SAME TASK PERFORMANCE, DIFFERENT INTERVENTION READOUTS: CROSS-RUN VARIATION UNDER IN-CONTEXT RULE ALIASING

Anonymous authors
Paper under double-blind review

ABSTRACT

Causal edits can yield different readouts from models with nearly identical task behavior. We study 26.1M-parameter transformers on a synthetic assignment language where RECENCY and RARITY select the same answer on every training document. All 75 runs across 25 configurations reach accuracy at least 0.999, yet under an answer-preserving multiplicity edit, the sign fraction—the fraction of valid pairs with positive log-odds displacement—spans more than 0.3 across seeds in 13 cells, reaching 0.879; there a strong positive tail contrasts with mostly near-zero negatives. Replacing the probe batch moves this fraction by at most 0.034 in four extreme runs. These are differences in intervention sensitivity rather than answer policy: on generated documents where the rules disagree, rarity-answer rates are at most 0.003 in the main grid. Supervision that distinguishes the rules changes the readout. A RARITY-labeled arm supplies a one-sided trained control for positive displacement; in the RECENCY-labeled arm, observed dispersion is lower. The dispersion comparisons remain exploratory and are not adjusted for cell selection. Strong directional readouts occur before runs meet the retrieval diagnostics. Intervention readouts should be reported with behavioral tests, independent training runs, checkpoint, probe population and gating policy rather than treated as configuration-level mechanism labels.

1 INTRODUCTION

A document may assert that a slot holds one value and later assert that it holds another. Answering a query can exploit several correlated cues. In ordinary text, the value stated later is often also the value stated once rather than many times. When those cues agree, accuracy does not determine how a model will respond when they diverge—for example in a stale retrieval corpus, a partially edited document, or an inconsistently propagated update. Which cue pairs are coextensive on a training distribution is difficult to establish in natural text without an audit. We therefore construct a corpus in which one pair is coextensive by identity, so that the relation is exact and can be perturbed under controlled supervision.

Documents in our synthetic assignment language are sequences of (entity, attribute, value) statements followed by a query; the answer is the most recent value assigned to the queried slot. Bindings are resampled per document and the stream is infinite, so answers cannot be recovered by memorizing cross-document bindings. Two statistics of the cue structure vary across a 5×5 grid, with the task, its Bayes-optimal policy and the training recipe fixed in every cell: R_{old}, the number of times a superseded value is repeated, and ΔD, the number of statements from the last update to the query.

The current value occurs once despite clipping, and the tie-break makes RARITY and RECENCY agree on every training query. These query-answer labels cannot distinguish the two rules. We invert the competing values' multiplicities while preserving the ground truth, token count, answer position and all other tokens. The resulting log-odds change, Δ, is positive in the rarity-directed direction and negative in the occurrence-count-directed direction. It measures intervention response rather than answer policy: on generated disagreement documents, rarity-answer rates are zero in 70 of 75 runs and at most 0.003 in the other five (§5.4, App. A.4.4). Within each cell, seeds share the

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.

training objective but use independent document streams. Equation 1 holds throughout the generator's support.

We then relabel disagreement documents according to either RECENCY or RARITY. In the widest cell, the RECENCY-labeled arm shows lower cross-seed standard deviation, from 0.268 over nine state-passing baseline runs to 0.132 over nine state-passing runs at a 1% inversion rate, and to 0.127 over ten runs at 30%. Exploratory, selection-unadjusted label permutations give $p = 0.070$ and $p = 0.042$; the lowest-dose result remains descriptive. A second cell gives $\sigma = 0.165$ at 1% over ten runs, but its baseline contains only three seeds. Conversely, the RARITY-labeled arm produces large positive readouts and a behavioral rarity preference in its state-passing runs, providing a trained positive control. In the original grid, 13 of 25 cells span more than 0.3 across seeds; the filler-slot control median stays within 0.015 nats (App. A.4).

What reproduces more reliably is behavior and coarse timing. Before meeting the retrieval criteria, some models occupy a positional shortcut whose ceiling follows in closed form from the token layout, and escape time is ordered by redundancy at the row level. Pooled over three seeds, the terminal sign fraction also rises with redundancy through $R_{\text{old}} = 12$. Predominantly negative pre-escape readouts occur in every redundancy row, but we report them as training-dynamics diagnostics rather than retrieval-rule attributions (§6.3).

Accuracy saturates while intervention readouts vary across runs. Generated disagreement documents show that the readout does not identify answer policy. Supervision supplies a trained positive control and exploratory dispersion comparisons. Checkpoint and schedule tests delimit reproducibility. These findings establish the phenomenon here, without identifying necessary or sufficient conditions for non-replication.

2 Related Work

Behavioral studies show that knowledge-conflict resolution tracks entity popularity and evidence coherence (Longpre et al., 2021; Xie et al., 2024), update plausibility (Kortukov et al., 2024), agreeing-source count (Chen et al., 2022; Jin et al., 2024a), repetition (Yan et al., 2024), and source authority (Schuster et al., 2026). They identify which cue wins, but fixed, generally unobserved corpora prevent attribution to training statistics. Mechanistic work localizes conflict-carrying components (Jin et al., 2024b), and end-to-end circuits capture non-trivial behaviors, although completeness remains difficult to certify (Wang et al., 2023). Prefix-matching induction heads are a canonical copy-based in-context-learning mechanism (Olsson et al., 2022), with detailed formation studies (Singh et al., 2024) and transient dynamics (Singh et al., 2023). Controlled-corpus studies identify distributional conditions for in-context learning (Chan et al., 2022; Raventós et al., 2023; Bratulić et al., 2025; Kim et al., 2026; Han et al., 2023), while a parallel line studies parametric knowledge storage and extraction (Allen-Zhu & Li, 2024).

Reddy (2024) is the closest timing antecedent: it traces an induction head's abrupt appearance to the data distribution and identifies progress measures preceding in-context learning, as does our loss-decrease-rate candidate peak (§6). Gibson et al. (2026) extends this to algorithmic phases implemented by distinct multi-layer subcircuits. Complementary work finds stability: Tigges et al. (2024) report higher-level circuit algorithms persisting across training and scale despite component changes, while van der Wal et al. (2025) quantify stability and outliers across fifty language-model pre-training runs. These studies ask when mechanisms emerge, persist, or replicate. We hold successful in-context retrieval fixed and ask whether independently trained models respond reproducibly to a fixed multiplicity intervention. This intervention-sensitivity measurement does not identify a unique answer-rule implementation.

The broader framing is predictive multiplicity and underspecification: the same training-and-validation criterion can admit models with equivalent aggregate performance while leaving other behaviors undetermined (Marx et al., 2020; D'Amour et al., 2022). Recent mechanistic work establishes that explanations need not be unique. Enumerating candidate explanations on Boolean functions and small networks yields systematic non-identifiability: several circuits can replicate a behavior, and one circuit can admit several interpretations (Méloux et al., 2025). Structurally near-disjoint circuits can be simultaneously faithful, sparse, and complete on the same task (Chen et al., 2026). Holding the task fixed while varying input frequency can likewise yield structurally different dis-

covered circuits that transfer across conditions and implement the same computation (Makou et al., 2026), while circuit-level conclusions can fail to survive defensible variation in the analysis pipeline itself (Mahale, 2026). These works primarily concern multiplicity within a trained model, across discovery runs, or across analytic specifications. We study a complementary between-training-run question under exact training-support equivalence: two candidate answer rules are coextensive on the training support, and we compare that construction with supervision that makes their targets differ. The experiment tests this construction rather than claiming that training-support coextensiveness is necessary or sufficient for explanation non-uniqueness.

Our positional rule is a shortcut (Geirhos et al., 2020; McCoy et al., 2019; Liu et al., 2024), unusual in that the token layout is ours to fix, so its payoff follows in closed form along a controlled axis rather than being diagnosed after the fact. Abrupt reorganizations are often summarized as phase diagrams (Power et al., 2022; Nanda et al., 2023; Doshi et al., 2024; Goddard et al., 2025); since discontinuous metrics can manufacture transitions (Schaeffer et al., 2023) we report the transition unthresholded alongside the gate. Park et al. (2025) is closest, a phase diagram for a synthetic in-context task with phases framed as competition between algorithms, but their axes are properties of the task distribution and phases are assigned by matching outputs against candidate algorithms. Ours are statistics of the cue structure within a document, with the task and its Bayes-optimal policy identical in every cell, so the in-distribution optimum is attained everywhere and behavioral matching cannot distinguish the phases at all.

3 SETUP

3.1 A LANGUAGE WITH AN ALIASED RULE PAIR

Documents are sequences of assignment statements followed by a query. Each statement binds a value to an (entity, attribute) pair, which we call a *slot*, and is emitted as four tokens `[ent] [attr] [val] [SEP]`; the query appends `[QUERY] [ent] [attr] [ARROW]` and the target is the value currently bound to the queried slot. Identifiers come from disjoint vocabularies of size 200, 8 and 512. Bindings are resampled per document, so memorized slot–value associations cannot solve the task.

Every document contains one queried slot: a value v_{old} is bound to it and repeated R_{old} times, a later statement rebinds the slot to v_{new}, and the query follows ΔD statements after that rebinding. Remaining positions are filled by statements from other slots, each independently receiving an update with probability $p_{\text{update}} = 0.5$. Values are drawn without replacement within a document, which makes "the answer value occurs exactly once before the query" structural rather than probabilistic and removes a baseline noise term.

The alias. For the queried slot s in document d, let $I_s(d)$ index its statements and let $V_s(d)$ be the set of values that actually occur in that slot. Define

$$c_s(v) = \big|\{i \in I_s(d) : \text{val}(s_i) = v\}\big|, \qquad t_s(v) = \max_{\substack{i \in I_s(d) \\ \text{val}(s_i)=v}} \text{pos}(s_i).$$

The pair $(c_s(v), -t_s(v))$ is ordered lexicographically, so RARITY first minimizes count and then breaks ties toward the latest occurrence. For every training document,

$$\underbrace{\text{val}\Big(s_{\arg\max_{i \in I_s(d)} \text{pos}(s_i)}\Big)}_{\text{RECENCY}} = \underbrace{\arg\min_{v \in V_s(d)} \big(c_s(v), -t_s(v)\big)}_{\text{RARITY}} \qquad \text{for all } d. \tag{1}$$

The current value occurs once and is latest. If only one old copy survives placement, the counts tie and the recency tie-break still selects the current value; otherwise it is the unique least frequent value. Thus clipping does not break the identity. Training input–target pairs cannot distinguish these two answer rules. This does not imply that arbitrary neural implementations have identical token-level losses or that they are connected by a flat path in parameter space. Five further candidates are separable observationally; per-cell rates and the closed forms they match are in App. A.3.

3.2 THE TWO AXES

Neither axis changes the target function. Both change statistics of the within-document evidence and may thereby affect optimization, but the experiments do not identify a mechanism-specific implementation cost. *Redundancy* $R_{\text{old}} \in \{3, 5, 8, 12, 16\}$, fixed per cell, sets how often the language demonstrates that a value appearing more than once is not the current one. Two lower values are excluded: at $R_{\text{old}} = 1$ the counts always tie and the edit has no domain, and at $R_{\text{old}} = 2$ four of five runs fail to meet the retrieval criteria (App. D.6). The upper end is structural: once $2R_{\text{old}}$ approaches the placement window the queried slot becomes identifiable without reading the query, which caps it at 16.

Update-to-query distance ΔD is the number of statements between the final rebinding and the query. Increasing ΔD inserts more statements between the final rebinding and the query. This changes a different corpus statistic from repeated-value redundancy, but the experiment does not isolate their neural implementation costs. For nominal $d = 2, 3, 5, 8, 16$, respectively, we sample ΔD uniformly from the integer intervals $[1, 3]$, $[2, 4]$, $[2, 8]$, $[4, 12]$, and $[8, 24]$. These are the intervals produced by round-to-nearest with ties to even.

Cell names and tables use the nominal centre d as the ΔD label. At constant ΔD the answer sits at a fixed token offset from the end, which makes "copy the value at offset k" fully correct with no slot matching. §6.1 derives the bound and compares it with observed plateau accuracy. We call an $(R_{\text{old}}, \Delta D)$ pair a *cell* and a cell trained at one seed a *run*, so the grid is 25 cells and 75 runs.

3.3 INVARIANTS AND TRAINING

We vary two designated corpus statistics and audit potential confounds on 1500 documents per cell. The prespecified checks pass within their stated tolerances except for partial matching of antecedent distance; additional quantities that covary with the axes are reported in App. A.2 and examined by the controls in App. F. Document length is sampled from $U[45, 55]$ statements independently of both axes. We do not equalize the tail-update fraction; we report it as a covariate and stratify on it.

We train 26.1M-parameter, 8-layer decoder-only transformers with $d_{\text{model}} = 512$, 8 heads, rotary embeddings, pre-norm blocks and per-head QK RMS normalization initialized at gain 2. Each 16000-step run uses batch size 256 and bfloat16 AdamW: peak learning rate 10^{-3}, $(\beta_1, \beta_2) = (0.9, 0.95)$, $\epsilon = 10^{-8}$, 500-step linear warmup, cosine decay to 10^{-4}, weight decay 0.1 on non-embedding matrix parameters, and gradient clipping at 1.0. Loss covers all non-padding tokens and documents are generated online. Arm-specific overrides are stated locally.

The budget is set by the slowest cell: the retrieval criteria are first met at step 10000 at $(R_{\text{old}} = 3, \Delta D = 16)$ in seed 1 and at 12000 in one run of the ten-seed extension (App. C.5), so 16000 steps leaves every analyzed run several thousand post-gate steps. In one tested high-redundancy cell the sign fraction remains within 0.995–1.000 over an eightfold budget range, whereas a boundary cell moves by 0.55 (App. D.7.3).

4 MEASURING INTERVENTION SENSITIVITY

4.1 THE MULTIPLICITY INVERSION

To probe the distinction between aliased candidate rules, we apply paired edits to held-out documents. The logic is that of an interchange intervention (Geiger et al., 2021; 2024), applied at the input rather than to an internal variable. The edit changes an input statistic under which the candidate predictions diverge while leaving the ground truth unchanged. Writing $m(x; v^\star) = \log p(v^\star \mid x) - \log p(v_{\text{truth}} \mid x)$, the readout is the paired difference

$$\Delta = m(x_{\text{edit}}; v^\star) - m(x_{\text{base}}; v^\star), \tag{2}$$

where both terms use $v^\star = v_{\text{old}}$, the post-edit prediction of RARITY, while the ground truth remains v_{new} on both sides.

For realized old-copy count R, the edit changes queried-slot counts from $\{v_{\text{old}} : R,\ v_{\text{new}} : 1\}$ to $\{v_{\text{old}} : 1,\ v_{\text{new}} : R\}$ without moving statements. For example, in queried-slot occurrence order (not necessarily adjacent in the document), $[v_{\text{old}}, v_{\text{old}}, v_{\text{old}}, v_{\text{new}}]$ becomes $[v_{\text{old}}, v_{\text{new}}, v_{\text{new}}, v_{\text{new}}]$.

The latest value and ground truth remain v_{new}, while RARITY changes to v_{old} and FREQUENCY changes oppositely (App. A.4.3). RECENCY remains v_{new} and PRIMACY remains v_{old}; a positive Δ is a rarity-directed log-odds displacement and need not change the predicted answer. Retaining the earliest old copy makes a fixed earliest-occurrence rule cancel, but does not rule out responses to other rewritten positions (App. A.4).

4.2 What we report

The primary readout is the fraction of valid edit pairs with $\Delta > 0$, using 400 held-out documents at the final checkpoint. Valid pairs satisfy the edit-domain and invariant checks. Terminal comparisons require answer accuracy > 0.99 and accuracy > 0.95 on repeated same-slot value tokens (the copy diagnostic; App. A.5). A two-sided exact binomial test against 0.5 quantifies document-sampling uncertainty within a fixed run, not variation across training runs. We also report the median and interquartile range of Δ. Contrast-pair mass is a separate diagnostic: restricting it changes the evaluated population and does not bound Δ or its median. Table 8 reports ungated seed-0 summaries, and Table 9 reports sign fractions for all three seeds.

5 Results

5.1 Behavior saturates everywhere

All 75 analyzed runs achieve in-distribution answer accuracy at least 0.999, including on the stratum where no update follows the queried rebinding, despite substantial variation in their intervention readouts. This concerns answer accuracy, not equality of confidence, loss or every in-distribution statistic. The regime is not guaranteed by the language: in the excluded $R_{\text{old}} = 2$ column, three of five runs terminate in the POSITIONAL state, attaining approximately 98%, 98%, and 70% of the fixed-offset ceiling (App. D.6).

5.2 Cross-run variation within cells

Readouts vary across seeds within cells (Figure 1). Thirteen of 25 cells span more than 0.3, and eight straddle 0.5. The widest cell, $(3, 8)$, spans 0.879: seeds 0 and 2 have sign fractions 0.098 and 0.977, but unrestricted medians -0.024 and $+1.083$ nats (Table 6). Both within-run binomial tests give $p < 10^{-50}$. Precise within-run estimates therefore coexist with cross-run variation and asymmetric displacement magnitudes.

Pooled over three seeds, the sign fraction rises with redundancy through $R_{\text{old}} = 12$ before falling at 16 (Table 9). Only two of the four steps hold in every seed individually. The fall at 16 is almost entirely seed 1's, so this pooled trend is descriptive rather than a reproducible per-cell ordering.

The spread is not explained by the measured probe-batch term or by the fixed-seed rerun check. The widest-cell range also does not narrow over the available checkpoints (App. D.7.4), and the audited rerun's losses match (App. A.4).

Magnitude thresholds show that the spread is not solely a zero-crossing artifact but is asymmetric. In the widest cell, cross-seed ranges of $S^+(\tau) = \Pr(\Delta > \tau)$ remain 0.948, 0.761 and 0.528 at $\tau = 0.1, 0.5, 1$ nat: the high-positive run retains a substantial tail, while the two low runs are over 98% within ± 0.5 nat. Across the grid, 14, 16 and 16 cells retain S^+ ranges above 0.3; the mass-restricted counts are identical (Table 5). Thus this cell's sign spread combines a strong positive tail with mostly small negative displacements, rather than symmetric large effects.

The shift constrains one account at row level, not per cell: pooled by row, the mass-restricted median is positive at every $R_{\text{old}} \geq 5$ in all three seeds, inconsistent with an idealized count-only response that favors higher occurrence count (App. A.4.3).

5.3 Covariates

Several quantities covary with the axes and are not fully separated by the available controls; we report them per cell (Table 2). Realized redundancy falls with ΔD because copies outside the

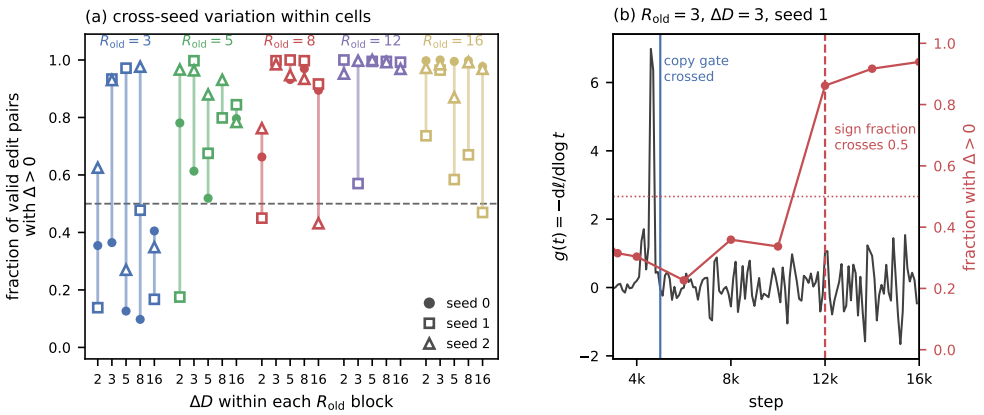

Figure 1: **(a)** Fraction of valid edit pairs with $\Delta > 0$ at 16000 steps for 25 cells under three seeds, joined per cell; filled circles seed 0, open squares seed 1, open triangles seed 2. Line length is the seed range, reaching 0.879; batch exchange moved four extreme runs from two cells by at most 0.034 (App. D.7.1). This is a displacement-sign summary, not a RARITY-answer rate. **(b)** The $R_{\text{old}} = 3$, $\Delta D = 3$ run of seed 1 (§6.3): loss-decrease rate $-\mathrm{d}\ell/\mathrm{d}\log t$ (left) and the same sign fraction (right). Its candidate peak is near step 4600, the copy diagnostic reaches 1.000 at step 5000, and the readout crosses near step 12000 at unchanged accuracy.

document canvas are discarded, and edit domain tracks it, so the cells with the weakest readout condition on the smallest subset.

