Cumulative Machine Culture: Distilled Inheritance in Two Preregistered Studies

Two preregistered confirmatory studies of knowledge transmission between stateless language-model generations

Abraham September 2026

Abstract

Can knowledge accumulate across generations of a language model that never updates its weights, purely through what one generation writes down for the next? We built a lineage kernel in which each "generation" of a frozen 8-billion-parameter model learns a hidden rule world from worked examples, writes notes, and hands a bounded notebook to the next generation, which is evaluated on held-out items. Seven inheritance conditions were compared, paired by lineage, including token-matched contrasts and unmatched controls. Two preregistered confirmatory studies were run. In a single-operator world (C2; 20 lineages × 5 generations), lineages inheriting a rewritten textbook outperformed lineages inheriting a raw log of their predecessors' attempts at generation 5 by 5.8 accuracy points (95% CI 2.5-9.2; Holm p = 0.009) and solved more items per token (Holm p = 0.016); the living-minus-frozen generation-2 notebook contrast was equivalent within ±12.5 points, without a detected superiority effect. In a harder world with one- or two-operator programs (C3; 18 lineages × 5 generations), the living lineage pulled ahead of its frozen generation-2 self by 6.9 points (95% CI 3.2-10.9; Holm p = 0.007), the preregistered ratchet-positive endpoint, and the distilled-versus-raw advantage replicated (+7.9 points; Holm p = 0.002). All gains are relative to a baseline near floor; absolute accuracies stay under 20%. The results are specific to one checkpoint and one task family.

1. Research question

Human capability accumulates because each generation inherits a compressed record of what earlier generations learned, improves it, and passes it on. This study asks whether an artificial system with no memory across runs and no weight updates can show the same dynamic when the only channel between generations is text.

Three questions were preregistered:

1. **Distillation.** At a matched budget of inherited tokens, does a generation that inherits a *rewritten textbook* (its predecessors' consolidated notes) outperform one that inherits a *raw log* of its predecessors' attempts?

2. **Ratchet.** Does a lineage that keeps rewriting its notebook keep improving over its own frozen past, or does re-authoring lose information?

3. **Economy.** Does distillation change correct solves per token spent?

The study does not ask whether the model "understands" the rules or whether the notebooks are true; it measures held-out accuracy under fixed conditions.

2. The world and the kernel

Task family. Hidden-rule record transformation: a JSON record with seven fields is transformed by a program of opaque operators (18 in the catalog, with saturating arithmetic and list operations); the model must output the transformed record. Operator names are salted aliases, so semantics can only be learned

from worked examples and inherited text. Items are generated deterministically from a study salt and registered in a cross-study contamination registry. Learn items (24 per generation) come with the correct answer for note-taking; eval items (24 per generation) do not.

Worlds. C2 used single-operator programs; C3 used one- or two-operator programs after calibration. This is a difficulty manipulation within one task family, not proof that C2 exhausted all learnable information.

Lineages and generations. A lineage comprises five generations. Each executed condition receives 24 learning tasks and 24 held-out evaluation tasks. The distilled arm writes learning notes and consolidates them into a notebook of at most 3,000 tokens for subsequent generations. Other arms follow their specified inheritance rules rather than all authoring textbooks. Conditions within each lineage share task streams, enabling paired contrasts.

Conditions (seven).

Condition	What generation g inherits
no_memory	nothing
placebo_matched	irrelevant prose (public-domain text), token-matched to the distilled notebook
raw_log	the previous generation's raw attempt log, truncated to the distilled notebook's size
raw_natural	raw history under its separately configured 7,000-token cap
distilled_textbook	the previous generation's consolidated notebook
static_notebook	the lineage's *generation-2* notebook, frozen
static_matched	the frozen generation-2 notebook, size-matched to what distilled inherited

static_matched is the token-matched frozen-history comparator. An endpoint difference tests continued updating against that comparator; absence of a significant difference does not prove that rewriting adds nothing.

Model and decoding. One fixed quantized Llama 3.1 8B checkpoint was used. Per-lineage generation-1 determinism probes governed whether early static cells could share observations. In C3, lineage 16 had 43/48 identical probe rows, below the 0.90 threshold; four static cells were therefore executed rather than copied.

