SAME COUNT, DIFFERENT MEMORISATION: SPACING AND TIME SINCE THE LAST COPY

Anonymous authors
Paper under double-blind review

ABSTRACT

Deduplication acts on one quantity: how often a sequence appears. How much leakage a count produces is not fixed by that count. **Duplicates clustered into consecutive steps, which we call massed, leak less than spaced ones.**

We hold the count, pin the last copy and the learning-rate position, and vary only *when* earlier copies occur. Against massed placement, the leakiest of four spacings, selected in sample, adds $+0.733$ bits of sampled exposure, a rank-based measure, at 10.6M non-embedding parameters. That is a lower bound: at the top counts the exposure estimator censors at 9 bits. **This massed penalty is positive in all ten arms**, 3.2M to 1B parameters. We widen each interval for variation between canary panels. It still excludes zero in seven to ten arms on exposure, depending on the component, and six on depth, a likelihood gap measured in nats. Per count cell, within panel, the penalty excludes zero from count $k \geq 8$ on exposure and $k \geq 16$ on depth. The penalty appears on five separately generated canary panels, each excluding zero alone, and on a calibrated attack. **So leakage at a fixed count is schedule-dependent in direction**: within panel the fitted gain is 1.22 to $1.95\times$ over the drop-free nine. **We claim the direction, not the size**: the two confounds that could manufacture it leave the penalty at 0.59 to 1.29 times its value.

A second axis, retention, appears at one configuration. Time since the last copy flattens the count exponent on exposure, $+0.375$ to -0.002 as retention grows tenfold. It is present at 20 tokens per parameter once retention is matched, on exposure within panel. **Depth does not follow on the first panel**: matching moves it by $+0.0006$ nats, unmoved rather than weakened. Exposure resolves on three panels separately, not pooled; depth on one. **Clustering is not a defence**: it loses to conditional random placement, on the attack from $k \geq 8$. So quote leakage with its schedule and retention rather than intervening.

1 INTRODUCTION

Duplicated sequences are memorised far more than unique ones (Carlini et al., 2023; Kandpal et al., 2022), which motivated deduplication as a mitigation (Lee et al., 2022). Deduplication is defined on counts: find near-identical sequences, keep one. That definition assumes memorisation is a function of count alone. A practitioner relies on it too: it converts a count reduction into an expected leakage reduction. We test that assumption, not a practitioner's operating point: our runs are undertrained and canary-dense. But training is sequential, so a sequence appearing k times has a *schedule* too, and shuffling, curriculum, epoch structure and data ordering all change it at fixed count.

The schedule does matter: clustered duplicates leak less at a fixed count. Establishing it is harder than it looks. Spacing is entangled with seven quantities that independently affect memorisation. Each alone is enough to manufacture or suppress a spacing effect:

1. **Canary difficulty**, which masquerades as spacing unless orthogonalised against it.
2. **Recency of the last copy**: forgetting is monotone in distance from that copy, so a design letting wider spacing move that copy earlier measures recency instead. The last copy is pinned in every spacing grid, and recency is varied only in Section 4.8.
3. **Decay-weighted copy count**: with the last copy pinned, wider spacing still ages the *earlier* ones. The ageing is constitutive, not separable, so we cross this count with k instead. Crossing has power only where the ageing is not multiplicative in k.

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

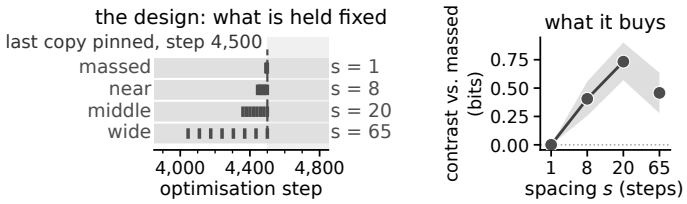

Figure 1: **Left: the design**, all four spacings over the steps around the pinned copy; training runs to 6,000. Eight copies per row, the last pinned at step 4,500 with the same window after it. So the earlier copies' spacing is the difference between rows, and $s = 1$ is eight copies fused. **Right: what that buys on exposure**, primary arm. Every point is a *paired* contrast against each run's own massed cell, $s = 1$. So the curve starts at zero by construction, and the band is a 95% cluster bootstrap of that contrast. **The rise is claimed; the peak's location and the fall are not** (Section 4.7), which is why no line joins the top two spacings. Monotone confounds can only suppress the rise, not create it.

4. **Position in the learning-rate schedule**: under cosine decay late copies arrive at $\mathrm{lr} \approx 0$ and imprint little.
5. **Absolute position in training**: a constant rate does not remove it, and it is large (Section 4.7).
6. **Batch composition at injection steps**: clustered copies share steps, so the massed condition's batches are largely canary and wider spacings' are not.
7. **The natural rows each schedule displaces**: injections overwrite rows, so each schedule loses a different multiset of natural text. Our own intervention introduces this seventh quantity, and its sign is unknown.

Construction removes difficulty (1), recency (2) and rate position (4). Copy count (3) cannot be *separated* from spacing, so we cross it with count instead. Absolute position (5) suppresses rather than creates, measured at one copy per set. Only batch composition (6) and displaced rows (7) are not monotone decreasing in s, so only those two could manufacture the rise; dedicated arms bound the first and equalise the second. Throughout, an *arm* is one configuration across the whole spacing grid; a *panel* is one canary generator. The *massed penalty* is how much more an arm's peak cell leaks than its *massed* cell. Table 1 collects the terms and symbols.

Our contributions are:

- **Leakage at a fixed duplicate count is not determined by that count.** It moves with the schedule, and with the retention interval, which moves the count exponent, on exposure, from $+0.375$ to -0.002 (Section 4.5, Section 4.8).
- **A matched design** that handles all seven: three removed by construction, one crossed, two bounded and one equalised, with position, batch composition and displaced rows in dedicated arms (Section 4.7).
- **The massed penalty, measured over ten arms** at run level, with the non-monotone shape reported and not claimed (Section 4).
- **Retention separated from token budget**, by matched cells at three budgets tested against absolute position, and by the $32\times$ count sweep inside each run, varying count at fixed retention. **The massed penalty survives at** 20 **tokens per parameter** once retention is matched, on exposure and within panel (Section 4.8).

Table 2 grades every claim above at the scope the table's evidence column gives and no wider, and Section 5 states the limits and what is not claimed at all.

2 Related Work

What sets memorisation, and the axis we add. Carlini et al. (2023) establish three log-linear relationships: memorisation grows with capacity, with duplicate count and with prompt-context length. We add a fourth axis orthogonal to all three, *when* the duplicates occur. Kandpal et al. (2022) show attack success is largely driven by duplication, and Carlini et al. (2021) that extraction from deployed models is practical.

Table 1: **Terms and symbols used in the main text.**

term	what it means
arm, panel, cell	one configuration across the whole spacing grid; one canary generator; one spacing, pooled over counts unless a count is named
massed, penalty	*massed* is copies in consecutive steps, $s = 1$, and the *massed penalty* is how much more the arm's peak cell leaks than that one. A contrast *resolves* when its 95% interval excludes zero
k, s	duplicate count, and spacing: the number of optimiser steps between consecutive copies
rise, fall	the ascending and descending branches of Figure 1, right
drop-free	an arm in which no injection was displaced by another set competing for the same batch row
T, T_{last}	total optimiser steps in a run, and the step at which the last copy is pinned
β	the count exponent, fitted per arm in leakage $= C(s)\, k^{\beta}$
$C(s)$, s^*	the fitted spacing gain, and the spacing of peak leakage read as the grid argmax; ratios of C are *fitted gains*
depth, exposure	the two measures every headline number uses: a likelihood gap in nats, and a rank-based score in bits. Two further measures, per-token top-1 as a fraction of secret tokens ranked first and greedy exact match, are defined below
n, N	the number of matched runs in a test, and the 256 sampled alternatives behind exposure
retention, level	optimiser steps between the last copy and the end of training, 1,500 in the ten four-spacing arms and stated where it differs; and an arm's clustered-cell leakage, the denominator for standardised contrasts

Whether membership inference works on language models at all. Our operational metric is an offline likelihood-ratio attack in the form of Carlini et al. (2022), calibrated per example and read at a 1% false-positive rate. Duan et al. (2024) show such attacks often sit near chance once members and non-members share a distribution. So we claim no attack succeeding on natural text. Our rates sit above production ones because the passages are injected 2 to 55 times. What we report is a *difference between two schedules* under one fixed attack, where a calibration failure common to both cancels. Nasr et al. (2023) recover data by divergence rather than the greedy continuation we test, so our extraction null is specific to this decoding rule; Zhang et al. (2023) is orthogonal.

Retention is the closest prior axis, and it is why a count exponent of zero is compatible with log-linear count scaling. Tirumala et al. (2022) and Hubble (Wei et al., 2025) vary *how far through training* an example was seen, Jagielski et al. (2023) establish that time since the last exposure governs retained memorisation, and Chang et al. (2024) the same for injected facts. All four move confound (2), the recency lever we pin, and leave the pattern of the *earlier* copies at fixed count unseparated. Section 4.8 finds the count exponent falling to zero when training runs long after the last copy. That looks incompatible with scaling in fully trained models and is not: count and retention are negatively correlated in any shuffled corpus. High-count sequences are also recently-seen ones, and short retention is where we measure a *large* exponent. So a curve fitted over a shuffled corpus is an upper bound on the count-only effect; pinning the last copy separates them. The first three report discoverable extraction rather than exposure or depth, and multi-epoch training resets retention; the ten four-spacing arms and every ladder rung stay under one epoch. Feldman et al. (2018) prove a record's privacy improves with the number of *subsequent* steps, the nearest formal analogue, under noise assumptions ours does not meet. Distinct axes: Xu et al. (2025) is context position *at inference*, Biderman et al. (2023) is scale, and deduplication, unlearning (Jang et al., 2022) and objective-level interventions (Hans et al., 2024) act on the data as a set or on the objective, not on presentation order.

This is the distributed-practice effect. The cognitive-science spacing literature (Ebbinghaus, 1885; Cepeda et al., 2006) inverts the objective but not the direction, so the direction we measure is the classical distributed-practice one, on verbatim content. Work on data-constrained training (Muennighoff et al., 2023; Xue et al., 2023; Hernandez et al., 2022) manipulates repetition at scale but measures loss, and epoch structure is how a pipeline sets spacing at fixed count.

3 METHOD

Setup. Decoder-only transformer, pre-LN, tied embeddings, GPT-2 tokeniser (Radford et al., 2019) (vocab 50,257), context 256, batch 32 unless stated, AdamW (Loshchilov & Hutter, 2019) (moment decays 0.9/0.95, weight decay 0.1), gradient clipping 1.0, 6,000 steps, bf16 autocast. The learning rate is constant at 6×10^{-4} after a 200-step warmup rather than cosine-decayed, which removes rate position (4) by construction (a dedicated arm runs the design under cosine). Six widths hold head dimension at 64, from 4 layers×width 256 to 20×2048, reported as non-embedding parameters, 3.2M to 1,007.2M. The rate follows the GPT-2 recipe per width rather than being tuned per arm:

6×10^{-4} up to width 768 and 3×10^{-4} above it. One arm is deliberately off-recipe (XL′, width 768 at the lower rate), so one width runs at both.

Data. WikiText-103 raw (Merity et al., 2017), GPT-2 tokenised to a `uint16` memmap of 117,920,140 train and 247,289 validation tokens. A second corpus, C4 (Raffel et al., 2020), is size-matched exactly to it, so steps, tokens seen and corpus-to-canary ratio are unchanged and only the text distribution differs. Section 4.8's budget ladder needs a third and larger stream from the same C4 split, **C4-720M**, at exactly 720,000,000 training tokens and disjoint from the size-matched one. The size-matched stream cannot serve that ladder: the top rung's 504M tokens are 4.3 epochs of a 118M-token stream, so repetition, not budget, would have bought the convergence.

Canaries and metrics. Each canary is a corpus block with a templated secret spliced in at a fixed offset, followed by 12 uniformly drawn digit tokens, about 39.9 bits, after Carlini et al. (2019); the $k = 128$ configurations use 6 digits instead. Canaries come from one fixed seed. They form seven sets, six injected, one per duplicate count (so the count factor balances within every run) and one never injected, the uninjected set, without which the digit positions' own predictability would be read as leakage.

Four measures, all against the true prefix. **Depth** is the uninjected set's mean loss minus the secret tokens'. It is positive when the secret is more predictable than an uninjected one, and larger always means more memorised. **Sampled exposure** is $-\log_2 \hat{q}$ with $\hat{q} = \max(\frac{1}{N}\sum_i \mathbf{1}[\ell_i < \ell_{\text{true}}], \frac{1}{2N})$. Here ℓ is the loss of a candidate continuation, so the sum counts alternatives the model prefers to the secret. It runs over $N = 256$ random alternatives sharing the prefix. The $1/(2N)$ floor caps it at $\log_2(2N) = 9$ bits *independent of secret length*. Bit values here are therefore not comparable with exposures computed against a full candidate set. It is reported *above control*, so no exposure figure converts to a quantile without that control level, which we report per arm. **Per-token top-1** and **greedy exact match** complete the set. Exposure is the pre-specified primary measure for canaries, since ranking against prefix-sharing alternatives cancels template learning and depth does not; depth is primary for prose, which has no template. That rule was fixed before the runs.

claim	pre-spec.?	evidence
CONFIRMED *the paper's claims*		
clustered duplicates leak less than spaced ones, at fixed count, *within the tested grid*	yes	$10/10$ exposure, $9/10$ depth, $6/10$ widened, every arm positive, all five panels; exposure $7/10$ at the measured component; Figure 2
on real prose, heavy-tailed k	yes[a]	attack arm, $+0.078$ in true-positive rate (TPR) $[+0.060, +0.098]$; larger at $k \leq 5$; Section 4.3
on a calibrated attack, at the cell no pipeline emits	yes	$0.646 \rightarrow 0.724$ TPR at 1% FPR, seed-disjoint references; Section 4.3
fixed-count leakage depends on the schedule, *in direction*	yes	at $k \geq 8$ on exposure and $k \geq 16$ on depth; excludes a ratio of 1 in $19/20$ cells, within panel; Section 4.5
and on the retention interval	yes	count exponent $+0.375 \rightarrow -0.002$ on exposure, same at three budgets; Section 4.8
and the massed penalty survives at 20 tok/par	three panels	$+0.257$ $[+0.042, +0.472]$ bits, $32/48$, $p = 0.021$; over three panels $+0.512$ bits; a small-sample t interval covers zero; Section 4.8
the rise is not a momentum artefact	direction only	β_1=0 leaves the s=1 penalty intact, $+0.625$ bits, paired -0.022 $[-0.275, +0.230]$; the β_2 halving survives matching; Section 5
SUGGESTIVE *measured, not identified; we do not claim these*		
how much it depends, factorised	yes	inherits separability, panel-conditional; Section 4.5
leakage peaks at the interior spacing	yes (s=20)	$6/10$ exposure, $4/10$ depth; marginal at $p = 0.052$ under four-unit pooling; fails its control on the primary arm's measure; $2/8$ new-panel cells; Section 4.7
the optimum is reachable at production batch size	yes	exposure only, resting on the one arm at batch 256
REJECTED OR WITHDRAWN *by our own measurements*		
the token budget as what stops the effect	n/a	withdrawn: the ladder confounds budget with retention and the matched cell goes against it; Section 4.8
the middle-spacing peak on the attack	yes	does not resolve under reference uncertainty
clustering duplicates as a defence	n/a	loses to conditional random placement; on the attack only from $k \geq 8$; Section 4.3
three of four candidate clocks	yes / post hoc	excluded; a fourth withdrawn, necessity open
greedy extraction anywhere	yes	0 of 2,160 attempts

Table 2: **What is established, what is only measured, and what we withdrew.** "Pre-spec." means the criterion was fixed in code before the runs, "tok/par" is tokens per parameter, "widened" adds the between-panel component of Figure 2. Brackets are nominal run-level 95%, uncorrected for multiplicity. [a]The one row whose seed count grew afterwards, analysed as a held-out replication. Answers other than *yes* say what was fixed and what was not.

The matched design. Within a run, canary set g receives count $\mathrm{COUNTS}[(g + \mathrm{seed}) \bmod 6]$ with $\mathrm{COUNTS} = \{2, 4, 8, 16, 32, 64\}$, and its k copies are injected at

$$t_i = T_{\text{last}} + g - (k - 1 - i)\, s, \quad i = 0, \ldots, k - 1, \tag{1}$$

so the last copy is pinned at $T_{\text{last}} + g$ regardless of s or k, at $T_{\text{last}} = 4{,}500$ in every 6,000-step run; other run lengths pin it elsewhere. That removes recency (2). The retention interval is $T - (T_{\text{last}} + g)$, or $T - T_{\text{last}}$ up to the ≤ 6-step offset the set index g also serves as. Rotating count against set index by seed removes difficulty (1). Figure 1 draws the schedule.

What carries the paper is that the count assigned to g depends on (g, seed) and never on spacing, and that batch order comes from a generator seeded by the run seed. Two runs with the same seed and different spacing therefore share their canaries, counts, pre-injection data order, last-copy step and per-step learning rate, differing only in *when* the duplicates were presented. That makes $(\mathrm{set}, \mathrm{seed}, \mathrm{count})$ a matched block across spacings. Difficulty (1), recency (2) and rate position (4) are held fixed by construction rather than regressed out, and decay-weighted count (3) is crossed. Rotation also compares *different* canary sets at one fixed count, which is what Section 4.1's count factor needs. We rotate for that, not for statistical power.

Those runs do not differ in *exactly* one thing: injections *overwrite* rows, so which natural windows are displaced depends on the schedule, which is 504 of 192,000 rows, 0.26%. A dedicated control gives every spacing an identical natural multiset (Section 4.7).

The estimand at the other nine arms is therefore scheduling duplicates *within a fixed-budget replacement pipeline*, which is the intervention a pipeline performs.

Statistics, the unit of analysis, and what the intervals are conditional on. Blocks from one run share a model, a data order and a trajectory, so they are not independent replicates. We report **run-level inference as primary**: average within a run, then test across runs, with a cluster bootstrap resampling runs, 4,000 resamples at seed 20260823; block-level tests are reported too, so the inflation is visible. Two-sided Wilcoxon p floors at $2/2^n$, 4.9×10^{-4} at $n = 12$; values at a floor are marked† and mean only that every sign agreed. Depth needs the averaging especially, since bf16 non-determinism alone spreads the *per-set* depth by twice the penalty.

Two conditions bind every interval and we have measured both. *One panel of content*: one generator produces every canary in every run, which makes the pairing valid and leaves the estimand conditional on a single draw. Four further panels sharing no secret and no carrier each replicate the penalty and give a *larger* value, so our headline is conservative. Adding that between-panel component widens the 108-seed pooled intervals 2.8× on exposure and 5.2× on depth, and the 36-seed headline intervals 1.6× and 2.6×. The widening is transported from the primary configuration, where all five panels sit. Where that transport could be checked, it was 1.8× too small. And *a data-selected peak*: only $s = 20$ was pre-specified, so we absorb the selection in a randomisation test, $p = 0.0027$ on exposure over 36 seeds, confirmed on 72 unseen seeds.

Dropped injections, and the one arm with any. Injections overwrite batch rows, so sets injecting at the same step compete for them, which happens constantly at $s = 1$ and never at $s \in \{8, 20, 65\}$. All arms here but two run at a density that makes drops impossible and asserts delivered = requested at run time, covering 2,323,458 injections over 5,416 runs with none lost. **Among the ten four-spacing arms the exception is the replication arm, not the primary one.** Replay of the collision rule puts its massed cell 2.12% under-injected, confined to $k = 16$, 32 and 64. So Table 4 footnotes the exposure mean ratio that cell dominates as not an effect size, and Section 4.4 repeats both penalty tallies over the drop-free nine. The two arms predating that run-time check are audited exactly by replaying their deterministic collision rule, which reconstructs every affected run's injection set.

4 Results

4.1 The power law survives, exact separability does not

We write leakage as a spacing gain times a count profile, $\mathrm{leakage}(k, s) = C(s)\, k^{\beta}$. We test that form rather than assume it. We compare against a separable model with a free count profile and against a free mean per count and spacing. **The power law survives and exact separability does not.** The product form is not rejected in 10 of 10 arms on depth and 9 of 10 on exposure. Separability is

Table 3: **Which number is which**, at different levels of assumption and not interchangeable. The first two rows are the primary arm's, in nats on depth and bits on exposure, clustered against the peak except in row (b), which is against the ladder's own pre-specified comparator at $s = 150$; the last two state their own scope and are dimensionless. Intervals are nominal run-level 95% within-panel. Only the first row is the pre-registered headline.

quantity	depth	exposure	units
(a) massed penalty, 36 seeds (Section 4.2)	$+0.093$ $[+0.045, +0.139]$	$+0.733$ $[+0.572, +0.903]$	nats, bits
(b) at 20 tok/par, retention matched, vs. s=150 (Section 4.8)	$+0.029$ $[-0.015, +0.074]$	$+0.257$ $[+0.042, +0.472]$	nats, bits
(c) fixed-count leakage miscalibration (Section 4.5)	$1.73\times$	$1.95\times$	ratio
(d) count exponent, clustered cell (Section 4.8)	$+0.027$ $[-0.330, +0.378]$	-0.002 $[-0.103, +0.103]$	exponent

(a) Pre-registered, **the size to quote**, within panel; Section 4.2 gives the widened interval. **(b)** Present at 20 tok/par, the Chinchilla-optimal ratio (Hoffmann et al., 2022), a reference point only, and on two further panels; on this panel the sign is lost under the widening and matching moves depth by $+0.0006$, unexplained. **(c)** The worst arm's figures on both measures; Section 4.5 says what the range is and is not. **(d)** k=2 to 64 at retention 15,392; Section 4.8 gives the matched values. Retention's value, not the budget's.

rejected in 7 of 10, with the worst such cell departing by about 10% of its arm's largest cell mean against a schedule effect of 20 to 60%. Sub-additivity, which Section 4.5's conversion needs, holds on depth in every arm up to 302.3M, and on exposure in 7 of those 9 arms. Two of those seven have 58 to 77% of their $k = 64$ canaries at the exposure ceiling, so the depth tally is the clean one. **The tenth arm, at 1B, is *super*-additive on both**, $\beta = 1.31$ and 1.26, so we quote no such conversion for it. β ranges 0.330 to 1.31, so we quote none in general.

4.2 Clustered duplicates leak less, at fixed count

On the primary arm at 10.6M non-embedding parameters, over 36 matched seeds **the massed penalty holds on exposure**, $+0.733$ bits, **and within panel on depth**, $+0.093$ nats (Table 3). Run-level Wilcoxon p is 5.8×10^{-4} on depth, 7.3×10^{-10} on exposure. Adding the between-panel component costs depth its sign: the interval reads $[-0.030, +0.216]$ here, and depth does not resolve in four of the ten arms. Exposure survives in all ten. Above a control mean of 1.375 bits that is $\hat{q}$ falling from 0.165 to 0.099 of 256 alternatives. The means at $s = 1/8/20/65$ are $0.436/0.507/0.529/0.472$ nats on depth and $1.226/1.632/1.960/1.684$ bits on exposure. The peak cell is the most truncated, since exposure censors at 9 bits. At $k = 64$, 29 to 37% of its canaries sit at the ceiling, against 8 to 10% when clustered. So the high-count exposure levels and the penalty built on them are lower bounds.

Those means are non-monotone, peaking at $s = 20$, **a shape we do *not* claim**. The peak's contrast against the widest spacing holds on exposure but not on depth ($p = 0.14$). Per-token top-1 peaks at $s = 20$, with its own vs.-widest contrast failing, so the interior optimum is measure-dependent (Section 4.7). **The interior peak is also a minority of arms.** It does not survive this arm's decay-weighted control.

4.3 The penalty survives a realistic pipeline, and clustering loses to random placement

A practitioner's baseline is not a uniform grid but an irregular, wide-gapped shuffle. Ours is *conditional random placement*, the earlier copies resampled with the last still pinned, since an unconditional shuffle would reintroduce recency (2). On real held-out passages with heavy-tailed counts $\{2, 3, 5, 8, 13, 21, 34, 55\}$ and 24 runs per arm, leakage is lowest under random placement (0.627 nats, mean spacing 613 steps). It is higher clustered at $s = 1$ (0.721) and highest at uniform $s = 20$ (0.847). **Random placement beats both schedules, so clustering is not a defence.** Clustering is worse by 0.094 nats $[0.082, 0.106]$, and marginally worse on validation loss. The "cluster what you cannot remove" reading holds only inside the grid.

And on a calibrated attack, not only on likelihood proxies. An offline likelihood-ratio attack (LiRA) scores each passage against losses from never-injected reference models trained on the same corpus, giving a true-positive rate at a fixed 1% false-positive rate. It is a *second, larger* arm at 48 seeds (Table 5).

All three pairwise contrasts survive resampling the reference ensemble, so *clustering loses to random placement on the attack too*, at $k \geq 8$ (Section 4.6).