We split the valid documents separately within each run at that run's median realized redundancy. The high-redundancy stratum has a larger absolute median displacement in 65 of 75 runs (Table 10). This association also changes edit size ($R-1$ rewritten tokens); it neither explains cross-run variation nor identifies a mechanism. Related gradients appear in the tested 4- and 12-layer arms (App. E.2); exploratory aliased-grid ablations yield correlated layer-level profiles across two runs, but no head meets the original joint criterion (App. F.6).

5.4 WHERE THE MEASUREMENT DEGRADES

Contrast-pair mass diagnoses how much probability remains on the two compared values after the edit; it does not bound their log-odds difference. Applying a per-document mass restriction of 0.5 shifts a cell median by as much as 63% where mean mass falls below 0.85, so mass-restricted medians are conditional summaries of the retained documents rather than lower bounds on cross-cell effects. The directional result is robust: recomputing the sign fraction on the mass-restricted subset moves it by at most 0.026 across the 75 runs, changes the widest cross-seed range from 0.879 to 0.877, and leaves the same 13 cells above 0.3 (App. B).

The readout is a sensitivity, not an implemented rule. On inverted documents generated rather than edited, where RECENCY and RARITY select different values, no run exhibits a stable rarity-answer policy: the rate of rarity answers is 0.000 in 70 of the 75 runs and at most 0.003 in the rest, against 1.000 in the arm that forces the objective toward RARITY (App. A.4.4). This is consistent with Δ rather than in tension with it. All $28{,}563$ base argmaxes over the value vocabulary are v_{truth}; after editing, 98.68% remain there, 0.03% move to $v^{\star}$, and 1.30% to third values (Table 4). Thus cross-seed spread measures log-odds movement, not answer policy.

5.5 VARIATION ACROSS TRAINING RUNS

Equation 1 establishes observational equivalence of two candidate answer rules on the training support. It does not imply that arbitrary neural implementations have identical token-level loss or share a flat parameter path. We therefore treat the cross-run dispersion as an empirical finding and test several alternative sources.

Replacing the document batch under a fixed model moves the sign fraction by at most 0.034 in four extreme runs across the two widest cells, while exchanging the two models on one shared batch recovers the cell's seed range within binomial error (App. D.7.1). Fixing initialization and varying the training-document stream gives $\sigma = 0.300$ over four streams, against the online ten-seed 0.271 (App. D.7.2). Across post-gate checkpoints, one run reaches a within-run standard deviation of 0.332 (App. D.7.3). Doubling the cosine period moves a cell by up to 0.55, and removing annealing moves one run by 0.748 while another moves by 0.012 (App. D.7.5). These comparisons show that sensitivity depends on training run, checkpoint and schedule, while document resampling does not explain the largest differences. Terminal checkpoints share a step budget but not post-gate duration (4000–14000 steps in the widest cell; App. E.3).

We also tested whether the multiplicity-inversion readout identifies a parameter direction unusually weakly constrained by the training loss. The measurement did not distinguish it from seven control readouts, including two constrained by the training targets (App. F.7). We therefore do not use this measurement as evidence for a flat rule-trading direction.

5.6 Supervision and a one-sided trained control

Equation 1 holds in every cell, so the grid contains no case where the training labels distinguish the two rules. We create one by relabeling: in a fraction p_{break} of documents, the queried-slot multiplicities are inverted at generation time and the two rules disagree. We call an arm RECENCY-labeled or RARITY-labeled according to which rule supplies the target on those documents. Filler slots invert at the same rate, or "the latest value is repeated" would identify the queried slot. At fixed seed the two arms are token-identical except for the target token of inverted documents. Readouts use non-inverted documents, matching the main grid's queried-slot base multiplicities; filler slots may still be inverted. Construction and structural checks are in App. A.6; per-dose values are in Table 11.

Among state-passing runs at $\Delta D = 8$, baseline cross-seed standard deviation is 0.268 ($n = 9$). Under RECENCY labeling, it is 0.132 at $p_{\text{break}} = 0.01$ ($n = 9$) and 0.127 at 0.30 ($n = 10$; Table 11). Exploratory one-sided Brown–Forsythe-type label permutations give $p = 0.070$ and 0.042; only the latter is below the nominal 0.05 level. Neither comparison is confirmatory or locates a minimum dose; the relabeled runs' near-zero medians leave dispersion confounded with magnitude. At $\Delta D = 5$, the 1% value is $\sigma = 0.165$ over ten runs, against only three baseline seeds.

Training runs are the units of comparison. Because the $\Delta D = 8$ cell was selected by its original three-seed range, the full-sample comparisons are neither selection-adjusted nor independent confirmation. Enumeration is exact for the specified label-assignment distributions, not a guarantee of finite-sample calibration for the composite equal-dispersion null; its randomization interpretation requires exchangeability (App. C). New-seed-only comparisons give $\sigma = 0.184$ versus 0.139 at 1% over five shared seeds and 0.179 versus 0.158 at 30% over six; both remain descriptive.

The RARITY-labeled arm supplies a one-sided trained positive control. At $p_{\text{break}} = 0.30$, all three state-passing runs at $\Delta D = 8$ and two of three trained runs at $\Delta D = 5$ have positive intervention readouts and behavioral rarity-answer rates of 1.000. Their mean mass-restricted medians are 25.92 and 21.89 nats. This shows that the probe responds when supervision induces a strong RARITY policy. Their near-ceiling sign fractions make the small observed ranges weak evidence about general stability. Across doses, exposure to the inverted configuration changes together with supervision, so the dose comparison does not isolate rule distinguishability from that exposure.

6 Shortcut Dynamics

Training routes differ by cell. We call the first observed crossing of 0.95 by the copy diagnostic *escape*, and the early loss-decrease-rate maximum a *candidate peak*; neither identifies an internal circuit. A late gate with a valuable shortcut accompanies thousands of steps on a positional plateau; an early gate accompanies a low-accuracy regime without shortcut saturation.

6.1 FIXED-OFFSET CEILING AND OBSERVED PLATEAU ACCURACY

Statements occupy exactly four tokens, so the answer's one-based offset measured backward from the end is $4\Delta D + 6$, independent of document length. A rule that ignores the query and emits the value at a fixed offset is therefore correct exactly when ΔD takes one particular value, so

$$\text{posCeil} \;=\; \max_{d} \Pr[\Delta D = d] \;=\; \frac{1}{|\text{supp}(\Delta D)|} \tag{3}$$

since ΔD is uniform on its support. Table 15 compares this to plateau accuracy in five cells: they occupy 98%, 98%, 73%, 70% and 1% of a ceiling that is itself falling (App. D.3).

Saturating Eq. 3 shows only that a positional rule suffices. As an operational retrieval diagnostic, we measure accuracy on value tokens that repeat an earlier value of the same slot and therefore require reliable same-slot copying. Throughout the plateau it reads 0.001–0.046 against answer accuracies of 0.077–0.326 in the four plateau cells (Table 15). These checkpoints have not demonstrated reliable copying under this diagnostic; the measurement does not prove that every retrieval-related internal computation is absent.

6.2 CANDIDATE PEAKS REFINE LOW-REDUNDANCY TIMING

Escape is abrupt on the evaluation grid: the copy diagnostic moves from near chance to above 0.95 between consecutive evaluations in every cell that escapes, and answer accuracy rises to 1.000. To avoid locating this change with a threshold on a 1000-step grid (Schaeffer et al., 2023), we define the loss-decrease rate $g(t) = -\mathrm{d}\ell/\mathrm{d}\log t$ from loss recorded every 100 steps and take its maximum over the first four fifths of training as a candidate peak. It usually precedes the copy gate; if the maximum in the final fifth is higher, we flag the candidate and rely on the gate rather than treating it as a reliable escape-time estimate.

At 100-step resolution the candidate peak separates four redundancy levels in seed 0 where the gate separates three (Table 14; Figure 2). The pooled height–time correlation ($\rho = 0.933$) is coordinate-dependent because $g(t) = t(-\mathrm{d}\ell/\mathrm{d}t)$ and cells differ by design; we do not treat it as independent evidence of sharper late transitions. Across the 75 runs, seven are flagged, all at $R_{\text{old}} \geq 12$: three in seed 0 and four in seed 2. At $R_{\text{old}} = 3$, candidate peak times span 1900 at $\Delta D = 8$, seed 2, to 9600 at $\Delta D = 16$, seed 1, a factor of 5.1. Under the fixed 16000-step budget, neither axis alone orders escape; 19 of the 20 seed-0/1 candidates at $R_{\text{old}} \geq 12$ lie at step 400. The ordering is compatible with a race between loss reductions, not an isolated mechanism identification.

6.3 PRE-RETRIEVAL READOUTS CAN REVERSE DIRECTION

Predominantly negative pre-escape readouts occur in every redundancy row of the main grid, not only in the excluded cells. At the last probe point preceding escape, the sign fraction is below 0.20 in 32 of 75 runs and below 0.05 in 22, spanning all five rows pooled over seeds. Of these 32 runs, 28 have a checkpoint-level mean-mass record, and 17 of those means are at least 0.5. Document-level mass-restricted pre-escape sign fractions are not reported, so these means do not establish that the direction survives restriction.

Their direction is frequency-type: after the edit makes v_{truth} frequent and v_{old} rare, $\log p(v_{\text{old}}) - \log p(v_{\text{truth}})$ decreases, so $\Delta < 0$. It is not interpretable as a retrieval-rule attribution, because these checkpoints have not met the retrieval criteria. A positional account is plausible but not identified (App. D.6). Five pre-escape readings exceed 0.6, but all lie within one evaluation interval of escape and therefore cannot be separated from probes that caught the transition (Table 13).

We condition terminal cross-run readouts on the retrieval-state criteria and report each run as RETRIEVAL, POSITIONAL, or NEITHER (App. A.5). Pre-escape readouts are retained only as training-dynamics diagnostics and are not treated as retrieval-rule attributions. Terminal readouts use a fixed budget of 16000 steps; post-escape trajectories drift in both directions (App. D.7.3).

Gate passage is required by the analysis policy but does not guarantee temporal stability. In the $R_{\text{old}} = 3$, $\Delta D = 3$ cell of seed 1 the copy diagnostic reaches 1.000 at step 5000 and the readout stays frequency-type for a further 7000 steps (sign fractions 0.23, 0.36, 0.34 at steps 6000, 8000,

10000) before crossing to 0.86 at 12000 and settling at 0.94. Accuracy is 1.000 at every one of those points and the loss-decrease rate has no peak near step 12000; its four largest values occur at steps 4600, 4700, 4300 and 15400 (Figure 1, right). A retrieval-gate-only rule would have admitted this run at step 6000 and returned the opposite direction: the sign fraction later moves by 0.71 without a corresponding accuracy change or a second observed loss-decrease-rate peak.

7 DISCUSSION

This construction separates reproducibility of task behavior from reproducibility of an intervention-based sensitivity measurement. Models with nearly identical in-distribution accuracy and similarly low rarity-answer rates can respond very differently to an edit that preserves the correct answer. The largest differences survive probe-batch exchange and document-level mass restriction, while the training-stream realization, checkpoint and schedule also change the readout. Exact answer-rule equivalence on the training support makes these comparisons controlled; it does not prove a flat parameter direction or a general criterion for when explanations are identified.

Three reporting practices follow from this case. First, report training-run variation alongside uncertainty from probe documents: a precise estimate within one run need not reproduce in another (van der Wal et al., 2025). Second, state the behavioral and retrieval-state criteria used to interpret a readout. Predominantly negative readouts occur before these criteria are met; meeting them does not ensure temporal stability. Third, specify checkpoint, schedule, probe population and gating policy. Cross-run range and within-run temporal standard deviation summarize different sources of variation and should be named separately.

The trained control produces positive readouts under a strong RARITY policy. Lower dispersion under RECENCY labeling remains exploratory, with smaller new-seed-only contrasts. They establish neither a minimum sufficient dose nor generalization beyond this task family.

8 LIMITATIONS

Scope. The main result uses one synthetic task family and an 8-layer, 26.1M-parameter transformer. Depth experiments cover 4 and 12 layers, and a 102.2M width experiment reproduces opposite readout directions, but these are not independent task-family replications. Rendering the placement as English and fine-tuning a 0.5B model both shift the readout toward the sign-fraction ceiling. Their passing-run ranges are ceiling-compressed, so neither arm tests cross-run dispersion outside the grid (App. E.4). The rendered and fine-tuned conditions also change vocabulary and token geometry. The initial QK-normalization gain affects which tested runs meet the retrieval criteria (App. E.1).

Sampling, cell selection, and training horizon. Several extensions use only two or three seeds. The $\Delta D = 8$ cell was selected by its original three-seed range and those seeds remain in the expanded baseline, so the full-sample comparisons are exploratory and not selection-adjusted. New-seed-only contrasts remain descriptive, and the second cell lacks a comparable baseline. Removing cosine annealing lowers the ten-seed online standard deviation from 0.271 to 0.200, still 3.0 times that arm's median and 1.3 times its largest matched late-window within-run value. Longer-run convergence remains open.

Interpretation. The edit measures a log-odds change, and its trained positive control does not make every intermediate value a classifier of answer policy. Contrast-pair mass diagnoses the output distribution but does not bound log-odds magnitude or guarantee mechanism specificity. The direct geometry measurement did not distinguish its intended controls, and the aliased-grid head ablations produced no head passing the original joint threshold; relaxed-threshold findings are exploratory.

Construction and confounds. Copy dispersion, slot count, document length and shortcut payoff are only partly separated by the available controls. Query replacement changes the prediction in at least 83.8% of documents in every run, but does not prove that every prediction uses the same retrieval computation. The fixed-band and p_{update} arms each contain a run with no interpretable readout (Apps. F.3, F.2), and the non-aliased generator fails its structural matching gates at high redundancy (App. F.4). Across relabeling doses, exposure to the edited configuration changes with supervision. Failed runs and unsuccessful measurements therefore remain part of the reported evidence.

AI USE STATEMENT

Generative AI tools, including GPT-family models, were used to critique and draft manuscript text, edit prose, provide feedback on experimental design and interpretation, assist with mathematical derivations and their presentation, and draft and review analysis and verification code. Synthetic corpora and model outputs were produced by the reported code and trained models rather than by a generative-AI service. The authors ran the experiments, inspected the underlying logs, reviewed and tested all AI-assisted code, checked suggested analyses against machine-readable outputs and primary sources, and manually verified the final claims and citations. The authors take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The language, axes, training recipe, edit and invariants are described in §3.1–§4 and App. A.4. Generator invariants (App. A.1), the state classification used to determine eligibility for terminal cross-run comparisons (App. A.5), and per-cell collision rates (App. A.3) are measured rather than asserted. Code and available machine-readable records are included in the supplementary material; the README documents reproduction procedures and coverage.

REFERENCES

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.1, knowledge storage and extraction. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 1067–1077. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/allen-zhu24a.html.

Jelena Bratulić, Sudhanshu Mittal, David T. Hoffmann, Samuel Böhm, Robin Tibor Schirrmeister, Tonio Ball, Christian Rupprecht, and Thomas Brox. Unlocking in-context learning for natural datasets beyond language modelling, 2025. URL https://arxiv.org/abs/2501.06256.

Stephanie Chan, Adam Santoro, Andrew Lampinen, Jane Wang, Aaditya Singh, Pierre Richemond, James McClelland, and Felix Hill. Data distributional properties drive emergent in-context learning in transformers. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 18878–18891. Curran Associates, Inc., 2022. doi: 10.52202/068431-1371. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/77c6ccacfd9962e2307fc64680fc5ace-Paper-Conference.pdf.

Hung-Ting Chen, Michael Zhang, and Eunsol Choi. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 2292–2307, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.146. URL https://aclanthology.org/2022.emnlp-main.146/.

Xi Chen, Mingyu Jin, Jingcheng Niu, Yutong Yin, Jinman Zhao, Bangwei Guo, Dimitris N. Metaxas, Zhaoran Wang, Yutao Yue, and Gerald Penn. All circuits lead to rome: Rethinking functional anisotropy in circuit and sheaf discovery for llms, 2026. URL https://arxiv.org/abs/2605.12671.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor

Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022. URL http://jmlr.org/papers/v23/20-1335.html.

Darshil Doshi, Aritra Das, Tianyu He, and Andrey Gromov. To grok or not to grok: Disentangling generalization and memorization on corrupted algorithmic datasets. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 3782–3810, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/105fdc31cc9eb927cc5a0110f4031287-Paper-Conference.pdf.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 9574–9586. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/4f5c422f4d49a5a807eda27434231040-Paper.pdf.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In Francesco Locatello and Vanessa Didelez (eds.), *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pp. 160–187. PMLR, 01–03 Apr 2024. URL https://proceedings.mlr.press/v236/geiger24a.html.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, November 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-00257-z. URL http://dx.doi.org/10.1038/s42256-020-00257-z.

Cole Gibson, Wenping Cui, and Gautam Reddy. Distinct mechanisms underlying in-context learning in transformers, 2026. URL https://arxiv.org/abs/2604.12151.

Chase Goddard, Lindsay M. Smith, Vudtiwat Ngampruetikorn, and David J. Schwab. When can in-context learning generalize out of task distribution? In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 19585–19599. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/goddard25a.html.

Xiaochuang Han, Daniel Simig, Todor Mihaylov, Yulia Tsvetkov, Asli Celikyilmaz, and Tianlu Wang. Understanding in-context learning via supportive pretraining data. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12660–12673, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.708. URL https://aclanthology.org/2023.acl-long.708/.

Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, Xiaojian Jiang, Jiexin Xu, Qiuxia Li, and Jun Zhao. Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 16867–16878, Torino, Italia, May 2024a. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.1466/.

Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. Cutting off the head ends the conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 1193–1215, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.70. URL https://aclanthology.org/2024.findings-acl.70/.

Minsung Kim, Dong-Kyum Kim, Jea Kwon, Nakyeong Yang, Kyomin Jung, and Meeyoung Cha. How training data shapes the use of parametric and in-context knowledge in language models. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23242–23257, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1064. URL https://aclanthology.org/2026.acl-long.1064/.

Evgenii Kortukov, Alexander Rubinstein, Elisa Nguyen, and Seong Joon Oh. Studying large language model behaviors under context-memory conflicts with real documents. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=xm8zYRfrqE.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7052–7063, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.565. URL https://aclanthology.org/2021.emnlp-main.565/.

Ajay Pravin Mahale. Explanation multiplicity: Circuit-level interpretability evidence does not survive defensible analytic variation, 2026. URL https://arxiv.org/abs/2608.13754.

Alireza Bayat Makou, Jingcheng Niu, Subhabrata Dutta, and Iryna Gurevych. Many circuits, one mechanism: Input variation and evaluation granularity in circuit discovery. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=Ur42Xx1QPv.

Charles T. Marx, Flavio du Pin Calmon, and Berk Ustun. Predictive multiplicity in classification, 2020. URL https://arxiv.org/abs/1909.06677.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3428–3448, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1334. URL https://aclanthology.org/P19-1334/.

Maxime Méloux, Silviu Maniu, François Portet, and Maxime Peyrard. Everything, everywhere, all at once: Is mechanistic interpretability identifiable? In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 32900–32923, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/51157713f1f1f572a05647e1aef1b1ed-Paper-Conference.pdf.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=9XFSbDPmdW.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads, 2022. URL https://arxiv.org/abs/2209.11895.

