the deployed embedder is the least degenerate of the three and §4.4 measured the most favourable case. A lexical index does not restore a draw; it replaces a similarity rule by a first-occurrence one.

V E-multivalue: the arm bought for row 1k

§4.1 left one question open with evidence on both sides, and this arm was registered, amended twice and run to close it. Three cells at f_0=0.9, twenty seeds, five models, the configuration otherwise equal to E-bimodal2 field for field (asserted by the runner before any call is made): near plants three falsehoods that are one-digit corruptions, far plants three at different magnitudes plus a leading-digit change, and sem plants three different real capitals. near against far is the load-bearing contrast and holds the world fixed; sem changes the world too and is a bridge to E-real, which is that cell at one falsehood.

The manipulation check, first, because a criterion evaluated on a manipulation that did not move is not evidence. It moved the planted values and it did not move the store.

Read the last column of each block first. The store starts with three distinct falsehoods and re-concentrates: the distinct count retrieved falls from 3.75 at the first step to 2.65 over the tenure, and the largest block rises from 3.00 to 5.08. Whichever falsehood is copied first grows a block, and the initial scatter is transient. That is a mechanism this arm measured and did not predict, and it is the reason the two criteria below come out as they do.

Both graded criteria came out against the hypothesis that bought the arm. P-T1-V1″ asked for far to exceed near by 0.10 and got -0.030: failed. Paired over the twenty seeds the two cells share, far is fuller on 0 and near on 3, at a sign-test p of 0.25 where twenty seeds could have reached 2e-06. P-T1-V2″ asked whether the count matters and predicted it does not: near sits -0.002 from the published one-falsehood arm, supported. The bridge cell also lands

Table 28. The same comparison as the out-of-sample capture comparison of §4.1, under four estimators, with a seed-clustered bootstrap band on the simulated rate. Arm A is the published curve (three topics); arm B re-measures on the forty-four it predicts. The bracket columns coincide wherever no cell needs filling, which is most of arm A.

arm	model	obs.	cells	thin	below	above	isotonic	≥ 30 only	bootstrap 95%
A	gemma-3-27b-it	0.682	9/9	3	0.081	0.081	0.081	0.000	[0.000, 0.151]
	granite-4.1-8b	0.705	9/9	2	0.499	0.499	0.499	0.301	[0.000, 0.842]
	ministral-8b-2512	0.818	9/9	2	0.545	0.545	0.549	0.365	[0.227, 0.763]
	phi-4	0.727	9/9	4	0.399	0.399	0.401	0.000	[0.000, 0.966]
	qwen3-8b	1.000	4/9	2	1.000	1.000	1.000	0.665	[0.993, 1.000]
	mean |sim−obs|, seed units				12.4	12.4	12.3	22.9	
B	gemma-3-27b-it	0.682	9/9	0	0.606	0.606	0.606	0.606	[0.412, 0.807]
	granite-4.1-8b	0.705	9/9	0	0.789	0.789	0.789	0.789	[0.645, 0.921]
	ministral-8b-2512	0.818	9/9	0	0.836	0.836	0.836	0.836	[0.705, 0.955]
	phi-4	0.727	9/9	0	0.807	0.807	0.807	0.807	[0.659, 0.906]
	qwen3-8b	1.000	8/9	2	1.000	1.000	1.000	1.000	[1.000, 1.000]
	mean |sim−obs|, seed units				2.3	2.3	2.3	2.3	

Table 29. Three rankers, four store compositions. k_{eff} is distinct on-topic values in the top-k, over the number available. "k oldest" marks a ranker whose top-k is exactly the k oldest on-topic records, which is what happens when every on-topic score ties and insertion order decides.

store	ranker	k_{eff}/avail	truth in top-k	sep(T,F)	k oldest	purity
t=0 9 false / 1 true	bm25	1/2	no	0.0000	yes	1.00
	char3	1/2	no	0.0000	yes	1.00
	minilm	2/2	yes	0.0052	no	1.00
t=70 captured, 79/1	bm25	1/2	no	0.0000	yes	1.00
	char3	1/2	no	0.0000	yes	1.00
	minilm	2/2	yes	0.0052	no	1.00
t=70 escaped, 9/71	bm25	1/2	no	0.0000	yes	1.00
	char3	1/2	no	0.0000	yes	1.00
	minilm	1/2	yes	0.0052	no	1.00
t=70 captured + 3 novel	bm25	1/5	no	0.0000	yes	1.00
	char3	1/5	no	0.0000	yes	1.00
	minilm	4/5	yes	0.0052	no	1.00

at 0.01 beside §4.5's 0.019 rather than beside far's 0.0, so the world carries an effect that distinguishability does not.