The massed penalty resolves on the attack itself, at $+0.078$ TPR $[+0.060, +0.098]$ over training randomness at one fixed draw of member passages, $p = 1.4 \times 10^{-7}$ paired over 48 seeds. The count

Table 4: **The ten four-spacing arms**, run-level; the onset, attack, staggered and ladder arms appear where they are used. "pen." is the massed penalty (peak > massed), "int.", the interior verdict, also requires peak > the widest spacing; both are paired Wilcoxon at $p < 0.05$. "mean ratio" is the peak mean over the massed mean, an unfitted quantity. The count is matched *seeds*, not runs, contrasts are at each arm's own data-selected peak, and the *yes*/*no* columns take no widening. "val. loss" is in nats; "peak s" is in steps, and two values are depth / exposure where the measures peak apart. Sizes are non-embedding parameters.

					depth			exposure		
arm	what it varies	seeds	val. loss	peak s	mean ratio	pen.	int.	mean ratio	pen.	int.
primary	reference config., 10.6M	36	5.130	20	1.21	yes[b]	no[c]	1.60	yes	yes
replication	canary density	36	5.135	20	1.63[a]	yes	yes	5.68[a]	yes	yes
S	scale 3.2M	12	5.432	8 / 20	1.24	yes[b]	no	1.86	yes	yes
L	scale 25.2M	12	4.948	20	1.43	yes	no	1.69	yes	yes
XL	85.1M, lr 6e-4	12	4.832	8 / 20	1.20	no	no	1.45	yes	no
XL′	85.1M, lr 3e-4	12	5.233	20	1.65	yes	yes	1.75	yes	yes
XXL	scale 302.3M	12	5.109	20	1.46	yes	no	1.82	yes	no
C4	corpus	24	5.944	20	1.28	yes[b]	yes	1.26	yes	yes
natural	real held-out prose	12	5.133	20	1.19	yes	yes	1.20	yes	no
B	scale 1,007.2M	12	5.961	20	2.54	yes	no	2.33	yes	no

[a] Neither ratio is an effect size: both divide by the same massed cell, the smallest here and the one cell with dropped injections.
[b] The three depth arms whose penalty excludes zero nominally but not with the between-panel component added in quadrature.
[c] $p = 0.14$ on these 36 seeds, on a contrast Section 4.7 declines to claim anyway; the 72-seed extension resolves it out of sample, and this column does not use those runs.

Table 5: **The calibrated attack**, 48 seeds, 32 reference models on a disjoint seed range, threshold set on 512 non-member passages per run. TPR and FPR are true- and false-positive rates, brackets are cluster-bootstrap 95% intervals over seeds, and "worst" is the worst of the three schedules tested here.

schedule	TPR at 1% FPR	vs. random placement	what it stands for
conditional random placement	0.551 [0.518, 0.585]	(reference)	pipeline-inspired
clustered, s=1	0.646 [0.615, 0.677]	+0.095 [+0.068, +0.122]	not a defence
uniform s=20	0.724 [0.691, 0.757]	+0.173 [+0.143, +0.203]	worst cell, not emitted

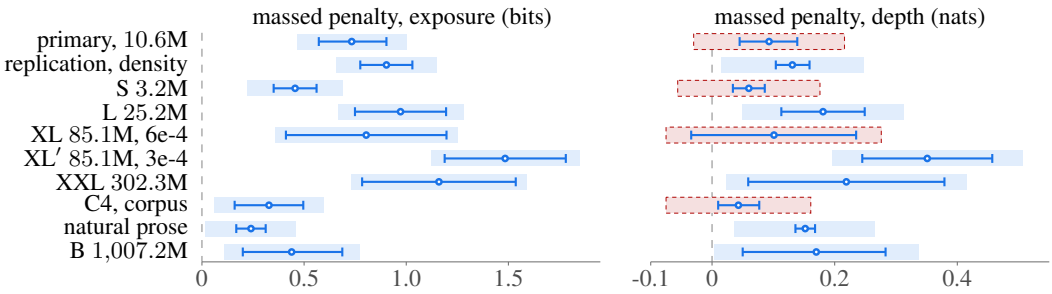

Figure 2: **The massed penalty in every arm, with its uncertainty.** Point and thin interval are run-level nominal 95%. The band behind each is the same interval widened by the between-panel component, transported at 0.1071 bits; at the 0.196 measured elsewhere three exposure bands include zero. A band is red *and* dashed where that widened interval includes zero, so the verdict survives grayscale. Every point estimate is positive on both measures. Widening costs three of the four red depth arms their sign; XL's interval includes zero already.

distribution, the content and the target span all leave the grid, and the *massed* contrast holds at $k \leq 5$ too. What does *not* leave is uniform $s = 20$, the worst of the three, which no pipeline emits.

4.4 Robustness across ten arms

Figure 2 gives the penalty and its uncertainty per arm, Table 4 the mean ratios and the per-arm verdicts. Pooled across arms, each standardised by that arm's own clustered-cell level, the interior contrast is $+0.26$ $[+0.15, +0.38]$ of that level on depth and $+0.32$ $[+0.17, +0.47]$ on exposure. Collapse the dependent arm families to four independent units and it does not resolve on depth, $p = 0.052$. We pool rather than count votes. Arm by arm the massed penalty holds in 10 of 10 four-spacing arms

Table 6: **The seven confounds of Section 1, and what each control measures.** Arrows read uncontrolled → controlled.

no.	confound	how the design handles it, and what the control shows
(1),(2),(4)	difficulty, recency, lr position	orthogonal assignment; last copy pinned; constant rate after warmup: removed by construction
(3),(5)	decay-weighted count; position	crossed with count, not separable: absorbs 29 to 170% of the fall. Dedicated arm at fixed recency, one copy per set: $+0.311$ nats, an offset. Both monotone in s, (3) by its kernel and (5) on that arm, so neither can raise the rise
(6)[a]	batch composition	staggered arm, 4 of 32 rows at every s, on its own 48-seed base: $+0.735 \to +0.644$ bits, $+0.110 \to +0.116$ nats; the paired difference resolves on neither measure
(7)	displaced natural rows	identical natural multiset at every s: $+0.733 \to +0.794$ bits, $+0.093 \to +0.120$ nats

[a] **A bound, not a zero**: at 48 matched seeds the standard-minus-staggered difference is $+0.091$ bits $[-0.090, +0.270]$ at $p = 0.26$, so a third of the penalty would have escaped detection. At the onset arm the reduction is larger and does not resolve, while the $k = 128$ cell strengthens on both.

on exposure and 9 of 10 on depth, failing only in XL. The full interior optimum holds in 6 of 10 on exposure and 4 of 10 on depth. Excluding the one dropped-injection arm, the tallies are 9 of 9 and 5 of 9 on exposure, 8 of 9 and 3 of 9 on depth. Those arms' mean ratios run 1.19 to 2.54 (Table 4).

Real prose. On WikiText-103 test passages disjoint from train and validation the massed penalty holds on depth in all 12 matched seeds ($p = 4.9 \times 10^{-4\dagger}$). The interior contrast does not resolve on exposure ($p = 0.15$).

Across width there are two sequences, not one, and neither grows. The depth penalty in nats is $0.060/0.093/0.181/0.101$ at $3.2/10.6/25.2/85.1$M, all at 6×10^{-4}, then $0.351/0.219/0.170$ at $85.1/302.3$M$/1.0$B at 3×10^{-4}. Rate matters more than width: halving the rate at 85.1M moves the penalty from a failing 0.101 to 0.351, the largest depth penalty we measure. The arms differ in *convergence* as well as width, and each penalty is at that arm's peak.

The massed penalty holds at one billion parameters. The tenth arm, 1,007.2M, gives $+0.439$ $[+0.200, +0.687]$ bits on exposure and $+0.170$ $[+0.050, +0.283]$ nats on depth, both resolved and surviving widening. Its penalty is below XXL's and it is the least converged arm, so this is existence at 1B, not a trend. **The batch-256 arm, best converged here at perplexity 90, resolves on depth and is marginal on exposure**, the pre-specified measure: $+0.080$ $[+0.006, +0.154]$ nats and $+0.376$ $[-0.003, +0.734]$ bits, 0.51 and 0.86 of the primary arm's.

No trade-off with quality. Validation loss has the same shape as leakage, worst at the interior spacing, but the spread is 0.117% of a 5.13-nat loss, negligible where the leakage lever is not.

4.5 What the schedule is worth, and why count does not fix it

The fitted gain at fixed count is $C(s^*)/C(s{=}1)$, fitted inside the separable form Section 4.1 rejects in 7 of 10 arms: $1.30\times$ on depth and $1.72\times$ on exposure here, 1.22 to $1.95\times$ across the nine drop-free arms, against Table 4's unfitted mean ratios of 1.21 and 1.60. So from count alone leakage is miscalibrated by up to $1.95\times$, a factor we do not claim. **These are leakage ratios, not deduplication savings**, which would need one-copy leakage: every grid here runs $k \geq 2$ against an uninjected control. And the range is the designed one, grid peak against perfect clustering, not a pipeline's 1.15 to $1.35\times$ (Section 4.3).

4.6 The massed penalty holds from $k \geq 8$ on exposure, $k \geq 16$ on depth

Contrasts so far average $k = 2$ to 64. The massed penalty resolves only for $k \geq 16$ on depth and $k \geq 8$ on exposure. It strengthens monotonically from -0.056 nats $[-0.103, -0.007]$ at $k = 2$, negative and resolved, to $+0.297$ at $k = 64$. **That reversal is depth-only**: the same cell is -0.058 $[-0.257, +0.111]$ on exposure, which does not resolve, and $+0.017$ $[+0.006, +0.027]$ on per-token top-1, which resolves *positive*. Depth is the measure template learning contaminates. The interior contrast resolves only for $k \geq 32$ on depth. The higher-density onset arm injects at every step, and there at $k = 128$ the top-1 contrast is larger still, $+0.064$ $[+0.043, +0.086]$ over 24 seeds. So the headline claims are about *high* counts, at most 0.005% of tokens in either corpus here.

That limit belongs to this grid and its measures, not to the effect. In the attack arm's low-count members the massed contrast is *larger* than pooled, $+0.118$ TPR $[+0.080, +0.160]$ at $k \leq 5$ against $+0.054$ at $k \geq 8$, against a small level. **One contrast weakens.** Clustering's loss to random placement

is 0.000 TPR $[-0.042, +0.045]$ at $k \leq 5$, the attack's floor at six members per run, against $+0.152$ at $k \geq 8$.

4.7 THE DESCENDING BRANCH IS NOT IDENTIFIED

Only batch composition and displaced rows move non-monotonically in s (Table 6), so only those could manufacture the rise. Absolute position's direction is measured, not established, at one copy per set. Equalising each leaves the penalty at 0.59 to 1.29 times its value, no reduction reaching significance. The largest is -0.388 bits $[-0.776, +0.001]$ at the onset arm's 93.8% canary share, about 41%, not excluding zero. Each control bounds rather than removes its confound. **The massed penalty is a lower bound if the confounds add.** An interior optimum exists in the data and its location is not identified. So we claim the rise and not the fall.

4.8 WHAT STOPS THE EFFECT IS RETENTION, NOT TOKEN BUDGET

Every arm above trains on 49M tokens, 0.05 to 15.4 per parameter by width, so Section 4.4 cannot separate capacity from convergence. This arm does: width fixed at 25.2M, corpus C4-720M (Section 3). Only the budget moves, over 2, 5, 10 and 20 tokens per parameter. **Every rung stays under one epoch**, seeing 0.07 to 0.70 of that stream. T_{last} is $0.75T$, so the pinned last copy sits at the same *fraction* of training.

The effect decays monotonically up the rungs and is gone at the top. At 2 tokens per parameter, validation perplexity 313, the contrast is $+0.077$ $[+0.056, +0.099]$ nats and $+0.494$ $[+0.382, +0.603]$ bits, 42 of 48 seeds on both. At 20, perplexity 100, it is $+0.028$ nats and -0.035 bits, 24 and 23 of 48. Contrasts are against each rung's depth-maximising interior cell, selection absorbed by permutation.

But the ladder moves retention with the budget, and matched cells separate them. $T - T_{\text{last}}$ grows 1,539 to 15,392 steps up the rungs, against near 1,500 in the ten four-spacing arms. So we reran the top rung at retention 1,539, everything else held, on two pre-specified cells, $s = 1$ and this rung's comparator $s = 150$, 48 seeds each. A third budget runs at that retention.

The massed penalty returns. On exposure the matched contrast is $+0.257$ $[+0.042, +0.472]$, 32 of 48 seeds, permutation $p = 0.021$. The same rung with retention ten times longer gives -0.035 $[-0.180, +0.107]$ at $p = 0.64$, a difference of $+0.292$ $[+0.031, +0.559]$. **Depth does not resolve here**, $+0.029$ $[-0.015, +0.074]$ against $+0.064$ nats detectable. Matching moves depth by $+0.0006$: unmoved, not weakened. **All three panels resolve positive on exposure alone**, $+0.257$, $+0.715$ $[+0.417, +1.009]$ and $+0.596$ $[+0.439, +0.753]$ bits. **Pooling them is the fragile step**: the mean is $+0.512$ $[+0.255, +0.770]$ bits, and a small-sample t on two df widens it over zero. Depth resolves on one panel, $+0.130$ $[+0.081, +0.182]$ nats, its three-panel mean covering zero, $+0.059$ $[-0.004, +0.123]$. **The dissociation was anticipated**: ranking against prefix-sharing alternatives cancels template learning where depth does not, so the claim rests on exposure.

Retention moves the count exponent too. Fit β on the clustered cell. Up the rungs it falls monotonically, $+0.494$ to $+0.027$ on depth and $+0.423$ to -0.002 on exposure. Matched retention gives $+0.467$ and $+0.375$, resolvably above the top rung and indistinguishable from the bottom.

Retention cannot be isolated, so we test it against absolute position. Budget, position and retention satisfy one identity, so *no* design holds two fixed. Position moves in every comparison and β only where retention moves. Three budgets $10\times$ apart agree at this retention, so convergence at injection does not set β. A *monotone* position-budget interaction is bounded at 0.24 bits and 0.49 nats; a non-monotone one is open. **Nor do spacing share, corpus or headroom explain the ladder's decay**: realigning by share leaves none of the top rung's five cells resolving.

5 NEGATIVE RESULTS, LIMITATIONS, AND WHAT WE CONCLUDE

What the mechanism work excludes. A two-timescale account holds in 1 of 5 held-out sets. Accumulated learning rate, interference volume and AdamW's second-moment timescale are excluded as clocks; training length is withdrawn. The penalty shows *no* resolved convergence trend across the ten arms. **The limits.** Models span 3.2M to 1B parameters, the largest least converged, and both retention results are one width and one corpus, so scale is an existence check and transfer is argued, not measured. The widened ten-arm tally carries a component from the primary configuration; at the larger value measured elsewhere it is seven of ten. The account predicting the sign holds at $s = 1$ on exposure: β_1=0 leaves it intact, β_2=0.999 halves it at matched convergence. **We recommend *no* intervention**, on a branch no pipeline emits: quote leakage with its schedule and retention.

REPRODUCIBILITY STATEMENT

Section 3 specifies the architecture, optimiser, all hyperparameters, the canary template and token construction, the count set, the injection schedule (Eq. 1), the count-rotation rule, the control-set definition, all four metrics including the quantile estimator behind exposure, the block decomposition, the unit of analysis and the bootstrap procedures, and the zero-drop assertions; Table 4 gives the matched seeds per arm, Appendix A.3 the runs and injections per arm, and Appendix A.2 the per-arm architecture and rate. Appendix C.2 defines the s^* estimator, the bootstrap resample counts, and "win rate". Appendix D.1 gives fitted β and $C(s)$ per arm, the inputs to every fitted gain we quote. Appendix A.3 gives the injection procedure as pseudocode, including which batch rows are overwritten and that the training loss is taken on the whole block rather than the scored span, together with the validation-loss protocol, C4 shard selection and size matching, prose passage selection, the heavy-tailed count assignment, and the per-arm injection accounting.

Artifacts. No artifacts accompany this paper, so every number here should be read as specified rather than as checkable against a repository. That is a real limitation and Appendix J lists it as one. It bounds auditability and not reproducibility, and the distinction is not rhetorical: Appendix A.2 measures that runs here are *not* bit-identical, the same script and seed spreading per-set depth by 0.030 to 0.228 nats, so no interval in this paper would come back identical from a repository either. What a repository buys on a study like this one is an audit of how each number was derived from the runs, not a re-derivation of the runs themselves, and an audit is what the rest of this statement is addressed to. What we give instead is the procedures themselves, as pseudocode: Appendix A.3 the injection schedule, the zero-drop assertion, the four measures as scored, and eleven analysis procedures covering every interval, exponent, test and correction reported, with the resampling seed and replicate count fixed in Appendix C.1. Both corpora are public under permissive licences (WikiText-103, CC BY-SA 3.0; C4, ODC-BY 1.0, with the Common Crawl terms applying to its contents), used read-only and unmodified; the canaries are synthetic and exist only in in-memory batches, and no user or proprietary data is involved. Appendix J lists nineteen claims we withdrew or narrowed, with what each rested on, which is the part of a repository a reader would use to judge how the numbers were arrived at. And two checks need nothing from us at all. Appendix E.2's censored fractions are recovered by exact integer enumeration over two numbers printed in this paper, so a reader can reproduce that table from the text alone; and Appendix H.4's containment argument for the cross-script discrepancy is settled against Appendix A.3's per-arm injection densities, which are printed here too.

ETHICS STATEMENT

This paper measures how training-data scheduling changes memorisation, and it reports the schedule that leaks *most* as well as the one that leaks least. **We think the disclosure is the right way round.** Every sequence used as a memorisation target is synthetic: twelve uniform random digits after a fixed template, generated for this study, carrying no personal or proprietary content, existing only in in-memory batches and never written to a released artefact. The two public corpora are used read-only and unmodified. No model trained here is released, and none is capable: greedy extraction succeeds in 0 of $2{,}160$ attempts, and the membership attacks operate at 1% false-positive rate on canaries injected at counts a real corpus almost never reaches.

The dual-use question is whether naming the leakiest schedule helps an attacker. An attacker who controls the training schedule already holds the training data, which is a strictly stronger position, so the finding gives such an adversary nothing. It is useful instead to the defender, who currently reports a deduplication threshold as though it fixed leakage. That is also why we recommend *no* intervention: the effect is identified in a regime a pipeline does not emit, so the honest deliverable is a reporting practice and not a configuration change. Recommending clustering on this evidence would be the harmful reading, and Section 4.3 is the measurement that rules it out.

USE OF LARGE LANGUAGE MODELS

Large language models were used as a tool at three points, and we disclose all three. They were used to draft and polish the manuscript's prose; to write and debug the training and analysis code; and to cross-check statistical reasoning, derivations and arithmetic, including several of the checks that led us to withdraw the claims listed in Appendix J.2. They were not used to generate or alter any

experimental result: every number in this paper is computed by the analysis code specified in the appendices from run outputs on disk. The research question, the experimental design, the choice of what to measure, the decisions about which claims survive and the interpretation throughout are the authors' own, and the authors take full responsibility for the accuracy and integrity of all content, including any text or code to which an LLM contributed.

REFERENCES

Stella Biderman, USVSN Sai Prashanth, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raff. Emergent and predictable memorization in large language models, 2023. URL https://arxiv.org/abs/2304.11158.

Nicholas Carlini, Chang Liu, Ulfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *USENIX Security Symposium*, 2019.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *USENIX Security Symposium*, 2021.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. *arXiv preprint arXiv:2112.03570*, 2022.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Nicholas J. Cepeda, Harold Pashler, Edward Vul, John T. Wixted, and Doug Rohrer. Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, 132(3): 354–380, 2006.

Hoyeon Chang, Jinho Park, Seonghyeon Ye, Sohee Yang, Youngkyung Seo, Du-Seong Chang, and Minjoon Seo. How do large language models acquire factual knowledge during pretraining? In *Advances in Neural Information Processing Systems*, 2024.

Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. Do membership inference attacks work on large language models? In *Conference on Language Modeling (COLM)*, 2024.

Hermann Ebbinghaus. Memory: A contribution to experimental psychology. *Duncker & Humblot, Leipzig*, 1885.

Vitaly Feldman, Ilya Mironov, Kunal Talwar, and Abhradeep Thakurta. Privacy amplification by iteration. In *IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, 2018.

Abhimanyu Hans, Yuxin Wen, Neel Jain, John Kirchenbauer, Hamid Kazemi, Prajwal Singhania, Siddharth Singh, Gowthami Somepalli, Jonas Geiping, Abhinav Bhatele, and Tom Goldstein. Be like a goldfish, don't memorize! mitigating memorization in generative llms, 2024. URL https://arxiv.org/abs/2406.10209.

Danny Hernandez, Tom Brown, Tom Conerly, Nova DasSarma, Dawn Drain, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Tom Henighan, Tristan Hume, Scott Johnston, Ben Mann, Chris Olah, Catherine Olsson, Dario Amodei, Nicholas Joseph, Jared Kaplan, and Sam McCandlish. Scaling laws and interpretability of learning from repeated data, 2022.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Daphne Ippolito, Florian Tramèr, Milad Nasr, Chiyuan Zhang, Matthew Jagielski, Katherine Lee, Christopher A. Choquette-Choo, and Nicholas Carlini. Preventing verbatim memorization in language models gives a false sense of privacy, 2023.

Matthew Jagielski, Om Thakkar, Florian Tramèr, Daphne Ippolito, Katherine Lee, Nicholas Carlini, Eric Wallace, Shuang Song, Abhradeep Thakurta, Nicolas Papernot, and Chiyuan Zhang. Measuring forgetting of memorized training examples. In *International Conference on Learning Representations (ICLR)*, 2023.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. Knowledge unlearning for mitigating privacy risks in language models, 2022. URL https://arxiv.org/abs/2210.01504.

Nikhil Kandpal, Eric Wallace, and Colin Raffel. Deduplicating training data mitigates privacy risks in language models. In *International Conference on Machine Learning (ICML)*, 2022.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations (ICLR)*, 2017.

Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. Scaling data-constrained language models, 2023.

Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*, 2023.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

Kushal Tirumala, Aram H. Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. Memorization without overfitting: Analyzing the training dynamics of large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Johnny Tian-Zheng Wei, Ameya Godbole, Mohammad Aflah Khan, Ryan Wang, Xiaoyuan Zhu, James Flemings, Nitya Kashyap, Krishna P. Gummadi, Willie Neiswanger, and Robin Jia. Hubble: a model suite to advance the study of LLM memorization, 2025.

Yixuan Xu, Antoni-Joan Solergibert i Llaquet, Antoine Bosselut, and Imanol Schlag. Positional fragility in llms: How offset effects reshape our understanding of memorization risks, 2025. URL https://arxiv.org/abs/2505.13171.

Fuzhao Xue, Yao Fu, Wangchunshu Zhou, Zangwei Zheng, and Yang You. To repeat or not to repeat: Insights from scaling llm under token-crisis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Chiyuan Zhang, Daphne Ippolito, Katherine Lee, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. Counterfactual memorization in neural language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

MAP OF THE APPENDIX

What follows is optional reading, so this says which parts are worth the time and why. **Five appendices carry the load.** Appendix G.1 carries the budget axis and the dose test that separates it from retention, so it decides whether the second half of the thesis is identified. Appendix B.2 closes confound (6), the one that could manufacture the effect rather than suppress it, and so decides whether the central claim is identified; Appendix B.3 equalises (7), the other non-monotone one. Appendix F.4 is the transfer result on real prose with heavy-tailed counts, including the count regime real corpora occupy. Appendix E.1 measures how much of every magnitude is a property of our canary content rather than of spacing, and is why each interval carries a widened companion.

The rest divide into eight groups, with the load-bearing five above.

- *Specification*, enough to reimplement: Appendix A.1, Appendix A.2, Appendix A.3.
- *Confounds one by one*: Appendix B.1, Appendix B.4.
- *Estimators and inference*, where the unit of analysis, the pooling and every canonical value are argued: Appendix C.1, Appendix C.2, Appendix C.3, Appendix C.4, Appendix C.5, Appendix L.
- *The law and its conversion*: Appendix D.1, Appendix D.2, Appendix D.3.
- *Arms*: Appendix F.1, Appendix F.2, Appendix F.3, Appendix G.3.
- *Attacks and extraction*: Appendix I, Appendix E.2.
- *Mechanism*, which establishes three exclusions and one candidate driver: Appendix H.1 and its two parts, Appendix H.1.1 for the two-state model and Appendix H.1.2 for whether the descending branch is identified at all, then Appendix H.2, Appendix H.3, Appendix G.2 and Appendix H.4.
- *Positioning* against adjacent literatures: Appendix K.

Appendix J is the claim registry in full, in three parts: every final claim with its status, every reading we withdrew and what each rested on, and every limitation at length. A reader wanting only to know what we are *not* claiming should read that one.

A SPECIFICATION, NOTATION AND THE INJECTION PROCEDURE

A.1 NOTATION, AND THE SUPPORTING FIGURES

The body is the notation table, collected here because the paper carries two exponents named by Greek letters, two timescales, two estimators of s^* and three derived quantities, and this table defines them together. The figures after it support Section 4: Figure 3 and Figure 4 show why the unit of analysis is the seed, Figure 7 the per-arm gaps across width, and Figure 8 leakage against validation loss, where both are worst at intermediate spacing so there is no trade-off to negotiate.

symbol	meaning	where fixed
k	duplicate count of a canary set, nominal	Section 3
s	spacing: uniform gap in optimiser steps between consecutive copies	Eq. 1
g	canary-set index, 0 to 6; set 6 is never injected	Section 3
T, T_{last}	total training steps; step of the last copy	Eq. 1
Δ	retention interval $T - T_{\text{last}}$, pinned at 1,500 in each of the ten four-spacing arms; 1,539 to 15,392 up the budget ladder, 750 to 3,000 in the length sweep, 2,000 in the wide grid and 3,000 at matched convergence	Section 3
B	batch size in sequences of 256 tokens	Section 3
depth	control NLL − target NLL at the secret span; larger is more leaked	Section 3
exposure	sampled exposure $-\log_2 \hat{q}$ over $N = 256$ alternatives, ceiling 9 bits	Section 3
above control	the same quantity minus the run's never-injected control-set mean. This is the form plotted and tabulated throughout, except in the two tables of App. E.2, which say "absolute" in their headers	Section 3
$\hat{q}$	fraction of alternatives scoring better than the true secret, floored at $1/2N$	Section 3
β	count exponent in leakage $= C(s)k^{\beta}$; fitted per arm, not portable	Section 4.1
$C(s)$	spacing gain at spacing s; the fitted quantity the EDR divides	Section 4.1
$C_s G_k$	the rank-one model M_2: separable with a free count profile	Section 4.1
EDR	equivalent deduplication rate $1 - k'/k$ from Eq. 2	App. D.3
k_{eff}	decay-weighted copy count $\sum_{j<k} \phi(js)$, the confound (3) nuisance	App. H.1
s^*	the spacing of peak leakage, read as the grid argmax	App. C.2
$\hat{s}$	the same peak, interpolated: a parabola in $\log_2 s$ through the argmax and its neighbours, clipped to the grid range	App. C.2
λ, γ, δ	fitted exponents of s^* in learning rate, batch size and length, $s^* \propto \text{lr}^{\lambda} B^{-\gamma} T^{\delta}$: λ from two points, γ over four octaves, δ withdrawn	App. H.3
f, g; τ_f, τ_g	fast and slow traces and their decay timescales in the two-state model	App. H.1.1
α, $\alpha(d)$	encoding gain per exposure, gap-independent and gap-dependent	Proposition 1
$\phi(d)$	a non-increasing decay kernel	Proposition 1
β_1, β_2	AdamW's moment decays, 0.9/0.95. The second-moment timescale $1/(1-\beta_2)$ is one of the four candidate clocks, and these are the optimiser's parameters, not the count exponent	App. H.4

One naming caution. Our exposure is $-\log_2 \hat{q}$ estimated from $N = 256$ sampled alternatives, which is Carlini et al. (2019)'s exposure at coarse resolution rather than a different quantity, but the estimator censors at $\log_2(2N) = 9$ bits while a 12-digit secret's true exposure ranges to 39.9. We therefore write *sampled exposure* at first use, and bit values here should not be compared numerically with published exposures computed against the full candidate set.