Core Francisco Park, Ekdeep Singh Lubana, and Hidenori Tanaka. Competition dynamics shape algorithmic phases of in-context learning. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu

(eds.), *International Conference on Learning Representations*, volume 2025, pp. 66381–66433, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/a635abfdd007c9a698410c8bbe5ebc33-Paper-Conference.pdf.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets, 2022. URL https://arxiv.org/abs/2201.02177.

Allan Raventós, Mansheej Paul, Feng Chen, and Surya Ganguli. Pretraining task diversity and the emergence of non-bayesian in-context learning for regression. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 14228–14246. Curran Associates, Inc., 2023. doi: 10.52202/075280-0626. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/2e10b2c2e1aa4f8083c37dfe269873f8-Paper-Conference.pdf.

Gautam Reddy. The mechanistic basis of data dependence and abrupt learning in an in-context classification task. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=aN4Jf6Cx69.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 55565–55581. Curran Associates, Inc., 2023. doi: 10.52202/075280-2425. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/adc98a266f45005c403b8311ca7e8bd7-Paper-Conference.pdf.

Jakob Schuster, Vagrant Gautam, and Katja Markert. Whose facts win? LLM source preferences under knowledge conflicts. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29430–29459, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1357. URL https://aclanthology.org/2026.acl-long.1357/.

Aaditya Singh, Stephanie Chan, Ted Moskovitz, Erin Grant, Andrew Saxe, and Felix Hill. The transient nature of emergent in-context learning in transformers. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 27801–27819. Curran Associates, Inc., 2023. doi: 10.52202/075280-1208. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/58692a1701314e09cbd7a5f5f3871cc9-Paper-Conference.pdf.

Aaditya K Singh, Ted Moskovitz, Felix Hill, Stephanie C.Y. Chan, and Andrew M Saxe. What needs to go right for an induction head? A mechanistic study of in-context learning circuits and their formation. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 45637–45662. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/singh24c.html.

Curt Tigges, Michael Hanna, Qinan Yu, and Stella Biderman. LLM circuit analyses are consistent across training and scale. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1287. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/47c7edadfee365b394b2a3bd416048da-Abstract-Conference.html.

Oskar van der Wal, Pietro Lesci, Max Müller-Eberstein, Naomi Saphra, Hailey Schoelkopf, Willem Zuidema, and Stella R Biderman. Polypythias: Stability and outliers across fifty language model pre-training runs. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 86180–86204, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/d611d06e3207330555fbc10810e70163-Paper-Conference.pdf.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 35623–35646, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/99261adc8a6356b38bcf999bba9a26dc-Paper-Conference.pdf.

Jianhao Yan, Jin Xu, Chiyu Song, Chenming Wu, Yafu Li, and Yue Zhang. Understanding in-context learning from repetitions. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 49968–49989, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/d88f6f81e1aaf606776ffdd06fdf24ef-Paper-Conference.pdf.

A Construction, edit, and statistical conventions

The appendices specify the generator, edit, analysis populations, and state criteria and collect the evidence behind the main claims. For a short audit path, the three-seed main grid is in Table 9; disagreement supervision and its fully enumerated dispersion comparisons are in Table 11 and App. C; timing is in Table 14; and failed or scope-limiting controls are in Apps. E.1, F.2, F.4, F.6, and F.7. Generator tables use 1500 pre-training documents per cell; other sample sizes are stated locally. We distinguish construction invariants from trained-model measurements.

A.1 Invariants and failure modes

bindMax: the largest number of distinct documents in which any single (slot, value) triple appears. Bounded by 4 throughout, consistent with chance collision at $200 \times 8 \times 512$ possible triples and $75{,}000$ statements sampled, so bindings are not memorizable across documents. Note that counting raw occurrences rather than documents would return $\geq R_{\text{old}}$ by construction, since the queried slot holds R_{old} copies of one value within a single document.

r_{len}: Pearson correlation between realized ΔD and document token count, $|r| \leq 0.043$ in every cell. Without this the ΔD axis would confound distance with length. The length distribution is $U[45, 55]$ in every cell rather than being adapted per cell: its lower bound is set by the most demanding cell (App. D.3) and then used everywhere, since a cell-dependent length distribution would itself be a covariate.

ansLast: documents in which the answer value equals the last value token before the query, which would admit a trivial copy rule. Zero in all 25 cells, by construction rather than by chance — values are drawn without replacement within a document.

updDens: density of update statements in the final decile of the pre-query region, divided by the global density, computed over filler slots only. Within $[0.946, 1.102]$ throughout. The queried slot's own rebinding is excluded because it sits at a fixed offset $m - \Delta D - 1$ and would land in the final decile at small ΔD, reporting the construction as a violation.

adj, ant: same-slot adjacency rate and mean distance between consecutive occurrences of a slot, reported separately for the queried slot (q) and pooled filler slots (f). Thus ant_q averages all consecutive queried-slot gaps. It is related to but not identical to the *q-gap* used in App. F.1: q-gap is only $p_{\text{final}} - \max\{i < p_{\text{final}} : s_i \text{ has the queried slot}\}$, in statement positions. The two tables also use different document samples. Adjacency is matched within 0.042. Antecedent distance is not: ant_q exceeds ant_f by 2–22%, largest at $(R_{\text{old}}, \Delta D) = (3, 2)$ where $13.32/10.92 = 1.22$. Both quantities are treated as measured positional covariates rather than interchangeable statistics.

Table 1: Generator invariants, 1500 documents per cell. Columns are defined in the text below: *bindMax* cross-document binding reuse, r_{len} correlation between realized ΔD and token count, *ansLast* trivial-copy documents, *updDens* pre-query update density in the final decile, *adj*/*ant* same-slot adjacency rate and antecedent distance for the queried (q) and filler (f) slots. The fifth invariant holds only partially: ant_q exceeds ant_f by 2–22%.

R_{old}	ΔD	bindMax	r_{len}	ansLast	updDens	adj_q	adj_f	ant_q	ant_f
3	2	3	−0.024	0	1.029	0.056	0.078	13.32	10.92
3	3	3	−0.023	0	1.020	0.058	0.077	13.21	10.95
3	5	3	+0.028	0	1.022	0.056	0.074	13.09	11.00
3	8	3	−0.035	0	1.021	0.060	0.077	12.55	10.92
3	16	3	−0.002	0	0.984	0.074	0.076	11.49	11.03
5	2	3	−0.003	0	1.074	0.087	0.104	9.57	8.29
5	3	3	−0.003	0	1.071	0.088	0.105	9.53	8.31
5	5	3	+0.009	0	1.013	0.090	0.107	9.30	8.34
5	8	3	+0.007	0	1.015	0.091	0.106	9.22	8.35
5	16	4	−0.013	0	1.008	0.092	0.106	8.54	8.36
8	2	3	+0.031	0	1.041	0.135	0.152	6.93	6.30
8	3	3	+0.031	0	1.031	0.135	0.151	6.91	6.32
8	5	3	−0.039	0	0.994	0.135	0.150	6.88	6.34
8	8	3	−0.003	0	0.981	0.132	0.150	6.81	6.29
8	16	3	−0.008	0	1.006	0.138	0.155	6.52	6.22
12	2	3	+0.007	0	1.102	0.180	0.197	5.35	5.00
12	3	3	+0.007	0	1.101	0.181	0.196	5.33	5.01
12	5	3	−0.036	0	1.083	0.183	0.198	5.36	5.00
12	8	3	+0.006	0	0.946	0.189	0.200	5.28	4.92
12	16	3	−0.043	0	1.056	0.188	0.206	5.16	4.84
16	2	3	+0.022	0	1.065	0.213	0.236	4.71	4.22
16	3	3	+0.022	0	1.059	0.213	0.237	4.71	4.23
16	5	2	+0.001	0	1.097	0.212	0.249	4.74	4.16
16	8	3	−0.018	0	1.052	0.213	0.248	4.69	4.16
16	16	2	−0.024	0	1.048	0.218	0.260	4.54	4.02

Why the label is deterministic. Randomly choosing between the old and new values would define a calibrated mixture target rather than identify an answer rule. We therefore keep the ground truth deterministic and vary rule distinguishability only in the non-aliased control.

A.2 Covariates

Columns of Table 2 use 1500 documents per cell. *slots* is the number of distinct (entity, attribute) pairs. R_{real} is the number of surviving v_{old} copies in the queried slot. *domain* is the empirical fraction with at least two such copies, before edit-invariant re-checking. Decimal entries are rounded to the shown precision; a displayed domain of 1.00 does not assert zero excluded documents unless an integer count is supplied. *support* and *posCeil* are the realized ΔD interval and $1/|\text{supp}(\Delta D)|$.

tail-0 is computed directly as the fraction satisfying $n_{\text{tail updates}} = 0$, where this count covers update statements strictly after the queried rebinding and before the query. A constant-density approximation would average $(1-\rho)^{\delta}$ over the realized values of δ, rather than substitute the nominal band centre, but that approximation is not used to generate Table 2. Tail-0 is a corpus property, not the accuracy measured on the same stratum (§5.1, App. A.5).

Where the alias holds through the tie-break. On a document in which only one copy of v_{old} survives placement, the queried slot holds one copy of each value, RARITY's counts tie, and Equation 1 holds through the tie-break toward the more recent occurrence rather than through a multiplicity gap. The training answer labels then do not distinguish RECENCY from RARITY as defined here; counts alone do not select a unique answer on those documents, so an additional tie-breaking rule is required. Before rounding, the empirical tie fraction is $1 - \text{domain}$. It is approximately 0.34 at $R_{\text{old}} = 3$, $\Delta D = 16$ and at most about 0.02 in the displayed high-redundancy rows, so it co-

varies with redundancy and is not separable from that axis here. The intervention readout itself is computed only where at least two old copies survive, so no reported Δ comes from a tied-count document. Cross-seed spread also remains when the printed tie fraction is near zero: among the nine cells whose domain rounds to 1.00, four span more than 0.3 ($(12, 3)$ at 0.428, $(16, 5)$ at 0.411, $(16, 8)$ at 0.327 and $(8, 2)$ at 0.313; Table 9). Thus rounded full coverage does not explain the non-replication, but it is not asserted to be exact full-domain coverage.

Table 2: Covariates that shift with the axes, 1500 documents per cell. Columns are defined in the text above.

R_{old}	ΔD	support	posCeil	slots	R_{real}	domain	tail-0
3	2	$[1, 3]$	0.333	27.6	2.52	0.92	0.66
3	3	$[2, 4]$	0.333	27.6	2.47	0.91	0.52
3	5	$[2, 8]$	0.143	27.6	2.39	0.89	0.37
3	8	$[4, 12]$	0.111	27.6	2.23	0.81	0.19
3	16	$[8, 24]$	0.059	27.9	1.91	0.66	0.04
5	2	$[1, 3]$	0.333	20.2	3.99	0.99	0.71
5	3	$[2, 4]$	0.333	20.3	3.89	0.99	0.61
5	5	$[2, 8]$	0.143	20.4	3.72	0.97	0.48
5	8	$[4, 12]$	0.111	20.5	3.48	0.96	0.31
5	16	$[8, 24]$	0.059	20.6	2.87	0.87	0.10
8	2	$[1, 3]$	0.333	14.6	5.82	1.00	0.82
8	3	$[2, 4]$	0.333	14.6	5.71	1.00	0.73
8	5	$[2, 8]$	0.143	14.8	5.44	0.99	0.62
8	8	$[4, 12]$	0.111	14.8	5.08	0.99	0.46
8	16	$[8, 24]$	0.059	15.0	4.06	0.95	0.19
12	2	$[1, 3]$	0.333	10.8	7.84	1.00	0.86
12	3	$[2, 4]$	0.333	10.8	7.69	1.00	0.79
12	5	$[2, 8]$	0.143	11.0	7.25	1.00	0.70
12	8	$[4, 12]$	0.111	11.0	6.79	0.99	0.59
12	16	$[8, 24]$	0.059	11.1	5.42	0.98	0.34
16	2	$[1, 3]$	0.333	8.7	9.05	1.00	0.90
16	3	$[2, 4]$	0.333	8.7	8.87	1.00	0.85
16	5	$[2, 8]$	0.143	8.8	8.37	1.00	0.75
16	8	$[4, 12]$	0.111	8.9	7.81	1.00	0.65
16	16	$[8, 24]$	0.059	9.1	6.34	0.98	0.41

A.3 COLLISION RATES

Table 3 reports how often each candidate rule selects the ground truth. A rule whose answer coincides with the target on every document cannot be separated from it by input–output observations on that support, because no document exists on which their answers disagree. Internal measurements may still differ, but behavior alone does not identify how such differences should be interpreted. This motivates the intervention of §4.

RECENCY and RARITY both collide at 1.000 in every cell. This is Equation 1 restated empirically: the identity holds document by document, so the target labels alone cannot distinguish the two rules on this training support.

FREQUENCY and PRIMACY collide at 0.001–0.309 and 0.000, respectively. Primacy continues to select v_{old}. Frequency also selects it when more than one old copy survives, but when the counts tie its recency tie-break returns v_{new}. Its empirical collision rate therefore approximately tracks $1 - \text{domain}$. The two tables use independent document samples and rounded outputs, so their cell values are not claimed to coincide exactly.

GLOBAL LAST UPDATE varies from 0.042 to 0.900 and is close to the directly counted tail-0 fraction on a separate sample. We treat this as an empirical consistency check, not as an exact cellwise equality generated from the closed-form approximation.

Table 3: Fraction of documents on which each candidate rule selects the ground truth, 1500 documents per cell. REC recency, RAR rarity, FRQ frequency, PRI primacy, GLU global last update, LTK last value token before the query, POS best fixed offset from the end.

R_{old}	ΔD	REC	RAR	FRQ	PRI	GLU	LTK	POS
3	2	1.000	1.000	0.085	0.000	0.651	0.000	0.341
3	3	1.000	1.000	0.099	0.000	0.528	0.000	0.345
3	5	1.000	1.000	0.120	0.000	0.359	0.000	0.163
3	8	1.000	1.000	0.163	0.000	0.183	0.000	0.127
3	16	1.000	1.000	0.309	0.000	0.042	0.000	0.070
5	2	1.000	1.000	0.024	0.000	0.722	0.000	0.339
5	3	1.000	1.000	0.028	0.000	0.612	0.000	0.343
5	5	1.000	1.000	0.026	0.000	0.487	0.000	0.163
5	8	1.000	1.000	0.043	0.000	0.303	0.000	0.122
5	16	1.000	1.000	0.130	0.000	0.101	0.000	0.071
8	2	1.000	1.000	0.005	0.000	0.831	0.000	0.353
8	3	1.000	1.000	0.005	0.000	0.741	0.000	0.353
8	5	1.000	1.000	0.009	0.000	0.626	0.000	0.155
8	8	1.000	1.000	0.013	0.000	0.447	0.000	0.119
8	16	1.000	1.000	0.049	0.000	0.195	0.000	0.069
12	2	1.000	1.000	0.001	0.000	0.884	0.000	0.351
12	3	1.000	1.000	0.002	0.000	0.823	0.000	0.351
12	5	1.000	1.000	0.006	0.000	0.720	0.000	0.153
12	8	1.000	1.000	0.005	0.000	0.567	0.000	0.122
12	16	1.000	1.000	0.021	0.000	0.333	0.000	0.070
16	2	1.000	1.000	0.002	0.000	0.900	0.000	0.337
16	3	1.000	1.000	0.002	0.000	0.861	0.000	0.337
16	5	1.000	1.000	0.002	0.000	0.755	0.000	0.161
16	8	1.000	1.000	0.003	0.000	0.649	0.000	0.122
16	16	1.000	1.000	0.025	0.000	0.442	0.000	0.074

LAST VALUE TOKEN collides at 0.000 everywhere. Values are drawn without replacement within a document, so the last value token before the query is never the answer. This removes a trivial copy baseline by construction rather than by chance, and is one of the five invariants of App. A.1.

POSITIONAL collides at 0.069–0.353, tracking $1/|\text{supp}(\Delta D)|$ across the four support sizes: 0.337–0.353 against 0.333 at $|\text{supp}| = 3$, 0.153–0.163 against 0.143 at 7, 0.119–0.127 against 0.111 at 9, and 0.069–0.074 against 0.059 at 17. The measured values sit 0.4–2.0 percentage points above the analytic bound because the empirical mode of a discrete uniform sample slightly exceeds $1/|\text{supp}|$. This is Equation 3 arrived at by counting rather than by derivation, and the agreement is a check on the derivation that involves no trained model.

A.4 THE EDIT AND ITS CHECKS

The multiplicity inversion of §4 defines the paper's primary intervention readout. This appendix reports what the edit preserves, what its controls measure, pipeline checks for numerical and sampling artifacts, and two measurements supporting the interpretation of Δ as a sensitivity rather than an implemented answer rule.

A.4.1 THE INVARIANTS AND THE CONTROLS

The five edit invariants. Token count, answer position, ground-truth answer, statement count and q-gap are checked document by document. Here q-gap is the distance from the final queried rebinding to the latest earlier queried-slot statement. The primary multiplicity inversion changes values but not slot identities or statement positions, so it preserves q-gap automatically; the shared check remains necessary for other edit families that move statements. The primary inversion additionally requires at least two surviving copies of v_{old}, which is why its domain can be below 1.00.

Idealized predictors. A predictor assigning logit κ to the target rule's prediction returns $\Delta = 2\kappa$; one assigning it to the ground truth returns $\Delta = 0$ identically. At $\kappa = 8$ we measure $+16.0$ and $|\Delta| < 10^{-9}$ in every cell. This is a check on the construction rather than a result: the values follow from Equation 2, and measuring them confirms the edit has no side effect on the contrast pair.

Pre-gate checkpoints. Before the retrieval criteria are met, Δ is $+0.000$ with 0.49 of documents positive at step 2000 and -0.002 with 0.40 positive at step 4000 ($R_{\text{old}} = 3$, $\Delta D = 2$). These measurements check that the edit does not automatically induce a large displacement at these checkpoints. They do not establish that the trained network contains no relevant internal computation.

The filler-slot control. The same form of multiplicity inversion is applied to a non-queried slot, with the queried slot untouched. The number of rewritten tokens is not matched document by document to the targeted edit. Its median readout is -0.009 to $+0.015$ across the grid — below 1% of the targeted readout wherever that readout exceeds 1 nat. In the five $R_{\text{old}} = 3$ cells of seed 0, where the median is -0.011 to -0.087 nats, the control is opposite in sign in all five and at most 5% of the median, so it is not producing a spurious effect in the same direction there. Two cells elsewhere in the grid do not admit that argument: at $R_{\text{old}} = 8$, $\Delta D = 2$ in seed 1 the control is -0.009 against a mass-restricted median of -0.006, and at $8, 16$ in seed 2 it is -0.005 against -0.012, the signs agreeing in both. Both are among the four runs whose mass-restricted median is negative at $R_{\text{old}} \geq 5$. Control overlap weakens the interpretation of these two negative medians, but does not establish that all four are edit noise. The most extreme case is outside the grid: -0.020 against a median of -0.055 with the signs agreeing, at $R_{\text{old}} = 8$, $\Delta D = 2$ in the p_{update} arm, whose readout we therefore report as uninterpretable (App. F.2). Per-cell control values are in Table 8.

Contrast-pair mass as a quality diagnostic. The inversion makes the answer value occur more than once before the query, which never happens during training. For the main-grid and disagreement-supervision probes, probabilities are normalized over the value vocabulary, so contrast-pair mass is conditional on that vocabulary. We therefore report the probability mass on $\{v^\star, v_{\text{truth}}\}$ and recompute conditional summaries on documents with mass at least 0.5. The primary sign fraction remains defined over all documents with edit domain; the mass-restricted results are sensitivity analyses on a different evaluated population, not a validity label applied retroactively to the primary statistic. Mean mass is 0.59–1.00 across the analyzed grid and falls as the displacement grows (§5.4). In one cell the restriction changes the median by 1% at 4000 steps and 16% at 32000 because the largest responses become concentrated in a subset of documents (App. D.7.3).