3. Preregistration and analysis

Each study was calibrated before its confirmatory design, sample size and analysis were frozen and reviewed. Generation outputs are hash-chained. C2 has 620 executed cells plus 80 shared copies; C3 has 562 executed cells plus 68 shared copies. Shared observations are not additional independent samples.

C2 primary (decision pair). distilled_textbook - raw_log at generation 5: H1 accuracy and H2 correct solves per 1,000 solve tokens, lineage-paired, exact sign-flip tests (all 2^20 sign patterns enumerated), Holm across {H1, H2}, $\alpha = 0.05$; TOST at ±0.125 only if H1 was not significant; a five-branch decision table fixed in advance. Ratchet (distilled - static_matched) as a preregistered secondary. Sized for 80% power at a 12.5-point effect from calibration variance (20 lineages).

C3 primary (ratchet pair). R1: distilled_textbook - static_matched accuracy at generation 5; R1-slope: per-lineage slope of that gap over generations 3-5; Holm across the pair; TOST only if R1 was not significant; a six-branch decision table. The C2 decision pair was preregistered as a replication secondary. Sized at $\alpha = 0.025$ for the Holm pair (18 lineages).

Efficiency is defined as correct solves per 1,000 solve tokens, zero when a lineage solves nothing, so it is always defined.

4. Results

4.1 C2: single-operator world (20 lineages)

Decision branch 1: superiority of the distilled textbook.

Contrast at generation 5 (distilled - raw_log)	Estimate	95% CI	Holm p
H1 accuracy	+5.8 points	+2.5 to +9.2	0.009
H2 correct solves per 1,000 tokens	+0.044	+0.011 to +0.079	0.016

Ratchet (distilled - static_matched, generation 5): +3.1 points, $p = 0.15$; equivalent within ±12.5 points (TOST $p < 0.001$). No detected endpoint superiority; equivalence applies only within the stated margin.

Pooled accuracy by generation:

Arm	g1	g2	g3	g4	g5
no_memory	.077	.075	.079	.085	.079
placebo_matched	.077	.073	.083	.077	.077
raw_log	.077	.133	.113	.115	.121
raw_natural	.077	.135	.138	.163	.144
distilled_textbook	.077	.175	.181	.190	.179
static_notebook	.077	.175	.185	.173	.175
static_matched	.077	.175	.156	.156	.148

Descriptively, much of C2's distilled gain appears by generation 2. Placebo and no-memory endpoint accuracies are similar, not identical. These observations do not establish that the task world was exhausted or that irrelevant context can never affect performance.

4.2 C3: one-or-two-operator world (18 lineages)

Decision branch 1: ratchet-positive.

Contrast	Estimate	95% CI	Holm p	Decision
R1: distilled - static_matched, generation 5	+6.9 points	+3.2 to +10.9	0.007	reject

Contrast	Estimate	95% CI	Holm p	Decision
R1-slope: gap growth per generation, g3-g5	+2.2 points/gen	+0.1 to +4.3	0.070	retain

Replication of C2 (secondary; Holm across the pair): distilled - raw_log at generation 5, +7.9 accuracy points (95% CI +4.4 to +11.1) and +0.065 correct solves per 1,000 tokens (95% CI +0.035 to +0.093); Holm p = 0.002 for both.

Pooled accuracy by generation:

Arm	g1	g2	g3	g4	g5
no_memory	.044	.049	.083	.042	.058
placebo_matched	.044	.058	.067	.039	.053
raw_log	.044	.076	.088	.074	.062
raw_natural	.044	.086	.106	.076	.083
distilled_textbook	.044	.097	.134	.118	.141
static_notebook	.044	.097	.123	.118	.102
static_matched	.044	.097	.109	.102	.072

C3's living lineage exceeds the frozen comparator at generation 5. The positive gap slope is not significant under the registered test. The trajectories alone do not identify a shift in the task distribution or the mechanism behind the endpoint gap.

4.3 Secondary observations

Schema violations. C2 distilled lineages had more record-domain schema violations at generation 5 than raw-log lineages (approximately 24% versus 12%). The mechanism is not established by these rates.

Efficiency. The reported metric counts correct evaluation solves per 1,000 evaluation solve tokens, excluding learning and notebook-authoring costs. It supports the distilled-versus-raw comparison, not total-cost superiority. In a separate post-hoc C2 accounting, recorded learning, evaluation and authoring tokens totaled 5,374,356 for distilled and 3,986,335 for raw_log.