P-T1-V3: which variable is γ a function of? No bar, by design: a twenty-seed arm should not decide a definition. Both $\varphi = j/k$ and the largest single-value non-truth block j^* are one-dimensional lookup tables over $0 \dots k$, fitted per model and held out per seed over 44,989 steps. φ scores -7,791.3 and j^* scores -9,282.6, so φ wins by 1,491.3 and Def. 1 names the right variable. That is a result in the primitive's favour and it is also what makes the two nulls inevitable: a reader responding to the count of non-truth records cannot respond to how those records disagree with one another. The scripted largest-block reader of §4.1 is a refuted model of this apparatus's reader, and row 1j withdraws the bound drawn from it.

W Runs, steps and cost

Every experiment in this paper, counted from results/. A "step" is one model call that produced a parseable answer and therefore one write to the store; steps whose answer did not parse are absent from the log rather than zero-filled, which is one of the invariants in §O. The total is 673,008 logged steps.

Table 30. What the manipulation did. The first two columns are the planted values' own spread. The rest is the largest single-value non-truth block among the k=8 retrieved records, and the number of distinct values retrieved, averaged over the first W steps. The two cells are indistinguishable at every window: three planted values give a block of three whether they are near or far apart, which is what makes the contrast a test of whether the model reads near values as agreeing.

	planted max/min		largest block, first W					distinct, first W				
cell	median	max	1	3	5	10	150	1	3	5	10	150
near	1.46×	6.85×	3.00	3.14	3.26	3.50	5.08	3.75	3.85	3.81	3.73	2.65
far	100.01×	100.16×	3.00	3.14	3.21	3.50	5.54	3.75	3.80	3.82	3.73	2.57

Table 31. E-multivalue's three cells against the published v=1 arm. The band is $[0.35, 0.50]$, the interval §4.1 calls empty.

cell	n	share in the band	capture	$\bar{f}_{\text{end}}$
near	100	0.030	0.870	0.849
far	100	0.000	0.970	0.924
sem	100	0.010	0.070	0.129
bimodal2 (v=1)	220	0.032	—	—

Table 32. The experimental record, counted at build time.

experiment	runs	models	logged steps
bimodal2	220	5	33,000
budget	90	5	13,461
cl_n10	9	3	270
cl_n20	9	3	540
cl_n40	9	3	1,080
cl_n80	9	3	2,160
converge	30	5	4,500
e1e2	144	1	5,759
ea	30	5	4,500
f0	60	1	8,991
frontier	40	2	6,000
gate	30	5	4,500
label	30	5	4,499
multivalue	300	5	44,989
prov2	60	5	9,000
prov_observed_apparatus_failure	30	5	4,500
ratchet	42	3	6,300
retriever	420	3	63,000
sched	45	5	6,750
temp	200	2	30,000
transient	135	5	20,229
transient2	180	5	26,983
transient3	1,980	5	296,997
verify	500	5	75,000
total	4,602		673,008

X The scorecard, every graded row

This paper's claims are graded here rather than in its sections, so that the sections can be about what was found and this appendix can be complete. Criteria were frozen in prereg/ before the corresponding data existed, and one generator writes all three tables from one source, so the index below cannot disagree with the rows it counts.

Table 33. The scorecard in eight lines, counted from the scorecard. Every graded row of Table 35 and Table 36 falls in exactly one class below. Where a family's status names two classes at once, the row is counted in the more serious one, which is the only direction of error a summary of one's own scorecard may have. It is an index, not a substitute, and the rows are all still there: 27 claim families in Table 35, 63 sub-rows in Table 36. What the unflattering rows are for. 46 of the 87 rows carry a positive grade and 37 an unflattering one, and the second number is not a defect count: each of those rows narrows a claim made above it, names a number of ours that moved, or marks a question this apparatus cannot decide. None of them is an experiment left undone: work nobody has run is not graded here at all, and is bounded where it is stated.