The temporal invariant constrains the edit. Update flags are non-decreasing within a slot and same-valued occurrences are contiguous, so a document in which the second of three v_{old} copies is replaced while the first and third remain is outside the language. This is why the multiplicity inversion of §4 retains the earliest copy rather than an arbitrary one: retaining a middle copy would interleave generations.

The choice also determines which positional response is controlled. Because v_{old} occupies the earliest occurrence of the queried slot in both the base and edited documents, a fixed rule that always selects that occurrence contributes equally to both terms of Equation 2 and cancels. This addresses the specific failure mode in which attention degrades to the slot's earliest occurrence under off-distribution input. It does not make every position-sensitive account cancel, because later occurrences are rewritten by the edit.

Why coverage falls with ΔD. The edit requires at least two copies of v_{old} to survive into the document. Two quantities bound it, on two different samples. On the 1500-document generator sample, the fraction of documents with at least two surviving copies is 0.66–1.00 across cells, which is an upper bound on coverage because the five edit invariants are not yet re-checked (Table 2). On the 400 probe documents of seed 0, the yield after re-checking is 0.69–1.00 (the yield column of Table 8). The two ranges are not nested, since they are measured on different document samples; within either sample, re-checking can only remove documents. At the lowest cell, $R_{\text{old}} = 3$, $\Delta D = 16$, the three seeds yield 274, 269 and 258 of 400 (App. A.4.4), so 0.645 is the lowest yield among the 75 main-grid terminal probes. Coverage falls with ΔD because copies placed outside the document canvas are discarded, so cells at low R_{old} and high ΔD condition on a subset with higher realized redundancy. §5.3 tests this with a within-cell split.

A.4.2 PIPELINE CHECKS

Two checks bear on whether the numbers in this paper are properties of the models or of the measurement path.

A rerun matching recorded losses. Retraining a cell with the same seed under a later version of the training script, which differs only in the density of probe evaluations, reproduces the training loss exactly: 88 logged steps agree to all recorded digits, including steps after the shortcut transition. In this audited rerun, changing probe-evaluation density did not alter any recorded training loss. This supports replayability for the tested configuration and shows no observed perturbation from probe scheduling; it is not an exhaustive exclusion of numerical nondeterminism across all runs, kernels or environments.

Two probe densities, one scale. The probe evaluated during training uses 200 documents and the terminal probe 400. Across the 50 runs of seeds 0 and 1 the two agree to 0.012 on average and 0.060 at worst in sign fraction, so trajectories (§6.3, App. D.7.3) and terminal values (Table 8) are reported on the same scale. Both are computed over all documents with edit domain and neither is gated on contrast-pair mass.

A.4.3 WHAT THE SIGN RULES OUT, AND WHERE

The multiplicity inversion raises the occurrence count of the correct answer from one to the realized old-copy count $R \geq 2$ and lowers the superseded value to one. An idealized count-only response, such as a context-frequency prior that monotonically favors increasing count, predicts a negative readout; position-dependent match aggregation need not predict that sign. Pooled by row the mass-restricted median is positive at every $R_{\text{old}} \geq 5$ in all three seeds. Four individual runs have negative mass-restricted medians, all close to zero; their small magnitudes limit conclusions about a negative effect. At the high end the direction of the shift is thus inconsistent with that count-only response and consistent with suppression — a statement about how the edit displaces the log-odds, not about which value the model emits (§5.4).

This opposes the majority-rule preference reported under knowledge conflict (Jin et al., 2024a; Yan et al., 2024), though not tightly: those settings repeat evidence across documents rather than within a slot. What the sign constrains is a family of response hypotheses rather than our two named answer rules, and it does so at the level of the row.

A.4.4 THE READOUT IS A SENSITIVITY, AND THE BEHAVIORAL DISCRIMINATOR AGREES

Two measurements support §5.4's claim that Δ records log-odds displacement rather than identifying answer policy. Both are pure inference on the 75 grid checkpoints and involve no retraining.

Behavioral choice on generated disagreement documents. The relabeling generator of §5.6 produces documents whose queried slot holds $\{v_{\text{old}} : 1,\ v_{\text{new}} : R\}$, so RECENCY and RARITY select different values there and the model's $\arg\max$ over the value vocabulary provides a behavioral discriminator, with no contrast pair, mass restriction or log-odds arithmetic. We generate 800 such documents per cell at $p_{\text{break}} = 1$ and read the $\arg\max$ at the answer position. Every document has two distinct candidates in every run, so the discriminator has full domain. The rate of rarity answers is 0.000 in 70 of the 75 runs and at most 0.003 in the remaining five, against 1.000 in the arm that forces the objective toward RARITY (Table 11). The recency rate is 0.749 to 1.000; the residual is documents answering neither candidate, which reaches 0.250 in the least stable runs. All five runs with a non-zero rarity rate are among those with the highest residual, so the occasional rarity answer tracks off-distribution instability rather than a stable rarity answer policy.

These inputs are off-distribution for grid models, which train at $p_{\text{break}} = 0$; they are off-distribution differently from the edit, being produced by the generator rather than by rewriting a base document, and that independence is the point. The measurement shows that no grid run expresses a stable RARITY answer policy on these generated disagreement documents, while the forced arm shows that such a policy is learnable in this architecture on this language. The aliased objective does not produce that observed behavior in these runs.

Table 4: Complete base-to-edited argmax transitions across the 75 main-grid runs. *old* is the superseded contrast value, *new* is the preserved ground-truth/current value, and *other* is any third value. The first panel contains all 28,563 valid edit pairs; the second contains 26,916 pairs with contrast-pair mass at least 0.5. Cells give pooled document counts (percentage within panel). Pooling is descriptive because the training run, not the document, is the replication unit.

subset	base \ edited	old	new	other
all	old	0 (0.000%)	0 (0.000%)	0 (0.000%)
	new	8 (0.028%)	28,185 (98.677%)	370 (1.295%)
	other	0 (0.000%)	0 (0.000%)	0 (0.000%)
mass ≥ 0.5	old	0 (0.000%)	0 (0.000%)	0 (0.000%)
	new	4 (0.015%)	26,912 (99.985%)	0 (0.000%)
	other	0 (0.000%)	0 (0.000%)	0 (0.000%)

Baseline margin and $\arg\max$ movement on the edited documents. The behavioral result above is measured on generated documents, while Δ is measured on edited ones, so we also measure the margin directly on the population where Δ is computed. Using the same document stream, the same non-inverted filter and the same seed as §4.2, we record $\log p(v^\star) - \log p(v_{\text{truth}})$ on the base and edited documents and the $\arg\max$ on both. Yield is 258 to 400 documents per run, tracking the edit domain of Table 2 — 274, 269 and 258 of 400 at $R_{\text{old}} = 3$, $\Delta D = 16$ against a measured domain of 0.66 there.

Across the 75 runs, the run-level baseline-margin summaries range from -14.18 to -11.34 nats, and the unrestricted median displacements range from -0.09 to $+5.33$ nats. Every base-document $\arg\max$ over the value vocabulary is v_{truth}. Table 4 enumerates all base-to-edited transitions rather than only transitions to $v^\star$.

In the unrestricted population, 28,185 of 28,563 edited argmaxes (98.68%) remain on v_{truth}, eight (0.03%) move to $v^\star$, and 370 (1.30%) move to a third value; the pooled value-vocabulary argmax change rate is 1.32%. The eight contrast-value transitions occur in eight of the 75 runs, at most one per run; the largest run-level rate is $1/334 = 0.30\%$. Third-value transitions are more concentrated across runs, reaching $69/400 = 17.25\%$ at $R_{\text{old}} = 12$, $\Delta D = 2$, seed 1.

With contrast-pair mass at least 0.5, 26,912 of 26,916 argmaxes (99.99%) remain on v_{truth}, four (0.01%) move to $v^\star$, and none move to a third value. Thus all 370 third-value transitions and four of the eight $v^\star$ transitions lie below the mass threshold. The restricted panel is a sensitivity analysis on a different evaluated population, not a validity filter applied retroactively to the primary statistic.

The eight runs containing a $v^\star$ transition have unrestricted median displacements from $+0.087$ to $+5.269$ nats, while the three largest run-level medians contain none. The complete matrix therefore makes the scope precise: substantial log-odds displacement rarely makes the contrast value the prediction, and high-mass pairs almost always retain the preserved truth, but the value-vocabulary argmax is not invariant over all unrestricted pairs.

Magnitude-threshold sensitivity. To test whether sign differences are confined to a neighborhood of zero, we fixed $\tau \in \{0, 0.01, 0.1, 0.5, 1\}$ and computed, separately in every run, $S^+(\tau) = \Pr(\Delta > \tau)$, $S^0(\tau) = \Pr(|\Delta| \leq \tau)$ and $S^-(\tau) = \Pr(\Delta < -\tau)$. These thresholds are descriptive magnitude scales rather than universal practical-significance cutoffs. The primary population contains all valid edit-domain documents; mass at least 0.5 defines a sensitivity-analysis population. Training runs remain the replication units.

At $\tau = 0.5$, the two low runs of the $(3, 8)$ cell have $(S^0, S^-) = (0.982, 0.018)$ and $(0.985, 0.015)$, whereas the high run has $(0.239, 0.000)$. The corresponding pairs in the $(3, 5)$ cell are $(0.887, 0.113)$, $(0.262, 0.000)$ and $(0.997, 0.003)$. The mass restriction leaves the grid-wide cell counts in Table 5 unchanged at $\tau = 0.1, 0.5, 1$ and preserves substantial positive tails in the two high runs. Thus thresholding reveals persistent magnitude heterogeneity but an asymmetric pattern: strong positive tails are contrasted mostly with near-zero or weaker negative displacements. Increasing ranges with τ do not mean that effects grow with the threshold; low-magnitude runs move into

Table 5: Magnitude-threshold sensitivity across the 25 cells. *max range* is the largest within-cell cross-seed range of $S^+(\tau)$; *cells* counts ranges above 0.3.

	all edit-domain		mass ≥ 0.5	
τ	max range (cell)	cells	max range (cell)	cells
0	0.879 $(3, 8)$	13	0.877 $(3, 8)$	13
0.01	0.928 $(3, 8)$	13	0.926 $(3, 8)$	13
0.1	0.948 $(3, 8)$	14	0.943 $(3, 8)$	14
0.5	0.988 $(12, 3)$	16	0.986 $(12, 3)$	16
1	0.993 $(12, 3)$	16	0.990 $(12, 3)$	16

Table 6: Unrestricted displacement distributions in the two cells with the largest original sign-fraction ranges. The last columns report $S^+(\tau) = \Pr(\Delta > \tau)$.

cell	seed	n	$q_{.05}$	median	$q_{.95}$	$S^+(.1)$	$S^+(.5)$	$S^+(1)$
$(3, 8)$	0	328	-0.321	-0.024	$+0.007$	0.000	0.000	0.000
	1	331	-0.206	-0.001	$+0.044$	0.012	0.000	0.000
	2	343	$+0.096$	$+1.083$	$+8.117$	0.948	0.761	0.528
$(3, 5)$	0	364	-0.802	-0.087	$+0.025$	0.005	0.000	0.000
	1	351	$+0.048$	$+1.964$	$+12.350$	0.915	0.738	0.624
	2	348	-0.122	-0.013	$+0.057$	0.011	0.000	0.000

S^0 while strong positive runs remain in S^+. Full (S^+, S^0, S^-) values for all 75 runs are supplied in machine-readable form.

These unrestricted medians should be compared with ungated summaries on a matched probe population. Table 8 reports ungated seed-0 medians; Table 9 reports sign fractions and contains no median column. Mass-restricted and unrestricted medians are distinct conditional summaries, without an assumed ordering between them. The seed-0 sign pattern nevertheless agrees across code paths: all five $R_{\text{old}} = 3$ cells are negative in both measurements.

A.5 Analysis populations and state gates

A document enters the primary sign-fraction population when the edit has a domain and all five edit invariants pass. This primary statistic uses every such document and is not gated on contrast-pair mass. Restricting to documents with contrast-pair mass at least 0.5 defines a separate sensitivity-analysis population. Within-run exact binomial tests quantify document-sampling uncertainty only. Cross-run standard deviations are computed from terminal run-level sign fractions with an $n - 1$ denominator. Terminal cross-run comparisons additionally apply the state criteria below.

For terminal analyses, a run is classified RETRIEVAL if the copy diagnostic exceeds 0.95 and in-distribution accuracy exceeds 0.99 at the final checkpoint; POSITIONAL if the copy diagnostic stays below 0.5 and accuracy is at least half the positional ceiling (Eq. 5), with the attained fraction reported alongside the label; and NEITHER otherwise. The three states are disjoint on the runs that enter the main results: across the analyzed grid and the $R_{\text{old}} = 2$ column, the copy diagnostic at each run's final checkpoint is either at most 0.046 or at least 0.999, so any threshold in $[0.05, 0.95]$ returns the same labels. The $\gamma = 1$ runs of App. E.1 fall inside that gap (0.132 and 0.244), so their labels depend on where the threshold sits; they are outside the analyzed grid, and the label we report for them is the attained fraction of the positional ceiling rather than a threshold-insensitive state.

A note on the escape threshold. We use 0.95 on the copy diagnostic; the transitions are sharp enough that the choice does not matter within the analyzed grid — the diagnostic moves from below 0.05 to above 0.99 between consecutive evaluations in every cell, so any threshold in $[0.05, 0.99]$ gives the same escape step. It does matter in one excluded cell: at $R_{\text{old}} = 2$, $\Delta D = 16$ the diagnostic reads 0.839 at step 4000 and 0.998 at 5000, so a 0.5 threshold would report escape at 4000 and our 0.95 threshold reports 5000. We use 5000 in App. D.6 and note the ambiguity there.

All 75 runs across the three seeds terminate in the RETRIEVAL state. Accuracy is 1.000 to four decimals in every cell except $R_{\text{old}} = 3, \Delta D = 16$ of seed 0 (0.9998) and $R_{\text{old}} = 8, \Delta D = 16$ of

seed 2 (0.99975), which is why the main text states the bound as ≥ 0.999; accuracy restricted to documents with no update after the queried rebinding is 1.000 in every cell; and the copy diagnostic is 1.000 in every cell. Escape steps on the 1000-step gate are the `gate` column of Table 14, which reports both seeds and both localizations.

The gate is not doing selection work within the analyzed grid. All 75 runs pass, so the classification excludes nothing here. It excludes four of the five $R_{\text{old}} = 2$ runs (App. D.6) and timestamps the readouts: because escape is as late as step 10000 at $R_{\text{old}} = 3$ in these runs, a probe applied at step 4000 would have returned a directional intervention readout before the retrieval criteria were met (§6.3).

A.6 Disagreement-supervision construction

Four generator parameters are added. A fraction p_{break} of documents is *inverted*: the queried slot receives k copies of v_{old} and r copies of v_{new} with $r > k$, against R_{old} and 1 in an ordinary document. We use $k = 1$ and $r = R_{\text{old}}$, which keeps the inverted slot at $R_{\text{old}} + 1$ statements, so window placement, ΔD, token count and the expected statements per slot are unchanged; the inversion is a strict exchange of multiplicities. On inverted documents RARITY selects v_{old} while RECENCY selects v_{new}, so Equation 1 fails there by construction. A fifth parameter selects which of the two supplies the label.

Filler slots invert at the same rate. If only the queried slot could show a repeated latest value, that property would identify it without reading the query. Every filler slot with an update is therefore inverted with the same probability p_{break}. The matched quantity is the fraction of *update-bearing* slots that are inverted: a filler slot with no update has one generation, and inversion is a no-op on it, so it does not enter the denominator.

The two arms are token-identical. The label rule is applied to an already-sampled value and consumes no randomness, and the inversion flag is drawn before the slot is built. At fixed seed the two arms therefore produce the same documents in the same order, differing only in the answer token of inverted documents. Since the probe reads only the prefix up to the answer position, the probe input distribution is identical in the two arms, and the readout sets coincide exactly: at $p_{\text{break}} = 0.30$ the edit domain contains 223, 218 and 238 documents in seeds 0, 1 and 2 under *both* labeling rules. Any difference in readout comes from the objective.

Backward compatibility. At $p_{\text{break}} = 0$ the inversion draw is short-circuited and consumes no randomness, so the generator is bit-identical to the one used for the main grid, and $p_{\text{break}} = 0$ in Table 11 is the published cell rather than a rerun. The one change this arm required inside the probe — taking v_{old} from the generation history rather than from the second-to-last queried statement, which differ only when multiplicities are inverted — is a no-op at $p_{\text{break}} = 0$: rerunning the readout on the published seed-0 checkpoints reproduces all 25 rows of Table 8 bit-identically.

A.7 Pre-training structural checks

All six checks pass at $R_{\text{old}} = 3$ in both cells (Table 7). Two deserve comment.

The *degenerate* row counts inverted documents on which the inversion failed to break the alias. Window overflow can discard copies of v_{new}; if only the one pinned at the rebinding survives, post-inversion counts are $\{v_{\text{old}} : 1, v_{\text{new}} : 1\}$ and the tie-break returns v_{new} again. Under RARITY-labeling, such a document receives a value selected by neither of the two labeling rules, creating label noise rather than a weaker signal. The retry loop that keeps every distinct value alive bounds this at zero in both cells.

The *alias* row is 0.007–0.009 below its nominal value, which is the realized inversion rate running slightly above p_{break} in this sample: `collide` draws its own document stream and measures 0.107–0.109 against a nominal 0.100, within 1.2 standard errors at $n = 1500$ (0.0077). The independent stream used by the covariate and pairing checks reads 0.086–0.098 on the same nominal value. Both are consistent with sampling variation in the per-document draw.

Table 7: Structural checks run on CPU before training, 1500 documents per cell for the first four rows, 3000 for *pairing* and 1200 for *degenerate*. Thresholds were fixed before training. *alias* is the fraction of documents on which RARITY and RECENCY still agree, which should equal $1 - p_{\text{break}}$. *inv-6* is the difference in inversion rate between the queried and filler sides. *ant* is the ratio of antecedent distances inside inverted documents, the break-internal version of the fifth generator invariant (App. A.1). *clamp* is the fraction of inverted documents falling back to the clamped placement path.

check	quantity	$\Delta D = 5$	$\Delta D = 8$	threshold
alias	RARITY = RECENCY	0.891	0.893	0.900 ± 0.03
inv-6	|queried − filler| inversion rate	0.003	0.005	< 0.05
ant	$\text{ant}_q/\text{ant}_f$ within inverted docs	1.045	0.950	$(0.80, 1.25)$
clamp	clamped placement, inverted docs	0.000	0.000	< 0.20
pairing	prefix identical across the two arms	3000/3000	3000/3000	exact
degenerate	inverted docs where the alias survives	0.000	0.000	< 0.02

B CORE MAIN-GRID RESULTS

Table 8 reports every summary we computed for each cell of seed 0, including the ones §4.2 argues against using. Seeds 1 and 2 enter through Table 9, which is the comparison the paper's central claim rests on; its per-cell medians and controls are quoted in App. A.4.3 and §5.4 where they bear on an argument. Medians and interquartile ranges are over the documents on which the edit has a domain; the sign fraction is the fraction of those documents on which $\Delta > 0$ (the rarity-directed sign), with p from a two-sided exact binomial test against 0.5. *ctrl* is the median displacement under the control edit on a non-queried slot. *mass* is the mean probability assigned to $\{v^\star, v_{\text{truth}}\}$ under the edit and *mOK* the fraction of documents meeting the 0.5 mass threshold.

Table 8: Per-cell readouts, main grid at 16000 steps, seed 0, 400 held-out documents per cell. n is the number with edit domain and yield is $n/400$; frac+ is the sign fraction. Median, IQR endpoints, mean, trim and ctrl are in nats, with trim the 5%-trimmed mean.