A.2 Model architectures, and the rest of the specification

The learning-rate column is what makes the width sweep readable, because the sequence is not at a fixed rate: the recipe gives 6×10^{-4} up to width 768 and 3×10^{-4} above, so **XXL and B share XL′'s rate and not XL's**, and the two sequences must be read separately. Both contrasts are given because they are different quantities: "gap at peak" is the peak-versus-massed contrast Table 4 reports, and "gap s=20−1" is the fixed-spacing difference Figure 7 plots, with its run-level Wilcoxon p in brackets. They coincide in the five arms whose peak is $s = 20$ and differ in S and XL, whose depth peak is $s = 8$; those are also the two arms where the fixed s=20 gap does not resolve, which is worth seeing beside the sequence. The B row is the 1B arm of Appendix F.3, which is the only one that microbatches its step; every other entry in this table trains the 32-row batch in one pass.

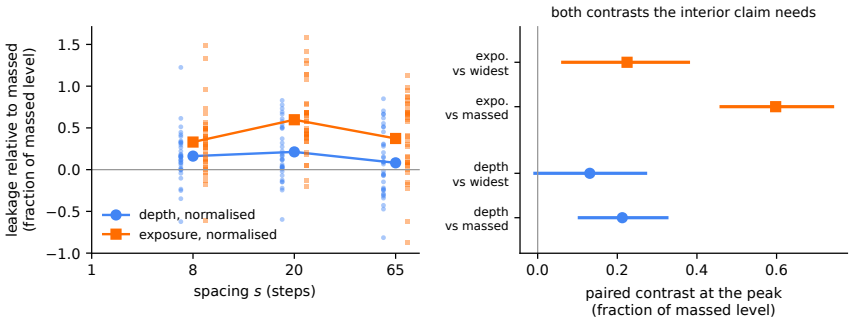

Figure 3: The primary-grid result *per run*, which Figure 1 summarises as a mean curve. **Left:** leakage relative to massed, each series divided by its own measure's mean massed level so the two share an axis without disturbing the pairing. **Right:** the two contrasts the interior claim requires, and the reason we separate them; vs.-widest on depth is the weak one, and it is the cell Section 4.7 declines to claim.

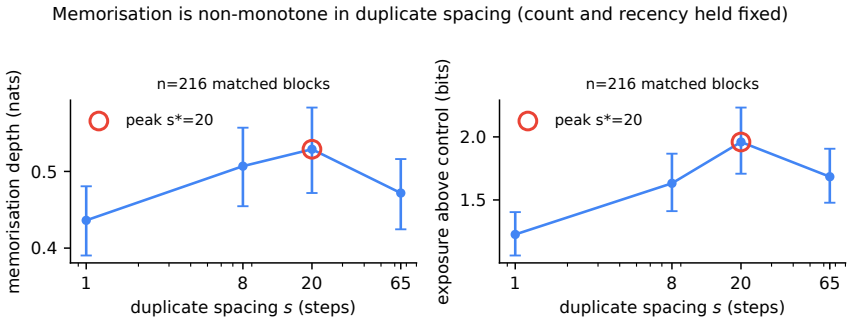

Figure 4: The same primary-grid result plotted *unpaired*: per-spacing means with bootstrap intervals over blocks. Between-run variation in overall leakage is large compared with the spacing effect, so it dominates an unpaired interval while cancelling exactly in the within-run paired comparison the tests actually use (Figure 3). An unpaired reading of this data would have led us to report no effect, which is why the matched design was necessary.

scale	layers	heads	width	total	non-emb.	lr	val. loss	gap at peak	gap s=20−1
S	4	4	256	16.1M	3.2M	6e−4	5.432	+0.060	+0.039 (0.092)
M	6	6	384	30.0M	10.6M	6e−4	5.130	+0.093	+0.093 (5.8e−4)
L	8	8	512	51.1M	25.2M	6e−4	4.948	+0.181	+0.181 (9.8e−4)
XL	12	12	768	123.9M	85.1M	6e−4	4.832	+0.101	+0.087 (0.380)
XL′	12	12	768	123.9M	85.1M	3e−4	5.233	+0.351	+0.351 (4.9e−4†)
XXL	24	16	1024	354.0M	302.3M	3e−4	5.109	+0.219	+0.219 (0.021)
B	20	32	2048	1110.6M	1007.2M	3e−4	5.961	+0.170	+0.170 (0.027)

The details a reimplementation needs. Every block is pre-LN: $x \leftarrow x + \mathrm{attn}(\mathrm{LN}(x))$ then $x \leftarrow x + \mathrm{MLP}(\mathrm{LN}(x))$, with a final LayerNorm before the head. Attention is PyTorch's `nn.MultiheadAttention` with a causal $-\infty$ mask, head dimension 64 at every scale. The MLP is $\mathrm{Linear}(d, 4d)$, **GELU**, $\mathrm{Linear}(4d, d)$, so the expansion ratio is 4. Positional encoding is a **learned absolute** embedding of 256 positions added to the token embedding, not sinusoidal and not rotary. The output head is a bias-free linear layer with its weight **tied** to the token embedding. There is **no dropout** anywhere, which matters for a memorisation paper and is why we state it: nothing in these runs regularises against memorising a canary. Initialisation is PyTorch's per-module default at the versions below, with no GPT-2-style residual rescaling. The optimiser is `torch.optim.AdamW` with lr as tabulated, $\beta = (0.9, 0.95)$, $\varepsilon = 10^{-8}$ (the default), decoupled weight decay 0.1 applied to *all* parameters with no exclusion for LayerNorm gains or biases, and gradient clipping at global norm 1.0. Forward and backward run under `torch.autocast` in **bf16** with fp32 master weights; the loss is fp32 cross-entropy over the reshaped logits. Software: Python 3.11.15, PyTorch 2.13.0+cu130, CUDA 13.0, NumPy 2.4.6, SciPy 1.17.1, `tiktoken` 0.14.0, on A100 40GB cards.

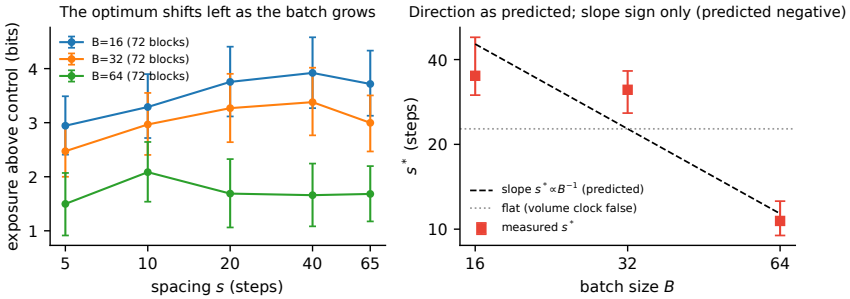

Figure 5: Left: the leakage-versus-spacing curve shifts left as batch size grows. Right: interpolated peak locations with bootstrap intervals over the first two octaves, against a B^{-1} reference slope anchored at the geometric mean rather than an endpoint so only the slope is asserted. That reference slope is *excluded* once the axis is extended to four octaves (Appendix G.2), and is drawn here as the prediction these three arms were read against rather than as a fit to them.

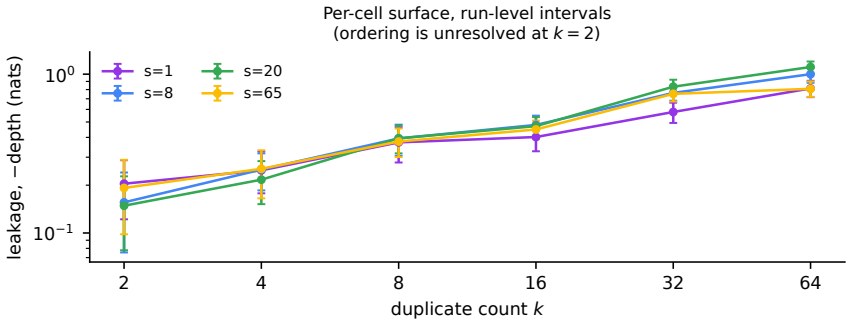

Figure 6: The per-cell surface: leakage against count on log axes, one line per spacing, with run-level intervals. Roughly parallel lines are what a shared exponent with a spacing-dependent gain looks like, and the departures are the non-separability that Section 4.1 rejects. Note the left-hand end: at $k = 2$ the intervals overlap and the ordering is not resolved, which is why Section 4.6 states the count regime in which each headline contrast holds.

Runs at a fixed seed are not bit-identical, and here is the size of it. bf16 autocast with non-deterministic reduction order means two runs of the same script at the same seed diverge. We measured it rather than asserting it was small: re-running one cell of the learning-rate arm gave validation loss 5.11771 against the stored 5.11789, a difference of 1.8×10^{-4} nats, and per-set depths agreeing to 0.0003 to 0.024 nats.

Neither reading of that pair is the per-cell floor: it is one draw, and the spread across cells is an order of magnitude wider. Appendix H.4 measures three repeats of a *different* cell and finds per-set depth spreads of 0.030, 0.075, 0.164, 0.185 and 0.228 nats, an order of magnitude larger, and the largest of those is twice the primary arm's depth massed penalty rather than well inside it. Three repeats across five counts sample the quantity better than one repeat across two runs, so the per-cell depth reproduction floor is 0.03 to 0.23 nats and the 0.024 figure is one draw from the low end of that range.

What survives that spread is narrower. Validation loss reproduces to 10^{-4} nats, so the runs are the same run in every aggregate sense; the discrepancy is confined to per-cell depth. The paired design is unaffected, because pairing holds the batch *contents* fixed across spacings and only the floating-point reduction order varies. And the reported *intervals* remain consistent with the larger floor, because depth is always averaged over 12 seeds and five counts, which is the reason Section 4.2's intervals are far narrower than a single cell's instability.

A.3 Injection procedure and the remaining protocol details

The procedure is specified here at the level a reimplementation needs, as pseudocode where a prose statement would leave a choice open. Everything below is a statement about what the runs did, not a description of intent.

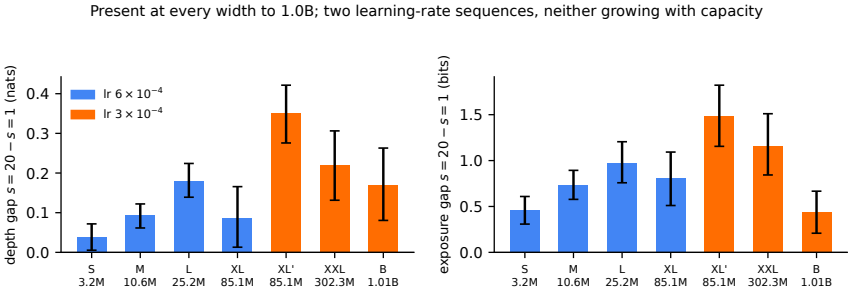

Figure 7: The spacing gap at each width, and the rate it was measured at. Bars are the *fixed* s=20 minus s=1 contrast with cluster-bootstrap intervals over runs, not the peak-versus-massed contrast Section 4.4 narrates and Table 4 tabulates; the two coincide in the arms peaking at $s = 20$ and differ at S and XL, which peak at $s = 8$ (Appendix A.2 gives both). Colour is the learning rate, and is why there are two sequences: the GPT-2 recipe gives 6×10^{-4} up to width 768 and 3×10^{-4} above, so XXL and B share XL′'s rate. The 3×10^{-4} sequence has three points (85.1M, 302.3M, 1,007.2M) and the 6×10^{-4} four (3.2M, 10.6M, 25.2M, 85.1M); both 85.1M arms are shown. Neither grows with capacity, and along 3×10^{-4} the absolute gap *falls at both* steps while the peak-over-massed *ratio* rises (Section 4.4, Table 4).

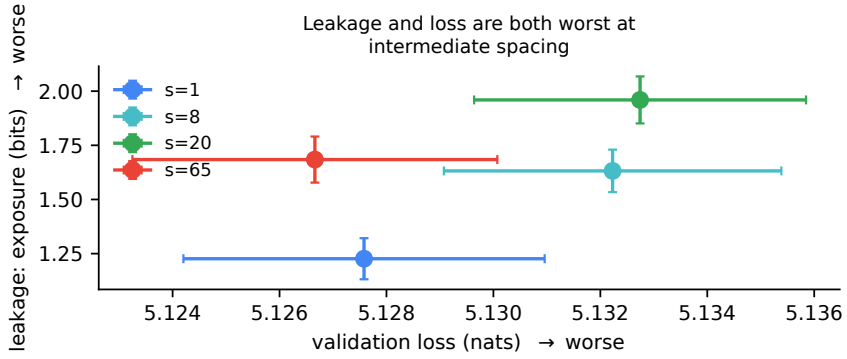

Figure 8: Leakage against validation loss by spacing. Both are worst at intermediate spacing, so there is no trade-off to negotiate between them. Both sets of bars are *unpaired* per-spacing standard errors, so the loss bars overlap across all four spacings even though the paired run-level test on the same data gives $p < 10^{-4}$; the loss differences are small against the between-seed spread and large against the within-seed one. Figure 3 and Figure 4 make the same point for leakage.

Canary construction. One generator seeded at `CANARY_SEED` = 777 produces every canary in every run, drawn in a fixed order: for set $g = 0, \ldots, n_g$ and index $c = 0, \ldots,$ cpg $- 1$, assign uid $= 10000 + 100g + c$, tokenise the prefix `"\nThe access code for user {uid} is "`, draw 12 digit tokens uniformly, or 6 in the onset and overfitting arms (Appendix E.2), draw a start offset uniformly in the corpus, take the 256 tokens from there as the carrier block, and overwrite positions 24 onwards with prefix followed by secret. The set with index n_g is the never-injected control. Because the order of draws is fixed by $(n_g,$ cpg$)$, the canary with a given uid is byte-identical across every run of an arm, which is what makes cross-run pairing valid; in the membership arm the control uids are offset to 90000+ so no uid is shared between a member and a non-member.

Injection into the batch. At step t, the batch is 32 blocks of 256 tokens drawn by a generator seeded from the run seed. If t is an injection step, the sets scheduled at t are visited in index order and their canaries overwrite batch rows starting at row 0 and increasing; no other row is touched, and the number of rows needed is cpg times the number of sets at that step. The assertion that this never exceeds 32 is checked statically from the schedule before training and again as a delivered-versus-requested count at the end, so a run that would drop an injection fails rather than producing a number.

```
for g in 0..n_g-1:  k[g] = COUNTS[(g + seed) mod n_g]
sched = {} ; for g, i in 0..k[g]-1:  sched[T_LAST + g - (k[g]-1-i)*s].append(g)
assert max_t |sched[t]| * cpg <= 32
```

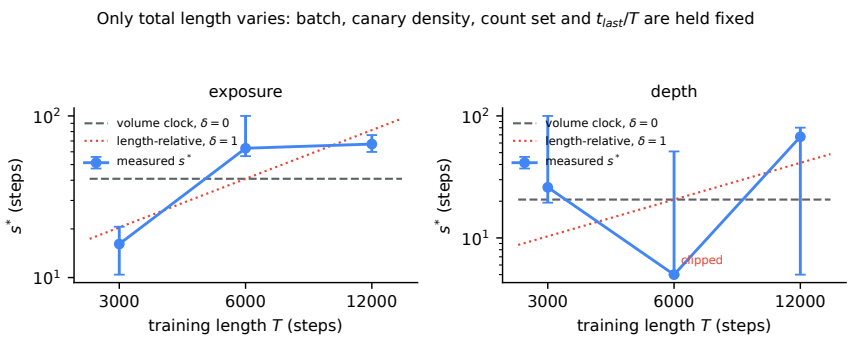

Figure 9: The arm that separates the two mechanism clocks: only total training length varies, with batch size, canary density, count set and the fraction of training after the last copy all held fixed. **The length exponent δ this arm was built to measure is withdrawn**: the arm varies convergence along with length, so what it fits is confounded, and Appendix H.3 and Appendix G.3 retract it. Read it as the evidence against a length clock rather than as a live estimate of one. Both candidate clocks are drawn anchored at the geometric mean of the measured optima, so only their slopes are asserted, which is the convention Figure 5 uses. A clipped estimate is marked, because clipping biases the fitted exponent towards the volume clock and would therefore flatter it.

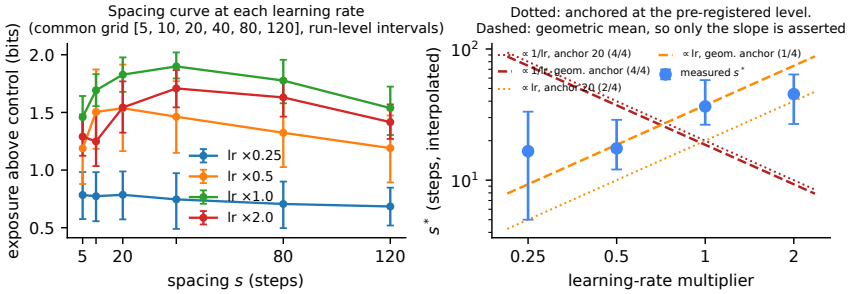

Figure 10: The learning-rate clock, on the four arms' common grid $\{5, 10, 20, 40, 80, 120\}$ with the interpolated estimator and run-level intervals. Each candidate scaling is drawn under both anchoring conventions, dotted at the pre-registered level and dashed at the geometric mean of the measured optima, and annotated with the number of arms whose interval excludes it under that convention: $1/\mathrm{lr}$ is excluded in 4 of 4 under both, and $s^* \propto \mathrm{lr}$ in 2 of 4 and 1 of 4. The $\mathrm{lr} \times 0.25$ arm doubles the range to $8\times$ and sits flat against $\times 0.5$, which is why Appendix C.2 reports it as having widened the slope interval rather than tightened it. s^* rises with learning rate but matches neither scaling, the basis for excluding both in Appendix H.3.

```
for t in 0..T-1:
  lr = LR * (t+1)/200 if t < 200 else LR
  batch = 32 random 256-token windows of the corpus
  if t in sched:  row = 0
    for g in sched[t]: for c in canaries[g]: batch[row] = c.block ; row += 1
  loss = cross_entropy over ALL 255 next-token positions of ALL 32 rows
assert delivered == requested
```

The analysis side, which is where a reimplementation would otherwise guess. The training procedure above is specified to a level a reimplementation can hit. The *analysis* is the half where a prose description leaves choices open, so the procedures every reported number passes through are given below as pseudocode: four for the core statistics, four for the batch axis, the selection-absorbing test, the between-panel component and the censoring recovery, and three for the measures themselves, the realistic arm's placement rule and the batch-composition control.

```
# 1.  cluster bootstrap, which produces every interval in this paper
rng = default_rng(20260823)       # stated, not arbitrary:  the submission date
for b in 1..4000:  idx = rng.integers(0, n_runs, n_runs)    # resample RUNS with replacement
  stat[b] = mean over idx of (run mean at s_peak − run mean at s_ref)
```

```
interval = percentile(stat, [2.5, 97.5])  ;  p = wilcoxon(per-run differences)

# 2.  nested model comparison of Sec 4, wild cluster bootstrap with Webb weights
fit M1:  leakage = C[s] * k**beta  (least squares in LINEAR space, all cells, no discard)
fit M2:  leakage = C[s] * G[k]  ;  fit M3:  free mean per (s,k) cell
F_obs = F(M1 vs M2) from residual sums of squares
for b in 1..2000:  # impose the null, then perturb whole runs by a Webb weight
  w = rng.choice([-sqrt(1.5),-1,-sqrt(0.5),sqrt(0.5),1,sqrt(1.5)], size=n_runs)
  y*[run] = fitted_under_null[run] + w[run] * residual[run]  # one weight per RUN
  refit both models on y*, record F[b]
p = mean(F >= F_obs)  # Rademacher signs recomputed as a check; changed no verdict

# 3.  the s* estimator
i = argmax over grid of mean leakage
if i is the first or last grid point:  report grid[i] as a BOUND, flag clipped
else:  fit a parabola through log2(grid[i-1..i+1]) vs mean[i-1..i+1]
      s* = 2 ** clip(vertex, log2(grid[0]), log2(grid[-1]))
# intervals resample runs and re-run the whole procedure, so clipping is inside the
bootstrap

# 4.  offline LiRA scoring
for each reference run r:  level[r] = mean over ALL its canaries of nll
mu[uid], sd[uid] = mean, std(ddof=1) over r of (nll[r][uid] - level[r])
for each target run t:  level[t] = mean nll over its NEVER-INJECTED control set
  lira[uid] = (mu[uid] - (nll[t][uid] - level[t])) / max(sd[uid], 1e-6)  # not `lambda`:
                # that is a Python keyword and will not parse
  thr = quantile(lira over non-members, 1 - fpr)  ;  TPR = mean(lira[members] > thr)
# AUC and TPR are computed WITHIN a run and then averaged over runs, never pooled

# 5.  the batch-axis exponent, its interval, and the reachability crossing
SEED = 20260823  ;  NBOOT = 4000
def stream(label):  return default_rng([SEED, crc32(label)])  # one stream PER SITE, so a
    # result cannot shift when unrelated code is inserted ahead of it in the script
for each arm a:  # a is a batch size at fixed canaries-per-group
  prof[s] = mean over ALL blocks of every seed of arm a, spacing by spacing
  log2s[a], clipped[a] = vertex(prof)  # procedure 3 above
  rng = stream("sstar|" + a)
  for b in 1..NBOOT:  # resample SEEDS, carry every block of a selected seed
    sel = [seeds[i] for i in rng.integers(0, n_seeds, n_seeds)]
    boot[a][b] = vertex(mean over all blocks of sel)
use = [a for a in arms if not clipped[a]]  # a clipped argmax is a bound, not a location
gamma    = -slope of log2s[use] on log2(B[use])     # s* ∝ B**(-gamma)
gamma[b] = -slope of boot[use][b] on log2(B[use])  ;  interval = percentile(gamma[b],
[2.5,97.5])
# leave-one-arm-out repeats the fit on every 3-subset of use
tok[a] = 2**log2s[a] * B[a] * 256          # s* in TOKENS, the reported quantity
# crossing:  solve tok(B) = 2e6 anchored on the LARGEST arm, so the extrapolation is
shortest
cross[b] = B[anc] * (2e6 / tok_boot[anc][b]) ** (1 / (1 - gamma[b]))  if gamma[b] < 1
# gamma[b] → 1 makes tok almost invariant in B and the crossing diverges; those replicates
# are DROPPED and their count is reported, so the percentiles are conditional on that filter

# 6.  the selection-absorbing randomisation test of Appendix C.3
T(m) = max over s in {8,20} of min(m[s] - m[1], m[s] - m[65])  # conjunction, max INSIDE T
T_obs = T(mean over seeds)
for b in 1..20000:
  for each seed:  permute the four spacing labels of that seed's four runs  # exchangeable
                # under the null of NO spacing effect, which is what is permuted
  T[b] = T(mean over seeds of the permuted assignment)
p = mean(T >= T_obs)  # one-sided:  T > 0 already requires an interior peak beating both ends

# 7.  the between-panel variance component, five panels
w[i] = 1 / SE[i]**2  ;  Q = sum w[i] * (y[i] - weighted_mean(y, w))**2
C = sum(w) - sum(w**2) / sum(w)
tau2 = max(0, (Q - (n_panels - 1)) / C)  # DerSimonian-Laird.  n_panels = 5, NOT the
duplicate
                # count k used everywhere else in this paper
sigma = sqrt(tau2)    # the POOLED estimate reweights by 1/(SE**2 + tau2); 1/SE**2 above is
the
                # fixed-effects weight used only to form Q
widened_half = 1.96 * sqrt(SE_analytic**2 + sigma**2)  # SE_analytic =
pen.std(ddof=1)/sqrt(n)
# HKSJ and a sign-flip permutation are reported beside DL because DL under-covers at this k

# 8.  recovering the censored fraction from two stored numbers per cell
# m[i] = N * qhat[i] is an integer in 0..N; only sum(m) and sum(e) are retained per cell
target_m = round(n * N * mean_qhat)  ;  target_e = n * mean_exposure
e(m) = log2(2*N) if m == 0 else log2(N / m)  # the ceiling is hit exactly when qhat == 0
solutions = every multiset {m[1..n]} with sum(m) == target_m  # exact, integers
            and abs(sum(e(m)) - target_e) <= 1e-9  # floats:  tolerance, never ==
censored = number of zeros in the multiset  # unique in most runs; where not, report the
RANGE
```

Four things in that listing are the ones a reimplementation would otherwise have to guess. Procedure 5 resamples *seeds* and not blocks, and takes one independent stream per site: a shared mutable generator makes any number depend on how much code ran before it, which is a silent coupling rather than a wrong formula. Its clipped arms are excluded from the fit rather than entered at the grid edge, because an argmax at an endpoint is a bound. Procedure 6 puts the maximum over interior cells *inside* the statistic, which is what absorbs the selection; computing T at a peak chosen from the same data would not. And procedure 8 is exact integer arithmetic on two published numbers per cell, so it is the one procedure here a reader can run against this paper alone.