grade	rows	what the grade means in this paper
withdrawn / superseded	7	asserted, then taken back or re-assigned; the sequence is in the post-mortems
not supported / failed	11	a frozen criterion was missed, on our own data
self-correction	7	our own arithmetic or our own criterion was wrong and is repaired here
not decidable / not evaluable	8	the apparatus or the band cannot separate the outcomes: not a null
acknowledged limit	4	a boundary of the result, stated where the result is
formal	1	proved from the definitions; no measurement can move it
measured	39	an artifact on disk, and a reader may dispute the artifact
argued / reported	6	no grade is claimed: an argument, or a number printed without a criterion
other	4	status words outside the eight classes above
total graded rows	87	every numbered row of both tables

Table 35. Every registered claim, one row each. Criteria were frozen in prereg/ before the corresponding data existed; failures are graded, not removed. Grades: measured (criterion met), failed/not supported (criterion missed), refuted (predicted direction wrong), withdrawn (we retract it), untested (apparatus floored), unresolved/undecided (sample insufficient), split (two admissible observables disagree), derived (not measured here), underpowered (the criterion's own bar is met pooled and missed on at least one of its seeds), diagnosed (a mechanism identified for a graded row, not itself a registered claim).

Claim	Status	Evidence
1. Outcome distribution bimodal, not a decaying mean	measured; below: 2 not supported, 1 withdrawn	8/7/0 split; gap 6.4× mean gap
2. $\gamma(\varphi)$ alone orders capture rate, zero fitted parameters	measured, ordinal; below: 1 not supported, 1 withdrawn	$\rho = +0.82$ at three seeds; $+1.00$, $p = 0.017$ at forty-four
3. The outcome is decided in the first few steps	measured; below: 1 withdrawn	12/15 at t=2 of 150

Table 35, continued.

Claim	Status	Evidence
4. Early copy rate orders capture rate	not supported at forty-four seeds; measured as published	ρ +0.92/+0.98 at three seeds, +0.31 to +0.80 at forty-four, $p \geq 0.13$
5. The mean-field sign criterion locates the boundary	measured falsification of our own	4/5 drift down, capture anyway
6. Score $= C \cdot s$: contamination costs determinism, not information	measured	+0.383 pass@8 − pass@1
Rows 7–7o are one axis in sixteen registered steps, not sixteen findings: they record how one claim was narrowed, floored, split and finally declared unresolvable after we bought the sample meant to resolve it.		
7. At matched budget, grounding timing dominates fraction	measured; below: 3 not supported	P-S1 5/5; 0.20/0.33/0.40 vs 0.47
16. The seed selects the topic and dominates the variance	measured; invalidates our own error bars	2.0% at three seeds, 68.8% at nine; groups 5.0σ apart
8. A majority admission gate erases model differences	measured; its own 5/5 is one collapse, not five confirmations	0.729–0.932 → 0.994; treated spread 0.0006 against control 0.9250
9. A cost-matched self-fed state slot helps at low f_0	underpowered 3 seeds, criterion met on 2 of them	P-A1 4/5 pooled; 5/3/5 per seed against a bar of 4
10. A source marker lowers contamination (D13′)	failed; below: 1 not supported	P-P2 2/5; P-P3 3/5
11. An uninformative marker sits nearer the lying end than the midpoint	measured	0.856 vs midpoint 0.664, 5/5
12. Damage from an intervention scales with unaided cleaning ability	withdrawn; below: 1 withdrawn	P-R2: gap 0.015 vs bar 0.30
13. A marker's effect is a single function of label–truth correlation	withdrawn; below: 1 withdrawn	P-L1: 0/5 monotone
14. The n/n_0 collapse is a discovery about contamination	withdrawn to an apparatus check; below: 1 withdrawn	a scripted truth-only reader satisfies it exactly
15. Provenance bits are needed in this regime	vacuous here	$\alpha_{\min} = 0$ for 4/5 models
17. The timescale attack: φ, not f, carries the decision, via an absorbing boundary	not supported, P-T1-D2/D3	219/219 runs leave that boundary; §5(ii′) kept word for word

Table 35, continued.