R_{old}	ΔD	n	yield	median	IQR low	IQR high	frac+	p	mean	trim	ctrl	mass	mOK
3	2	364	0.91	−0.011	−0.043	+0.007	0.35	3e-08	−0.109	−0.034	+0.000	1.00	1.00
3	3	362	0.91	−0.046	−0.213	+0.055	0.36	3e-07	−0.170	−0.115	+0.001	1.00	1.00
3	5	364	0.91	−0.087	−0.219	−0.030	0.13	4e-51	−0.218	−0.150	+0.004	1.00	1.00
3	8	328	0.82	−0.024	−0.075	−0.009	0.10	1e-54	−0.073	−0.055	+0.000	1.00	1.00
3	16	274	0.69	−0.023	−0.104	+0.038	0.41	2e-03	−0.094	−0.058	+0.001	1.00	1.00
5	2	397	0.99	+0.066	+0.008	+0.137	0.78	2e-30	+0.087	+0.076	−0.006	1.00	1.00
5	3	393	0.98	+0.010	−0.013	+0.037	0.61	8e-06	+0.029	+0.014	+0.002	1.00	1.00
5	5	395	0.99	+0.003	−0.033	+0.049	0.52	0.48	+0.176	+0.014	−0.001	0.98	0.98
5	8	380	0.95	+0.147	+0.027	+0.333	0.80	2e-33	+0.293	+0.199	−0.000	0.99	0.99
5	16	343	0.86	+0.090	+0.012	+0.255	0.80	2e-29	+0.298	+0.154	+0.004	0.98	0.99
8	2	397	0.99	+0.037	−0.015	+0.159	0.66	9e-11	+0.199	+0.107	−0.003	1.00	1.00
8	3	397	0.99	+0.604	+0.350	+1.080	0.99	6e-113	+0.888	+0.776	−0.002	0.99	1.00
8	5	397	0.99	+0.385	+0.156	+0.872	0.93	4e-78	+0.749	+0.552	+0.002	0.98	0.98
8	8	396	0.99	+0.338	+0.191	+0.626	0.97	3e-97	+0.547	+0.439	+0.004	0.99	0.99
8	16	389	0.97	+0.243	+0.106	+0.532	0.89	9e-62	+0.526	+0.371	+0.004	0.99	0.99
12	2	400	1.00	+3.523	+1.888	+6.091	1.00	8e-121	+4.258	+4.067	+0.001	0.80	0.82
12	3	399	1.00	+0.643	+0.393	+1.030	0.99	2e-111	+1.024	+0.791	+0.001	0.98	0.98
12	5	398	0.99	+5.329	+2.953	+7.951	1.00	1e-117	+5.527	+5.465	+0.014	0.66	0.66
12	8	398	0.99	+2.678	+1.567	+4.416	1.00	3e-120	+3.391	+3.178	+0.010	0.89	0.91
12	16	396	0.99	+1.867	+0.900	+3.367	0.98	2e-103	+2.468	+2.239	−0.001	0.91	0.93
16	2	400	1.00	+1.827	+1.095	+3.153	1.00	3e-118	+2.210	+2.075	+0.012	0.98	0.99
16	3	400	1.00	+3.625	+2.088	+5.923	1.00	8e-121	+4.278	+4.066	−0.004	0.82	0.83
16	5	399	1.00	+1.575	+0.935	+3.057	0.99	1e-115	+2.454	+2.209	+0.006	0.91	0.91
16	8	398	0.99	+2.216	+1.538	+3.240	1.00	1e-117	+2.580	+2.423	+0.004	0.95	0.98
16	16	395	0.99	+0.891	+0.364	+2.011	0.98	1e-101	+1.463	+1.250	+0.001	0.98	0.98

Table 9: Sign fraction per cell under three seeds at 16000 steps, 400 held-out documents each, ordered by the range across seeds. Thirteen of 25 cells span more than 0.3; the binomial standard error is at most 0.031 at the smallest edit domain ($n = 258$), and the largest range is 48 standard errors of the difference between the two proportions. All 75 runs of these three seeds reach in-distribution answer accuracy at least 0.999 and terminate in the RETRIEVAL state; this does not assert equality of loss, confidence or every held-in statistic. These ranges quantify within-cell cross-seed variation; they are not a universal threshold for between-cell comparisons. Adding the third seed widened the maximum from 0.845 at $(3, 5)$ to 0.879 at $(3, 8)$ and moved the cell carrying it from $(3, 5)$ to $(3, 8)$. Ranges are computed before rounding, so some rows differ by 0.001 from the difference of the printed seed values.

R_{old}	ΔD	s0	s1	s2	range	R_{old}	ΔD	s0	s1	s2	range
3	8	0.098	0.477	0.977	0.879	16	2	0.998	0.736	0.972	0.261
3	5	0.126	0.972	0.270	0.845	3	16	0.405	0.167	0.349	0.238
5	2	0.781	0.175	0.967	0.791	5	8	0.800	0.798	0.931	0.133
3	3	0.365	0.934	0.930	0.569	8	5	0.932	1.000	0.947	0.068
16	16	0.977	0.469	0.969	0.508	8	8	0.970	0.997	0.934	0.063
3	2	0.354	0.138	0.625	0.487	5	16	0.796	0.844	0.784	0.060
8	16	0.895	0.916	0.432	0.485	12	2	1.000	1.000	0.953	0.047
12	3	0.990	0.570	0.998	0.428	16	3	1.000	0.965	0.985	0.035
16	5	0.995	0.584	0.870	0.411	12	16	0.980	0.992	0.969	0.023
5	3	0.613	0.997	0.964	0.384	8	3	0.992	0.997	0.985	0.013
5	5	0.519	0.675	0.880	0.361	12	8	1.000	0.995	0.992	0.008
16	8	0.997	0.670	0.992	0.327	12	5	0.997	0.995	1.000	0.005
8	2	0.662	0.450	0.762	0.313						

Mass and the mass-restricted median. Contrast-pair mass is 1.00 with mOK 1.00 in all five $R_{\text{old}} = 3$ cells of seed 0 and 0.98–1.00 at $R_{\text{old}} = 5$ and 8. It falls below 0.85 in ten cells across the three seeds, lowest at $R_{\text{old}} = 12$, $\Delta D = 2$ in seed 1 (0.59), at $12, 5$ in seed 0 and $3, 5$ in seed 1 (both 0.66), and at $12, 5$ and $12, 3$ in seed 2 (0.69 and 0.70), so mass tracks the realized effect rather than either axis. The largest shift the mass restriction produces is at $R_{\text{old}} = 3$, $\Delta D = 5$ in seed 1, where the median falls from $+1.964$ to $+0.718$ at mass 0.66. The $R_{\text{old}} = 12$ row of seed 0 shows the consequence for the median: values run $+3.52$, $+0.64$, $+5.33$, $+2.68$, $+1.87$ in ΔD and the low cell is the one with the highest mass (0.98 against 0.66 and 0.80 in its neighbours), so the median ordering covaries with contrast-pair mass and does not by itself establish a mechanism gradient. The sign fraction is 0.98–1.00 across the row. The primary sign fraction uses all documents with edit domain; its separately recomputed mass-restricted version is reported below. The median column of Table 8 is ungated; mass-restricted medians are quoted in the main text where an argument depends on them (App. A.4.3, §5.4). The widest pair of §5.2 is $(R_{\text{old}}, \Delta D) = (3, 8)$ across seeds 0 and 2: medians -0.024 and $+1.083$ at exact binomial p of $1{\times}10^{-54}$ and $5{\times}10^{-88}$, a 45-fold gap in magnitude with opposite signs, both much smaller in magnitude than the run-level baseline-margin summaries, which range from -14.18 to -11.34 nats (App. A.4.4).

Mean against sign fraction. The mean and the sign fraction respond differently at two places in the grid. At $R_{\text{old}} = 5$, $\Delta D = 5$ the mean is $+0.176$ — larger than the mean at $\Delta D = 3$ in the same row ($+0.029$) — while the sign fraction is 0.52 against 0.61 and the binomial test returns $p = 0.48$. The mean is carried by four documents above $+8$ nats in a distribution otherwise concentrated within ± 0.5. At $R_{\text{old}} = 12$, $\Delta D = 3$ the mean is $+1.02$ against $+5.53$ at $\Delta D = 5$, a fivefold difference where the sign fraction differs by 0.01. Both are cells where the 5%-trimmed mean sits closer to the median than to the mean, which is the signature of a heavy tail rather than a shifted centre.

The sign fraction under the mass restriction. The primary sign fraction is computed over all documents with edit domain, whereas a mass restriction changes the evaluated population rather than certifying individual readouts as valid or invalid. We therefore recomputed the sign fraction on the mass-restricted population in all 75 runs as a sensitivity analysis. It moves by at most 0.026 — at $R_{\text{old}} = 3$, $\Delta D = 3$ in seed 1, where mOK is 0.72 — and by a median of 0.000, against a median within-run standard deviation of 0.074 over post-gate checkpoints (App. D.7.3). The cross-seed range of the widest cell reads 0.877 against 0.879, no cell's range moves by more than 0.014,

the movement is not one-directional (two ranges widen by up to 0.006), and thirteen cells span more than 0.3 either way. Where mass is low, the restriction preferentially removes documents with large $|\Delta|$. In the worst run all 103 removed documents were positive, so both the numerator and denominator change; nevertheless the sign fraction moves by at most 0.026 across runs, whereas the corresponding median can move by 63% (§5.4).

B.1 WITHIN-RUN REDUNDANCY STRATIFICATION

Table 10 reports the within-run stratification of §5.3. Valid documents are split separately in each run by realized redundancy, and the two strata are summarized separately. Because the edit rewrites $R-1$ value tokens, the split also changes edit size and is descriptive rather than an isolated redundancy dose. Because redundancy is discrete, the strata need not contain equal numbers of documents. The high-redundancy stratum has a larger absolute median displacement in 65 of 75 runs.

Table 10: Within-run redundancy stratification, three seeds. Each run is split separately by realized redundancy. medHi and medLo are ungated medians over the two strata; the median in Table 8 is also ungated but summarizes the full edit-domain population. Neither is a mass-restricted median. The gradient runs in the direction of the cell's own effect in 65 of the 75 runs. ^{a} marks the six exceptions whose pooled median is below 0.15 nats in absolute value, where the pooled median is small. ^{b} marks the four that invert a gradient while carrying an effect: the inversion is 8% and 3% of the pooled median in seeds 0 and 1, and 17% and 10% in the two contributed by seed 2. The two-seed version of this table carried no inversion above 8%, so the claim the third seed removes is that inversions are confined to gradients negligible relative to their cell; what survives is that the split tracks the size of the effect, at Spearman $\rho = 0.865$ over the 50 runs of seeds 0 and 1. Because cells differ by design, this coefficient pools between-cell and within-cell association; no within-cell coefficient is reported.

		seed 0		seed 1		seed 2	
R_{old}	ΔD	medHi	medLo	medHi	medLo	medHi	medLo
3	2	-0.016	-0.005	-0.051	-0.041	$+0.029$	$+0.002$
3	3	-0.046^{a}	-0.047	$+2.879$	$+2.592$	$+1.472$	$+0.467$
3	5	-0.120	-0.069	$+7.458$	$+0.430$	-0.017	-0.011
3	8	-0.041	-0.015	-0.001^{a}	-0.001	$+1.486$	$+0.870$
3	16	-0.022^{a}	-0.024	-0.032	-0.017	-0.032	-0.007
5	2	$+0.083$	$+0.018$	-0.043	-0.013	$+0.481$	$+0.322$
5	3	$+0.014$	$+0.003$	$+1.847$	$+1.129$	$+0.242$	$+0.199$
5	5	$+0.013$	-0.009	$+0.059$	$+0.006$	$+0.300$	$+0.175$
5	8	$+0.190$	$+0.098$	$+0.073^{a}$	$+0.136$	$+0.596$	$+0.158$
5	16	$+0.115$	$+0.044$	$+0.433$	$+0.250$	$+0.172$	$+0.025$
8	2	$+0.086$	$+0.003$	$+0.000^{a}$	-0.019	$+0.113$	$+0.081$
8	3	$+0.618$	$+0.572$	$+6.420$	$+4.106$	$+1.022$	$+0.936$
8	5	$+0.476$	$+0.250$	$+2.326$	$+1.752$	$+0.997$	$+0.974$
8	8	$+0.356$	$+0.285$	$+5.651$	$+4.318$	$+2.054$	$+0.327$
8	16	$+0.295$	$+0.142$	$+0.507$	$+0.287$	-0.016	-0.008
12	2	$+3.747$	$+3.136$	$+7.589$	$+2.943$	$+0.833$	$+0.318$
12	3	$+0.681$	$+0.599$	$+0.003^{a}$	$+0.023$	$+4.017^{b}$	$+4.447$
12	5	$+6.119$	$+3.987$	$+1.042^{b}$	$+1.074$	$+5.089$	$+2.802$
12	8	$+3.164$	$+2.098$	$+2.301$	$+1.076$	$+1.987$	$+0.610$
12	16	$+2.218$	$+0.997$	$+1.406$	$+0.728$	$+1.111$	$+0.455$
16	2	$+2.274$	$+1.351$	$+0.117$	$+0.111$	$+0.780^{b}$	$+0.923$
16	3	$+3.556^{b}$	$+3.857$	$+1.129$	$+0.543$	$+1.448$	$+0.608$
16	5	$+1.857$	$+1.171$	$+0.040$	$+0.034$	$+0.444$	$+0.213$
16	8	$+2.344$	$+1.967$	$+0.099$	$+0.048$	$+1.420$	$+0.945$
16	16	$+1.096$	$+0.624$	-0.028	$+0.024$	$+0.955$	$+0.652$

C DISAGREEMENT SUPERVISION AND THE TRAINED POSITIVE CONTROL

This appendix reports the aliased baseline, disagreement-supervision doses, run-level outcomes, fully enumerated label-permutation comparisons, state-gate failures and the trained RARITY-labeled positive control. Construction checks are in App. A.6; construction limits and model interventions are in App. F.

Table 11: Disagreement-supervision summary at $R_{\text{old}} = 3$ for $\Delta D = 8$ and the $\Delta D = 5$ replication. $p_{\text{break}} = 0$ is the aliased baseline and n counts RETRIEVAL-passing runs. Three-run rows list sign fractions; larger samples point to their run lists. Ranges and sample standard deviations use unrounded terminal fractions. Behavioral columns are arg-max rates on 400 generated inverted documents per run; they and the final column of run-level mass-restricted medians are summarized across runs by the median. Individual recency-answer rates in the expanded 1% rows span 0.993–1.000 (App. C.6). Positive-control ranges are ceiling-limited.

training labels	p_{break}	n	sign fraction	range	σ	p_{rec}	p_{rar}	median Δ
aliased baseline	0.00	3	0.098, 0.477, 0.977	0.879	0.441	—	—	−0.001
	0.00	9	App. C.5	0.879	0.268	—	—	—
RECENCY-labeled	0.01	3	0.688, 0.498, 0.395	0.293	0.149	1.000	0.000	−0.002
	0.01	9	App. C.6	0.311	0.132	1.000	0.000	+0.043
	0.10	3	0.530, 0.334, 0.471	0.196	0.101	1.000	0.000	−0.027
	0.30	3	0.161, 0.252, 0.303	0.141	0.072	1.000	0.000	−0.061
	0.30	10	Table 12	0.454	0.127	1.000	0.000	−0.070
RARITY-labeled	0.30	3	1.000, 0.995, 1.000	0.005	0.003	0.000	1.000	+26.25
Replication at $\Delta D = 5$								
aliased baseline	0.00	3	0.126, 0.972, 0.270	0.845	0.452	—	—	−0.013
RECENCY-labeled	0.01	3	0.726, 0.520, 0.748	0.228	0.126	1.000	0.000	+0.042
	0.01	10	App. C.6	0.445	0.165	1.000	0.000	+0.021
	0.10	3	0.405, 0.237, 0.729	0.492	0.250	1.000	0.000	−0.042
	0.30	3	0.345, 0.234, 0.469	0.236	0.118	1.000	0.000	−0.052

Fully enumerated dispersion comparisons. Tests reconstruct terminal sign fractions from machine-readable integer counts n_+/n at step 16000. In seed order, the two nine-run inputs are

$$\begin{aligned} \text{aliased } (0,1,2,3,4,5,6,7,9): &\ 32/328,\ 158/331,\ 335/343,\ 239/326,\ 259/335,\\ &\ 184/341,\ 237/324,\ 105/343,\ 239/339;\\ 1\%\ (0,1,2,3,5,6,7,8,9): &\ 225/327,\ 164/329,\ 130/329,\ 128/335,\ 217/328,\\ &\ 206/332,\ 235/339,\ 229/335,\ 145/329. \end{aligned}$$

The ten 30% ratios are listed in Table 12. The baseline, 1%, and 30% groups have mean absolute deviations from their group medians of 0.199, 0.112, and 0.089, respectively; the corresponding baseline-minus-dose differences are 0.087 and 0.110. Because the passing sets differ, the primary comparisons are unpaired and describe condition-specific populations of state-passing runs.

For each dose we pool the conditions and enumerate every fixed-size assignment: $\binom{18}{9} = 48{,}620$ at 1% and $\binom{19}{9} = 92{,}378$ at 30%. Group medians are recomputed for every assignment of

$$T_{\text{BF}} = \frac{1}{n_x}\sum_i |x_i - \widetilde{x}| - \frac{1}{n_y}\sum_j |y_j - \widetilde{y}|,$$

where x is the aliased baseline and the one-sided alternative is greater aliased dispersion. Applying the supplied verifier to the integer inputs gives Brown–Forsythe-type values 0.070115 and 0.041742 and variance-ratio sensitivities 0.042842 and 0.004536. Fractions printed elsewhere to three decimals are display values and do not by themselves reproduce these six-decimal outputs. The manuscript therefore reports the respective values as 0.070, 0.042, 0.043 and 0.005, with the Brown–Forsythe-type comparison primary and all threshold language explicitly nominal.

These values are exact tail proportions under the specified label-permutation reference distributions. Complete enumeration of that algorithm is distinct from claiming an exact finite-sample test for the

composite null of equal dispersion under arbitrary differences in distributional location or shape. Its randomization interpretation requires exchangeability under the reference null; the observed conditions also differ in location. We therefore treat these as exploratory comparisons and use 0.05 only as a nominal reference level.

As a seed-matched sensitivity analysis, we exhaustively swap condition labels within the shared seed sets. At 1%, the shared seeds are $\{0, 1, 2, 3, 5, 6, 7, 9\}$ and all $2^8 = 256$ assignments give Brown–Forsythe-type/variance-ratio $p = 0.051/0.047$. At 30%, the shared seeds are $\{0, 1, 2, 3, 4, 5, 6, 7, 9\}$ and all $2^9 = 512$ assignments give $p = 0.047/0.004$. These are sensitivity analyses rather than randomized paired tests. No Monte Carlo sample, add-one correction or permutation seed is used. Thus the 1% result remains descriptive, while the 30% comparison falls below the nominal one-sided Brown–Forsythe-type reference level; neither comparison is confirmatory or locates a minimum sufficient dose.

C.1 Measured confounds

Queried-slot structure. The retry loop makes an inverted queried slot contain $R_{\text{old}} + 1$ realized statements, against approximately $1 + 0.75R_{\text{old}}$ at $\Delta D = 8$ and $1 + 0.80R_{\text{old}}$ at $\Delta D = 5$ otherwise. A model could exploit this cue instead of reading the query. We test that possibility by replacing the query with another update-bearing slot from the same document while holding token count and answer position fixed. The replacement changes the prediction on at least 99.3% of documents in 68 of 75 main-grid runs and at least 83.8% in every run; in the nine passing $\Delta D = 5$ RECENCY-labeled runs from the original three-seed, three-dose design, the unchanged fraction is at most 0.0072. Conversely, it is 0.987–0.995 in three positional runs, its expected behavior for models that do not read the query. This validates the diagnostic without proving that every successful prediction uses one retrieval computation.

Training exposure. At large p_{break}, training includes the multiplicity configuration later presented by the edit. This exposure may affect both location and dispersion. Yet the token-matched RARITY-labeled arm receives the same exposure and gives approximately $+26$-nat medians and near-unit sign fractions. Exposure alone therefore does not determine direction, although the design cannot separate supervision from label-dependent effects of that exposure on variance.