Three properties of that listing are load-bearing and easy to get wrong. The bootstrap resamples runs and carries all of a run's blocks together, which is what makes it a *cluster* bootstrap; resampling blocks would understate every interval. The wild bootstrap perturbs by *sign* rather than resampling residuals, because drawing whole runs' residuals with replacement leaves the interaction present in every draw and collapses every p-value toward 0.5. And the LiRA level is removed per run before comparing, the reference runs by their own canary mean and the target by its control set; without it the statistic divides by more noise than the nuisance it removes (Appendix I).

Three procedures remain that a prose statement leaves open: the measures every number in the paper is computed from, the placement rule of the arm that carries the refutation of our own recommendation, and the control that closes the one confound able to manufacture the effect.

```
# 9.  the four measures, exactly as scored, all against the TRUE prefix
nll(uid) = total cross-entropy of the SECRET tokens only, given the true prefix
depth(uid) = mean nll over the never-injected control set − nll(uid)
# sampled exposure is a RANK against N = 256 alternatives that share the prefix
for i in 1..N:  l[i] = nll of alternative i
qhat = max( |{i :  l[i] < l_true}| / N ,  1/(2N) )    # STRICT <, a tie is not "better"
exposure(uid) = − log2 qhat              # the 1/(2N) floor caps this at log2(2N) = 9 bits
# that cap depends on N and NOT on secret length, so these bits are not comparable with
# an exposure taken over a full candidate set.  Exposure is reported ABOVE CONTROL, so no
# value here converts to a quantile without the control level Appendix E.2 lists per arm
top1(uid) = fraction of secret positions whose argmax is the true token
greedy(uid) = 1 if greedy decoding from the true prefix reproduces EVERY secret token
# every per-canary value is averaged WITHIN a run before any test is taken across runs

# 10.  conditional random placement and the heavy-tailed counts (the realistic arm)
k[d] = COUNTS[(d + seed) mod 8]      # COUNTS = {2,3,5,8,13,21,34,55}, the grid arms' rule,
                                     # so count still does not depend on the schedule
t_last[d] = T_LAST + d              # last copy pinned exactly as in Eq. 1
rng = default_rng(seed)              # the RUN seed only, so one seed's clustered and
                                     # random-placement arms differ in placement alone
random:    t[d] = [t_last[d]] + rng.integers(T_FIRST, T_LAST, k[d]-1)   # T_FIRST = 700
clustered:  t[d] = [t_last[d] - (k[d]-1-i)     for i in 0..k[d]-1]
uniform:   t[d] = [t_last[d] - (k[d]-1-i)*20  for i in 0..k[d]-1]
# sampling is WITH replacement over steps; two exposures of one document drawn to the
# same step are resolved by the row-overwrite rule, so the delivered count is unchanged

# 11.  the staggered arm that equalises batch composition, the sixth confound
offset(g) = 67 * g                   # standard arm:  offset(g) = g
# 67 makes two sets sharing a step impossible at EVERY spacing in {1,8,20,65}:
#   at s=1 a set's window is at most 63 steps wide, so offsets 67 apart cannot overlap
#   at wider s a collision needs 67d = 0 mod s for d <= 5; 67 = 3 mod 8, 7 mod 20, 2 mod 65
assert per run:  no step carries canaries from two different sets
# so every injection step carries 4 of 32 canary rows at EVERY spacing, and the only
# thing left differing from the standard arm is whether six sets land together at s=1
# a set keeps its offset at every spacing, so the between-set spread in trailing steps
# cancels in the paired contrast; only within-arm contrasts compare across the two arms
```

The loss is taken on the whole block. Training loss is the mean next-token cross-entropy over all 255 predicted positions of all 32 rows, with no masking or reweighting of the secret span. So a canary contributes to the gradient exactly as a natural block of the same length does, and the secret span is 12 of the 256 positions within it. Only the *measurements* are restricted to the secret span.

Validation loss. Forty batches of 32 blocks drawn from the validation memmap with the same generator, full-block cross-entropy, model in eval mode. Drawn rather than a fixed sweep, so validation loss carries sampling noise of order 0.003 nats, which is why the 0.0047-nat spacing effect on validation loss reported under the quality trade-off in Section 4.4 is paired by seed rather than an absolute level comparison. The pairing is exact, and for a specific reason: the validation generator is the *same* `torch.Generator` instance the training loop used, seeded from the run seed, and by the time validation is reached it has been advanced by the same number of draws in every run of a seed, because every run of a seed takes the same number of training steps and draws one batch per step. So the forty validation batches are byte-identical across spacings within a seed and the paired comparison is a comparison on the same data. A cheaper design would use a fixed sweep and not need this argument, and a future version should.

C4 shard selection and size matching. The English C4 split is streamed in its default shard order, documents tokenised and concatenated, and the stream stopped at exactly $117{,}920{,}140$ training tokens and

247,289 validation tokens, which are the WikiText-103 memmap's sizes. So steps, tokens seen and the corpus-to-canary exposure ratio are identical between the two corpus arms and only the text distribution differs. No filtering, deduplication or reordering is applied beyond what the released split already has.

Section 4.8's ladder reads a second, larger stream, **C4-720M**, from the same split and built the same way, but from shards 1 to 8 and truncated to exactly 720,000,000 training tokens out of the 1,361,372,435 those eight shards tokenise to. The size-matched stream above lies wholly inside shard 0, which this one never reads, so the two are disjoint. Its validation split is the C4 validation split unchanged, which is what makes the rungs' perplexities comparable to each other and to the C4 arm's. Each run records its own corpus path, tokens seen and epoch fraction, and all 624 agree per rung: 50,438,144, 126,091,264, 252,190,720 and 504,381,440 tokens at 25,219,072 non-embedding parameters, which is 0.0701, 0.1751, 0.3503 and 0.7005 of an epoch.

Prose passage selection. The natural-passage arms draw from the WikiText-103 *test* split, which is disjoint from both train and validation; injecting validation text would contaminate a reported outcome. Passage starts are drawn without replacement from one `CANARY_SEED` generator, oversampled by a factor of four and accepted greedily in sorted order only if at least 256 tokens clear of the last accepted one, with an assertion that enough disjoint windows were found. Each passage is a 256-token block: tokens 0 to 127 are the prefix and tokens 128 to 191 the scored target span. Depth is the target span's NLL against a never-injected control set of passages drawn the same way, and the rank statistic ranks the true span against target spans taken from other held-out passages given the same prefix.

Heavy-tailed count assignment in the realistic arm. Counts are $\{2, 3, 5, 8, 13, 21, 34, 55\}$ assigned to document d as COUNTS$[(d + \text{seed}) \bmod 8]$, the same rotation rule as the grid arms so count still does not depend on the schedule. The last copy is pinned as in Eq. 1; the random-placement arm places the other $k - 1$ exposures uniformly in $[T_{\text{first}}, T_{\text{last}})$ from a generator seeded only by the run seed, so the random-placement and clustered arms of one seed differ in placement alone. $T_{\text{first}} = 700$, so the window is 3,800 steps wide; it is not the first step after warmup, which is 200. Sampling is *with* replacement over steps, and a collision, meaning two exposures of one document drawn to the same step, is resolved by the same row-overwrite rule as everywhere else, so the delivered count is unchanged; collisions occur in 0.7% of placements at the largest count.

This is therefore *not* a permutation of a production corpus. It is a random-placement baseline *conditioned* on a pinned final exposure, which is what the matched design requires: an unconditional shuffle would move the last copy and reintroduce confound (2), the one the whole design exists to remove. We call it a *conditional random-placement baseline* below and in Section 4.3. It is the right comparison for isolating placement at fixed recency and the wrong one for claiming equivalence to what a pipeline emits, and no result here rests on the latter.

Per-arm injection accounting. The total quoted in Section 3 is the sum of the last column, and every run in every listed arm asserts delivered equals requested before writing its result file. Each row's per-run figure is read from that arm's run files rather than transcribed, so "per run" is a fact about the arm and not an average, and an arm whose runs disagreed on it would print a range. **The set of rows, though, is curated and not a dump of the disk, and two revisions of this table were short because of it**: one row per arm the paper reports, added when that arm is first used. Twice an arm was analysed and reported without its row being added, the extra canary panels and two of the matched-retention cells, so the total moved after the fact. The check stated below catches this kind of omission. The third it found was the B=256 arm's new cells, 36 runs at $s = 1$ and 24 added at $s = 20$: the old row covered both batch arms at their original 60 runs each, so the extension had no home. Both arms now carry their own row. The injection total alone cannot tell the two readings apart, because both arms run at 252 injections per run, so the row label had to be checked against the runs themselves rather than against the sum. There are 6,345 result files in all, and this table accounts for 5,416 of them. **The difference is not simply unused runs.** It is three things. The two arms predating the run-time check, the replication grid (144) and the older $k = 128$ rotation (72), *are* used and are excluded from this table only because they have no audited injection count; they are audited by replay instead. The 207 pilots of the rejected design are not used. The remainder is reference models for the calibrated attacks, which inject nothing by construction and so have no row in an *injection* ledger although the attack intervals rest on them, together with single-run probes. **So this table is the per-arm injection ledger for arms that inject, not a census of runs**, and a reader auditing it should read the row set as the claim and the file count as its bound.

arm	runs	injections per run	total
primary	144	504	72,576
primary extension	288	504	145,152
S 3.2M	48	504	24,192
L 25.2M	48	504	24,192
XL 85.1M	48	504	24,192
XL′ 85.1M	48	504	24,192
XXL 302.3M	48	504	24,192
C4	96	504	48,384
natural prose	48	504	24,192
realistic	72	282	20,304
overfit	24	1,488	35,712
wide grid	72	180	12,960
batch sweep	240	252 to 504	75,600
learning rate	288	310	89,280
matched tokens	180	180	32,400
length sweep	288	180	51,840
membership inference	48	504	24,192
membership inference, calibrated	112	0 to 504	48,384
absolute position	48	24	1,152
schedule (cosine, WSD)	96	504	48,384
onset	48	1,488	71,424
composition, standard	192	504	96,768
composition, staggered	192	504	96,768
onset composition, standard	24	1,488	35,712
onset composition, staggered	24	1,488	35,712
displacement control	144	504	72,576
β_2 sweep, 6 cells, with s=1	456	310	141,360
β_1=0 at $s \in \{1, 20\}$, two β_2	48	310	14,880
β_2=0.999 rate-matched, with calibration	30	310	9,300
matched-convergence length	60	180	10,800
SGD with momentum	36	310	11,160
independent canary panels	192	504	96,768
calibrated attack, real schedules	176	0 to 282	40,608
attack, high-memorisation arm	40	0 to 1,488	35,712
1B parameters	48	504	24,192
batch sweep, B=256	120	252	30,240
batch sweep, B=128	60	252	15,120
canary fraction, B=128	60	1,008	60,480
matched tokens, B=128	60	180	10,800
single-run probes and checks	18	180 to 310	5,190
token budget, 2 tok/par	144	504	72,576
token budget, 5 tok/par	156	504	78,624
token budget, 10 tok/par	144	504	72,576
token budget, 20 tok/par	180	504	90,720
retention-matched, 5 tok/par	96	504	48,384
retention-matched, 10 tok/par	96	504	48,384
retention-matched, 20 tok/par	96	504	48,384
retention-matched, panels 778 and 779	192	504	96,768
total	**5,416**		**2,323,458**

Two arms predate the run-time assertion and are absent from the table: the replication grid (144 runs) and the older $k = 128$ rotation (72). Both are auditable exactly, because the collision rule is deterministic given the rotation, the count set and the canaries per set, all of which each run records. Replay gives 2.12% dropped at $s = 1$ and 0.00% at $s = 8, 20, 65$ in the replication grid, with the loss confined to $k = 16, 32, 64$ at 3.12%, 2.60% and 2.08%; and 0.00% everywhere in the rotation, whose six sets carry distinct spacings so at most two share a step. So the replication grid's massed cell is under-injected by a stated 2.12% rather than an unknown amount. A further 207 runs on disk are pilots of the rejected design of Appendix L under cosine decay; no result here uses any of them, and the emitter lists them so they are excluded by declaration rather than omission. Four rows show a range rather than a product. The batch sweep's arms differ in canary density by design; the two calibrated membership arms' reference models inject *nothing*, which is what makes them reference models, so their per-run figures start at 0 and their totals are exact sums. The primary arm is listed as its pre-registered 144

runs plus the 288-run extension of Appendix F.2, separately, because every headline number is computed on the first.

The learning-rate row includes the 36 runs that extend that arm's grid to $s = 120$ (Appendix C.2) and the length-sweep row the 72 that extend all three arms' grids rightwards: $s = 130$ at $T = 3{,}000$, $s = 150$ and 250 at $T = 6{,}000$, and $s = 150$, 250 and 400 at $T = 12{,}000$, six added cells at 12 runs each (Appendix C.2 tabulates every column).

B THE SEVEN CONFOUNDS, AND THE ARMS THAT CONTROL THEM

B.1 WHAT EACH CONFOUND DOES IF NOT REMOVED

Confound (4), the learning-rate schedule, and why we abandoned cosine decay. Appendix F.1 re-runs the grid under cosine and warmup-stable-decay and reports what survives; this paragraph is about why the primary design does not use them. Under cosine decay the single-exposure forgetting curve measures as a non-monotone inverted U: $+0.132$ nats at distance 200, meaning a copy seen 200 steps before the end appears *less* memorised than copies further back. That is mechanistically absurd, and it is an artefact of late copies arriving at $\mathrm{lr} \approx 0$ and imprinting nothing. With a constant rate after warmup the same measurement is cleanly monotone (-3.719 nats at distance 200). This is why the learning rate is constant throughout, and it is a case where the confound did not merely add noise but inverted the sign of a curve we would otherwise have interpreted.

Injection collisions, in detail, and what they do to batch composition. Injections overwrite batch rows, so two sets injecting at the same step compete for rows, and sets whose indices differ by $g - g'$ collide exactly when $(g - g') \equiv 0 \bmod s$. At $s = 1$ that holds for every pair at every step, which is why the massed arm alone lost injections; at $s \in \{8, 20, 65\}$ with six sets it never holds. The density used here (24 rows needed against a batch of 32) makes the collision impossible rather than merely unlikely, and the run-time assertion is what turns that from an argument into a check.

Making drops impossible is not the same as making the treatment identical. Delivering many canary rows in one batch and delivering 4 rows in each of several batches inject the same copies, but they are not the same update.

The composition at $s = 1$ is a *per-seed* quantity rather than a constant, and the difference is the arithmetic Appendix B.2 exists to control. Set g occupies steps $[T_{\text{last}} + g - (k[g] - 1),\ T_{\text{last}} + g]$, so it injects at T_{last} only when $k[g] \geq g + 1$, and because count is rotated against set index by seed the number of sets meeting that condition depends on the rotation. Enumerating the six rotations of $\mathrm{COUNTS} = \{2, 4, 8, 16, 32, 64\}$ gives $6, 5, 4, 4, 5, 6$ sets co-scheduled at T_{last} and a union of windows spanning 64 to 69 distinct steps. The rows below are read off the logged per-step canary counts of the instrumented arm of Appendix B.2, which records them at every step of every run, rather than re-derived from the same arithmetic:

s	sets sharing a step	canary rows of 32	% canary	injection steps per run
1	4 to 6	16 to 24	50.0 to 75.0%	64 to 69
8	1	4	12.5%	126
20	1	4	12.5%	126
65	1	4	12.5%	126

So the worst case is a $6\times$ difference in canary share and the typical case is smaller: averaged over the six rotations the mean canary share at an injection step is 23.7% of the batch at $s = 1$ against 12.5% at every wider spacing, because the wide steps at the edges of the massed window carry few sets. Either figure is perfectly correlated with the massed condition, and either is a property of running six sets per run to balance the count factor rather than of clustering a document's duplicates: a real pipeline that clusters one document's copies does not thereby co-locate six unrelated duplicated documents. Appendix B.2 measures what it does and removes it, and does so at the *worst* case rather than the mean, since the staggered arm equalises every step at 4 rows.

confound	measured effect if not controlled	our control
canary difficulty	spread across sets at fixed k spans 0.31 nats, $3.3\times$ the 0.093-nat spacing effect	rotate count against set by seed
recency of the last copy	about $6\times$ larger and *monotone* in s	pin t_{last} in all conditions
LR-schedule position	forgetting curve inverts to a U, $+0.132$ versus -3.719 nats at $d = 200$	constant LR after warmup
decay-weighted copy count	effective exposures fall about $4\times$ from $s = 1$ to $s = 65$ at $\tau \approx 10^3$	cross count with spacing: this estimates β separately but does *not* identify it from a gain shift (Appendix H.1)
absolute position in training	moving one exposure from step 500 to 6000 at fixed recency moves leakage 0.311 nats, $5.4\times$ the decline it could explain	*neither removed nor crossed*: measured and bounded only (Appendix B.4)
batch composition at injection steps	75% canary rows at $s = 1$ against 12.5% at every wider spacing	a staggered arm equalising it at 12.5% everywhere (Appendix B.2)

The recency contrast deserves its own line: with recency confounded the effect is about $6\times$ larger and monotone in spacing. A study that does not pin the last exposure will report a large, clean and wrong result, which is our best explanation for why this question has not been settled.

B.2 Batch composition at injection steps, the sixth confound

This appendix covers the one confound in Section 1 that acts in the direction that would *manufacture* our robust claim rather than suppress it, so the argument the others rest on does not reach it. Every confound in Section 1 except (6) and (7) is monotone decreasing in s, which is why Section 4.7 can argue that none of them produces the rise from clustered to intermediate spacing. Batch composition is not monotone in s: it is extreme at $s = 1$ and flat afterwards (Appendix B.1), so the argument does not reach it and it needs its own measurement.

The mechanism being tested. Training clips the gradient norm at 1.0 and the loss is a mean over all 32 rows. If a batch that is three-quarters canary has a larger gradient norm than one that is one-eighth canary, then where the clip binds the whole update is rescaled, each copy imprints less, and it happens at the massed end only, which inflates the massed penalty. A second route needs no clipping at all: AdamW's second moment responds to the norm, and its response is not a pure rescaling.

The arm. We re-ran the primary configuration with two changes and nothing else. First, every run records the pre-clip total gradient norm at every step together with the number of canary rows in that step's batch, which makes the clip rate at injection steps directly observable. Second, a `stagger` parameter moves set g's window to $T_{\text{last}} + 67g$ instead of $T_{\text{last}} + g$. At 67 no two sets share a step at *any* spacing in $\{1, 8, 20, 65\}$: at $s = 1$ a set's window is at most 63 steps wide so offsets differing by 67 cannot overlap, and a collision at wider s would need $67d \equiv 0 \bmod s$ for some $d \leq 5$, which fails for all three ($67 \equiv 3 \bmod 8$, $7 \bmod 20$, $2 \bmod 65$). The script verifies the property per run rather than trusting the argument. Staggered runs therefore carry 4 of 32 canary rows at every spacing, so composition is equalised and the only thing that differs from the standard arm is whether six sets land together at $s = 1$.

Staggering widens the spread in how much training follows a set's last copy, from 6 steps to 335 of the 1,500 that remain. That is a between-set difference and not a within-set one: a set keeps its offset at every spacing, so it differences out of the paired contrast to first order, which is the same argument Section 3 already makes for the 6-step version. It is only first order: retention decays non-linearly, so a shift of up to 335 steps need not cancel exactly, and the residual is reported rather than argued away. Count is rotated against set index by seed, so every count meets every offset across seeds. What it does cost is comparability of *absolute* levels with the primary arm, so only the within-arm spacing contrast should be read across the two.

Result 1: clipping does not bind preferentially at the massed end. The premise is false in the direction that matters. Canary blocks are more predictable than fresh corpus text, so injecting them *lowers* the gradient norm rather than raising it: over 48 runs per cell the mean pre-clip norm at injection steps is 0.76 to 1.00 against 1.68 at ordinary steps, and the clip rate at injection steps is 6.0 to 8.0% at $s \in \{1, 8, 20\}$, specifically 7.99, 6.91 and 6.00%, against about 36% everywhere else. Composition does move the rate, 7.99% at the massed condition's 16 to 24 rows against 3.89% at 4 rows in the same condition staggered, so the route is real rather than absent. But massing compresses the same injections into about half as many steps, 64 to 69 against 126, and in absolute terms the clipped-step count is then the same: 5.3 at $s = 1$ standard, 4.9 at $s = 1$ staggered, 7.6 at $s = 20$. Those three products are computed on the logged injection-step count per run, which averages 66.5 at $s = 1$ rather than a flat 64 (Appendix B.1). The massed end does not receive more clipping, it receives the same clipping over fewer steps.

The clip rate by spacing in fact runs the other way, and it points at a confound we already report. At $s = 65$ it is 24.2%, three times the massed value, because with $k = 64$ and $s = 65$ a set's first copy lands at step 405 where gradient norms are still large, whereas at $s = 1$ every copy lands near step 4500 where they are small. In absolute terms the gap is wider than the rate suggests: 30.5 clipped injection steps at $s = 65$ against 5.3 at $s = 1$, a factor of 5.8. So the clip rate at injection steps tracks *when* those steps occur, which makes it a mechanism for confound (5), absolute position in training, measured in Appendix B.4 and reported there as neither removed nor crossed. If clipping suppresses imprinting it therefore suppresses it most at the *widest* spacing, which would depress leakage there: that is a candidate contribution to the *descending* branch, which Section 4.7 already reports as unidentified, and not to the ascending one. It is now the *third* candidate for that branch alongside decay-weighted copy count and absolute position, and the three are not separable in this design: clipping at wide spacing is a consequence of depositing early, which is what absolute position is, so it is better read as a partial mechanism for confound (5) than as an independent term.

Result 2: the penalty is unchanged under equalised composition. Forty-eight seeds per spacing per design, paired by seed, contrasts at each design's own peak, which is $s = 20$ under both. These are fresh runs at the primary configuration with the instrumentation added, not a re-analysis of the primary grid, so the standard column is an independent 48-seed sample of the same thing Section 4.2 measures on 36: it reads $+0.110$ against that section's $+0.093$ on depth and $+0.735$ against $+0.733$ on exposure, the second agreeing to three decimals.

measure	design	rows/step at s=1	massed penalty	p
exposure	standard	16 to 24 of 32	$+0.735$ $[+0.614, +0.864]$	2.3×10^{-13}
exposure	staggered	4 of 32	$+0.644$ $[+0.506, +0.779]$	7.1×10^{-11}
depth	standard	16 to 24 of 32	$+0.110$ $[+0.081, +0.141]$	5.8×10^{-9}
depth	staggered	4 of 32	$+0.116$ $[+0.080, +0.155]$	1.8×10^{-8}

Equalising composition changes the massed penalty by a factor of 0.876 on exposure and 1.053 on depth. Both staggered intervals exclude zero. **And the paired test is available here now, because both designs ran the same** 48 **seeds.** Taken within seed, the standard-minus-staggered difference is $+0.091$ bits $[-0.090, +0.270]$ at $p = 0.26$ on exposure and -0.006 nats $[-0.053, +0.041]$ at $p = 0.92$ on depth, with 25 of 48 seeds larger under the standard design on both measures. So the confound is bounded by an interval on the difference rather than by one estimate's resolution, which is the test Result 3 says containment is not. On depth the staggered penalty is the *larger* of the two, so if co-location does anything it makes our reported figure conservative. At this configuration the penalty survives equalised composition, and the confound is bounded by measurement rather than excluded by argument. *Bounded and not excluded, and the bound is now the difference*: its upper end is $+0.270$ bits, 37% of the exposure penalty, against a point estimate of 12%, so a contribution of about a third would still have escaped. The bound is looser than the $\pm 27\%$ the 12-seed resolution implied, and it is a bound on the right quantity. With Result 3 below finding a 41% reduction at 93.8% canary share that does not resolve, the two readings together are a dose-response in share, undetectable at the primary arm's worst-step 75% and material at 93.8%, both peak-step shares so the pair is on one basis.

Result 3: the same control at the configuration where the asymmetry is worst, which is partly adverse. The primary arm carries 24 of 32 canary rows at $s = 1$. The onset arm of Appendix E.2 carries 30 of 32, 93.8% against 18.8% at $s = 20$, and it is the arm carrying the $k = 128$ contrast Section 4.6 reports, so controlling the milder asymmetry and leaving the worse one uncontrolled would be the wrong way round. We re-ran that configuration too, 12 seeds per spacing per design, offset $129g$ so no two sets share a step at either of its spacings.

The clipping picture is the same: the clip rate at injection steps is 14.8% at $s = 1$ with 30 rows against 17.2% at $s = 20$ with 6, so it is again *higher* at the wider spacing, and equalising composition at $s = 1$ drops it to 4.7%. The contrast is not the same, and this is the one place in the paper where equalising composition costs us:

measure	standard	staggered	staggered / standard
per-token top-1	$+0.026$ $[0.007, 0.041]$	$+0.032$ $[0.013, 0.050]$	$1.231\times$
depth	$+0.351$ $[0.259, 0.447]$	$+0.262$ $[0.159, 0.374]$	$0.75\times$
exposure	$+0.940$ $[0.721, 1.170]$	$+0.552$ $[0.254, 0.889]$	$0.59\times$

All six intervals exclude zero, so the massed penalty survives at this configuration too. But on exposure the reduction is the largest we measure: the paired difference is -0.388 $[-0.776, +0.001]$ at $p = 0.05$, which does not exclude zero, so composition plausibly accounts for about 41% here without the reduction resolving. Whether one estimate falls outside the other's interval is not a test of their difference, and we do not read it as one. On depth the standard estimate stays inside the staggered interval and on top-1 the staggered penalty is larger, so the effect is measure-dependent. Every interval in this table is from the single bootstrap of Appendix C.1 at the stated seed, which is what a fixed seed and one canonical table exist to guarantee.

Where it comes from matters, and it is not where the paper's claim lives. Per count cell, the staggered contrast is *larger* at $k = 128$ on both measures, $+0.076$ against $+0.051$ on top-1 and $+1.636$ against $+1.168$ bits on exposure, and the reduction is entirely at the low counts, where the staggered contrasts at $k = 8$ and 16 turn slightly negative. That is intelligible: at 93.8% canary the massed batch has two rows of natural text, an update unlike anything else in the paper, and its effect is proportionally largest where the signal is smallest. The consequence we accept is that this arm's *pooled* contrast overstates the spacing effect, while the $k = 128$ cell Section 4.6 actually quotes does not, and is if anything conservative. It is a further reason the paper's claims are scoped to high count.