Claim	Status	Evidence
18. The retrieval falsifier fired: an embedding ranking displaces capture	measured, P-T1-R0/R1; below: 1 not supported	-0.500 at f_0=0.1 against a zero band of 0.10
19. The four intervention criteria, re-read against a 44-seed control	measured: no verdict changes	8 criteria; control capture moves by up to +0.68; P-P2 2/5 → 3/5, still failed
20. …and at three seeds no clustered sign test can be significant at all	derived	minimum attainable two-sided p is 0.25 at 3 seeds, 0.016 at 7
21. The two-arm disagreement's within-arm half is decided, and only just: multi-valued runs hold the dead band at 0.1220 against 0.0193	measured, at the edge of its own resolution	difference 0.1027; facts resampled with replacement give [0.006, 0.241] with 2.0% of draws at or below zero
22. "The recommended remedy violates append-only" was too strong: a 2026 remedy attacks the same failure and removes nothing	self-correction	candidate conclusions as separate entries with Noisy-OR probabilities (Liao et al., 2026); the claim holds of forgetting, not of remedies
23. The abstract said the consistency gate drives every model to 0.994; the measurement is 0.993	self-correction	every cell lies in $[0.9934, 0.9935]$; 0.9938 is the ceiling, a different quantity two sentences earlier in the same abstract
24. Eleven of this paper's tables are emitted from their artifacts, and generating them found three defects	code-verifiable	residual of hand-typed three-decimal values 184 → 117; the three are the two rows below and a var(n_f) cell that read 0.347 where its grader computes 0.477

One row per claim family. Each status cell names the grades of that family's sub-rows, so a failure is counted here even though its row is elsewhere: all 63 sub-rows are in Table 36 in full, of which 5 failed, 5 self-correction, 4 withdrawn, 3 split, 3 acknowledged limit, 2 undecided, 1 untested, 1 unresolved, 1 refuted, 1 not supported. Nothing is dropped, and nothing is summarised into a kinder word than it had.

Table 36. Supporting refinements, moved from Table 35 by the stated rule. The direction is deliberate: a selection rule for a scorecard should be biased against its authors, so unflattering rows stay in the main table and supporting detail moves here. claims_all.tex is the source both tables are generated from.

Claim	Status	Evidence
The 63 rows below were moved here from Table 35 by the rule in papers/split_claims.py: a lettered sub-row whose status is supportive moves; every top-level claim and every row carrying a withdrawal, failure or self-correction stays. Nothing is dropped.		
1a. The interval survives a 15× sample, sparse rather than empty	measured at n=220	P-T1-B1′: 3.6% vs bar 10%, uniform 20.6%; P-T1-B2 gap 0.799
1b. The balance between the modes	withdrawn	8/7 becomes 39/173
1c. Its mean describes 8.2% of the runs	measured, frozen first	P-T1-M1: 18/220 within ±0.10 of $\bar{f} = 0.792$

Table 36, continued.

Claim	Status	Evidence
1d. Its median describes 68.2%, so the objection is to the mean	reported; narrows 1c	P-T1-M2: 150/220; best centre 74.5%
1e. Bimodality at frontier scale, against our registered prediction	measured, bar 0.50	P-T1-F1: pooled 0.850; sonnet-4.5 20/20
1f. Empty at $f_0 = 0.1/0.3/0.5$ too	measured, bar ≤ 0.20	P-T1-F0-1: 0/20 three times
1g. f_0=0.9 is not the adversarial extreme	measured, post hoc	capture 0.500/0.700/0.750/0.682, peak at f_0=0.5
1h. At f_0=0.1 the mean falls inside a strictly empty interval	reported, not registered	mean 0.459, 0/20 in $(0.306, 0.512)$
1i. The empty interval is not an artefact of a two-valued answer space	measured, scripted control	a state-independent reader on the same urn puts 113/300 runs in the dead band at p=0.3
1j. …and the scripted bound we drew from a largest-block reader	withdrawn: that reader is refuted as a model of ours	held out over 44,989 steps, $\gamma(\varphi)$ beats $\gamma(j^*)$ by 1491 in log-likelihood
1k. …and our two arms still disagree about the scope, with two candidates now excluded	split: reported, not resolved; both splits post hoc	synthetic highest-q model 0/44 in the band; E-real multi-valued 5/41 against 6/311, within one model 4/39 against 1/141
1l. …the number of planted falsehoods does not move the interval	measured, P-T1-V2″, predicted against our own first prereg	v=3 near sits -0.002 from v=1, bar ± 0.10
1m. …nor does their distinguishability	failed, P-T1-V1″, the arm's load-bearing criterion	far − near = -0.030 against a bar of $+0.10$; 0/20 seeds fuller, sign-test $p = 0.25$
1n. …because an initial scatter is transient: the store re-concentrates	measured, not predicted	largest block 3.00 → 5.08; distinct retrieved 3.75 → 2.65 over the tenure
1o. …and γ's independent variable is φ, not the largest single-value block	measured, P-T1-V3, no bar by design	held-out log-likelihood favours φ by 1491 over 44,989 steps
2a. The published γ out of sample	failed, one-sidedly	12.1 seed units, worst z 8.50; under-predicts 4/5, the fifth at the ceiling in both
2a$^{\dagger\prime}$. …and how far that failure is established	self-correction: the z credits the simulation with exactness	seed-clustered bootstrap: outside the band for 2/5, not 5/5; widest band 0.966
2b. γ re-measured on the topics it predicts	measured	2.2 seed units, worst z 1.32, 1/5