C.2 Per-run values

Table 12: Run-level outcomes at $p_{\text{break}} = 0.30$, $R_{\text{old}} = 3$, $\Delta D = 8$, all at step 16000 and passing the RETRIEVAL gate. *esc* is the first step with copy diagnostic above 0.95; n_+/n gives positive-Δ documents over the edit-domain denominator; *mass* is mean contrast-pair mass and *ctrl* the non-queried-slot control median. Medians are mass-restricted, whereas frac+ uses all edit-domain documents; the exact ratios n_+/n supply the run-level fractions used for Table 11.

objective	seed	esc	acc_{nb}	acc_{inv}	p_{rar}	n_+/n	mass	ctrl	median Δ	frac+
RECENCY-labeled	0	5000	1.000	1.000	0.000	36/223	1.00	+0.003	−0.194	0.161
	1	4000	1.000	1.000	0.000	55/218	1.00	+0.000	−0.061	0.252
	2	5000	1.000	1.000	0.000	72/238	1.00	+0.000	−0.037	0.303
	3	5000	1.000	1.000	0.000	98/227	1.00	+0.001	−0.003	0.432
	4	4000	1.000	1.000	0.000	17/239	1.00	+0.000	−0.169	0.071
	5	6000	1.000	1.000	0.000	70/229	1.00	−0.000	−0.041	0.306
	6	6000	1.000	1.000	0.000	53/229	1.00	+0.001	−0.268	0.231
	7	4000	1.000	1.000	0.000	62/222	1.00	+0.001	−0.079	0.279
	8	13000	1.000	1.000	0.000	76/237	1.00	+0.002	−0.140	0.321
	9	2000	1.000	1.000	0.000	127/242	1.00	+0.001	+0.007	0.525
RARITY-labeled	0	5000	1.000	1.000	1.000	223/223	0.92	+0.003	+26.26	1.000
	1	6000	1.000	0.999	1.000	217/218	0.99	−0.002	+25.25	0.995
	2	6000	1.000	1.000	1.000	238/238	0.99	+0.001	+26.25	1.000

Escape intervals overlap at 30%: 2000–13000 steps under RECENCY-labeling and 5000–6000 under RARITY-labeling. The run escaping at 13000 reads 0.321, near the RECENCY-labeled mean of 0.288, and does not itself drive dispersion. Excluding inverted documents reduces the edit-domain count from 328 in the published cell to 218–242 here, with comparable yield and control magnitude. State

gating uses non-inverted accuracy because mixed accuracy is bounded by $1 - p_{\text{break}}$ before the inverted stratum is learned.

C.3 Replication at a second band width

At $R_{\text{old}} = 3$, $\Delta D = 5$, all nine original RECENCY-labeled runs—three seeds at each of three doses—pass the retrieval gate. Their behavioral recency-answer rates are 1.000, rarity-answer rates 0.000, and query-replacement unchanged fractions at most 0.0072.

At three seeds, σ falls from 0.452 in the aliased baseline to 0.126 at 1%, but the corresponding ranges are 0.228, 0.492 and 0.236 at doses 1%, 10% and 30%. Thus this cell reproduces low-dose contraction without a dose ordering; its baseline has only three seeds, so no matched-sample variance or threshold claim is made.

In the 30% RARITY-labeled positive control, two of three runs pass with medians $+23.76$ and $+20.02$ nats, sign fractions 1.00 and behavioral rarity-answer rates 1.000. The third ends in NEITHER at accuracy 0.006 and copy diagnostic 0.002 and is excluded rather than replaced by a rescue run. Its query-replacement value is uninformative because the model predicts almost nothing correctly.

C.4 Failed conditions and schedule-dependent rescue runs

At $p_{\text{break}} = 0.03$, two of three RECENCY-labeled runs terminate at the positional ceiling under the original cosine schedule. With the same 500-step linear warmup to 10^{-3} followed by a constant rate of 10^{-3}, they escape at steps 2000 and 5000; together with the third seed, that schedule gives sign fractions 0.653, 0.500 and 0.486. We report these schedule-rescue runs separately and exclude the dose from the curve because changing the schedule moves a matched run's readout by as much as 0.748.

At $p_{\text{break}} = 0.10$, one RARITY-labeled run reaches neither the retrieval nor positional state, ending with accuracy and copy diagnostic 0.024 and behavioral rarity-answer rate 0.007. All three RECENCY-labeled runs pass. This asymmetry is compatible with different optimization difficulty but one failed run does not identify its cause.

C.5 The widest cell at ten seeds

We extended the selected widest cell, $R_{\text{old}} = 3$, $\Delta D = 8$, from three to ten seeds. Seed 8 reaches copy diagnostic 0.998 but accuracy 0.985, fails the RETRIEVAL criterion, and is excluded from the terminal summary; its sign fraction is 0.391 and it remains in the explicitly ungated online reference of Table 21.

The nine passing terminal fractions, by seed, are s0 0.098, s1 0.477, s2 0.977, s3 0.733, s4 0.773, s5 0.540, s6 0.731, s7 0.306 and s9 0.705, giving sample standard deviation 0.268 and range 0.879. All six added passing runs lie inside the original endpoints. The extension therefore neither enlarges the range nor independently confirms it, because the cell and its original three observations were selected from the same grid; nine runs also do not resolve the number of modes.

C.6 The lowest dose at ten seeds, in both cells

We extended the 1% dose to ten seeds in both cells. Corpus, model, optimizer and training schedule match the original three-seed recipe; the only additional logged difference is the evaluation-only sample size described below.

At $\Delta D = 8$, nine runs pass and give $\sigma = 0.132$ over a range of 0.311, against $\sigma = 0.268$ in the nine-run aliased baseline. At $\Delta D = 5$, all ten pass and give $\sigma = 0.165$ over a range of 0.445; its aliased baseline has only three seeds. The label-permutation comparisons are specified above. The passing terminal fractions are, respectively,

$$0.688, 0.498, 0.395, 0.382, 0.662, 0.620, 0.693, 0.684, 0.441$$

and

$$0.726, 0.520, 0.748, 0.344, 0.484, 0.648, 0.303, 0.701, 0.394, 0.640.$$

At $\Delta D = 8$, seed 4 terminates at 71% of the positional ceiling (accuracy 0.079, copy diagnostic 0.003) and is excluded. At $\Delta D = 5$, seed 9 has an anomalous query-replacement unchanged fraction of 0.096, median $+0.363$, mass 0.97 and accuracy 0.999; excluding it gives $\sigma = 0.171$, so contraction does not depend on that run. Across the 19 passing runs, behavioral recency-answer rates are 0.993–1.000 and rarity-answer rates zero, including three individually non-decisive intervention fractions.

The original three seeds use 4000 state-diagnostic documents per evaluation and the seven additions 20,000. The fixed, training-disjoint evaluation stream changes diagnostic precision but not training or terminal-probe inputs; no gate decision lies within the resulting error.

C.7 Sample size, selection and gate failures

For normal-reference scale, the expected range of n draws is $d_n\sigma$, with $d_3 = 1.693$ and $d_{10} = 3.078$; ranges at different n are therefore not directly comparable. More importantly, comparing two variances at $n = 3$ requires a ratio of 19.0 to cross a one-sided 0.05 F threshold, and a single three-seed estimate has χ^2 interval $[0.52\hat{\sigma}, 6.28\hat{\sigma}]$. These calculations show that three-seed comparisons have little resolution except for very large variance ratios; a null result does not establish comparable dispersion.

The observed error is not one-directional: three seeds overstate baseline dispersion (0.441 versus 0.268 over nine passing runs) but understate 30% dispersion (0.072 versus 0.127 over ten). The 1% dose likewise gives 0.149 versus 0.132 at $\Delta D = 8$ and 0.126 versus 0.165 at $\Delta D = 5$. The fully enumerated full-sample comparisons are reported above; near-ceiling ranges are not interpreted as stability.

For selection sensitivity, we restrict both conditions to passing seed labels not used to select the $\Delta D = 8$ cell. At 1%, shared new seeds $\{3, 5, 6, 7, 9\}$ give $\sigma = 0.184$ in the aliased baseline and 0.139 under relabeling; at 30%, $\{3, 4, 5, 6, 7, 9\}$ give 0.179 and 0.158. These smaller descriptive contrasts show why the full comparison is not an independent confirmation.

Three reported runs clear the copy threshold but fail the accuracy criterion: baseline seed 8 ($0.998/0.985$), fixed-band $d = 16$ seed 2 ($1.000/0.127$), and rendered seed 1 ($0.992/0.975$), where each pair is copy/accuracy. Five original-schedule disagreement-supervision runs instead fail to reach RETRIEVAL: two positional runs at 3%, one NEITHER rarity-labeled run at 10%, one NEITHER $\Delta D = 5$ positive-control run, and $\Delta D = 8$, 1%, seed 4. Rescue runs are reported separately rather than replacing these outcomes. Because these failures come from heterogeneous cells, doses and labeling arms, their pooled count is not interpreted as a per-arm probability of passing.

C.8 Behavioral validation and the trained positive control

Behavior and the intervention readout answer different questions. On generated inverted documents, p_{rec} and p_{rar} record which value the model emits; Δ records a log-odds change under an edit of a non-inverted document. At $p_{\mathrm{break}} = 0.01$, the first three RECENCY-labeled runs have sign fractions 0.688, 0.498 and 0.395 while their behavioral rarity-answer rates are all zero. An intermediate causal sign fraction is therefore not a classifier of answer policy.

The trained RARITY-labeled arm provides a one-sided positive control. At $p_{\mathrm{break}} = 0.30$, all three passing runs at $\Delta D = 8$ have behavioral rarity-answer rate 1.000 and mass-restricted medians 26.26, 25.25 and 26.25 nats. At $\Delta D = 5$, two of three runs pass and have rarity-answer rate 1.000 with medians 23.76 and 20.02 nats; the third ends in NEITHER. None of the five passing runs combines a behavioral rarity preference with a non-positive median readout. This validates the trained-control direction, not a classifier for intermediate values.

D Training dynamics and run variation

This appendix reports the pre-retrieval shortcut, retrieval-gate and candidate-peak timing, the positional ceiling, and measured sources of variation across training runs. State and document-population definitions are centralized in App. A.5; pre-escape readouts remain here as training-dynamics diagnostics rather than mechanism attributions.

D.1 PRE-ESCAPE READINGS

Table 13 uses the last probe before escape in all 75 runs. The five readings above 0.6 are 0.74, 0.67, 0.67, 0.65 and 0.62; they occur only 813, 365, 365, 365 and 365 steps before the next 1000-step gate crossing. They therefore do not distinguish a positive pre-retrieval plateau from a probe that caught the transition.

Table 13: Pre-escape diagnostics. *Left*: selection denominators and counts of checkpoints with a recorded mean contrast-pair mass; the 0.5 threshold is applied to that checkpoint-level mean, not to individual documents. *Right*: the 32 runs with sign fraction below 0.20, by row and seed. The cutoff is descriptive; 22 of the 32 are below 0.05. Missing mass records are excluded from the mass counts. Online-to-terminal direction comparisons are run-level rather than document-paired.

runs trained	75
of which cross the gate	75
of which have a probe point before escape	75
of which that point is at step ≥ 200	75
sign fraction < 0.20	32
sign fraction < 0.05	22
of the 32, carrying a mass record	28
of those 28, mass ≥ 0.5	17
of the 32, terminal sign fraction > 0.5	27
of the 32, terminal sign fraction < 0.5	5

R_{old}	s0	s1	s2	row
3	4	1	2	7
5	0	2	0	2
8	3	0	1	4
12	4	0	5	9
16	4	1	5	10
	15	4	13	32

D.2 LOCATING ESCAPE

At each interior loss record t_i, both figures use the centered estimate $\widehat{g}(t_i) = [\ell(t_{i-1}) - \ell(t_{i+1})]/[\log t_{i+1} - \log t_{i-1}]$, which estimates $g(t) = -\mathrm{d}\ell/\mathrm{d}\log t$.

In seed 0 the candidate-peak row means separate four redundancy levels where the gate row means separate three. Across the 50 runs of seeds 0 and 1, height and time correlate at Spearman $\rho = 0.933$, but this pools designed cell differences and is coordinate-dependent because $g(t) = t(-\mathrm{d}\ell/\mathrm{d}t)$. We do not use it as independent evidence for sharper late transitions. Nineteen of the 20 high-redundancy runs peak at step 400 and one at 500, inside warmup. Warmup alone does not explain the contrast: under the same schedule, loss around the peak falls by 1.16 nats at $(R_{\text{old}}, \Delta D) = (16, 8)$ but only 0.03 at $(3, 8)$, and the ordering also holds in integrated learning rate.

Seed 2 gate row means are 4800, 2600, 1000, 1000 and 1000 steps as R_{old} increases. Its usable candidate-peak means are 4220, 1960, 820, 400 and 367; four runs are excluded because the final-fifth maximum is higher: R12D8s2, R12D16s2, R16D5s2 and R16D16s2. Together with R16D2s0, R16D3s0 and R16D5s0 in Table 14, this gives seven of 75 runs. A seed-2 candidate at step 300 confirms that 400 is not the recording-grid floor.

Dataloader audit. We verified that successive passes draw fresh documents. A prior replay bug drove training loss below the task's analytic entropy floor while held-out loss rose, so such runs are identifiable as memorization and excluded from every result.

D.3 DERIVATION OF THE CEILING

Each statement occupies exactly four tokens, so statement j (zero-indexed) spans token positions $4j$ through $4j + 3$ with its value token at $4j + 2$. The query appends four tokens, giving a document of $4m + 4$ tokens for m statements.

If the query follows ΔD statements after the rebinding of the queried slot, that rebinding is statement $m - \Delta D - 1$ and its value token sits at position $4(m - \Delta D - 1) + 2 = 4m - 4\Delta D - 2$. Counting from the end of the sequence,

$$\text{one-based offset of the answer value from the right boundary} = (4m{+}4){-}(4m{-}4\Delta D{-}2) = 4\Delta D{+}6. \tag{4}$$

Table 14: Escape for seeds 0 and 1. *gate* is the first 1000-step evaluation with copy diagnostic above 0.95. *peak* and h are the position and height of the maximum loss-decrease rate $-\mathrm{d}\ell/\mathrm{d}\log t$ over the first four fifths of training, recorded every 100 steps. † marks an early candidate peak whose height is exceeded in the final fifth during cosine decay; its printed step is the candidate, which is not used as a reliable escape-time estimate. Ties use average ranks; seed 2 is summarized below.

		seed 0			seed 1		
R_{old}	ΔD	gate	peak	h	gate	peak	h
3	2	8000	7500	9.86	7000	6100	8.12
3	3	6000	5600	6.66	5000	4600	6.98
3	5	5000	4200	7.15	5000	4600	5.77
3	8	7000	6800	8.80	6000	5600	7.55
3	16	4000	3400	5.11	10000	9600	14.48
5	2	2000	1000	3.36	3000	2800	6.80
5	3	1000	1000	3.10	3000	2300	5.59
5	5	1000	800	2.71	3000	2000	4.90
5	8	4000	2900	6.40	4000	2900	7.99
5	16	4000	3000	9.37	4000	3300	9.30
8	2	1000	800	3.46	1000	900	4.13
8	3	1000	600	2.80	2000	1100	4.77
8	5	1000	800	3.42	1000	900	3.97
8	8	1000	700	3.19	2000	1600	6.36
8	16	1000	800	3.48	2000	1400	5.20
12	2	1000	400	2.62	1000	400	2.72
12	3	1000	400	2.80	1000	400	2.74
12	5	1000	400	2.75	1000	400	2.75
12	8	1000	400	2.63	1000	500	2.67
12	16	1000	400	2.64	1000	400	2.71
16	2	1000	400†	1.60	1000	400	2.38
16	3	1000	400†	2.27	1000	400	2.60
16	5	1000	400†	2.17	1000	400	2.30
16	8	1000	400	2.44	1000	400	2.28
16	16	1000	400	2.37	1000	400	2.63

Equation 4 does not contain m. The answer's position measured backward from the end depends only on ΔD; document length cancels.

Consider a rule that ignores the query tokens and emits the value token at a fixed offset k from the end. By Equation 4 it is correct exactly when $4\Delta D + 6 = k$, that is when ΔD takes one particular value. Maximizing over k,

$$\mathrm{posCeil} = \max_{d} \Pr[\Delta D = d] = \frac{1}{|\mathrm{supp}(\Delta D)|}, \tag{5}$$

the second equality holding because ΔD is uniform on its support.

D.4 Scope of the fixed-offset bound

Equation 5 applies to rules that choose one fixed token offset independently of the document. It follows from the marginal distribution of ΔD and the unique-value construction; whether another observable feature predicts ΔD does not change the bound for this fixed-k class. It does not bound document-dependent positional strategies or all query-ignoring predictors.

Feasibility requires $n_{\mathrm{stmts}} \geq R_{\mathrm{old}} + \Delta D_{\max} + 5$. The most demanding cell, $(16, 16)$, reaches this constraint at the lower boundary of $U[45, 55]$. Extending either axis without changing the canvas would induce placement truncation and alter the realized marginal distribution of ΔD, in which case the uniform-support simplification in Equation 5 would have to be recomputed.

The length and placement audits characterize possible cues for broader rule classes: the empirical correlation between realized ΔD and token count is at most 0.043 in each cell, the generator samples

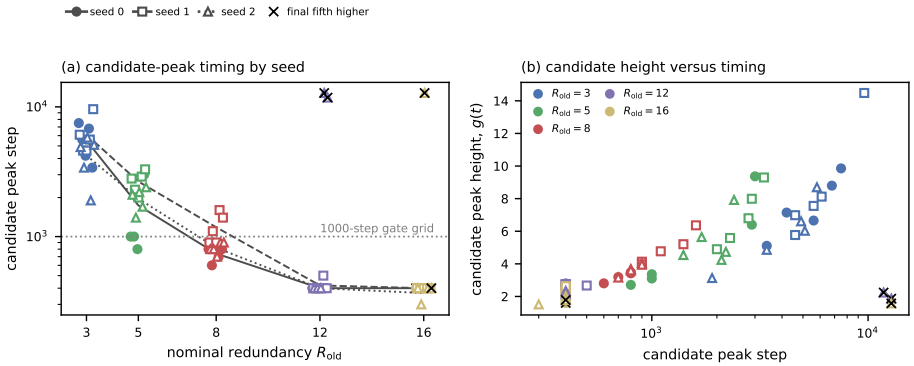

Figure 2: Candidate peaks of the loss-decrease rate $g(t) = -\mathrm{d}\ell/\mathrm{d}\log t$, recorded every 100 steps and selected over the first four fifths of training. Circles, squares and triangles denote seeds 0, 1 and 2; seed 0 is filled, seeds 1–2 are open, and color denotes R_{old}. A black $\times$ marks an early candidate whose height is exceeded in the final fifth. **(a)** Candidate step against R_{old}; each line joins a seed's mean over unflagged ΔD values, and the dotted line marks the 1000-step gate grid. **(b)** Candidate height against step over all 75 runs. Flagged candidates are not used as reliable escape-time estimates (Table 14).

ΔD uniformly, and queried and filler statements share vocabularies and format. These checks do not establish that every document-dependent cue is absent. In particular, queried-slot copies are 2–22% more dispersed than filler copies (App. A.1); this leaves open query-ignoring strategies outside the fixed-offset class and need not invoke the same computation measured by the copy diagnostic.

D.5 Observed plateau heights

Table 15: Positional ceiling against observed plateau accuracy. The $R_{\text{old}} = 2$ rows report terminal accuracy at 12000 steps; the $(R_{\text{old}}, \Delta D) = (3, 5)$ row reports the last evaluation before the copy diagnostic leaves chance.