What we would have concluded had it gone the other way. Recorded here because the arm was launched with this reading fixed. A materially smaller penalty under equalised composition would mean part of the measured effect is a property of injecting six sets at once rather than of spacing, and the headline effect size would have had to be restated with the standard-arm figure labelled an upper bound. Note the direction of the asymmetry: co-location at $s = 1$ can only argue for a *smaller* true penalty, so a staggered penalty at least as large is the outcome that costs us nothing, and it is the outcome we report either way.

B.3 Confound (7), natural-row displacement, equalised

Injections *overwrite* batch rows rather than extending the corpus. Within a seed the natural rows *drawn* are identical across spacings, because the data-order generator is seeded by the run seed alone and is consumed once per step whether or not that step injects; but which of those rows is *overwritten* follows the schedule. So each contrast in Table 4 is the total effect of two things: placing the k copies at different steps, which is the intervention, and losing a different multiset of natural rows, which is confound (7) and is unsigned. It is 504 of 192,000 rows, 0.26%, the same order as the whole canary intervention. This arm equalises it.

The design, and why equalising the rows removed is not enough. Let P_s be the set of (step, slot) positions that spacing s would overwrite and $U = \bigcup_s P_s$ over $s \in \{1, 8, 20, 65\}$. In every condition we overwrite all of U, write this condition's copies at the positions in P_s exactly as the published arm does, and write one *fixed* filler multiset F of held-out natural rows at the positions in $U \setminus P_s$. Every condition then trains on the identical multiset, the corpus minus U plus the copies plus F, so the only thing that varies is where in U the copies sit. Fixing F once is what makes this a control rather than half of one: drawing filler *per slot* would equalise the rows removed while leaving the conditions differing in what was *added*, since $U \setminus P_s$ is a different set in each. F is drawn from a seed depending on neither the condition nor the run seed, 20260904. Because $|P_s| = \sum \text{COUNTS} \times \text{cpg} = 504$ independent of s, $|F|$ is constant within a seed; both are asserted at run time rather than assumed, and every one of the 144 runs wrote its full filler complement and delivered 504 of 504 injections.

Result: every contrast survives, and every one is larger. Thirty-six matched seeds at the primary configuration, validation loss 5.129 against the published arm's 5.130.

	depth (nats)		exposure (bits)	
vs. clustered	this control	published	this control	published
$s = 8$	$+.082$ [$+.057, +.109$]	$+.071$ [$+.029, +.114$]	$+.443$ [$+.364, +.520$]	$+.405$ [$+.254, +.567$]
$s = 20$	$+.\mathbf{120}$ [$+.089, +.151$]	$+.093$ [$+.046, +.139$]	$+.\mathbf{794}$ [$+.680, +.911$]	$+.733$ [$+.572, +.903$]
$s = 65$	$+.088$ [$+.024, +.147$]	$+.036$ [$-.021, +.090$]	$+.550$ [$+.380, +.714$]	$+.458$ [$+.272, +.636$]

Leading zeros dropped; every value is nats or bits as headed.

The ratios, and the range they leave. At the headline $s = 20$ cell the controlled level over the published one is $0.120/0.093 = 1.29$ on depth and $0.794/0.733 = 1.08$ on exposure. Taken with Appendix B.2's five ratios, which run 0.59 to 1.27, the two controls together leave the massed penalty at 0.59 **to** 1.29 **times its uncontrolled value**, which is the range Section 4.7 quotes: the low end is Appendix B.2's onset exposure cell and the high end is this arm's depth.

Run-level win rates under the control are 31, 32 and 26 of 36 on depth and 35, 36 and 31 of 36 on exposure, so the sign tests give 1.3×10^{-5}, 1.9×10^{-6} and 0.011 on depth and 1.1×10^{-9}, $2.9 \times 10^{-11\dagger}$ and 1.3×10^{-5} on exposure. Both the massed penalty and the interior contrast hold, and **the one cell the published arm fails resolves here**: $s = 65$ on depth is $+0.088$ $[+0.024, +0.147]$ against the published $+0.036$ $[-0.021, +0.090]$, the only one of these three vs.-clustered cells that fails there. That is a contrast Section 4.7 declines to claim on other grounds, so this strengthens the evidence for claims already made rather than adding one.

Two costs and one residual, none of them argued away. U is about $3.2\times$ the rows any single schedule touches, 1,564 to 1,632 against 504. So every condition here displaces more natural text than any published arm, and *the levels are not comparable to the ten*. Only the within-arm spacing contrast is, and that is the quantity at issue. $|U|$ varies across seeds because the count rotation does, and the union is taken within a seed, which

is where the contrast is paired and is stricter than one union across the arm. And F's internal *order* cannot be identical when the free slots differ, since the copies and the filler share U by construction; that is a data-order difference within natural text rather than a difference in what was trained on, and it is the same class of thing the design isolates, so it is reported as the residual rather than removed.

B.4 Absolute position at fixed recency

A fifth confound is unremovable at fixed T_{last}: wider spacing necessarily places earlier copies earlier in absolute training, where the model's loss scale and gradient norm differ. Constant learning rate removes the schedule component of this but not the model-state component, and the single-exposure forgetting curve cannot separate them, because in a run of fixed length a copy at step p is both at position p and at distance $T - p$ from readout.

To separate them we hold the gap from exposure to readout fixed at 1500 steps and move the exposure through training, so total length is $p + 1500$ and recency is constant by construction. One exposure per set, so $k = 1$, and 12 runs per position:

position p	total steps	depth leakage	exposure above control	val. loss
500	2000	-0.046 $[-0.099, +0.007]$	$+0.130$ $[-0.006, +0.284]$	6.421
1500	3000	$+0.032$ $[-0.048, +0.111]$	$+0.280$ $[+0.035, +0.498]$	5.987
3000	4500	$+0.237$ $[+0.138, +0.332]$	$+0.343$ $[+0.084, +0.607]$	5.437
6000	7500	$+0.265$ $[+0.194, +0.331]$	$+0.658$ $[+0.472, +0.827]$	4.918

Position matters, and by a lot: run-level Friedman gives $p = 2.4 \times 10^{-6}$ on depth and 9.8×10^{-4} on exposure, and late minus early is $+0.311$ nats $[0.227, 0.399]$ and $+0.528$ bits $[0.308, 0.722]$. For scale, the primary arm's whole peak-to-widest decline is 0.0571 nats and 0.2753 bits, so this effect is $5.4\times$ and $1.9\times$ larger than the feature it could explain.

One limitation makes that an upper bound rather than a point estimate, and it is the reason we do not treat this as a measurement of deposition alone: because total length is $p + 1500$, the model doing the *readout* also differs across arms, from validation loss 6.42 to 4.92. So the table measures deposition and readout jointly. The bound is in the conservative direction for our purposes, since it can only overstate how much of the descending branch position could explain. Note also that in the main design this readout issue does not arise: every canary in a run is read out by the same final model, and only deposition positions differ across spacings.

The consequence for the paper is the same as for confound (3). Both are monotone decreasing in s at fixed k, so neither can produce the ascending branch, and the massed penalty stands as a lower bound. **For (5) that direction is measured at $k = 1$,** the count this arm runs, while the headline contrasts live at $k \geq 8$; Section 1 states it with that scope. Confound (3)'s monotonicity needs no such arm, being a property of its kernel at every k. **And the lower-bound reading needs a second condition we have not tested, so we state it rather than leave it implicit: that the confounds do not cancel.** Each is monotone decreasing in s on its own, which is what makes each individually unable to create the rise, and the bound then follows if their contributions to the measured contrast *add*, or at least share a sign. If two of them offset each other over the range compared, the sum need not be monotone even though every term is, and the reading would fail. Nothing here measures a cross term: the controls are run one confound at a time, by design, and a crossed arm is what would settle it. So *lower bound* throughout means: a lower bound if the remaining confounds add, with (5)'s direction measured at $k = 1$ and (3)'s given by its kernel. The cell that could test (5) at higher count is the $k = 2$ column, and the paragraph below is why it cannot: its sign is not consistent across arms. One more scale point, since the ratio above invites a misreading: $5.4\times$ is against the peak-to-widest decline, which is the branch Section 4.7 declines to claim, and against the massed penalty itself, the feature this confound must not create, the same $+0.311$ nats is $3.4\times$ its $+0.0927$. Both can produce the descending branch, so the interior optimum is not identified. A design that would separate them needs the readout model held fixed while deposition position varies, which at fixed T_{last} means varying k and s jointly in a way we have not attempted.

One candidate for that design, and why it is not one. The obvious place to look is the $k = 2$ cell of the primary grid. At $k = 2$ the last copy is pinned and there is exactly one other, so moving s from 1 to 20 moves a single earlier copy back by nineteen steps while count, recency, per-step learning rate, canary content and the readout model are all held. That looks like the deposition-position component on its own, and its primary-arm depth value is -0.056 $[-0.103, -0.007]$: negative, which is the suppressive sign the monotonicity argument wants. We looked at this seriously and it does not work, for two reasons.

The first is that the construction is circular. At $k = 2$ the gap between the two copies *is* the treatment, so the cell measures deposition position and the spacing effect together and cannot separate them; reading the whole of it as confound (5) assumes there is no spacing effect at $k = 2$, which is the proposition the confound analysis exists to establish. That the readout model is held fixed is a genuine improvement over the length arm above, and it is why the cell is worth reporting, but holding readout fixed does not by itself isolate deposition when deposition and gap are the same variable.

The second is empirical, and it is what settles it. If the cell did estimate a confound with a consistent sign, the sign would be consistent across arms. It is not. Over all ten arms and both measures the $k = 2$ contrast is suppressive and resolved in four cells, unresolved in thirteen, and *adverse* and resolved in three: the replication arm at $+0.255$ $[+0.123, +0.400]$ bits and L at $+0.281$ $[+0.174, +0.390]$ bits on exposure, and natural prose at $+0.024$ $[+0.008, +0.039]$ nats on depth. Three arms therefore put the $k = 2$ contrast in the direction that could manufacture the rise, resolved. Quoting the primary arm's depth cell as evidence for monotonicity would be selecting one favourable cell from twenty, so we report the tally instead and leave confound (5) bounded by the length arm rather than established by this one.

C Estimators, canonical values and run-level inference

C.1 The canonical interval table, and the bootstrap seed

This appendix is the single site for every interval in the paper, and it fixes the procedure that produces them. Every interval is a cluster bootstrap resampling *runs*, carrying all of a run's blocks together, at $4{,}000$ resamples, drawn from a NumPy `default_rng` seeded at **20260823**. Where an interval is quoted in more than one place, the value in the table below is the one that governs; the duplicates predate the fixed seed, which is the whole of why they differ. The spread between them bounds the resampling noise at **one unit in the second decimal**, which is smaller than every substantive difference the paper rests on but is not something a reader should have to infer.

The procedure is stated because a bound quoted to three digits is reproducible only with its resampling seed: the primary arm's depth massed penalty is $+0.093$ $[+0.045, +0.139]$, and every interval in this paper comes from the procedure above at the seed it names.

arm	measure, contrast	estimate	interval
primary, 36 seeds	depth, vs. massed	$+0.0927$	$[+0.0451, +0.1393]$, $p = 5.8 \times 10^{-4}$
primary, 36 seeds	depth, vs. widest	$+0.0571$	$[-0.0054, +0.1178]$, $p = 0.139$
primary, 36 seeds	exposure, vs. massed	$+0.7331$	$[+0.5720, +0.9032]$, $p = 7.3 \times 10^{-10}$
primary, 36 seeds	exposure, vs. widest	$+0.2753$	$[+0.0765, +0.4822]$, $p = 1.2 \times 10^{-2}$
onset comp., standard	exposure, vs. massed	$+0.9403$	$[+0.7208, +1.1700]$
onset comp., staggered	exposure, vs. massed	$+0.5524$	$[+0.2543, +0.8885]$

Fixing a seed makes the intervals reproducible from the same run files. It does not make them reproducible from a reimplementation, because the runs themselves are not bit-identical (Appendix A.2). And a fixed seed is not a free choice made after seeing results: it is the submission date, chosen for that reason, and no interval in this paper was selected by trying seeds.

C.2 The s^* estimator, and every reported value

Two estimators appear and every value is labelled with both estimator and grid. *Grid argmax* is the spacing with the largest mean effect, quantised to the grid. *Interpolated* fits a parabola in $\log_2 s$ through the grid argmax and its two neighbours and takes the vertex, clipped to the grid range; a grid-edge argmax cannot be bracketed and is reported as a bound. Bootstrap intervals resample *runs*, carrying all of a run's blocks together, as everywhere else in this paper and as stated above. Resampling blocks instead would understate every interval. The resample count is 2000 to 4000 resamples. "Bootstrap peak support" is the fraction of resamples in which a given spacing is the argmax. "Win rate" is the fraction of blocks, or of runs at run level, in which the paired difference is positive.

reported value	arm	grid	estimator
$s^* = 20$	primary	$\{1, 8, 20, 65\}$	grid argmax
$\hat{s} = 17.9$	primary	$\{1, 8, 20, 65\}$	depth
$\hat{s} = 23.8$	primary	$\{1, 8, 20, 65\}$	exposure
$\hat{s} = 14.3$	C4	$\{1, 8, 20, 65\}$	depth
$\hat{s} = 16.6, 17.5, 36.5, 45.3$	lr $\times 0.25, 0.5, 1, 2$	$\{5, 10, 20, 40, 80, 120\}$	exposure
$\hat{s} = 31.6, 31.0, 23.8, 34.6$	lr $\times 0.25, 0.5, 1, 2$	$\{5, 10, 20, 40, 80, 120\}$	depth
$\hat{s} = 16.1, 63.1, 67.0$	$T = 3{,}000/6{,}000/12{,}000$	$\{5, 10, 20, 40, 65, 100\}$	exposure
$\hat{s} = 35.0, 31.2, 10.7$	$B = 16, 32, 64$	$\{5, 10, 20, 40, 65\}$	exposure
$\hat{s} = 67.6$	wide grid	$\{1, 20, 65, 150, 300, 600\}$	depth
$\hat{s} = 88.6$	wide grid	$\{1, 20, 65, 150, 300, 600\}$	exposure
$\hat{s} = 22.3 \to 5.0$	matched tokens, B=32 $\to$ 64	$\{5, 10, 20, 40, 65\}$	depth
$\hat{s} = 65.0 \to 18.0$	matched tokens, B=32 $\to$ 64	$\{5, 10, 20, 40, 65\}$	exposure
the length-sweep cells Appendix G.2 quotes			
$s^* = 26.0, 5.0, 67.5$	$T = 3{,}000/6{,}000/12{,}000$	$\{5, 10, 20, 40, 65, 100\}$	interp., depth, 5.0 clipped
$s^* = 65.0, 29.8, 10.4$	$B = 16, 32, 64$	$\{5, 10, 20, 40, 65\}$	interp., depth, 65.0 clipped
$\hat{s} = 9.6$	B=64, canary-fraction control	$\{5, 10, 20, 40, 65\}$	depth
$\hat{s} = 18.7$	B=64, canary-fraction control	$\{5, 10, 20, 40, 65\}$	exposure
$s^* = 40.0$	lr $\times 2$, lowest-loss arm	$\{5, 10, 20, 40\}$	interp., depth, clipped
$\hat{s} = 14.3$	C4, highest-loss arm	$\{1, 8, 20, 65\}$	depth
$\hat{s} = 17.9$ and 29.8	primary and B=32, matched loss	$\{1, 8, 20, 65\}$, $\{5, 10, 20, 40, 65\}$	depth
$\hat{s} = 8.9$ and 10.4	XL and B=64, matched loss	$\{1, 8, 20, 65\}$, $\{5, 10, 20, 40, 65\}$	depth

The β_2 arm's interpolated values are 77.7 [31.0, 120] at $\beta_2 = 0.5$, 37.8 [25.2, 65.0] at 0.90, 24.4 [21.6, 61.6] at 0.95, 38.8 [33.1, 46.5] at 0.99 and 28.8 [18.0, 43.0] at 0.999, each from six spacings at 12 seeds (Appendix H.4), plus 14.1 and 5.0 for the two matched-convergence length cells on the three-point common grid (Appendix G.3). With those included, every value quoted anywhere in the paper appears in this appendix. Which comparisons are legitimate follows directly: the four learning-rate arms share $\{5, 10, 20, 40, 80, 120\}$, the three batch arms share $\{5, 10, 20, 40, 65\}$, the three length arms share $\{5, 10, 20, 40, 65, 100\}$, and the wide-grid value must not be compared with any of them, since its grid, count set, canary density and training length all differ. The learning-rate arms are read on their full common grid rather than on $\{5, 10, 20, 40\}$, and letting it go stale is exactly the failure mode this appendix exists to prevent. The learning-rate grid runs to $s = 120$: the point estimates are unchanged, because the argmax and its neighbours are, and the two upper intervals are [26.5, 58.3] and [27.6, 63.6] rather than clipped at 80, which is what makes the $\times 1$ against $\times 2$ comparison informative at all.

The canary-fraction control's s^* is 9.6 on depth and 18.7 on exposure, which is what Appendix G.2 concludes from, since 9.6 sits with the $B = 64$ value of 10.4 rather than the $B = 32$ value of 29.8. And one number needs its estimator named: the "6.5× shift in s^* across a 4× change in batch" is the *grid argmax* on depth, $65 \to 10$; the interpolated values in the same column give 6.25×, and the exposure triple gives 3.27×. All three exceed the 1.7× floor, so the conclusion is unchanged, and the table shows which column produces it.

Each of the three exponents rests on one adjacent pair. The mechanism exponents of Appendix H.3 and the training-length row of Appendix H.3 are log-log fits through three points, or four for the learning rate, and decomposing each into its adjacent-pair local slopes shows the same structure every time:

exponent	measured s^*	adjacent-pair ratios	slope
λ, learning rate	16.6, 17.5, 36.5, 45.3	1.05×, **2.09**×, 1.24×	+0.54
δ, training length	16.1, 63.1, 67.0	**3.92**×, 1.06×	+1.03
γ, batch size	35.0, 31.2, 10.7	1.12×, **2.92**×	−0.86

The last column is the raw slope $d \log_2 s^* / d \log_2 x$, so for batch size γ is its negation under the convention $s^* \propto B^{-\gamma}$: 0.86 on these three arms, and 0.23 [0.05, 0.38] once the axis is extended to four octaves (Appendix G.2). In each row one pair, in bold, carries the whole slope and every other is flat: 1.05, 1.24, 1.06 and 1.12 all sit inside the 1.7× estimator noise floor established above. The learning-rate row is the informative case, because we extended it deliberately. Adding a fourth arm at lr × 0.25 doubled the range to 8× and did *not* resolve the problem: the new point is flat against its neighbour at 1.05×, so the arm went from a two-point estimate with one uninformative point to a two-point estimate with two, and the slope interval widened from [+0.23, +1.12] to [+0.19, +1.22] rather than tightening. That is the outcome we report. So each exponent is effectively a two-point estimate with the remaining points uninformative, and none of the three is distinguishable from a threshold somewhere between the two sampled values that bracket the moving pair. That is a statement about functional form, not about precision, and it is not repaired by widening an interval: a power law is simply not

identified by three points one of whose gaps is flat. What the arms do establish is the sign and the order of magnitude, which is what Appendix H.3, both rows, claim. For δ that is no longer load-bearing for anything: what it was load-bearing *for* was the refutation of the token invariant, and Appendix G.3 withdraws both. δ is retained here only as a record of what the unmatched sweep measured.

The length row's middle interval reaches exactly 100.0 on the common grid, the grid edge, so $s^*(6{,}000) = 63.1$ read as a lower bound there, and a larger middle value would put that point further above the line through the endpoints, that is further towards saturation above $T = 6{,}000$. We ran the 12 runs extending that arm to $s = 150$, which is the measurement this arm owed. They resolve it in the favourable direction: at $T = 6{,}000$ the peak stays at $s = 65$ and exceeds the new right edge at $s = 150$ by $+0.114$ bits, bootstrap $P(\text{edge} \geq \text{peak}) = 0.030$, so that s^* is interior rather than censored and the middle point is where the common-grid fit put it. All three peaks in the sweep are interior on exposure.

Is the length arm's s^* interior to its own grid? The exponent $\delta = 1.03$ (the training-length row) rests on s^* rising to about 67 steps at $T = 12{,}000$ on a grid whose right edge is $s = 100$, and an argmax two grid points from an edge invites the objection that the peak is the grid's rather than the data's. We therefore extended the two longer arms rightwards, 12 further runs per added spacing. Leakage relative to $s = 5$, on exposure, in bits, blank where the grid does not extend:

T	s=5	10	20	40	65	100	150	250	400
3,000	0	+0.036	**+0.058**	−0.036	−0.077	−0.197			
6,000	0	−0.004	+0.014	+0.069	**+0.144**	+0.067	+0.030	−0.117	
12,000	0	+0.026	+0.068	+0.093	**+0.273**	+0.164	+0.113	+0.185	+0.080

The peak-against-right-edge test at each length, on the widest grid each arm has, with the bootstrap and the Wilcoxon side by side because they disagree in one cell:

T	peak vs. edge	margin (bits)	$P(\text{edge} \geq \text{peak})$	Wilcoxon p
3,000	s=20 vs. s=100	+0.255 [+0.140, +0.362]	0.000	0.002
6,000	s=65 vs. s=250	+0.261 [+0.167, +0.350]	0.000	0.001
12,000	s=65 vs. s=400	+0.192 [+0.027, +0.355]	0.010	**0.052**

All three peaks are interior to grids two to four times wider than the ones they were measured on, so the three values δ is fitted through are not grid artefacts. The $T = 6{,}000$ arm's edge is $s = 250$, and it resolves more strongly against the wider edge, $p = 0.001$, because leakage falls to -0.117 there. The widest test of the three, at $T = 12{,}000$ against $s = 400$, is **marginal on the Wilcoxon** ($p = 0.052$) while the bootstrap puts it at 0.010; Appendix I states that the Wilcoxon governs where the two disagree, so we report that cell as marginal and note that the two shorter arms are the resolved ones. And the $T = 12{,}000$ exposure tail is not monotone past the peak: $+0.273$, $+0.164$, $+0.113$, $+0.185$, $+0.080$, so $s = 250$ exceeds $s = 150$ by 64%. Twelve runs per cell at differences of this size cannot separate that bump from noise, and it does not touch the peak claim, but it is the reason we describe the descending branch as a fall and not as a shape.

The table above is the authority for this extension, and the five tests in the two tables here are the ones we report. The row holds 288 runs: a base of 216 plus six added cells at 12 each, over seven spacings at $T = 3{,}000$, eight at $T = 6{,}000$ and nine at $T = 12{,}000$.

One qualification, against us. This arm carries almost no signal on depth, as the training-length row states, and the extended grid shows it plainly: at $T = 6{,}000$ the depth profile has no interior peak worth the name, its largest entry sitting at $s = 10$ and falling monotonically thereafter to -0.065 at $s = 250$, and at $T = 3{,}000$ the depth peak is not resolved against its right edge ($p = 0.79$). Only at $T = 12{,}000$ is the depth peak interior and resolved, at $s = 65$ against $s = 400$ by $+0.089$ nats, $p = 0.009$. So the exponent is an exposure result and we do not claim it on depth. On exposure, by contrast, all three peaks are interior once the grid is extended, which is the condition δ needs.

C.3 Run-level inference for every arm

Table 4 in the main text gives the per-arm summary; this appendix gives the intervals and the block-level comparison behind it.

Does effect size fall as the model converges? Retention is pinned near 1,500 steps in all ten arms, so an association with validation loss here would be a second channel rather than the retention result again. Across the ten, loss tracks the raw peak-to-massed ratio at Spearman $\rho = +0.45$ on depth and $+0.32$ on exposure, and the paired penalty at $\rho = -0.22$ and -0.44: **the ratio rises with loss while the absolute gap falls**. The link is the level, which falls with loss at $\rho = -0.59$ and -0.50, the strongest pair here: a less converged arm leaks less in

absolute terms, so the same ratio buys a smaller gap. **None of the six resolves**, exact permutation p over all 10! orderings running 0.08 to 0.54, and the arms differ in width, corpus, rate and panel at once, so these are associations across heterogeneous arms and not an identified convergence effect. The pattern is scale-dependent, and in absolute units the effect is if anything larger in the better converged arms. The identified version needs an arm matched on convergence, which Appendix G.3 runs for the optimum's location rather than for the penalty.

Multiplicity, and what is not corrected. No family-wise or false-discovery correction is applied anywhere in this paper, and the secondary tables carry marginal cells, $p = 0.049$, 0.05 and 0.052 among them, that a correction over all arms, measures and counts would not leave standing. Two things carry the weight instead of a correction. The primary contrast, its measure and its comparator were fixed in code before the runs, so the headline is one test and not the best of many. **The family, stated so it can be checked**: the pre-registered test is the massed contrast on the primary arm, and everything else in the paper is either a replication of it or a different question. And the robust claim rests on sign agreement, 10 of 10 arms on exposure and five panels each resolving alone, which is a different kind of evidence from one threshold crossed once. Read the secondary cells as descriptive, and a marginal one as marginal.

The replication arm's exposure ratio of 5.68 [3.43, 19.40] is the one entry that should not be read as an effect size. Its massed cell, the denominator, is both the smallest in the table and the one cell in the paper with dropped injections (Section 3), so the ratio is dominated by its denominator; this is the concrete reason the drop-free nine are tabulated separately and why the abstract's range is taken over them. **Two ranges are quoted over those nine and they are different quantities**, which is worth stating plainly because the two are easy to swap. The *raw* peak-to-massed ratios run 1.19 (natural, depth) to 2.54 (B, depth) in Table 4, with every interval excluding 1 except XL on depth, the arm that fails that contrast. The *fitted* schedule gains $C(s^*)/C(s{=}1)$, which are what the abstract and Section 4.5 quote, run 1.22 (XL and natural prose, depth) to 1.95 (B, exposure). The replication arm's 2.66 on exposure lies above that range and is excluded from it for the reason just given. The upper end has the same denominator caveat as the replication arm's, in weaker form: the 1B arm's massed cell is the second smallest in the table, behind the replication arm's, so 2.54 is a ratio inflated by a small denominator and not the largest absolute penalty, which is XL′'s (Appendix F.3).