Table 36, continued.

Claim	Status	Evidence
2b$^{\dagger\prime}$. …and out of sample under a split-half refit	measured, not registered	1.15×/1.16×; signed z +0.32/ + 0.35 vs −5.15
2c. The one-sidedness is our fill rule, not the model	withdrawn (App. T)	the rule fires for qwen3-8b alone, the one model arm A predicts exactly
2d. …on a model family it was never fitted on	split	P-T1-F4, bar 2 seed units: gpt-4o 4.0, sign opposite to 2a
2e. …and no estimator repairs arm A	measured, four estimators	bracket width 0 on 4/5; isotonic 12.3 vs 12.4 seed units
3a. …because of (1)'s $1/n$	withdrawn (the reading, not the measurement)	Prop. 3: vacuous over the first 49 steps
4a. …because the predictor stops separating the models	diagnosed	cross-model spread of the early copy rate 0.333 → 0.061
7a. Front-loading's magnitude bar	failed	P-S3: $0.27 < 0.33$
7b. Pooled estimates stable to a re-run; verdicts identical	measured, by accident	two realisations, §L
7c. Ordering survives three budgets; at R=45 back-loading is worth nothing	measured	0.07/0.20/0.47 against control 0.47
7d. The schedule effect shrinks as the budget grows	failed	P-B1, P-B2: spread 0.20 → 0.40
7e. …and fails in the direction a prior derivation predicts	attribution, not confirmation	the confirming sweep is not run
7f. Free re-grounding is a single point, not a neighbourhood	derived, not measured	spread 0 at α=1, $\ln T$ at every $\alpha < 1$; the limits do not commute
7g. The sweep past the peak, on the frozen observable	untested (apparatus failure)	$f_{\max}(148)$=0.516, 0.016 above the threshold
7h. …and the discriminator at three seeds	undecided, inside the band	−0.088; interval unusable either way
7i. The scale-free retest on six fresh seeds	split	−0.167 against −0.060, bar −0.15
7j. …and the supported side sits on the floor clause's boundary	disclosed, not used to overturn	arms 0.700/0.733/0.700, span 1 seed unit
7k. The resolution P-T1-S3 called unbuyable, bought	unresolved at ≤ 1980 runs, marked permanently	+0.46σ against a 3σ bar
7l. …and the sample it needs grew when measured	measured, against us	effect 0.115 → 0.021; se 21% above the $1/\sqrt{n}$ projection; 3σ needs ∼1895 seeds
7m. …and the discriminator at 44 seeds	undecided at 1.14σ	−0.052; at three seeds −0.088, outside the band

Table 36, continued.