Sensitivity. At generation 5, all raw_natural lineages carry a cap flag. The implemented any-arm/any-flag exclusion therefore removes every lineage in both studies. This differs from the narrower design-flag wording in the preregistrations. The original sensitivity remains non-estimable. Separate, labeled post-hoc checks exclude design-flagged lineages; C3's distilled-static estimate is +7.2 points (N = 15, unadjusted p = 0.012). These checks cannot replace the planned robustness analysis.

5. Interpretation

The studies support a narrow finding: externally inherited distilled text improved held-out performance over bounded raw history for this checkpoint and task family. They do not establish that rewriting is necessary for every form of knowledge transfer, that notebooks faithfully encode rules, or that raw experience provides no benefit.

C2 did not detect an endpoint advantage over frozen history within its equivalence margin; C3 did detect a positive endpoint difference. This motivates further tests of continued updating, but it does not show that C2

learned everything possible by generation 2. C3's nonsignificant slope leaves sustained cumulative improvement unresolved.

6. What the results do and do not license

Licensed. For this checkpoint and task family, distilled textbook inheritance outperformed raw-log inheritance at generation 5 on accuracy and correct evaluation solves per kilotoken in C2 and C3. In C3, distilled also exceeded its token-matched frozen generation-2 comparator at generation 5, meeting the preregistered ratchet-positive endpoint rule. The gap slope was not significant. Both accuracy advantages are below the 12.5-point planning effect size.

Not licensed. That the notebooks are correct, that the model understands the operators, that the effect generalizes to other checkpoints, task families, or interactive agents, that culture is cheaper than no culture, or any claim about the ratchet's trajectory beyond generation 5. Absolute accuracies remain low (under 20%), so these are statements about relative transmission, not about competence.

7. Deviations and integrity notes

Structural near-duplicate registrations: single-operator items collide by construction across E002's own partitions in the contamination registry. Three recorded deviations (D-cal2-1, D-c2-1, D-c3-1) tolerate such collisions when both rows are the lab's own E002 rows in an explicit allow-list; exact, text, and trace collisions and any collision with another study still halt. D-c3-1 also admits E002 rows under the original task specification when their partition is allow-listed; as implemented the guard applies to any allow-listed partition, not only the two calibration partitions it was written for, and no observed result relied on it. All tolerated findings are itemized in the ledgers.

A replication-block field in C3's analysis file embeds a five-contrast Holm value (0.0043) where the preregistered two-test value applies; the correct pair value (0.0016) is reported here and in the file's holm_across_pair block. Fixed in code for future runs; the frozen output is unchanged.

One calibration preregistration (cal3) carried a copied prose line stating 12 eval items where the frozen manifest, thresholds, and report all used 24; dispositioned in the C3 preregistration.

Integrity checks verified the sealed execution inventories and ledger links. Operational interruptions are not treated as independent replications.

The independent review rechecked artifact integrity and replayed saved inference from sealed outputs using the analysis implementation. The main accuracy sign-flip tests were also recomputed separately. The reviewer had access to saved analysis; this was not an outcome-blinded independent implementation of every statistic.

8. Limitations and next tests

One quantized checkpoint; one synthetic task family; five generations; low absolute accuracy. Priority follow-ups are diversified task calibration, a second model family, and controlled expert-to-novice teaching. Additional generations could test trajectory claims. None of these extensions is established by the current findings.

9. Reproducibility statement

Studies E002c2_cumulative_machine_culture_confirmatory and E002c3_ratchet_two_op_confirmatory are frozen with byte-hash ledgers; calibration studies E002cal2a, E002cal2b, and E002cal3 likewise. Per-item outputs, hash-chained ledgers, analysis outputs, the labeled post-hoc sensitivity files, and the reviewer audits

are archived with the lab. The analysis is deterministic given the sealed outputs (exact enumeration for all sign-flip tests; bootstrap seeds fixed in the manifests).

Suggested citation

Abraham. (2026). Cumulative Machine Culture: Distilled Inheritance in Two Preregistered Studies. Abraham Labs, Research Report E002-C2/C3, version 1.0.0.