Block-level and run-level values for both required contrasts. Run level averages within a run before testing across runs; the interval is a cluster bootstrap resampling runs with all of a run's blocks carried together, 4000 resamples. The 12-seed arms are at the Wilcoxon floor 4.9×10^{-4}, marked† in the table.

The second column counts matched *seeds*, the unit the paired tests are taken over, whereas the per-arm accounting of Appendix A.3 counts training *runs*, which is the same arm at four spacings: the primary arm's 36 seeds here are the 144 runs there. Both columns are correct under their own definition, and they are not interchangeable.

		depth, vs. massed			depth, vs. widest		exposure, vs. widest	
arm	seeds	diff	p block	p run	p block	p run	p block	p run
primary	36	+0.093	4.8e-08	5.8e-04	3.9e-03	0.14	2.2e-03	1.2e-02
replication	36	+0.131	8.4e-23	1.6e-09	2.3e-06	2.3e-03	9.1e-08	2.8e-05
S	12	+0.060	4.5e-04	3.4e-03	1.5e-02	0.18	3.4e-04	4.9e-03
L	12	+0.181	4.5e-10	9.8e-04	2.0e-04	0.11	1.6e-03	9.3e-03
XL	12	+0.101	1.4e-02	0.20	2.1e-02	0.30	5.7e-03	0.27
XL′	12	+0.351	7.3e-11	4.9e-04†	2.9e-09	9.8e-04	4.2e-06	4.9e-03
XXL	12	+0.219	2.4e-05	2.1e-02	2.4e-02	0.27	3.8e-01	0.42
C4	24	+0.043	4.1e-04	3.4e-02	1.4e-06	6.4e-05	2.3e-07	3.0e-05
natural	12	+0.152	5.6e-13	4.9e-04†	1.0e-04	2.4e-03	3.5e-01	0.15
B	12	+0.170	1.1e-03	2.7e-02	7.0e-01	0.73	8.1e-01	0.57

Block-level values are Wilcoxon contrasts unless labelled Friedman; omnibus Friedman values differ from contrast values, and we quote contrasts throughout.

The peak is selected on the data, and here is the test that absorbs the selection. Contrasts above and in Table 4 are taken at each arm's observed peak spacing. Only $s = 20$ was pre-specified, from the primary arm; S and XL peak at $s = 8$, so for those two the cell was chosen after seeing the data, and selecting the largest interior cell before testing it against both endpoints makes an unadjusted pairwise p anti-conservative. The randomisation test that matches the claim uses

$$T = \max_{s \in \{8,20\}} \min\big(\bar{m}_s - \bar{m}_1,\ \bar{m}_s - \bar{m}_{65}\big),$$

the conjunction the interior claim requires, with the maximum over interior cells inside the statistic so the selection is absorbed. Its null distribution comes from permuting the four spacing labels independently within each seed, 20,000 draws: the four runs sharing a seed differ only in the spacing treatment, so under the null of no spacing effect they are exchangeable, which makes this the standard randomisation test for a matched-block design with four treatments.

Two statements about what the test does and does not license, because T is easy to over-read. The null being permuted is **no spacing effect at all**, not *monotone* spacing effect, and we fitted no monotone alternative as a comparator, so the p-value is not a likelihood comparison between an interior and a monotone profile. What earns it more than a generic omnibus is the shape of T: it is a *minimum* against both endpoints, so $T > 0$ requires the sample peak to be interior and to beat both ends, and a monotone profile in either direction drives T negative. So a significant positive T does bear on interiority rather than merely on the existence of some spacing effect. It remains a test against the no-effect null and should be read that way.

	depth			exposure		
arm	peak	T	perm. p	peak	T	perm. p
primary	20	$+0.057$	0.0066	20	$+0.275$	0.0027
replication	20	$+0.066$	0.0001	20	$+0.416$	< 0.0001
S 3.2M	8	$+0.060$	0.0027	20	$+0.344$	< 0.0001
L 25.2M	20	$+0.130$	0.0031	20	$+0.376$	0.0092
XL 85.1M	8	$+0.099$	0.0465	20	$+0.330$	**0.0523**
XL′ 85.1M	20	$+0.245$	< 0.0001	20	$+0.737$	0.0013
XXL 302.3M	20	$+0.112$	0.0389	20	$+0.146$	**0.2737**
C4	20	$+0.043$	0.0018	20	$+0.328$	< 0.0001
natural	20	$+0.069$	0.0067	20	$+0.103$	0.0336
B 1,007.2M	20	$+0.033$	**0.2985**	20	$+0.042$	**0.4001**

This must not be read as a conservative correction, because it is not one. The permutation p is *smaller* than the unadjusted vs.-widest Wilcoxon in most rows, and the interior optimum resolves in 16 of 20 arm-by-measure combinations against the 6 of 10 and 4 of 10 that Table 4 tallies. The two are different tests: the Wilcoxon column uses two of the four cells and only the signs of their differences, while T uses all four cell means and their magnitudes, so it is both the more powerful test and the selection-absorbing one. What that licenses is narrow and is the same thing Section 4.4's pooling said: the vote counts understate the evidence that the *measured curve turns over*, and 4 of 10 was in part a property of the test and not of the data. Two of the four failures are the 1B arm, on both measures at $p = 0.30$ and $p = 0.40$, so the tenth arm supports the massed penalty and gives no evidence at all of a turn; that is the direction the tally moved when it was added, and it moved against the interior reading. It licenses nothing about identification. Section 4.7 shows the descending branch is largely absorbed by decay-weighted copy count, and a sharper test of the shape is not evidence that the turn is a change in per-exposure gain. The two failures are XL on exposure, at the boundary, and XXL on exposure, which fails decisively; XL is the arm that fails the massed penalty too.

C.4 Pooled across-arm estimates

Pooling of the run-level paired difference across arms, each standardised by that arm's own clustered-cell mean, so the entries are *fractions of the clustered level* and $+0.357$ means 35.7% of it. Standardising is what makes nats and bits comparable and stops a single high-leakage arm dominating. Fixed effects would be the wrong model, because the arms differ deliberately in scale, corpus, learning rate and canary density.

Three estimators are reported rather than one, because the interval is doing real work here and DerSimonian-Laird is known to under-cover at high heterogeneity with few studies, which is exactly this case at I^2 of 75 to 97% over nine arms. HKSJ replaces DL's variance with one built from the observed dispersion of the arms and refers it to t on eight degrees of freedom, which is the recommended default in that regime. The permutation column flips the signs of whole arms, which needs no distributional assumption at all; with 9 arms it enumerates all $2^9 = 512$ sign vectors exactly, so its smallest attainable two-sided p is $2/512 = 3.9 \times 10^{-3}$ and all four rows sit at that floor, meaning the observed pooled effect is the most extreme of every sign assignment.

measure	contrast	DL pooled	HKSJ	perm. p	I^2	drop-free only
depth	massed penalty	$+0.357$ [0.236, 0.477]	[0.214, 0.499]	$3.9 \times 10^{-3\,\dagger}$	89%	$+0.308$ [0.212, 0.405]
depth	interior	$+0.263$ [0.146, 0.381]	[0.154, 0.373]	$3.9 \times 10^{-3\,\dagger}$	75%	$+0.256$ [0.129, 0.384]
exposure	massed penalty	$+0.886$ [0.588, 1.184]	[-0.018, 1.790]	$3.9 \times 10^{-3\,\dagger}$	97%	$+0.571$ [0.366, 0.776]
exposure	interior	$+0.322$ [0.172, 0.473]	[0.046, 0.598]	$3.9 \times 10^{-3\,\dagger}$	81%	$+0.266$ [0.151, 0.382]

The arms are not independent studies. Random-effects pooling assumes independence across studies and ours are not independent: primary and replication differ only in canary density, and S / L / XL / XL′ / XXL differ only in width inside one architecture family, one corpus and one codebase. Violating independence understates the interval, in the direction that flatters us. Collapsing each dependent family to one observation by inverse-variance averaging within it, which leaves four independent units, and taking HKSJ intervals on those:

measure	contrast	collapsed to 4 units, HKSJ	verdict
depth	massed penalty	$+0.298$ $[0.155, 0.442]$	non-zero, $p = 7.0 \times 10^{-3}$
depth	interior	$+0.252$ $[-0.003, 0.507]$	**marginal**, $p = 0.052$
exposure	massed penalty	$+0.486$ $[0.015, 0.956]$	non-zero, $p = 0.046$
exposure	interior	$+0.263$ $[0.037, 0.489]$	non-zero, $p = 0.034$

So the massed penalty survives every treatment but one: the nine-arm HKSJ interval on exposure is $[-0.018, 1.790]$ and straddles zero, at $I^2 = 97\%$, which is the regime HKSJ is deliberately conservative in and the reason we report it. Collapsing to four units *narrows* that same contrast to $[0.015, 0.956]$, $p = 0.046$, which looks backwards and is not: most of the heterogeneity is within families, and inverse-variance averaging inside a family removes it before the between-unit variance is estimated, so four mutually consistent units carry a tighter interval than nine mutually inconsistent ones. On depth the massed penalty is non-zero under every treatment. The interior contrast on depth is not: under the most conservative combination it straddles zero at $p = 0.052$. Section 4.4 quotes the DL interval and this straddling is the qualification on it. The point estimates move little, which is what one expects when the dependence affects precision more than location.

Both contrasts are non-zero in aggregate under this nine-arm treatment even though the interior contrast reaches significance in only 4 of the ten arms on depth: the vote count was understating the evidence, as one would expect when six arms of 12 seeds each share a sign against a Wilcoxon floor of 4.9×10^{-4}. Excluding the replication arm changes little. The pooling covers the nine arms by width, up to 302.3M and not the 1B arm, which would raise the massed row and lower the interior one.

Pooling does less than it appears to. It does not remove heterogeneity: I^2 between 75 and 97% means the pooled figure is an average over arms whose true effects genuinely differ, so it is a typical magnitude and not a constant. It does not make the arms independent, which is why the collapsed row is the one to quote. And it does not address identification: the interior contrast being reliably non-zero says the measured curve turns over, not that the turn reflects a change in per-exposure gain rather than the ageing of earlier copies (Appendix H.1), which points the other way. Those two facts are not in contradiction; they answer different questions, and Section 4.4 points forward to the second so a reader meeting the pooled estimate first is not misled.

C.5 Baseline contamination in depth

Depth is a never-injected control's NLL minus the target's (Section 3), and every canary shares the prefix template, so total injection load is a nuisance parameter. Halving canaries per set from 8 to 4 moved the control's NLL by 0.353 nats while the target's absolute NLL at $k = 64$ moved only 0.122:

grid	injections/run	control NLL	k=64 NLL	depth
8 canaries/set	1008	5.155	4.573	$+0.582$
4 canaries/set	504	5.508	4.695	$+0.813$

So most of the apparent change in memorisation is the baseline moving, and only within-grid comparisons of depth are valid. Exposure does not have this problem because it ranks the true secret against alternatives that share the prefix, so template learning cancels in the comparison.

D The separable law: fitted exponents, per-count cells and the exchange rate

D.1 Fitted β and $C(s)$ per arm

Inputs to Eq. 2. The rows below are emitted by the same script that computes the headline rates, so the two cannot drift apart: substituting any row into Eq. 2 reproduces the quoted rate to within 0.1 points. Values are given to four significant figures because two arms have $C \approx 0.008$, where four decimal places leave enough ambiguity to move the recovered rate by half a point. β is fitted per arm and is not portable across arms (Section 4.1).

The nested tests behind Section 4.1, in full. Stated here with the conversion they support.

Converting a spacing effect into a count of duplicates needs a form relating the two. We write leakage as a spacing gain times a count profile, $\text{leakage}(k, s) = C(s)\, k^{\beta}$. That form is tested, not assumed. Three nested models per arm. M_1: the product form. M_2: separable with a free count profile, i.e. the matrix of cell means over (s, k) has rank one. M_3: a free mean per cell. M_1 against M_2 tests the power law, M_2 against M_3 separability, with wild cluster bootstrap p-values over runs.

The power law survives and exact separability does not. M_1 is not rejected against M_2 in 10 of 10 arms on depth and 9 of 10 on exposure, while M_2 against M_3 gives $p \leq 0.05$ in 7 of 10 and 6 of 10. The departure is real, and its size decides whether it matters. The worst cell departs by 4.5 to 13.1% of its arm's largest cell, so the product form is good to about 10% while the schedule effect it describes is 20 to 60% (Figure 6). One caveat on that scale: normalising by the largest cell understates relative structure at small k, where the primary arm's $k = 2$ discrepancy is about 30% of its own value.

Sub-additivity, which the conversion depends on, has an exception and the 1B arm is it. Up to 302.3M the exceptions differ by measure, so no arm fails on both. $\beta < 1$ on depth everywhere and in 7 of 9 on exposure. Doubling the count therefore multiplies primary-arm leakage by $2^{0.48} = 1.39$, not 2. The 1B arm is *super*-additive on *both*, $\beta = 1.31$ and 1.26, the first arm to fail on both. That is what a model that has barely begun to memorise looks like. Its absolute leakage is the lowest here, so each further duplicate still buys more than the last. We quote no exchange rate for it, since the conversion's reading depends on $\beta < 1$. β is shared across spacings within an arm but not across arms, ranging 0.330 to 1.31 on depth, so we quote none for memorisation in general, which the exchange rate below divides by.

The exchange rate, its derivation and the three conditions it adds. The derivation is long and the three conditions it adds are what a reader checking the conversion needs, rather than what a reader reading the claim needs. Under leakage$(k, s) = C(s)k^\beta$ the absolute reduction from cutting k to k' is $C(s)(k^\beta - k'^\beta)$, proportional to $C(s)$, which is why the absolute benefit is schedule-dependent by the schedule gain and the relative reduction $1 - (k'/k)^\beta$ is not. Two limits bind the magnitude. Separability is rejected in 7 of 10 arms, though the worst cell departs by only about 10% of its arm's largest. And Section 4.6's $k = 2$ depth cell is negative, which is why every rate we quote is refitted on the resolving range as above. The same quantity restates as an *equivalent deduplication rate*, the fraction of duplicates one would delete to buy the same reduction, 55.9% on exposure at $k \geq 8$ under approximate separability and conditional on this panel. Inverting the count profile brings in all three conditions the gain avoids, so the gain is what we lead with (Figure 11).

The form cannot change sign, and is never fitted where it would have to. $C(s)k^\beta$ takes one sign for every $k > 0$, while depth is a difference of losses and does change sign: it is -0.056 nats at $k = 2$ and positive by $k = 64$. A power law through data that crosses zero would be misspecified. **Every rate we quote** comes from a fit on the resolving range, $k \geq 8$ on exposure and $k \geq 16$ on depth, where depth is positive in all ten arms, so the sign change falls outside the fitted range rather than inside it. The coefficient table above and Section 4.1's separability tests use every cell instead, and Appendix D.3 prints both fits side by side. Exposure never changes sign at any count. This is why the EDR is quoted at $k \geq 8$ and why quoting it for a $k = 2$ to 5 corpus would be a misuse of the number.

The schedule gain, which is what Section 4.5 leads with. $C(s^*)/C(s{=}1)$ is a ratio of fitted spacing gains at fixed count, so unlike the EDR it does not divide by β and inherits neither the exponent's censoring nor the $\beta > 1$ exception. Intervals resample *runs* with C and β refitted inside each resample, 2000 draws at the seed of Appendix C.1. The peak is the argmax of the arm's own fitted $C(s)$, which is $s = 20$ everywhere except S on depth and the 1B arm on both, where it is $s = 8$.

	depth		exposure	
arm	ratio	95%	ratio	95%
primary	1.30	[1.19, 1.43]	1.72	[1.56, 1.93]
replication	1.61	[1.49, 1.74]	2.66	[2.33, 3.14]
S 3.2M	1.32	[1.21, 1.47]	1.76	[1.52, 2.21]
L 25.2M	1.41	[1.27, 1.56]	1.64	[1.47, 1.91]
XL 85.1M	1.22	[0.99, 1.55]	1.50	[1.32, 1.72]
XL′ 85.1M	1.60	[1.42, 1.84]	1.70	[1.53, 1.91]
XXL 302.3M	1.45	[1.23, 1.71]	1.71	[1.49, 2.00]
C4	1.34	[1.20, 1.50]	1.40	[1.29, 1.53]
natural prose	1.22	[1.20, 1.25]	1.25	[1.18, 1.34]
B 1,007.2M	1.73	[1.26, 2.28]	1.95	[1.25, 3.44]

Nineteen of the twenty intervals exclude 1. The exception is XL on depth, [0.99, 1.55], which is the same arm and measure that fails the massed-penalty contrast in Table 4, so it is one arm failing on one measure and not a separate failure. These intervals are **within-panel** and are not widened by the between-panel component of Appendix E.1; a reader wanting the panel-marginal version should apply the same quadrature the per-arm table there applies to the penalties.

Why linear space, and why a wild bootstrap. The fits in the table above are least squares in linear space, using every cell; the quoted rates are refitted on the resolving range (Appendix D.3). The natural alternative, a log-space fit of the multiplicative model, is not usable on this design: leakage at small counts is small and noisy, so 35% of run-level cells at $k = 2$ and 26% at $k = 4$ are negative. A log fit must discard them, and the runs that

survive are those whose small-count cells happened to land high, which is selection on the noise. Fitting it that way moves the EDRs in *both* directions, down for the primary arm (54.4% to 49.0% on exposure) and up for the replication arm (57.4% to 70.1%), because the discard rate differs by spacing and so moves the $C(s)$ ratio as well as β. The largest movement is C4: with its cells restored, it is 25.5% [17.2, 37.8] on depth and 34.2% [24.7, 44.9] on exposure. Bootstrap intervals resample runs, not blocks, for the reason given in Section 4.2.

The price of linear space is heteroscedasticity, so the nominal F distribution does not apply and every p-value in Section 4.1 is a wild cluster bootstrap over runs with *Webb six-point* weights. Two-point Rademacher signs admit only 2^n distinct sign vectors, which at $n = 12$ runs is 4096 and gives a coarse reference distribution; the six-point weights are the standard remedy below roughly 12 to 15 clusters. Recomputing with Rademacher signs changes no tally and no verdict in any arm: no p_{23} crosses 0.05, the same seven arms on depth and six on exposure reject separability under both weightings, and among the rejecting arms the largest movement is the 302.3M arm on depth, 0.0135 against 0.0185. The largest movement anywhere is C4 on depth, 0.481 against 0.453, which is far from any threshold. We report Webb because it is the right choice at this cluster count, not because it changed anything. Signs rather than resampling matter here: drawing whole runs' residuals with replacement leaves the interaction present in every draw, so the reference distribution centres on the observed statistic and every p-value collapses toward 0.5. The nominal F-test errs the other way and is far too conservative, since the linear model carries no run blocking factor, so between-run differences in overall leakage swamp the interaction: nominal p reaches 0.95 where the bootstrap gives 5×10^{-4}.

Is $\beta < 1$ an artefact of the exposure ceiling? Exposure is bounded at $\log_2(2N) = 9$ bits, and a bounded metric must flatten as it saturates, which mimics sub-additivity. The most extreme arm is XL′ (85.1M at lr 3×10^{-4}), whose largest cell reaches 95.6% of the ceiling at $k = 64$. XL′'s exposure contrasts and EDR are therefore attenuated, and its depth figures are the ones to prefer. Appendix E.2 now quantifies the saturation rather than inferring it from the mean: 68.8–77.1% of that cell's canaries sit exactly at the ceiling, and XXL's corresponding cell is 58.3–66.7%. Those two arms' exposure β should be read as ceiling-limited outright, not merely as possibly attenuated. Across arms this is not what sets the exponent: β correlates with the fraction of headroom used at only -0.14 on exposure and -0.10 on depth. Within an arm it is real, and refitting on the counts whose mean exposure stays below half the ceiling raises β by up to 0.13. Depth has no ceiling, and $\beta < 1$ there in 9 of 10 arms, so sub-additivity is a property of the measurement's unbounded form too and not an artefact of the bounded one.

	depth					exposure				
arm	β	$C(1)$	$C(8)$	$C(20)$	$C(65)$	β	$C(1)$	$C(8)$	$C(20)$	$C(65)$
primary	0.480	0.1127	0.1368	0.1466	0.1224	0.693	0.1615	0.229	0.2782	0.2174
replication	0.695	0.02932	0.04391	0.04712	0.03504	1.150	0.01198	0.02453	0.03191	0.01842
S 3.2M	0.855	0.01903	0.02507	0.02407	0.01821	1.228	0.01166	0.01841	0.02057	0.01305
L 25.2M	0.493	0.1099	0.1356	0.1546	0.1226	0.628	0.2468	0.3364	0.4054	0.329
XL 85.1M	0.413	0.1638	0.1994	0.2005	0.1728	0.532	0.4006	0.4972	0.6016	0.5131
XL′ 85.1M	0.515	0.1313	0.1844	0.2103	0.1535	0.613	0.3608	0.5158	0.6118	0.481
XXL 302.3M	0.552	0.1062	0.131	0.1535	0.1194	0.708	0.1996	0.2566	0.3422	0.2963
C4	0.998	0.007753	0.0102	0.0104	0.006204	0.807	0.1071	0.1452	0.1502	0.08959
natural	0.330	0.326	0.3708	0.3987	0.36	0.358	0.4501	0.5249	0.5647	0.5134
B 1,007.2M	1.305	0.002812	0.004877	0.004814	0.003886	1.261	0.007544	0.01475	0.01365	0.0114

For the realistic arm the three schedules replace the spacing columns: on depth $\beta = 0.393$ with $C = 0.2291$ (random placement), 0.2593 (clustered), 0.3130 (uniform $s = 20$); on exposure $\beta = 0.370$ with $C = 0.4748$, 0.5479, 0.6165.

Separability, per arm. Nested model comparison from Section 4.1. p_{12} tests the power law (M_1 vs M_2) and p_{23} tests separability (M_2 vs M_3); "dev" is the worst cell's departure from the best separable fit as a percentage of the arm's largest cell, with a 95% bootstrap upper bound; "rk1" is the share of the cell-mean matrix's squared spectrum in its leading singular value, which equals 1 exactly when the arm is separable.

	depth					exposure				
arm	runs	p_{12}	p_{23}	rk1	dev	runs	p_{12}	p_{23}	rk1	dev
primary	36	0.091	0.0005	0.9956	6.9% (<8.5)	36	0.075	0.0005	0.9934	8.8% (<11.9)
replication	36	0.789	0.0005	0.9918	9.5% (<12.2)	36	0.001	0.0005	0.9771	12.2% (<15.1)
S 3.2M	12	0.839	0.0380	0.9912	8.3% (<13.4)	12	0.673	0.240	0.9949	6.6% (<11.0)
L 25.2M	12	0.204	0.232	0.9968	6.2% (<12.6)	12	0.630	0.180	0.9960	6.7% (<12.5)
XL 85.1M	12	0.142	0.0045	0.9931	9.7% (<14.8)	12	0.321	0.0055	0.9915	10.6% (<15.0)
XL′ 85.1M	12	0.144	0.0170	0.9958	5.9% (<10.1)	12	0.270	0.0040	0.9937	9.0% (<13.6)
XXL 302.3M	12	0.509	0.0185	0.9913	10.2% (<15.8)	12	0.345	0.0090	0.9880	12.1% (<16.7)
C4	24	0.193	0.453	0.9968	7.8% (<11.9)	24	0.929	0.0050	0.9956	6.2% (<11.1)
natural	12	0.784	0.0010	0.9989	4.5% (<5.8)	12	0.624	0.055	0.9968	7.5% (<13.8)
B 1,007.2M	12	0.415	0.531	0.9810	12.3% (<24.1)	12	0.542	0.377	0.9801	13.1% (<24.1)
wide grid	12	0.777	0.233	0.9956	9.0% (<16.3)	12	0.450	0.066	0.9928	11.6% (<17.3)

BIC prefers M_1 in 10 of 10 arms on depth and 9 of 10 on exposure, the exception being the replication arm on exposure, where it prefers M_2 and the power law is also rejected ($p_{12} = 0.001$). Exposure β exceeds 1 in three of the ten arms, so that is not what singles this one out. The wide-grid arm is listed below the rule because it uses a different spacing grid, count set and canary density, so it is not part of any tally in Section 4.1; it rejects neither model, and its departure bound of 17.3% is the loosest here apart from the 1B arm's 24.1%, which is what a 6×4 grid at 12 runs should be expected to give.

D.2 Contrasts per count cell

Run-level paired differences within each count cell, primary arm, with cluster-bootstrap intervals. Contrasts are taken at the *arm's* peak spacing, $s = 20$, rather than at each cell's own argmax: a per-cell argmax makes the contrast identically zero in any cell that happens to peak at the massed end, which is a definition rather than a measurement. "pred" is the massed penalty the arm's own fitted separable law implies at that count, $(C(20) - C(1))k^{\beta}$, included so that an unresolved cell yields a bound rather than a shrug. Where the interval's upper end lies below "pred", the design *excludes* an effect as large as the fit predicts. That is a finding and not a power failure.

	depth (nats)			exposure (bits)		
k	s^* − massed	p	pred	s^* − massed	p	pred
2	−0.056 [−0.103, −0.007]	4.4e-2	0.047	−0.058 [−0.257, +0.111]	8.2e-1	0.189
4	−0.031 [−0.085, +0.020]	3.2e-1	0.066	+0.006 [−0.209, +0.218]	7.7e-1	0.305
8	+0.022 [−0.034, +0.083]	5.6e-1	0.092	+0.260 [+0.081, +0.460]	1.7e-2	0.492
16	+0.069 [+0.004, +0.133]	4.3e-2	0.128	+0.457 [+0.187, +0.733]	4.4e-4	0.796
32	+0.256 [+0.203, +0.308]	1.5e-10	0.179	+1.595 [+1.325, +1.877]	5.8e-11	1.286
64	+0.297 [+0.216, +0.377]	3.7e-8	0.250	+2.138 [+1.770, +2.580]	2.9e-11†	2.078

On depth the interval excludes "pred" at $k = 2$, 4 and 8, and on exposure at $k = 2$, 4, 8 and 16, where its upper end of +0.733 lies below the predicted 0.796. So in this arm the fitted law over-predicts the massed penalty at low count, which is the same fact the model-free conversion of Appendix D.3 shows from the other side.