Claim	Status	Evidence
7n. …and the shape criterion at 44 seeds	failed on magnitude and significance	-0.074 vs bar -0.15, 1.62σ
7o. …and the ordering, which the axis rests on, at 220 runs per arm	reported, not graded	front < uniform < back at all three budgets
16a. …so each seed is worth 1.33 runs however many models run on it	derived from the measured ICC	design effect $1 + (m-1)\rho = 3.75$
16b. …and no verdict changes, every criterion being a point estimate against a bar	stated, checked row by row	§M
16c. …and 3.75 is a property of pooled levels; a paired contrast carries 1.14–1.47	measured at 44 seeds, both	levels 3.56–3.66; clustered/naive se 1.80–1.85 against 0.93–1.15
16d. …and one ρ for five models is an assumption that does not hold	measured, against Prop. 4's form	ρ +0.18 to +0.99 over ten pairs; implied design effect 1.73–4.95
9a. …and its high-f_0 direction	tie, reported as such	P-A2: 0 better, 1 worse, 4 tie
17a. …and φ is fast, which the prereg discounted in advance	measured, P-T1-D1	median first passage 2 steps
17b. …and Prop. 3's bound is also loose where the decision is taken	self-correction	11.4×; identity residual 10^{-16} over 2420 steps
18a. …and ranking polarises rather than raising collapse, as we predicted	refuted, P-T1-R2, registered first	0.533 → 0.000 at f_0=0.1; 0.933 → 1.000 at f_0=0.9
18b. …and the cause is ranking determinism, not lost topic routing	measured, P-T1-R5	keeping the oracle reproduces 1.00 of the displacement
18c. …and the mechanism is variance, not breadth	not supported, P-T1-R3: the registered mechanism fails, the one named as least sure survives	identical set on 59/59 steps vs 0/59; var(n_f) 0.240 vs 5.753
18d. …and k_{eff} misses the exact match it was defined to catch	self-correction: the wrong summary of the right idea	ranking at k=8 equals uniform at k=1 (0.000/1.000); k_{eff} 1.40 vs 1.00

Table 36, continued.

Claim	Status	Evidence
18e. …and the arm overran its spend guard, so one of three models is excluded	acknowledged limit: two models judged	$5.06 stop, $5.13 spent vs $2.06 registered; 33 runs dropped
18f. …and the determinism is ranking's, not MiniLM's	measured, three rankers, offline	max truth/false separation 0.0052; BM25 and char-trigram k_{eff} both 1 against MiniLM's 4
21a. …and the $\sqrt{\text{deff}}$ rule of thumb disagrees with the cluster bootstrap	measured	$[-0.029, 0.235]$ spans zero; both are printed and the bootstrap decides because it is the statistic's own sampling distribution
21b. …and the between-arm half is a construct difference, which no item count decides	acknowledged limit	the arms differ on whether a novel value is a variant of a numeral or a different city (Anonymous, 2026c, row 29)
21c. …and a decisive replication of the first half costs a fourfold corpus	derived	164 multi-valued runs per group at the measured design effect, 1,405 runs, 141 facts against the 36 we have
22a. …and Prop. 1 stands under it, because the store is still append-only	formal	the ceiling is about what the store contains; what changes is which functional the agent acts on
22b. …and whether the mutable part merely moved to the weight vector is not answered here	acknowledged limit	it needs the Noisy-OR update's behaviour under a growing false majority, which nobody in this paper has derived
23a. …and it was found by building this paper's first generated-macro namespace rather than by rereading	code-verifiable	every three-decimal literal in the abstract is now a macro read from results/ at build time; the audit fails if one is replaced
23b. …and four of the six graders behind those numbers persisted nothing	reported, against us	they recomputed on every run and printed, so the abstract held the only durable copy; FRONTIER.json and GATE.json exist as of this version
23c. …and the namespace's first name collided with this family's own notation	self-correction	\gnd is gnd in all eight preambles, \providecommand lost the collision silently, and the abstract typeset macro names for one build; the emitters use \newcommand now, so a collision fails the build
24a. The empty interval's cuts were stored rounded, so the run that defines the lower edge fell outside it	self-correction	0.306 and 0.512 against the definition's 0.30625 and 0.5125; the two rows of Table 6 were counted against different thresholds and Table 9's void column against a third
24b. …and the corrected count is the more favourable one, which is why it is in the body	measured	middle share 0.045 → 0.036 against a bar of 0.10; P-T1-B1′ supported under both, and the density ratio falls 0.18 → 0.15 against a bar of 0.25

Table 36, continued.

Claim	Status	Evidence
24c. …and four of the graders behind these tables persisted nothing before this version	reported, against us	FRONTIER, GATE, BIMODAL2, MEAN and TRANSIENT3 exist now; the papers held the only durable copies