R_{old}	ΔD	support	|supp|	posCeil	acc	copy	ratio
2	2	$[1, 3]$	3	0.333	0.3255	0.002	98%
2	3	$[2, 4]$	3	0.333	0.3253	0.001	98%
3	5	$[2, 8]$	7	0.143	0.1045	0.002	73%
2	8	$[4, 12]$	9	0.111	0.0773	0.046	70%
2	5	$[2, 8]$	7	0.143	0.0018	0.003	1%

Four rows are at $R_{\text{old}} = 2$ and therefore outside the analyzed grid. The two support-size-three cells reach 98% of the bound and agree in accuracy to 0.0002; the wider-support cells attain 70–73% or less. The $R_{\text{old}} = 2$ row is non-monotone: the $\Delta D = 5$ cell reaches neither the positional ceiling nor retrieval, while its neighbours occupy their ceilings or escape. Shortcut payoff therefore does not by itself order escape time.

D.6 The $R_{\text{old}} = 2$ cells

Four of these five seed-0 runs fail to meet the retrieval criteria within 12000 steps: three terminate POSITIONAL and one NEITHER. We retain them because they show both the positional plateau and the misleading readout that would result from ignoring retrieval state (Table 16).

The three positional rows give sign fractions 0.01–0.05 and negative readouts, but their copy diagnostics are only 0.001–0.046. Their readings therefore cannot be interpreted as choices between RECENCY and RARITY.

Table 16: The excluded $R_{\text{old}} = 2$ column at 12000 steps, seed 0. Four of the five never meet the retrieval criteria within the budget: three terminate POSITIONAL and one NEITHER. The probe reading on the three positional runs is the artifact analyzed in §6.3.

ΔD	support	posCeil	acc	copy	Δ / sign frac	state
2	$[1, 3]$	0.333	0.3255	0.002	-0.47 / 0.05	positional, 98% of ceiling
3	$[2, 4]$	0.333	0.3253	0.001	-0.44 / 0.01	positional, 98%
5	$[2, 8]$	0.143	0.0018	0.003	-0.00 / 0.22	neither, 1%
8	$[4, 12]$	0.111	0.0773	0.046	-1.30 / 0.02	positional, 70%
16	$[8, 24]$	0.059	1.0000	0.999	$+0.33$ / 0.72	retrieval, escaped at 5000

A fixed-offset account could produce the observed sign. The inversion rewrites $R_{\text{old}} - 1$ copies of v_{old} as v_{new} at the same statement positions; when a fixed offset lands on one of those positions, the output moves toward v_{new} and $\Delta < 0$.

The corpus does not identify the model's selected offset. At support size three, the empirical modal offset carries approximately one third of the mass, and its overlap with rewritten positions is 0.345 and 0.339 in the first two cells. These overlaps are compatible with a positional contribution in those cells, but they do not explain magnitude ordering: the $\Delta D = 8$ positional run has the largest absolute readout, 1.30 nats. At wider supports the corresponding overlaps are 0.161, 0.120 and 0.091, while the empirical mode changes between samples. We therefore neither identify the selected offset nor attribute readout magnitude to overlap alone; the readings remain state-dependent artifacts rather than rule attributions.

These cells enter no terminal cross-run comparison. The $\Delta D = 16$ row would pass the retrieval gate but remains outside the prespecified R_{old} grid; conversely, the 0.3255-accuracy positional run is excluded despite having converged to a different function. This is why the gate uses the copy diagnostic rather than loss or accuracy alone.

D.7 SOURCES OF RUN VARIATION

§5.5 asks which measured sources can account for the cross-run spread. The subsections below compare probe-batch resampling under fixed weights, training-stream order at fixed initialization, checkpoint and budget, and learning-rate schedule. These interventions act at different stages and do not form a complete variance decomposition; they test whether any one measured term explains the largest observed differences.

D.7.1 DOCUMENT BATCH AGAINST RUN

Each published readout pairs a trained run with its same-seed probe batch, so run and document batch are confounded in the grid. For the two extreme models from each of the two widest cells, we therefore evaluate both models on both batches, forming a 2×2 model-by-batch comparison.

The diagonal reproduces the published readings exactly: 0.098 and 0.977 at $(R_{\text{old}}, \Delta D) = (3, 8)$, and 0.126 and 0.972 at $(3, 5)$. Holding the model fixed and replacing the batch changes the sign fraction by 0.034 and 0.020 in the first cell and by 0.001 and 0.001 in the second. For the first pair, the corresponding binomial standard errors are 0.016 and 0.008 at edit-domain size $n = 328$. Holding the batch fixed while exchanging models changes the reading by 0.899 and 0.846, reproducing the published cross-seed ranges of 0.879 and 0.845 within sampling error. The two larger batch effects are 2.1 and 2.5 standard errors, so the batch term is bounded rather than called zero, but it does not explain the run difference.

D.7.2 INITIALIZATION VERSUS TRAINING-DOCUMENT STREAM

A grid seed controls both initialization and the sampled training-document stream. We hold initialization fixed at seed 0 and vary four data seeds; data-seed 0 reproduces the published online reading exactly (0.0943), anchoring the arm against environment drift.

Table 17: The training-stream arm. All four runs share one initialization (`--seed 0`). The `data-seed` 0 row is the published run. Readouts are at step 16000 on the online probe; the band is the within-run range of the sign fraction over 8000–16000.

`data-seed`	sign fraction	median Δ	mass	late band
0	0.0943	-0.075	0.999	0.119
100	0.2086	-0.577	0.999	0.086
101	0.6821	$+0.011$	0.996	0.127
102	0.6467	$+0.018$	0.997	0.078

The four streams span 0.588 with $\sigma = 0.300$, against 0.271 over ten ordinary seeds. Their own late-window standard deviations have median 0.040 and maximum 0.050, and all four have contrast-pair mass at least 0.996. The cross-stream standard deviation of 0.300 is therefore 7.5 times the median and 6.0 times the largest within-run value in this arm. Training-stream variation can therefore produce dispersion of the same order with initialization fixed.

With only four streams, a normal-reference calculation gives $[0.170, 1.119]$ for σ; we use it only to illustrate the imprecision of the estimate. The arm does not establish that stream variation exceeds the ordinary-seed term. It is also not a variance decomposition: initialization and stream are varied one at a time rather than crossed. Probe-batch replacement acts after training and moves the readout by at most 0.034, whereas this arm changes the data from which the weights are learned.

D.7.3 Budget sensitivity

Terminal grid readouts use a fixed budget. We vary two cells over three budgets, three cells over two budgets with escape held at step 1000, and one cell over 13 post-gate checkpoints. The saturated $(R_{\text{old}}, \Delta D) = (16, 2)$ cell remains at sign fraction 0.995–1.000, whereas the boundary cell $(3, 2)$ moves by 0.55.

Table 18: Readout against budget. The first two cells vary budget at fixed seed; the last three hold the gate crossing at step 1000, isolating post-gate trajectory differences. Medians use the contrast-pair mass restriction. The 12000-step row uses a finer evaluation grid.

cell	budget	escape	median	sign fraction	seed
$R_{\text{old}} = 16$, $\Delta D = 2$	4000	1000	$+2.363$	0.995	0
	16000	1000	$+1.811$	0.998	0
	32000	2000	$+2.433$	1.000	0
$R_{\text{old}} = 3$, $\Delta D = 2$	12000	2508	-0.000	0.43	0
	16000	8000	-0.011	0.354	0
	32000	6000	$+0.262$	0.907	0
$R_{\text{old}} = 8$, $\Delta D = 2$	16000	1000	-0.006	0.45	1
	32000	1000	$+0.338$	0.98	1
$R_{\text{old}} = 16$, $\Delta D = 16$	16000	1000	-0.009	0.47	1
	32000	1000	$+0.102$	0.79	1
$R_{\text{old}} = 16$, $\Delta D = 5$	16000	1000	$+0.040$	0.58	1
	32000	1000	$+0.013$	0.54	1

The saturated cell changes little across budgets, while the boundary cell moves because both escape and distance past the gate change. Among the three fixed-escape comparisons, two move substantially and one does not; doubling the cosine period is therefore not a uniform sharpening of the map.

Across 69 runs with at least five post-gate probes, within-run sign-fraction standard deviations range from 0.007 to 0.332 with median 0.074. The largest run crosses sign 7000 steps after gate crossing (§6.3). In the 32000-step boundary run, sign fraction has standard deviation 0.03 while the mean of Δ has standard deviation 0.49 nats despite accuracy and copy diagnostic remaining 1.000. Check-

point variation is therefore non-negligible, while the sign fraction remains less tail-sensitive than the mean. All standard deviations use the $n - 1$ denominator.

For $(16, 2)$, the mass restriction changes the median by 1% at 4000 steps and 16% at 32000. At 4000 steps only 2 of 400 documents exceed 8 nats, with maximum 10.86; the 32000-step upper tail reaches 14.26. The larger mass-restriction effect is thus associated with an upper tail rather than a uniform shift. The idealized value $2\kappa = 16$ is not a ceiling for the trained network.

D.7.4 The cross-seed range does not narrow where it is widest

We compare cross-seed ranges at common probe steps from 8000 to 16000; 20 of 25 cells provide at least four such points. The mean range falls from 0.335 in the first half to 0.305 in the second, with 13 of 20 cells narrowing (one-sided Wilcoxon $p = 0.077$). The widest cell does not narrow: $(R_{\text{old}}, \Delta D) = (3, 8)$ reads 0.882 early and 0.884 late. The $(3, 5)$ range falls from 0.949 to 0.899, whereas $(3, 3)$ widens from 0.491 to 0.537.

Across these 20 cells, observed narrowing is concentrated near the $[0, 1]$ boundary: cell mean sign fraction correlates with the late/early range ratio at Spearman $\rho = -0.70$. For three values with mean μ, the maximum possible range is $\min\{3\mu, 1, 3(1 - \mu)\}$. When the mean exceeds 0.99, the range is therefore bounded above by 0.03. Small ranges near this boundary can reflect ceiling compression and do not, by themselves, establish convergence across runs. The three widest cells remain at means 0.46–0.64, where the bound is 1.00, and occupy 79–88% of it. Thus the largest differences do not observably converge over the available checkpoints, but the experiment does not address arbitrarily longer schedules.

D.7.5 Constant learning rate

We test whether cosine annealing freezes runs at different positions by retraining the widest cell under a constant learning rate. After the same 500-step linear warmup to 10^{-3}, the rate remains at 10^{-3} rather than decaying to 10^{-4}. Corpus, model and all other training fields match the main grid; for the continued runs, total step count is the only additional logged difference. Ten seeds are evaluated at step 16000. The first three continue to 32000 and the remaining seven stop at 16000, but the constant schedule does not read the total budget, so the step-16000 state of a continued run is directly comparable to an independently terminated 16000-step run. All ten reach accuracy and copy diagnostic 1.000.

At step 16000 the constant-schedule arm spans 0.583 with $\sigma = 0.200$, against 0.894 and 0.271 for the same ten seeds under cosine on the same online probe. Relative to constant-schedule within-run standard deviations, the residual cross-run value is 3.0 times the median (0.066) and 1.3 times the maximum (0.153). Removing annealing therefore reduces but does not eliminate dispersion.

The schedule effect is not a uniform shift: five paired changes are positive and five negative, with mean $+0.045$, standard deviation 0.342 and sign-test $p = 1.000$. Individual changes range in magnitude from 0.012 to 0.748, and the cross-seed rank ordering changes.

At step 32000 the three continued runs read 0.711, 0.721 and 0.765, giving a cross-run range of 0.054 and sample standard deviation approximately 0.029. Over steps 24000–32000 their within-run standard deviations are 0.039–0.083. The small terminal range is not stable across nearby checkpoints: it is 0.33 at step 28000 and 0.28 at step 24000, while seed 1 moves from 0.50 to 0.72. This arm rejects the specific explanation that cosine decay alone freezes runs at different positions; it neither establishes stable convergence in the observed window nor rules out convergence under a substantially longer schedule.

E Architecture and model scope

This appendix tests QK-normalization gain, depth, width, and two departures from the main surface and training regime. These small arms sometimes change multiple factors, so they delimit scope rather than estimate general architecture effects. Removing QK normalization itself remains untested.

E.1 THE QK-NORMALIZATION GAIN

Attention applies per-head RMS normalization to queries and keys before RoPE, with learnable gain vectors initialized at $\gamma = 2$. Let P_p denote the orthogonal RoPE matrix at position p, let $D_q = \mathrm{diag}(w^{(q)})$ and $D_k = \mathrm{diag}(w^{(k)})$, and let $\hat{q}_p, \hat{k}_r$ have RMS at most one. Then

$$z_{pr} = \frac{(P_p D_q \hat{q}_p)^\top (P_r D_k \hat{k}_r)}{\sqrt{d_h}}, \qquad |z_{pr}| \le \|w^{(q)}\|_\infty \|w^{(k)}\|_\infty \sqrt{d_h}. \tag{6}$$

At initialization the right-hand side is $\gamma^2\sqrt{d_h} = 8\gamma^2$ for $d_h = 64$. Under an equal-distractor approximation, assigning attention fraction p to one of n positions requires

$$z_t - z_o = \log \frac{p(n-1)}{1-p}, \tag{7}$$

which is 7.6 at $n = 224, p = 0.9$. This is an attention-concentration benchmark under the stated approximation, not a criterion for task retrieval.

We set $\gamma = 1$ in a slow cell, $(R_{\mathrm{old}}, \Delta D) = (3, 5)$, and a fast cell, $(16, 2)$, using two seeds each and otherwise matching the main recipe.

Table 19: The four $\gamma = 1$ runs. *escape* is the first 1000-step evaluation with copy diagnostic above 0.95; *bound* is the maximum over layers of the terminal right-hand side of Equation 6. The two slow-cell runs do not receive intervention readouts because they fail the retrieval-state criteria and have mean contrast-pair mass 0.15.

cell	seed	γ	escape	acc	copy	mass	bound
$R_{\mathrm{old}} = 3, \Delta D = 5$	0	1	none	0.0925	0.132	0.15	23.1
$R_{\mathrm{old}} = 3, \Delta D = 5$	1	1	none	0.194	0.244	0.15	37.2
$R_{\mathrm{old}} = 16, \Delta D = 2$	0	1	1000	1.000	1.000	0.97	20.5
$R_{\mathrm{old}} = 16, \Delta D = 2$	1	1	1000	1.000	1.000	0.89	23.7

Neither slow-cell run meets the retrieval criteria within 16000 steps. Their accuracies are 0.0925 and 0.194; the latter exceeds the fixed-offset ceiling $1/7$ and so lies outside that rule class, while copy diagnostics are 0.132 and 0.244. Only 0.11 and 0.02 of documents meet the contrast-pair mass restriction, so these runs provide retrieval-state outcomes rather than interpretable intervention readouts. In the fast cell, both runs escape at step 1000 and end at accuracy and copy diagnostic 1.000; their ungated medians are $+2.10$ and $+2.14$ nats and their sign fractions 0.993 and 0.987.

The four model-wide bounds range from 20.5 to 37.2. These upper bounds do not establish the target–distractor gaps actually attained and therefore cannot by themselves explain retrieval success or failure. Endpoints also cannot exclude an effect of initial gain on early optimization. This experiment varies the gain at fixed QK normalization; removing QK normalization itself was not tested.

E.2 DEPTH

We repeated the $\Delta D = 5$ column at 4 and 12 layers for $R_{\mathrm{old}} \in \{3, 8, 16\}$, with two seeds per cell; the 8-layer values are from the main grid. Documents are paired by seed, but width is fixed, so depth and parameter count covary (13.4M, 26.1M and 38.7M).

All twelve new runs reach accuracy and copy diagnostic 1.000. The coarse escape ordering survives: the 1000-step gate falls at 4000–6000 steps for $R_{\mathrm{old}} = 3$, 1000–2000 for 8, and 1000 for 16. The within-cell redundancy gradient aligns with each cell's aggregate effect in all twelve runs, including the two cells with negative medians.

Pooled over seeds, the $R_{\mathrm{old}} = 3$ sign fraction is lowest at every depth: 0.59, 0.55 and 0.42 at 4, 8 and 12 layers. The ordering between 8 and 16 does not persist: their pooled fractions are 0.99 versus 0.97 at four layers, 0.97 versus 0.79 at eight, and 0.81 versus 0.99 at twelve. Thus only the coarse low-versus-high redundancy ordering replicates across depth.

The loss-decrease rate supports only the same coarse ordering (Table 20). Four of the twelve new depth runs are flagged because the final-fifth maximum exceeds the early candidate; including the six 8-layer reference runs, five of eighteen entries are flagged. Among the unflagged entries, $R_{\mathrm{old}} = 3$ peaks at 3400–5100 steps versus 400–1200 elsewhere, but 8 and 16 are not separated.

Table 20: Candidate loss-decrease-rate peak positions in the depth arm, on the 100-step recording grid. † marks an early candidate whose height is exceeded during the final cosine decay. The printed step is the early candidate, not the later decay peak, and is not used as a reliable escape-time estimate. The 8-layer rows are the corresponding main-grid runs from Table 14.

layers	R_{old}	seed 0	seed 1
4	3	3400	5100
4	8	400	500†
4	16	300†	400
8	3	4200	4600
8	8	800	900
8	16	400†	400
12	3	5100	4400
12	8	1200	1000
12	16	400†	400†

The readout itself has no monotone depth ordering. At $R_{\mathrm{old}} = 3$, sign fractions at 4, 8 and 12 layers are 0.91, 0.13 and 0.12 in seed 0, but 0.26, 0.97 and 0.72 in seed 1. The seed-0 median moves from $+2.858$ nats at four layers to -0.058 at twelve, changing sign rather than only magnitude. The largest movement with depth at fixed seed is 0.79 in sign fraction; the largest seed difference at fixed depth is 0.65 in this arm, against 0.845 in the corresponding main-grid cell.

Depth and parameter count covary, and the arm contains only one ΔD column. That column was selected when two seeds were available, when $\Delta D = 5$ carried the widest $R_{\mathrm{old}} = 3$ range; after adding the third seed it is second to $\Delta D = 8$ (0.845 versus 0.879, Table 9). The selection does not affect the escape comparison, but it makes the readout comparison deliberately a high-variation case. This arm therefore shows that the joint depth–capacity change can move the readout by the order of reseeding; it neither ranks depths nor establishes stabilization with depth.

E.3 A WIDTH ARM AT 102M PARAMETERS

At fixed eight-layer depth, we retrained the selected widest cell, $(R_{\mathrm{old}}, \Delta D) = (3, 8)$, at $d_{\mathrm{model}} = 1024$ with 16 heads (102.2M parameters versus 26.1M). Head dimension remains 64, so capacity grows through head count rather than head size; schedule, budget, batch size, learning rate and matched-seed data streams follow the main recipe.

All three seeds reach accuracy and copy diagnostic 1.000, with zero query-replacement unchanged fraction. Their sign fractions are 0.27, 0.86 and 0.67, with medians -0.022, $+0.251$ and $+0.118$ nats. The first two runs point in opposite directions and are individually decisive under exact binomial tests ($p = 5 \times 10^{-17}$ and $p = 3 \times 10^{-45}$). The cross-seed standard deviation is 0.301 over a range of 0.59. This is 4.1 times the main-grid median within-run standard deviation of 0.074, but does not exceed the largest observed within-run value of 0.332.

Cross-run variation therefore also occurs in this wider arm, but the experiment does not establish how its population dispersion changes with width. The corresponding three-seed main-grid range is 0.879, and comparing two three-run ranges is too imprecise to establish shrinkage. Escape occurs at steps 3000, 3500 and 3500, within the 2000–12000 span of the nine state-passing 26.1M runs in this cell. Thus the arm establishes neither reduced dispersion nor systematically earlier gate crossings with width.

E.4 TWO SURFACES OUTSIDE THE GRID

Two arms leave the main grid's four-token surface and from-scratch 26.1M training regime. One renders the same placement layer as English sentences; the other fine-tunes Qwen2.5-0.5B on that corpus. Both use $(R_{\text{old}}, \Delta D) = (3, 8)$.

In the rendered arm, each slot retains one of four English templates across its appearances, and values are adjective–noun pairs whose adjective injectively indexes the value. The rendered-from-scratch arm uses 150 values; the Qwen arm uses all 145 adjectives surviving its single-token filter, also injectively. Queried and filler statements both average approximately 7.5 tokens, differing by only 0.03 tokens.