The interior contrast, the peak against the *widest* spacing, is the second thing the interior claim needs, and Section 4.6 scopes it by count. It is tabulated here separately because the two measures do not scope the same way:

	depth (nats)		exposure (bits)	
k	s^* − widest	p	s^* − widest	p
2	−0.044 [−0.128, +0.040]	2.6e-1	−0.220 [−0.503, +0.055]	2.6e-1
4	−0.037 [−0.101, +0.031]	2.0e-1	−0.123 [−0.330, +0.089]	1.8e-1
8	+0.018 [−0.046, +0.080]	8.0e-1	+0.079 [−0.159, +0.334]	7.4e-1
16	+0.022 [−0.052, +0.093]	4.7e-1	+0.102 [−0.213, +0.430]	2.3e-1
32	**+0.082** [+0.017, +0.144]	2.3e-2	+0.227 [−0.079, +0.522]	2.3e-1
64	**+0.301** [+0.213, +0.390]	4.9e-7	**+1.587** [+1.201, +1.980]	1.3e-8

So in this arm the interior contrast resolves at $k = 32$ and 64 on depth and at $k = 64$ alone on exposure, and is negative but unresolved at $k = 2$ and 4 on both. The interior claim is therefore narrower in count than the massed penalty, and narrower on the pre-specified measure than on the other, which is consistent with Section 4.7 finding it unidentified there.

The low-count regime across arms. The primary arm is not the whole evidence, and it is the least sensitive at low count. The massed penalty at each arm's peak spacing, at the two smallest counts:

	$k = 2$		$k = 4$	
arm	depth	exposure	depth	exposure
primary (36 seeds)	−0.056 (4.4e-2)	−0.058 (0.82)	−0.031 (0.32)	+0.006 (0.77)
replication (36 seeds)	+0.006 (0.41)	**+0.255** (8.7e-4)	+0.035 (0.056)	**+0.321** (1.2e-4)
XL′ (12 seeds)	+0.054 (0.57)	−0.207 (0.30)	**+0.298** (3.4e-3)	**+0.863** (9.8e-4)
natural (12 seeds)	**+0.024** (2.7e-2)	−0.035 (0.27)	**+0.043** (3.4e-3)	+0.110 (0.054)

Bold marks the cells that are positive *and* resolve at $p < 0.05$, so the bolding is not the same thing as resolving: two of eight resolve positive at $k = 2$ and four of eight at $k = 4$, while at $k = 2$ a third cell resolves negative and is left unbolded. The honest summary is that the low-count effect is present and small where the arm can see it, rather than absent. The one negative cell that reaches nominal significance is the primary arm on depth at

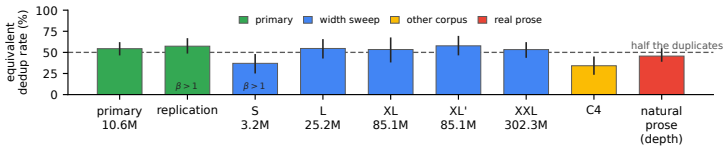

Figure 11: What the schedule axis is worth in deleted duplicates, per arm, each on its pre-specified measure, named in the tick label where it is depth: flat across five widths to 302.3M, lower on a second corpus and on real prose, and empty below $k \approx 8$. **The 1B arm is deliberately absent**, because it is super-additive on both measures and Section 4.1 quotes no exchange rate for it; Eq. 2 still evaluates there but its reading depends on $\beta < 1$. The two arms marked $\beta > 1$ are the remaining ones where the same caveat applies on this measure. Intervals bootstrap *runs* with β and C fitted jointly and are **within-panel**. We do not widen them by the between-panel component: that component is estimated on the penalties at the primary configuration, and a ratio of two fitted gains is a different estimand, so carrying it across would be an assumption and not a correction. Read these as orders, high-count-only and conditional on $\beta < 1$.

$k = 2$; it is one of 48 cell-level tests here, no other arm or measure in this table reproduces it, and the replication arm's $k = 2$ cell points the other way at a much smaller p. We therefore do not claim a low-count inversion, and we do not claim the low-count regime is empty either.

We also note the consequence for how the separability bound should be read. The worst-cell departure in Section 4.1 is normalised by the arm's *largest* cell, which systematically understates structure at small k: the nominal $k = 2$ discrepancy in the primary arm is 5.2% of the largest cell and so falls inside the 6.9% bound, while being of order 30% of that cell's own value. The bound is a statement about absolute leakage, not about relative structure at low count, and should not be read as the latter.

D.3 Is the EDR robust to what it rests on?

The equivalent deduplication rate (EDR) converts a ratio of leakages into a count of duplicates by equating $C(s_{\text{massed}})k^{\beta}$ with $C(s^*)k'^{\beta}$:

$$\frac{k'}{k} = \left(\frac{C(s_{\text{massed}})}{C(s^*)}\right)^{1/\beta}, \tag{2}$$

so $1 - k'/k$ is the schedule effect as a rate. Figure 11 plots it per arm.

The EDR is the most quotable number in this paper and the one most likely to be cited without its conditions, so the three things it depends on are each checked against the same runs.

(i) It is fitted where the contrast does not resolve. Eq. 2 is fitted over $k = 2$ to 64, while Section 4.6 shows the underlying contrast resolves only from $k \geq 16$ on depth and $k \geq 8$ on exposure. Refitting on the resolving cells only:

	exposure, resolving range $k \geq 8$		depth, resolving range $k \geq 16$	
arm	all cells	resolving only	all cells	resolving only
primary	54.4	55.9 [47.5,64.0]	42.2	46.6 [36.9,56.5]
replication	57.4	60.0 [50.6,69.0]	49.5	52.6 [44.6,62.3]
S 3.2M	37.0	38.2 [26.0,51.9]	27.6	27.5 [19.5,37.8]
L 25.2M	54.6	59.1 [48.7,68.7]	50.0	57.0 [47.7,65.1]
XL 85.1M	53.4	53.6 [38.4,68.2]	38.8	43.8 [19.3,71.8]
XL′ 85.1M	57.8	61.0 [48.5,73.8]	59.9	58.8 [48.5,70.8]
XXL 302.3M	53.3	56.9 [45.8,66.1]	48.7	47.9 [36.0,59.0]
C4	34.2	35.4 [25.2,47.4]	25.5	29.9 [19.3,48.8]
natural	47.0	50.3 [35.1,69.4]	45.6	57.5 [41.9,85.5]

The all-cell columns are point estimates only. Their intervals sit behind Figure 11 and are not repeated here, because they come from a separate bootstrap and the two sets of resamples would differ in the last digit. The rate quoted for this arm is the *resolving-only* pair from this table, 55.9% $[47.5, 64.0]$, so it is the refit and not the all-cell fit.

Restricting to the resolving range *raises* the rate, by 2.4 points on average on exposure (range 0.2 to 4.5) and 3.7 on depth (range -1.1 to 11.9), always well inside the intervals. So the all-cell fit is the *lower* of the two, and the headline is not an artefact of fitting through cells that carry no resolved signal. This is a favourable answer to a fair objection.

(ii) The $1/\beta$ exponent, isolated. The reported intervals come from a joint bootstrap over β and $C(s)$ together, which is correct but hides how much of the width is the division by β. Holding the $C(s)$ ratio at its point estimate and moving β to the ends of its own bootstrap interval gives, in the primary arm on exposure, 59.1% at $\beta = 0.607$ and 50.0% at $\beta = 0.785$: about 9 of the 14 points of full interval width. The arm where this dominates is the smallest model, whose β interval $[0.835, 2.163]$ spans a factor of 2.6 and whose rate accordingly ranges 49.3% to 23.1%. That arm's rate should be read as order-of-magnitude only.

(iii) The conversion assumes a product form that Section 4.1 rejects. A bound on cell-mean departure is not a bound on the error of a quantity that divides by β, so we also compute the exchange rate with no functional form: for each measured count k, find by interpolation in $\log k$ the count k' at which the *worst* schedule leaks as much as the clustered schedule does at k, and report $1 - k'/k$.

arm (exposure)	k=2	k=4	k=8	k=16	k=32	k=64	fitted
primary	-19.5	1.9	24.4	54.5	45.0	58.8	54.4
replication	out	out	out	58.7	70.7	59.7	57.4
S 3.2M	-122	out	28.5	67.0	48.9	38.0	37.0
L 25.2M	out	out	34.5	53.0	65.3	57.7	54.6
XL 85.1M	out	14.7	54.1	75.4	53.4	55.2	53.4
XL′ 85.1M	-13.6	41.2	50.6	63.2	74.8	59.8	57.8
XXL 302.3M	out	15.1	36.2	69.0	78.1	48.9	53.3
C4	-200	-25.9	16.3	30.4	34.3	39.1	34.2
natural	-11.9	30.1	49.0	15.8	30.3	69.7	47.0

"out" means the clustered leakage at that count is not bracketed by the worst schedule's measured count profile, so inverting would extrapolate the very surface being used to check an extrapolation. Read from $k = 16$ rightwards, the model-free rates straddle the fitted number in five of the nine arms; in the replication, S, XL and XL′ arms every such rate lies *above* it. The model-free estimator therefore corroborates the sign and the order of magnitude of the fitted figure and does not bound the form error it hides, which in most arms is larger than the bootstrap width. Read leftwards, the rates fall through zero and become large and negative, which is the same fact as Section 4.6 seen through this estimator: at $k \leq 4$ the conversion has no content, and the negative entries are noise around a quantity that is not identified there rather than evidence that spreading helps at low count.

The practical consequence is stated plainly. The EDR is a statement about duplicate counts of 8 and above. For a corpus whose duplicates sit at $k = 2$ to 5, this paper reports no usable conversion, and quoting "worth half the duplicates" for such a corpus would be misuse of the number.

Is $k \geq 8$ a regime that exists? We tried to measure it and cannot answer at this scale. The scope statement above invites the question of whether the regime the conversion applies to is populated at all, so we measured it directly on both corpora: non-overlapping 50-token spans, hashed and tallied, reported as the share of *tokens* living in spans of each multiplicity rather than the share of distinct spans, since a distinct-span histogram is all singleton tail. The answer on these corpora is that it barely exists. WikiText-103 has 99.86% of its tokens in singleton 50-token spans, with 0.14% at $k = 2$ to 5 and a maximum multiplicity of 6; the C4 subset is 99.96% singleton, 0.03% at $k = 2$ to 5, 0.005% at $k \geq 8$ and a maximum of 19.

That is a fact about 118M-token corpora, not about corpora, and the distinction is the whole of it: a span occurring eight times in a 10^{11}-token corpus occurs zero or once in a 10^{-3} fraction of one. So we also measured the scaling, on nested prefixes at $1/32$ to 1 of each corpus, to see whether the regime fills in with size. It does, and the rate is not estimable from a subset: the elasticity of the $k \geq 2$ token share with respect to corpus size is $+1.14$ on WikiText-103 and $+0.28$ on C4, and extrapolating those two slopes to 10^{11} tokens gives answers differing by a factor of about 400. We therefore report the measurement and decline the extrapolation. Two things follow that are worth having. The prevalence question needs a full-corpus measurement we have not done, and no claim in this paper rests on prevalence. And within these experiments the natural corpus contributes essentially no duplicate structure of its own, so the injected canaries are the only source of counts above one, which is a condition the design assumes and which we check here rather than assert.

E PANEL DEPENDENCE, AND THE EXPOSURE CEILING

E.1 DOES THE RESULT DEPEND ON THE ONE CANARY PANEL WE GENERATED?

Every synthetic arm in this paper draws its canaries from one generator seeded at `CANARY_SEED` $= 777$, so all of our uncertainty is conditional on a single draw of secrets and carriers. Fixed content is not *required* by the design: canaries could be resampled per panel and held fixed only across that panel's spacing conditions, which preserves pairing exactly while letting content vary. So we ran four further panels, seeds 778 to 781, on the primary grid at 12 seeds each, verified before launch to share no carrier block and no secret with the published panel or with each other.

panel	seeds	massed penalty s=20 vs. 1	interior s=20 vs. 65
exposure, bits			
777 (published)	108	$+0.802$ [$+0.724$, $+0.881$]	$+0.390$ [$+0.286$, $+0.499$]
778	12	$\mathbf{+1.012}$ [$+0.826$, $+1.184$]	$+0.119$ [-0.279, $+0.504$]
779	12	$\mathbf{+0.960}$ [$+0.755$, $+1.176$]	$+0.548$ [$+0.177$, $+0.944$]
780	12	$\mathbf{+1.083}$ [$+0.899$, $+1.297$]	$+0.122$ [-0.183, $+0.413$]
781	12	$\mathbf{+0.927}$ [$+0.657$, $+1.244$]	$+0.109$ [-0.255, $+0.448$]
depth, nats			
777 (published)	108	$+0.109$ [$+0.088$, $+0.131$]	$+0.083$ [$+0.050$, $+0.116$]
778	12	$\mathbf{+0.217}$ [$+0.155$, $+0.276$]	$+0.018$ [-0.059, $+0.092$]
779	12	$\mathbf{+0.216}$ [$+0.166$, $+0.271$]	$+0.084$ [-0.003, $+0.168$]
780	12	$\mathbf{+0.142}$ [$+0.087$, $+0.202$]	-0.004 [-0.106, $+0.106$]
781	12	$\mathbf{+0.214}$ [$+0.163$, $+0.268$]	$+0.096$ [$+0.017$, $+0.177$]

The massed penalty is not a property of our canaries. It resolves on all four new panels on both measures, at $p = 4.9 \times 10^{-4}$ in seven of the eight cells, which is the Wilcoxon floor at 12 seeds, and 9.8×10^{-4} in the eighth. It is also *larger* on all four new panels than on the published one, on both measures, so the panel we happened to publish gives the smallest of five estimates on each measure and the headline figure is conservative rather than cherry-picked. Five panels is the strongest content-generalisation evidence in the paper.

What this costs us, and which denominator the widening uses. Magnitude is content-dependent, and the size of that dependence is a variance component, not a spread. **The denominator matters here, and the obvious choice is the wrong one.** Dividing the between-panel spread by the *mean* of the panels' standard errors gives 2.5 and 4.5, and it is the wrong denominator: the four new panels have 12 seeds against the published panel's 108, so their intervals are wider for reasons of sample size rather than of content and averaging them in dilutes the factor. The intervals a reader needs widened are the published ones, so the denominator has to be the published panel's.

That leaves the question of how to estimate σ_{between} itself, and it is a real one, because the variance of a panel mean is $\sigma^2_{\text{between}} + \text{SE}^2_{\text{panel}}$: the naive standard deviation of the means contains sampling error and is biased upward. The natural repair is to subtract the mean within-panel variance. On three panels that repair gives 0.065 on exposure and 0.056 on depth against the same three panels' naive 0.109 and 0.062, which would have cut the exposure factor by a third. (The pair to compare it with is the three-panel one; 0.105 and 0.051 are the *five*-panel naive values, and crossing the two is the mistake the table below exists to prevent.) We report it and do not use it, because it is unweighted and the panels differ ninefold in precision: extended to five panels the same estimator subtracts an average within-panel variance dominated by the four small panels, and it returns $\sigma_{\text{between}} = 0.000$ on exposure. An estimator that reports no content variation at all, on data whose heterogeneity statistic is $Q = 9.7$ on 4 degrees of freedom, is not the one to quote, and one whose value collapses when panels are added is diagnosing its own weighting rather than the corpus.

The estimator that handles unequal precision is the standard random-effects one. Its DerSimonian-Laird step weights each panel by $1/\text{SE}^2$, which is a *fixed*-effects weight, to form Q and from it τ^2; the pooled estimate then reweights by $1/(\text{SE}^2 + \tau^2)$. On five panels it gives $\sigma_{\text{between}} = \mathbf{0.1071}$ bits on exposure ($Q = 9.69$, 4 df) and $\mathbf{0.0577}$ nats on depth ($Q = 27.34$, 4 df), with restricted maximum likelihood as a check at 0.099 and 0.048. So the naive figures, 0.109 and 0.062, are within a few percent of the weighted answer, and the two extra panels raise the degrees of freedom from two to four without moving the number. An interval covering panel-to-panel as well as run-to-run variation has width $\sqrt{\text{SE}^2 + \sigma^2_{\text{between}}}$, which for the published panel means multiplying by $\mathbf{2.8}\times$ on exposure and $\mathbf{5.2}\times$ on depth. These remain orders rather than sharp estimates (four degrees of freedom is not many), but they are no longer two, and the correction that would have shrunk them does not survive contact with the extra panels.

Which σ is which, and the inputs behind the two factors. Five σ pairs appear in this appendix and two of them reconcile with one of the two printed factors each, so the table below names all five and

which one governs. The governing pair and its denominator are given together to four figures, with the arithmetic written out. The published panel's massed penalty is $+0.8024$ bits with SE $= 0.04125$ over 108 seeds, and $+0.1094$ nats with SE $= 0.01124$. Then $\sqrt{0.04125^2 + 0.1071^2}/0.04125 = 2.781$ and $\sqrt{0.01124^2 + 0.0577^2}/0.01124 = 5.231$, which are the $2.8\times$ and $5.2\times$ quoted above and in Section 3.

estimator	panels	σ exposure	σ depth	used?
naive SD of panel means	3	0.1093	0.0618	no, three panels and biased up
method of moments	3	0.0651	0.0563	no, unweighted
naive SD of panel means	5	0.1045	0.0506	no, contains sampling error
method of moments	5	0.0000	0.0426	no, collapses on exposure
weighted random effects	5	**0.1071**	**0.0577**	**yes, governs**

The depth factor is the sensitive one: 0.0577 gives 5.23 and rounds to 5.2, while the same arithmetic on σ rounded to 0.058 gives 5.25 and rounds to 5.3. We quote 5.2 because it comes from the unrounded estimate, and print the estimate to four figures for that reason.

Carrying the widening through Table 4, arm by arm. A constant multiplier is the wrong instrument here. σ_{between} is a variance component and adds in quadrature, so an arm's widened standard error is $\sqrt{\text{SE}^2 + \sigma^2_{\text{between}}}$ and the resulting inflation depends on how precise that arm already was: it runs from $1.13\times$ for XL on exposure, whose own interval is wide, to $7.17\times$ for natural prose on depth, whose own interval is narrow. Multiplying every arm by the primary arm's ratio would inflate the imprecise arms by a factor never measured on them.

arm	penalty	nominal 95%	widened 95%	
exposure, bits				
primary	$+0.733$	$[+0.572, +0.903]$	$[+0.465, +1.001]$	survives
replication	$+0.903$	$[+0.775, +1.030]$	$[+0.657, +1.148]$	survives
S 3.2M	$+0.456$	$[+0.351, +0.561]$	$[+0.221, +0.690]$	survives
L 25.2M	$+0.972$	$[+0.749, +1.195]$	$[+0.666, +1.278]$	survives
XL 85.1M	$+0.804$	$[+0.411, +1.197]$	$[+0.358, +1.250]$	survives
XL′ 85.1M	$+1.484$	$[+1.188, +1.781]$	$[+1.121, +1.848]$	survives
XXL 302.3M	$+1.160$	$[+0.784, +1.536]$	$[+0.730, +1.591]$	survives
C4	$+0.328$	$[+0.160, +0.496]$	$[+0.059, +0.597]$	survives
natural prose	$+0.240$	$[+0.168, +0.312]$	$[+0.018, +0.462]$	survives
B 1,007.2M	$+0.439$	$[+0.200, +0.687]$	$[+0.107, +0.770]$	survives
depth, nats				
primary	$+0.093$	$[+0.045, +0.139]$	$[-0.030, +0.216]$	**straddles zero**
replication	$+0.131$	$[+0.104, +0.159]$	$[+0.015, +0.248]$	survives
S 3.2M	$+0.060$	$[+0.034, +0.086]$	$[-0.056, +0.176]$	**straddles zero**
L 25.2M	$+0.181$	$[+0.113, +0.249]$	$[+0.049, +0.313]$	survives
XL 85.1M	$+0.101$	$[-0.034, +0.235]$	$[-0.075, +0.276]$	**straddles zero**
XL′ 85.1M	$+0.351$	$[+0.245, +0.457]$	$[+0.196, +0.506]$	survives
XXL 302.3M	$+0.219$	$[+0.059, +0.379]$	$[+0.023, +0.415]$	survives
C4	$+0.043$	$[+0.010, +0.077]$	$[-0.075, +0.161]$	**straddles zero**
natural prose	$+0.152$	$[+0.136, +0.168]$	$[+0.037, +0.266]$	survives
B 1,007.2M	$+0.170$	$[+0.050, +0.283]$	$[+0.003, +0.337]$	survives, narrowly

So the widening costs four of the ten depth cells and none of the exposure cells, and one of the four is the primary arm, the cell a reader is most likely to quote. Two things bear on how much weight it should carry. It is a variance-component correction resting on four degrees of freedom, and against it stands the direct evidence: the massed penalty resolves on *all five* panels independently, on both measures, at or near the 12-seed Wilcoxon floor, with all four new panels larger than the published one. Five panels agreeing is a stronger warrant for the direction than an inflation computed from those same five panels is against it. The honest summary is that the direction is replicated and the depth *magnitude* in any single arm is not resolvable against content variation.

The *sign* is unaffected, since all five panels agree and each resolves on its own; what is conditional is the number. Every magnitude in this paper should be read as conditional on panel 777, and the pooled direction as the portable claim.

Second, the interior contrast does *not* replicate across panels, and this is where the interior column in Table 4 should be read as unreplicated rather than replicated. Over the four new panels it resolves in two of eight cells: on panel 779 on exposure only ($p = 0.02$; $p = 0.13$ on depth) and on panel 781 on depth only ($p = 0.03$; $p = 0.52$ on exposure). It resolves on neither measure on panel 778 ($p = 0.79$ and $p = 0.62$) or on panel 780

($p = 0.62$ and $p = 0.85$), and on panel 780 the depth peak sits at $s = 65$, so one panel in five does not place the depth optimum in the interior at all. So content identity is sufficient to remove the interior optimum while leaving the massed penalty intact, which is independent evidence for the split this paper already draws between its identified and unidentified halves, and further reason not to claim the shape.

Two limits on this arm. Four extra panels at 12 seeds each is enough to show the massed penalty survives a change of content and enough to show the between-panel spread exceeds the within-panel one, but five panels estimate a content variance component on four degrees of freedom, which fixes an order and not a value. A hierarchical model over panels, which is what would properly widen every interval in the paper, needs more panels than we have. And all five panels are run on the *primary* configuration, so σ_{between} is measured there and then added in quadrature to arms that differ in corpus, width, learning rate and measure. Content sensitivity could interact with any of those, and we have not crossed panels with arms to find out. The widened intervals for C4, natural prose and the 1B arm are therefore panel-*transported* rather than panel-marginal, and the 10 of 10 and 6 of 10 tallies should be read as what the primary configuration's content variance does to each arm's interval, not as a measured content-marginal result for that arm.

And the transport has since been checked once, against a larger component, so here is the sensitivity. Appendix G.1 measures $\sigma_{\text{between}} = 0.196$ bits on exposure at the matched-retention cell, $1.8\times$ the 0.1071 transported here. That value belongs to a different configuration at a level $3.5\times$ smaller, so substituting it is a second transport rather than a correction, and neither component is the arm-marginal one. What we can do is print the tally under both, which bounds the choice instead of hiding it. At 0.1071 the exposure penalty excludes zero in 10 of 10 arms; **at** 0.196 **it excludes zero in** 7 **of** 10, with C4 $[-0.092, +0.748]$, natural prose $[-0.151, +0.631]$ and the 1B arm $[-0.019, +0.896]$ losing their sign, and these are the three smallest penalties in the paper. On depth the two components are close, 0.0577 transported against 0.0523 measured, and the tally is 6 of 10 either way. **So the ten-arm exposure tally is the one figure in this paper that depends on which component is carried**, and the seven-arm reading is the conservative one. The nine-arm pooled estimate of Appendix C.4, which straddles zero on HKSJ, does not depend on it at all.

E.2 The extraction null, quantified

Zero greedy exact matches in 2,160 injected attempts, with duplicate counts to 128: 1,440 in the onset arm and 720 in the overfitting arm. The onset arm *scores* 1,728 canaries, 48 runs at 36, but 6 of each 36 are the never-injected control set, which cannot be extracted from training data by construction and so does not belong in an extraction denominator.

Which arms this null belongs to, and why the exponent is six. Both arms use 6-token secrets by construction, because 12 digits carry 39.9 bits and are out of reach in this budget, so an arm whose purpose is to find the regime where extraction *does* happen has to shorten the secret. Both arms use six, so the sixth power is the right one and the calculation stands. Twelve-token arms would need the twelfth.

Two further sets of runs reach $k = 128$ and are *not* in this denominator: the older $k = 128$ rotation (72 runs), which predates the zero-drop assertions and so has no audited injection count, and the high-memorisation attack arm (40 runs), scored for membership rather than for greedy extraction. Both use 6-digit secrets. Omitting attempts can only make a zero-success null *weaker*, so 2,160 is the conservative denominator. Note that there are two $k = 128$ *configurations* and four sets of runs that reach it, which is why we count configurations here.

The expectation has to be summed cell by cell rather than taken from a mean. Applying the best cell's *mean* top-1 of 0.1146 to all attempts understates it, because $p \mapsto p^6$ is convex so a mean of powers exceeds a power of the mean, and the best single run reaches 0.278. Summing $n\,p^6$ cell by cell instead gives 0.0105 expected successes in the onset arm and 0.0270 in the overfitting arm, 0.0375 together, against zero observed. So $P(\text{zero}) = e^{-0.0375} = 0.96$: the null is consistent with the model and weakly informative, which is neither reassurance nor a foregone conclusion. Summing $n\,p^{12}$ over the same cells bounds a 12-token secret's expected successes at 1.7×10^{-5}, using the largest single-run top-1 we observe, which is the sense in which the 12-token arms carry no information here.