The construction itself is preserved: RECENCY–RARITY agreement is 1.000 over 2000 documents per cell and the last-token rule reads 0.000. The surface nevertheless changes three load-bearing quantities relative to the grid: value-pool size, statement length, and positional geometry. The measured positional ceiling falls from 0.111 to 0.039 under whitespace tokenization and to 0.023 under Qwen tokenization. These arms therefore do not identify a surface-only effect.

Table 21: Rendered and fine-tuned runs at $(R_{\text{old}}, \Delta D) = (3, 8)$. For each displayed run, n counts valid edit pairs among 200 probe documents that meet the per-document contrast-pair mass restriction; sign fractions and medians use that subset, whereas *mass* is the unrestricted edit-domain mean. Rendered runs are evaluated at step 16000; fine-tuned rows summarize post-gate evaluations over steps 1000–2000. † marks a run failing the joint RETRIEVAL criterion. The main-grid reference is the ungated online mean over all ten seeds, including seed 8, which fails the terminal state criterion. The distinct nine-run state-passing terminal mean is 0.593.

arm	seed	acc	n	mass	sign fraction	median Δ
rendered, 26.1M	0	1.000	115	0.575	0.957	+4.69
	1†	0.975	121	0.601	0.983	+5.63
	2	1.000	98	0.474	1.000	+7.26
fine-tuned, 0.5B	0	1.000	110	0.550	0.927	+9.31
	1	1.000	121	0.605	0.951	+10.09
	2	1.000	123	0.615	0.991	+7.31
main-grid online ref.	0–9	0.985–1.000	—	—	0.580 (mean)	—

The two passing rendered runs average 0.978 against the online main-grid reference of 0.580, a descriptive displacement of about 0.40. The summaries differ in run and document gating, so this is not a paired surface effect. The fine-tuned runs average 0.957 against a zero-shot intervention sign fraction of 0.115. Before fine-tuning, the rarer answer wins on 0.096 of 125 decisive disagreement documents (95% interval $[0.044, 0.148]$), with median rarity–recency logit difference -1.81 nats. Its zero-shot copy diagnostic is only 0.726, so fine-tuning builds retrieval before the reported displacement. This does not demonstrate traversal of a flat rule-trading direction.

Neither arm supports a dispersion claim: their passing-run ranges, 0.043 and 0.064, are ceiling-limited and smaller than the corresponding within-run sign-fraction ranges, whose median is 0.088 and maximum is 0.112. A fine-tuned seed-0 learning-rate control changes the sign fraction from 0.927 to 0.962 while retaining a step-100 gate crossing, much smaller than the 0.84 zero-shot-to-fine-tuned displacement. The rendered arm still changes surface, value pool, statement length and positional ceiling together; the fine-tuned arm is one model family with three seeds. Rendered seed 1 fails the joint state criterion, and seed 2 has unrestricted mean mass 0.474, so its reported readout uses only the 98 documents meeting the mass restriction.

F CONSTRUCTION CONFOUNDS AND MEASUREMENT LIMITS

This appendix collects controls for collinear corpus quantities, conditions under which the disagreement-supervision control cannot be constructed, and direct measurements whose component-level or geometric interpretations are limited or negative. The two slot-count controls are informative jointly: shortening changes length and removes most of the copy-dispersion difference,

whereas changing filler update probability holds length fixed and largely preserves that difference. These results bound the main claims rather than supply additional confirmation.

F.1 Matching slot count by shortening

Shortening the $R_{\text{old}} = 8$, $\Delta D = 2$ documents matches their slot count to the $R_{\text{old}} = 16$ reference. Calibration on 2000 documents gives 8.94 versus 8.69 slots. It also reduces the queried-copy dispersion difference from 41% in the main grid to 11%, but changes document length from about 204 to 117 tokens and lowers realized redundancy from 5.84 to 3.98. Training and evaluation are otherwise unchanged.

Table 22: Slot matching by shortening. *q-gap* is the statement-position distance from the final queried rebinding to the latest earlier occurrence of the queried slot; *q-kept* is the number of surviving old-value copies. These differ from the all-pair antecedent averages in Table 1. Sign fractions and mass-restricted medians are listed by seed.

condition	slots	q-gap	q-kept	tokens	frac+ (s0/s1/s2)	mass-restricted median (s0/s1/s2)
$R_{\text{old}} = 8$ main	14.60	7.37	5.84	204.4	0.662/0.450/0.762	—
$R_{\text{old}} = 8$ shortened	8.94	5.82	3.98	117.3	0.950/0.989/0.983	$+0.337/+0.821/+1.351$
$R_{\text{old}} = 16$ main	8.69	5.23	9.04	205.0	0.998/0.736/0.972	—

All three shortened runs reach accuracy and copy diagnostic 1.000. Their sign fractions lie outside the main $R_{\text{old}} = 8$ range and move toward the $R_{\text{old}} = 16$ reference, although their mass-restricted medians still span $+0.337$ to $+1.351$ nats.

This movement opposes the within-run association between realized redundancy and displacement magnitude, but the combined changes do not isolate redundancy's causal contribution. Splitting the shortened runs at their own median redundancy, their low-redundancy halves have medians $+0.172$, $+0.432$ and $+1.109$, all above the main-cell seed-0 high-redundancy half at $+0.086$. The arm therefore establishes a directional effect of the combined change in length, slot sparsity and copy dispersion; it does not identify which member carries the effect.

F.2 Matching slot count at fixed length

Changing p_{update} moves filler-slot occupancy and hence slot count without changing document length. We set $p_{\text{update}} = 0.90$ at $R_{\text{old}} = 8$ and 0.28 at $R_{\text{old}} = 16$, matching the two arms at 11.51 and 11.55 slots and about 205 tokens. Unlike shortening, this manipulation leaves queried-copy dispersion 36% apart, compared with 41% in the main grid.

Table 23: Calibration and readouts for the fixed-length slot-count arm. $\rho = p_{\text{update}}/(1 + p_{\text{update}}R_{\text{old}})$ is the generator's expected fraction of filler statements that are updates; it is not q-gap or an empirical nearest-antecedent statistic. Medians use the contrast-pair mass restriction. The bracketed sign fraction is not interpreted: its filler-slot control is -0.020 against a median of -0.055.

condition	p_{update}	slots	q-gap	q-kept	tokens	ρ	frac+ (s0/s1/s2)	mass-restricted median (s0/s1/s2)
$R_{\text{old}} = 8$ main	0.50	14.60	7.37	5.84	204.4	0.100	0.662/0.450/0.762	—
$R_{\text{old}} = 8$ arm	0.90	11.51	7.48	5.94	205.2	0.110	[0.154]/1.000/0.997	$-0.055/+2.852/+1.159$
$R_{\text{old}} = 16$ arm	0.28	11.55	5.51	8.90	205.2	0.051	0.997/1.000/1.000	$+0.936/+1.235/+2.543$
$R_{\text{old}} = 16$ main	0.50	8.69	5.23	9.04	205.0	0.056	0.998/0.736/0.972	—

All six arm runs reach RETRIEVAL, but the $R_{\text{old}} = 8$ seed-0 readout is uninterpretable because its filler control is comparable to the targeted median. The all-run mean difference of 0.282 includes that observation. The all-run median difference is 0.003; taking a median reduces the influence of an extreme value but does not remove that observation from the sample. Excluding it leaves only two interpretable $R_{\text{old}} = 8$ runs and gives a median difference of approximately 0.0015 using the displayed fractions. These small, ceiling-limited samples do not determine whether fixed-length slot-count matching aligns the two cells.

The arm still records that slot count and expected filler update density move in opposite directions in its two legs, while copy dispersion remains 36% apart between matched cells. Together with

the shortening arm, this shows that length, slot sparsity, filler density and copy dispersion have not been isolated from one another. It neither establishes alignment nor identifies a single monotone explanation for the readout.

F.3 FIXED-WIDTH BANDS

Because the main ΔD axis also changes band width, we repeated the $R_{\text{old}} = 3$ row with $\Delta D \sim U[d, d+8]$ for $d \in \{1, 4, 8, 16\}$, holding the positional ceiling at $1/9$. The $d = 4$ cell reproduces the original $(R_{\text{old}}, \Delta D) = (3, 8)$ runs exactly. All generator invariants pass, RECENCY–RARITY agreement is 1.000, measured positional collision is 0.120–0.131, and the antecedent-distance cue weakens across the row.

Sign fractions are 0.870/0.104/0.204 at $d = 1$, 0.098/0.477/0.977 at 4, 0.319/0.359/0.418 at 8, and 0.584/0.486 at 16; the third $d = 16$ run has no interpretable readout. Three cells straddle 0.5, and the new $d = 1$ cell spans 0.766 with all three runs individually decisive: exact binomial $p = 1 \times 10^{-48}$ in the positive direction and $p = 3 \times 10^{-55}$ and 3×10^{-30} in the negative direction. Variation in the analytic ceiling therefore cannot be the sole source of cross-run spread. The sign-fraction row means are 0.393, 0.517, 0.365 and 0.535, so the readout is not ordered by d. The coarse gate means increase from 5000 to 5333, 6333 and 7667 steps, but the 100-step loss-decrease-rate candidate peaks do not give a stable between-cell ordering: at $d = 16$ they are 1400, 6500 and 10100 across seeds, a sevenfold within-cell span that exceeds the between-cell differences. The arm therefore does not isolate a reproducible ΔD-dependent escape gradient.

At $d = 16$, seed 2 has copy diagnostic 1.000 and a loss-decrease-rate candidate peak of 10.16 at step 10100, but only 0.127 in-distribution accuracy and 0.05 on the no-tail-update stratum. Mean contrast-pair mass is 0.12 and only 3% of documents meet the mass restriction, so the run is classified NEITHER and receives no intervention readout. This demonstrates why the state criterion requires both the copy diagnostic and task accuracy; the two other reported cases where these conditions separate are listed in App. C.7. Whether continued training would convert reliable copying into task accuracy is unresolved.

F.4 WHY THE ARM DOES NOT RUN AT LARGE R_{old}

The disagreement-supervision construction requires an inverted queried slot to remain structurally indistinguishable from filler slots. At $k = 1$ this fails in every measured cell with $R_{\text{old}} \geq 8$. The retry loop protects the single v_{old} copy only on the queried side. Because that copy is first in the slot's temporal order, it receives the smallest of the sampled positions and survives only when every sampled queried-slot position is nonnegative.

Let $p_{\text{final}} = n_{\text{stmts}} - 1 - \Delta D$ be the pinned rebinding index, w_{st} the width of the candidate placement window, and $m = \min\{p_{\text{final}}, w_{\text{st}}\}$ the number of nonnegative positions in that window. Since the R_{old} positions are sampled without replacement, the success probability of one retry is

$$\frac{\binom{m}{R_{\text{old}}}}{\binom{w_{\text{st}}}{R_{\text{old}}}} = \prod_{i=0}^{R_{\text{old}}-1} \frac{m-i}{w_{\text{st}}-i} \leq \left(\frac{m}{w_{\text{st}}}\right)^{R_{\text{old}}}.$$

The measured ratio m/w_{st} falls from about 0.72 at $R_{\text{old}} = 3$ to 0.37 at 16 as the filler budget shrinks. At $R_{\text{old}} = 16$, the with-replacement upper bound is therefore of order 10^{-7} per retry. Consistent with this calculation, the observed clamp fractions are 0.979 and 1.000 in the two measured $R_{\text{old}} = 16$ cells: the 64 retries almost always exhaust, and the fallback placement makes the queried slot structurally distinguishable.

Raising k to 8 largely removes clamping and restores the antecedent ratio at $\Delta D = 8$, but reduces the visible multiplicity margin to one. Independent copy loss then makes the queried side over-invert relative to the filler side: 0.107 versus 0.050 at $\Delta D = 8$ and 0.109 versus 0.058 at 16. The resulting differences, 0.057 and 0.051, fail the inversion-rate gate. Table 24 therefore records two distinct failures: structural marking at $k = 1$ and unmatched visible inversion rates at $k = 8$.

The construction is therefore restricted to low redundancy. It covers the grid's widest cell, $(R_{\text{old}}, \Delta D) = (3, 8)$, but cannot test whether dispersion contraction extends to the high-redundancy cells where substantial cross-run spread also occurs.

Table 24: Why the arm is confined to $R_{\text{old}} = 3$. Measured on the generator alone, 1500 documents per cell, $p_{\text{break}} = 0.10$. *clamp* is the fraction of inverted documents falling back to the clamped placement path; *ant* is the ratio of antecedent distances, queried over filler, inside inverted documents; *inv-6* is the queried–filler difference in inversion rate; *adj*Δ is the adjacency-rate difference. Bold fails its gate.

R_{old}	ΔD	k	r	clamp < 0.20	ant $(0.80, 1.25)$	inv-6 < 0.05	adjΔ < 0.05
3	5	1	3	0.000	1.045	0.003	0.016
3	8	1	3	0.000	0.950	0.005	0.013
8	8	1	8	**0.438**	**0.674**	0.005	0.002
8	16	1	8	**0.688**	**0.583**	0.010	0.005
16	8	1	16	**0.979**	**0.481**	0.016	0.008
16	16	1	16	**1.000**	**0.408**	0.030	0.026
16	8	8	9	0.075	0.812	**0.057**	0.027
16	16	8	9	0.188	**0.662**	**0.051**	0.015

F.5 Model interventions where a behavioral label exists

The disagreement-supervision arm is the only setting here where a model-level intervention can be checked against a behavioral rule label. We analyze three seeds of the token-matched supervision arms at $(R_{\text{old}}, \Delta D) = (3, 8)$ on 1200 held-out documents per seed. Behavioral rarity-answer rates are 1.000 under RARITY-labeling and 0.000 under RECENCY-labeling, with non-inverted accuracy 1.000.

Layer localization replicates. In each RARITY-labeled model, we replace the residual stream at the rare value's source position before layer ℓ with the corresponding activation from a non-inverted document. Patching before layer 3 lowers the rarity-answer rate from 1.000 to 0.000, 0.059 and 0.061 across the three seeds; patching before layer 7 leaves it unchanged to three decimals. The last layer at which patching changes the label is 4, 3 and 6. The decisive depth range therefore replicates approximately, although its endpoint does not.

Head identity does not replicate. The preregistered carrying-head criterion requires removal to lower the rarity-answer rate below 0.5 while costing at most 0.01 non-inverted accuracy. Seed 0 has two qualifying heads: removal lowers the rate to 0.205 and 0.234 at accuracy 0.992. Seed 1 has none; its largest label movement costs 0.090 accuracy. Seed 2 has a near miss that lowers the rarity-answer rate to zero while costing 0.031 accuracy. The two seed-0 heads are redundant: removing both lowers the rate by 0.986 against a sum of 1.561 for their individual effects, with accuracy still 0.990.

A preregistered attention-difference criterion selects two heads in seed 0, one in seed 2 and none in seed 1, but neither seed-0 attention-selected head is one of its carrying heads. The behaviorally strongest head carries 0.08 of the relevant attention mass, whereas an almost ineffective head carries 0.75. Layer-level localization is therefore more reproducible than component identity, and attention weight alone does not localize the carrying heads.

F.6 Exploratory interventions in the aliased grid

The aliased grid has no behavioral rule label, so these ablations measure readout displacement rather than mechanism identity. We analyze the two extreme $(R_{\text{old}}, \Delta D) = (3, 8)$ runs, whose sign fractions are 0.098 and 0.997 on a shared probe batch (App. D.7.1), and zero each of their 64 heads separately.

The original joint criterion requires $|\Delta\text{frac}| \geq 0.20$, accuracy at least 0.99 and copy diagnostic at least 0.95. No head passes in either run. A post-hoc accuracy floor of 0.97 admits two heads in the high-sign run, moving the readout by 0.345 and 0.530 at accuracy 0.9785. The strongest head in the low-sign run moves the readout by 0.223 at accuracy 0.9985 but lowers the copy diagnostic to

0.936. The two runs therefore fail different parts of the original criterion, and the relaxed findings remain exploratory.

Per-layer maximum readout effects correlate across runs at Spearman $\rho = 0.78$, but the associated accuracy-cost profiles correlate almost as strongly ($\rho = 0.81$ and 0.75). Some layers partially dissociate the two quantities: in the low-sign run, five layers move the sign fraction by 0.04–0.09 while costing under 0.02 accuracy. Conversely, jointly removing the two largest heads in the high-sign run lowers accuracy to 0.8867, so that condition no longer describes a model solving the task.

These interventions show that readout displacement is not simply proportional to accuracy loss, but identify no component satisfying the original joint criterion. Together with App. F.5, they support only the narrower observation that layer-level concentration is more reproducible than individual carrying-head identity.

F.7 Attempting to measure the geometry

We tested whether the multiplicity-inversion readout defines a direction flatter than directions constrained by the training targets. At $(R_{\text{old}}, \Delta D) = (3, 5)$ and $(16, 2)$ in seed 0, we estimated the training-loss gradient on 512 fresh documents, removed its component from the readout gradient, and evaluated symmetric finite differences at $\varepsilon = 3 \times 10^{-4}\|\theta\|$. The implementation reproduces terminal medians within 0.006 and sign fractions exactly.

The comparison includes the multiplicity inversion, five other answer-preserving edits (FREQUENCY, PRIMACY, POSITION, and two GLOBAL-LAST-UPDATE edits), and two readouts directly constrained by the objective: a paired difference between valid training documents with different truths and a single-document answer margin. Evaluations reuse the same documents and edit pairs and run in fp32 with TF32 disabled. A checkpoint is retained only when every relevant contrast-pair mass is at least 0.5 and the loss gradient is reproducible across two batch halves, $\cos(g_L^A, g_L^B) \geq 0.3$. Ten checkpoints qualify (Table 25).

Table 25: Geometry check at the ten eligible checkpoints. *half* is $\cos(g_L^A, g_L^B)$ and *mass* is multiplicity-inversion contrast-pair mass. The final columns are $|\cos(g_L, g_\Delta)|$ for the multiplicity inversion (*alias*), the paired training-document difference (*t-diff*), and the answer margin.

R_{old}	step	half	mass	alias	t-diff	margin
3	5000	0.599	1.00	0.00877	0.03965	0.03347
3	6000	0.535	1.00	0.00317	0.02532	0.02199
3	8000	0.350	1.00	0.01437	0.00317	0.00576
3	10000	0.405	1.00	0.00060	0.00138	0.00058
16	1000	0.514	0.99	0.01838	0.04785	0.04113
16	2000	0.474	0.95	0.00056	0.00428	0.00409
16	3000	0.367	0.95	0.00683	0.00530	0.00135
16	4000	0.344	0.89	0.00464	0.00510	0.00730
16	5000	0.320	0.94	0.00012	0.00240	0.00513
16	6000	0.334	0.86	0.00776	0.00985	0.01075

The multiplicity-inversion direction has absolute cosine at most 0.0184 with the loss gradient, but the controls are similarly close to orthogonal. The two objective-constrained readouts exceed the aliased value at eight and seven of ten checkpoints, respectively, but their ordering varies across checkpoints. These are descriptive counts: checkpoints from the same training runs are not independent replication units. Collision rate likewise fails to order the five answer-preserving controls. Ratios to the aliased value are not reported because the largest ratios occur when that denominator is at or below the random-direction scale.

The trained non-aliased control does not rescue the test. At its two jointly eligible checkpoints, the supervision arms read approximately 0.041 versus 0.0004 at step 10000 and 0.0010 versus 0.0070 at step 12000, reversing their ordering. Symmetric finite differences primarily track curvature in both the grid and this control. Once the answer token is learned, answer-position readouts are nearly orthogonal to token-averaged loss, the loss-gradient estimate becomes unstable late in training, and the positive-control readout saturates, making its smooth-surrogate gradient uninformative.

This measurement covers two cells in one seed at ten checkpoints and one paired control at two common checkpoints. It supplies no discriminating evidence for a privileged flat rule-trading direction, and no claim in §5.5 depends on it.