What that calculation assumes, since it is not nothing. Teacher forcing is harmless here: if every position is top-1 given the *true* prefix then greedy decoding stays on the true prefix at every step, so $P(\text{greedy exact match})$ equals $P(\text{all six positions top-1 under teacher forcing})$ identically, with no gap. The independence of positions is not harmless. Writing the joint as p^6 from a marginal p assumes the six digit positions succeed independently, and they do not: they share a model, a carrier block and a per-canary difficulty, so successes should be positively correlated within a secret. Positive correlation makes the true $P(\text{all six})$ *larger* than p^6 at fixed marginal, so 0.037 is an **underestimate** of the expected count and 0.96 is an **upper bound** on $P(\text{zero})$.

A larger expectation makes observing zero *less* probable, so the zero-match observation is *more* informative than $P(\text{zero}) = 0.96$ implies, not less. What it is informative about needs stating carefully, because two directions are in play. 0.037 understates the *theoretical* expectation, and the measurement is zero, so *every* marginal-based

figure over-predicts what we actually observed, the independent 0.037 and the larger value correlation implies alike. We therefore do not rest on that calculation: summing cell by cell handles heterogeneity between cells and nothing about correlation within a secret, whose size we cannot quantify from these data. What we quote instead is a descriptive bound: zero successes in 2,160 attempts gives at most about $3/2160 = 1.4 \times 10^{-3}$ per attempt by the rule of three. That bound is neither assumption-free nor run-clustered, since attempts share models, carriers and per-canary difficulty and positional successes are positively correlated within a secret, so the effective sample size is below 2,160 and the true bound is *larger* than 1.4×10^{-3}.

The two arms this appendix reports. Both push the count range past the main grid's $k = 64$ to $k = 128$ at $s \in \{1, 20\}$, the measured safest and worst spacings, with 5 injected sets of 6 canaries plus a control, so 30 of 32 batch rows at $s = 1$ and the zero-drop guarantee still holds; both assert delivered equals requested. They differ in one thing. The *onset* arm (48 runs, 24 seeds) asks whether spacing moves the count at which a secret becomes extractable, at the primary arm's corpus. The *overfitting* arm (24 runs, 12 seeds) trains the same model for the same 6,000 steps on a 2M-token subset, about 25 epochs, on the diagnosis that the extraction null is a property of an undertrained base model rather than of the design. Neither produced a single exact match, so the onset question has no answer here and that is the result. What both do show is the massed penalty, strengthening with count and at $k = 128$ larger than anywhere else in this work:

		per-token top-1		**absolute** exposure, bits	
arm	k	s=1	s=20	s=1	s=20
onset (24 seeds)	8	0.032	0.045	1.72	2.09
	16	0.038	0.064	1.84	2.34
	32	0.038	0.071	2.14	3.19
	64	0.048	0.091	2.43	3.95
	128	0.051	**0.115**	3.58	**4.68**
overfitting (12 seeds)	8	0.009	0.021	3.22	3.26
	16	0.023	0.065	3.19	5.06
	32	0.014	0.076	2.92	5.22
	64	0.028	0.157	4.62	7.03
	128	0.069	**0.144**	6.31	**7.42**

Run-level paired contrasts at $k = 128$: on top-1 the onset arm gives $+0.064$ $[+0.043, +0.086]$, $p = 1.9 \times 10^{-4}$ over 24 seeds, a ratio of 2.250, and the overfitting arm $+0.074$ $[+0.037, +0.109]$, $p = 6.8 \times 10^{-3}$ over 12 seeds, a ratio of 2.067; on exposure $+1.094$ $[+0.839, +1.356]$, $p = 3.6 \times 10^{-7}$ and $+1.101$ $[+0.538, +1.729]$, $p = 4.9 \times 10^{-3}$. Every count cell in both arms is positive on top-1 and every one but the overfitting arm's $k = 8$ resolves, so the count trend is monotone in the onset arm on both measures up to $k = 64$ and the two arms agree on a claim neither was designed to test.

Two qualifications follow. The exposure *contrast* peaks at $k = 64$ and is smaller at $k = 128$ in both arms, $+1.52$ against $+1.09$ bits in the onset arm and $+2.41$ against $+1.10$ in the overfitting arm, so $k = 128$ is the maximum of the top-1 contrast and $k = 64$ of the exposure contrast, which is how Section 4.6 quotes them. The exposure *levels* rise throughout in three of the four series, from 2.43 to 3.58 bits at $s = 1$ and 3.95 to 4.68 at $s = 20$ in the onset arm; the exception is the overfitting arm at $s = 1$, which dips 3.22 to 2.92 bits between $k = 8$ and $k = 32$ before rising to 6.31. Either way what compresses is the contrast and not the signal.

Why it compresses is partly a ceiling effect, and two arguments about that are wrong. The ceiling is not lower for a 6-token secret, since it is $\log_2(2N) = 9$ bits and depends only on $N = 256$. And a mean sampled quantile far above the floor does not show the cells are clear of it: a censored canary contributes exactly 0 to that mean rather than $1/(2N)$, so a cell with half its canaries at $\hat{q} = 0$ and half at 0.054 has mean 0.027, nominally $14\times$ the floor, with half the sample sitting on it. The right statistic is the fraction at the floor.

The censored fraction per cell, recovered exactly from the published runs. Recovering this looks at first as though it needs a re-score, because per-canary $\hat{q}$ is not retained. It does not. Two identities in the two numbers already stored per cell determine the multiset of per-canary quantiles. Since $\hat{q}$ is the fraction of N alternatives below the secret it is a multiple of $1/N$; writing $m_i = N\hat{q}_i \in \{0, \ldots, N\}$ for the n canaries of a cell,

$$\sum_i m_i \;=\; n\,N\,\bar{\hat{q}} \;\text{(an integer)}, \qquad \sum_i e_i \;=\; n\,\bar{e},$$
$$\text{where } e_i = \log_2 2N \text{ if } m_i = 0, \text{ and } \log_2(N/m_i) \text{ otherwise,}$$

because the ceiling is reached exactly when $\hat{q} = 0$ and the stored $\bar{\hat{q}}$ is the mean of the *raw* $\hat{q}$. Enumerating every multiset consistent with both, run by run, recovers the number of censored canaries; the solution is unique in most runs and we report the range over all solutions where it is not, which is partial identification and not an estimate. The enumeration is exact arithmetic on published numbers, so any reader can reproduce it.

arm	cell	n scored	abs. expo.	% of 9 bits	mean $\hat{q}$	censored
primary	s=1, k=64	144	4.10	45.5%	0.155	7.6–9.7%
primary	s=20, k=64	144	6.44	71.5%	0.046	**29.2–36.8%**
primary	$k \leq 16$, worst s	576	≤ 3.0	$\leq 33\%$	≥ 0.246	$\leq 2.8\%$
onset	s=20, k=128	144	4.68	52.0%	0.119	4.9–18.1%
overfitting	s=20, k=128	72	7.42	82.4%	0.027	**48.6–58.3%**
XXL	s=20, k=64	48	8.05	89.5%	0.009	**58.3–66.7%**
XL′	s=20, k=64	48	8.60	**95.6%**	0.002	**68.8–77.1%**

One note on reading this table against Appendix D.2, which looks inconsistent with it and is not. The absolute exposures here differ by $6.438 - 4.095 = 2.343$ bits between $s = 1$ and $s = 20$ at $k = 64$, while Appendix D.2 reports the same cell's paired contrast as $+2.138$. Both are right and the gap of 0.205 bits is exactly the movement of the *control* level, which is 1.252 at $s = 1$ and 1.457 at $s = 20$. The control set is never injected, but its measured exposure is not spacing-invariant: the injected canaries at different spacings train a different model, which moves where the control's own secrets rank. So the control does not cancel in a difference of absolute exposures, and the two numbers are the absolute contrast and the above-control contrast respectively.

So censoring is real and substantial at the top of the count axis, and *the bias runs against our own claims at every cell we checked*. Truncation from above means a censored cell's measured exposure understates its true value, and in the primary arm the censored fraction is itself non-monotone in s, *peaking at* $s = 20$: 7.6–9.7%, 14.6–22.2%, 29.2–36.8%, 10.4–15.3% at $s = 1, 8, 20, 65$, the same shape as the effect. The peak cell is the most truncated, so both the massed penalty and the interior fall are measured smaller than they are, and every exposure magnitude at $k = 64$ here is a lower bound.

One claim in this appendix narrows. We wrote that the contrast compression at the highest count is "real rather than instrumental". In the onset arm that survives: the censored fraction does not clearly rise from $k = 64$ to $k = 128$ at $s = 20$ (5.6–13.2% against 4.9–18.1%, overlapping). In the overfitting arm it does not: censoring rises from 26.4–44.4% to 48.6–58.3% and the $s = 20$ cell is the more censored of the pair at $k = 128$, so part of that arm's contrast compression is instrumental after all. We claimed both arms and can claim one.

β at the ceiling-limited arms. Appendix D.1 notes XL′ reaches 95.6% of the ceiling at $k = 64$; it is quantified as 68.8–77.1% of its canaries censored there, and XXL is 58.3–66.7%. Sub-additivity on exposure at those two arms should be read as ceiling-affected, which is why Appendix D.1 rests it on depth, where there is no ceiling and $\beta < 1$ in 9 of the 10 arms, the 1B arm excepted.

The ladder rungs and the matched-retention cells, where the retention claim actually lives. The table above covers the primary grid and the widest arms. It does not cover the convergence ladder or the matched-retention cells, and exposure is the measure those rest on, so the fraction there decides whether $+0.257$ and $+0.715$ bits are magnitudes or lower bounds of unknown size. Same enumeration, same 9-bit ceiling, four canaries per set, 48 runs per cell unless the count says otherwise. The worst cell is $k = 64$ everywhere, so that is the cell shown.

rung or cell	s	abs. expo.	% of 9 bits	mean $\hat{q}$	censored
ladder, 2 tok/par, retention 1,539	1	4.40	48.9%	0.122	6.8–10.4%
	20	5.83	64.8%	0.056	**15.6–24.0%**
	65	3.88	43.1%	0.173	5.2–8.9%
ladder, 5 tok/par, retention 3,848	1	2.97	33.0%	0.266	5.7%
	65	3.86	42.9%	0.168	4.2–7.8%
	150	3.15	35.0%	0.234	4.2%
ladder, 10 tok/par, retention 7,696	1	2.08	23.1%	0.394	0.5–1.0%
	65	2.58	28.7%	0.302	1.0–1.6%
	150	2.55	28.3%	0.311	0.5%
ladder, 20 tok/par, retention 15,392	1	1.79	19.9%	0.444	**0.0%**
	65	1.84	20.4%	0.432	0.5%
	150	1.90	21.1%	0.420	0.0–0.5%
matched, 10 tok/par, retention 1,539	1	4.57	50.7%	0.113	9.9–13.0%
matched, 20 tok/par, retention 1,539	1	4.09	45.4%	0.176	8.3–11.5%
	150	4.28	47.6%	0.154	8.9–12.0%
matched, 20 tok/par, second panel	1	3.19	35.4%	0.211	2.1–3.1%
	150	3.54	39.3%	0.175	3.1–3.6%

Three things follow and all three run against the objection this answers. **Censoring falls up the ladder and is gone at the top rung**, 15.6–24.0% at the bottom to 0.0–0.5% at the top. That top rung is where $\beta = -0.002$, so the exponent's collapse is not truncation: the ceiling bites hardest at the rung where β is largest and not at all where it is zero, which is the reverse of a ceiling artefact. **At the matched cells it is balanced across the two spacings compared**, 8.3–11.5% at $s = 1$ against 8.9–12.0% at $s = 150$, and a ceiling that truncates both arms

of a paired contrast by the same amount cannot produce the contrast. **And the levels are about a third of the primary arm's**, 8–12% against 29.2–36.8% at its headline cell and 68.8–77.1% in XL′; on the second panel they are lower again at 2.1–3.6%. So the matched-retention cells are among the least censored in this paper and their exposure magnitudes are magnitudes.

One number this costs us. The most censored cell in the set is the *bottom* rung's own peak, $s = 20$ at $k = 64$, at 15.6–24.0%. That rung's $+0.494$ bits is therefore a lower bound too, for the same reason the primary arm's is. It does not change a sign or a comparison, since both the bottom rung's cells are truncated and the attenuation up the ladder is measured against a comparator that is less censored, not more. We would rather print it than have it found.

The control levels, so that any exposure figure in this paper can be converted. Exposure is reported above control throughout, which makes a reported $+1.96$ bits easy to misread as $\hat{q} = 2^{-1.96} = 0.257$. (The two tables in this appendix are the exception to that convention and say so in their headers: both report *absolute* exposure, so nothing should be added to them.) It is not: the primary arm's control mean is 1.375 bits, so the absolute exposure is 3.335 and $\hat{q} = 0.099$. Since that misreading is evidently easy, the control means are here, one number per arm, in bits: primary 1.375, S 1.549, L 1.347, XL 1.268, XL′ 1.374, XXL 1.621, C4 1.424, natural prose 1.101, onset 0.845, overfitting 0.991. Add the relevant one to an exposure figure for an absolute value, then negate and exponentiate for $\hat{q}$. Note that these control means, being means of *exposure* in bits rather than of $\hat{q}$ itself, are not usable as evidence about censoring, for the reason given above; the censored fractions are the table above.

One thing the quantiles do still show, independently of censoring. The ratio $\hat{q}(s{=}1)/\hat{q}(s{=}20)$ falls from 2.00 at $k = 64$ to 1.70 at $k = 128$ in the onset arm and from 5.27 to 2.03 in the overfitting arm, so the clustered condition closes part of the gap at the highest count while both levels keep rising. We report that and do not explain it: two spacings and one count step is not enough to say whether it is saturation in the underlying quantity, the two-row natural-text batch that 30-of-32 injection forces at $s = 1$ (Appendix B.2, Result 3), or noise at 12 to 24 runs. It is a further reason Section 4.6 quotes the top-1 contrast at $k = 128$ and the exposure contrast at $k = 64$ rather than treating either as monotone to the end of the range. And these arms have two spacings, so they speak to the massed penalty and say nothing about the interior optimum.

What is informative is the distance to extraction, at either secret length. A 12-token uniform digit secret carries $12 \log_2 10 = 39.9$ bits, and even odds of reproducing it needs per-token top-1 of $0.5^{1/12} = 0.944$; a 6-token secret needs $0.5^{1/6} = 0.891$. The measured 0.1146 at $k = 128$ comes from a 6-token arm, so we give both, taking per-token top-1 to be a property of the model and the carrier rather than of the secret's length. Extrapolating top-1's growth, roughly $k^{0.33}$, gives $128 \times (0.944/0.1146)^{1/0.33} \approx 7.6 \times 10^4$ duplicates at 12 tokens and $128 \times (0.891/0.1146)^{1/0.33} \approx 6.4 \times 10^4$ at 6, so of order 10^5 at both: an extrapolation over nearly three decades, which we label as such rather than lean on. The cause is the base model. Validation loss 5.13 nats is perplexity about 169, and such a model does not reproduce 40 random bits verbatim at any duplicate count we can afford. A design aiming to *demonstrate* extraction rather than to measure a graded signal would use a much shorter secret; we chose the opposite trade, because exposure and depth are graded and a short secret would saturate them.

These runs did extend the count range to $k = 128$, where the massed penalty is stronger: $s = 20$ gives 2.25× the top-1 and $+1.09$ bits of exposure relative to $s = 1$. That is the strongest single piece of evidence for the massed penalty in this work, which is why Section 4.6 reports it rather than leaving it inside a null-result appendix.

F Robustness arms: rate schedule, further seeds, one billion parameters, realistic pipeline

F.1 Does the effect survive a decaying learning rate?

The design uses a constant rate after warmup to remove confound (4), which is a deliberate choice and also a restriction: production runs anneal, and a reader is entitled to ask whether the effect is an artefact of the flat schedule. So we re-ran the primary grid under two annealing schedules, changing only $\mathrm{lr}(t)$. Cosine decays from the warmup peak to zero over all 6,000 steps, reaching 0.156× the peak by T_{last}. Warmup-stable-decay holds the peak through 80% of training and then decays linearly to 10%, so the full rate is in force across the entire injection window and the annealing happens *after* the last copy. Reading the two against the constant arm separates the two things annealing does: (const, WSD) isolates annealing after the last copy, and (WSD, cosine) isolates depositing at a decayed and varying rate.

measure	schedule	seeds	val. loss	massed penalty	interior, p
depth	constant	36	5.130	$+0.093$ [0.045, 0.139]	$+0.057$, 0.14
depth	WSD	12	5.135	$+0.183$ [0.146, 0.223]	$+0.120$, 0.027
depth	cosine	12	5.759	$+0.448$ [0.357, 0.547]	$+0.076$, 0.30
exposure	constant	36	5.130	$+0.733$ [0.572, 0.903]	$+0.275$, 0.012
exposure	WSD	12	5.135	$+1.233$ [0.966, 1.519]	$+0.637$, 0.0093
exposure	cosine	12	5.759	$+1.617$ [1.058, 2.147]	peak at the widest point

The massed penalty survives both, and grows. It is resolved under every schedule on both measures, at $1.68\times$ the constant arm's size under WSD on exposure and $1.98\times$ on depth, and $2.21\times$ and $4.83\times$ under cosine. The "seeds" column counts matched seeds, the unit the paired tests use, and the constant row is the same 36-seed primary grid every other headline number in the paper is computed on, so the numbers in its row are the ones in Section 4.2. So the effect is not an artefact of the flat schedule; if anything the flat schedule is the conservative choice, which matters because it was chosen for identification rather than for effect size.

And this arm is the one that scopes Section 4.8's retention results by schedule. Those are measured at a constant rate, and retention is the mechanism most sensitive to annealing, since a decayed rate updates the weights less and therefore forgets less. WSD isolates exactly that: the peak is held through 80% of training, which is $4{,}800$ steps against $T_{\text{last}} = 4{,}500$, so deposition happens at the full rate in both arms and **all of the annealing falls inside the retention interval**. The penalty is $1.68\times$ larger on exposure and $1.98\times$ on depth under WSD, so **decaying across the retention window preserves leakage rather than erasing it**. The constant rate is therefore the schedule that forgets fastest, and the retention boundary Section 4.8 reports would be a lower bound on where the effect stops under a production anneal. That last step is a conjecture and not a measurement: this arm runs at retention $1{,}500$ and the boundary is at $15{,}392$, so it extrapolates an order of magnitude. What we do not have is the β collapse itself remeasured under WSD, which would need the ladder rerun on a second schedule.

The interior optimum does not survive cosine, and that is expected. Under WSD it holds on both measures. Under cosine the depth contrast against the widest spacing fails ($p = 0.30$) and on exposure the profile is monotone increasing to $s = 65$, so there is no interior optimum to test. This is what Section 1 predicts rather than a surprise: under cosine, wider spacing puts earlier copies at a *higher* learning rate, so confound (4) is reintroduced pointing in the opposite direction to confound (3), and it inflates the wide end. Appendix B.1 measures that directly, where cosine turns a single-exposure forgetting curve from -3.719 to $+0.132$ nats at distance 200. The cosine arm therefore does not test the shape; it tests whether the massed penalty survives having a confound put back, and it does.

What this arm cannot claim. The cosine arm is not convergence-matched: validation loss is 5.759 against 5.135, because a schedule that spends its last steps near zero learning rate reaches a worse loss at equal steps. Some of its larger penalty is therefore a property of a less converged model rather than of the schedule, and we do not decompose the two. And 12 seeds per schedule against the constant arm's 36 means the two annealed arms sit at the Wilcoxon floor of 4.9×10^{-4}, so their p values say "as small as 12 seeds can show" and not more.

The learning rate appears with three signs, and only one comparison varies it at fixed length. Three results in this paper move the rate and disagree about what happens to the penalty, so here is which grid and which length each belongs to. *One.* Section 4.4 halves the rate at width 768, 6×10^{-4} to 3×10^{-4}, at $T = 6{,}000$ on the grid $\{1, 8, 20, 65\}$: the depth gap moves from 0.101 to 0.351, so the penalty *grows* as the rate falls. *Two.* This appendix lowers the rate late rather than throughout, under cosine and WSD, at the same length and the same grid: the penalty *grows* again. *Three.* Appendix G.3 divides the rate by about 5.5 *and* quadruples the length, on the grid $\{5, 20, 65, 150\}$: the $s = 20$ contrast against $s = 5$ goes to zero on exposure. The first two vary the rate at fixed length on a grid whose massed cell is $s = 1$ and agree with each other. The third varies two things at once and its lowest cell is $s = 5$, so it measures the interior contrast rather than the clustered one, and it cannot separate a rate effect from a length effect. Nothing here is evidence that lowering the rate removes the penalty; what Appendix G.3 bears on is stated there.

F.2 The primary configuration at 72 further seeds

The primary grid is 36 seeds and every headline number in this paper is computed on it. We later ran the identical configuration at seeds 36 to 107, a further 288 runs, and report them here rather than folding them into the headline. The reason is that holding the pre-registered sample fixed is what makes the extension an *out-of-sample* test rather than a restatement: it uses seeds the original grid never touched while sharing its canaries, count-rotation rule and schedule exactly, so it is the held-out confirmation that the data-selected peak of Section 3 otherwise lacks.

This paper treats added seeds two different ways, and the distinction is worth stating. Here we keep the 36-seed figure as the headline and hold the extension out. In Section 4.3 we do the opposite: the realistic arm

was enlarged from 12 seeds to 48 after its contrast had failed to resolve at 12, and the 48-seed value is what Table 2 carries. The difference is what the peak selection depends on. The primary grid's contrast is measured *at a spacing chosen with those runs in hand*, so adding seeds to it would fold the selection into the estimate, and an out-of-sample draw is the only thing that removes it. The realistic arm has no selection to absorb: its three schedules are fixed by construction, the comparison is between named conditions, and no spacing was picked from the data, so extra seeds buy precision on a pre-specified contrast and nothing else. We would not extend the primary grid the same way, and Table 2 marks the two rows differently for that reason.

measure	sample	seeds	vs. massed	vs. widest
depth	pre-registered	36	$+0.093$ $[0.045, 0.139]$	$+0.057$ $[-0.005, 0.118]$, $p = 0.14$
depth	extension	72	$+0.118$ $[0.096, 0.140]$	$+0.096$ $[0.060, 0.135]$, $p = 1.2 \times 10^{-5}$
depth	pooled	108	$+0.109$ $[0.088, 0.131]$	$+0.083$ $[0.050, 0.116]$, $p = 7.8 \times 10^{-6}$
exposure	pre-registered	36	$+0.733$ $[0.572, 0.903]$	$+0.275$ $[0.077, 0.482]$, $p = 0.012$
exposure	extension	72	$+0.837$ $[0.749, 0.925]$	$+0.448$ $[0.324, 0.577]$, $p = 5.5 \times 10^{-9}$
exposure	pooled	108	$+0.802$ $[0.724, 0.881]$	$+0.390$ $[0.286, 0.499]$, $p = 9.1 \times 10^{-10}$

The row that matters is depth against the widest spacing. It is the one contrast the pre-registered grid does *not* resolve, reported as such throughout this paper, and on 72 seeds it never saw the same design resolves it at $p = 1.2 \times 10^{-5}$, with the pooled estimate at $p = 7.8 \times 10^{-6}$. Measured peak-to-massed ratios agree across the three samples, 1.213, 1.283 and 1.259 on depth and 1.598, 1.662 and 1.641 on exposure, so the extension is a replication and not a different result.

What this does and does not change. It shows the interior optimum on depth in the primary arm is real as a feature of the measured curve, and that the pre-registered grid was underpowered for it rather than contradicting it. It does *not* touch identification. Section 4.7 shows the descending branch is absorbed by decay-weighted copy count, and resolving the contrast more sharply says the curve turns over more clearly, not that the turn is a change in per-exposure gain. Everything this paper declines to claim about the interior optimum, it still declines to claim.

Why the headline is still the 36-seed grid, stated as a choice rather than a default. Adopting the pooled estimate as primary means re-deriving β, $C(s)$, the EDR, the separability tallies and the per-count table on all 108 seeds, because every one of those is fitted on the same cells. Re-deriving all of them at once is how a derived number comes to disagree with the text that quotes it. Against that, the pre-registered sample is the smaller one and reporting a $p = 0.14$ where we hold data giving 1.2×10^{-5} understates our own result. We judge the indexing risk the larger of the two and we record the trade rather than presenting the choice as obvious. The re-derivation is the largest piece of work left, and the direction it would move every number is known: the pooled massed penalty is larger on both measures, so the quoted figures are the conservative ones.

F.3 One billion parameters

Scale was the one objection to this work that no arm answered. The ladder stopped at 302.3M non-embedding parameters, the effect's magnitude is non-monotone inside the sampled range so there was no trend to extrapolate, and the withdrawal of δ (Appendix G.3) left open, until the batch axis was extended (Appendix G.2), whether the whole structure sits inside a single production batch. This arm measures the one contrast that matters at three times the largest previous width.

Configuration. 20 layers, 32 heads, width 2048, head dimension 64 as everywhere else, giving 1,007.2M non-embedding and 1,110.6M total parameters with tied embeddings. Learning rate 3×10^{-4}, which is the recipe rate at this width and makes this the *third* point of the 3×10^{-4} sequence, which has three points with it and two without. Everything else is inherited unchanged: Eq. 1, the count rotation, constant learning rate after warmup, the fixed panel seed, and the zero-drop assertion, which holds at 24,192 of 24,192 injections over 48 runs.

At this width the 32-row batch does not fit in 40GB alongside fp32 AdamW state, so each step is split into four microbatches of eight whose gradients accumulate before one optimiser step. **Each microbatch's mean cross-entropy is divided by the number of microbatches before `backward`**, i.e. the code accumulates $\nabla(L_i/4)$ and not ∇L_i. That division is what makes the accumulated gradient equal the full-batch one, and equal microbatch sizes are a second, separate requirement: with unequal sizes a plain mean of per-microbatch means would not be the full-batch mean in the first place. Equal sizes alone therefore do not establish the equivalence, and the distinction is not pedantic here: clipping is at 1.0 and Appendix B.2 reports pre-clip norms of 0.79 to 1.00 at injection steps, so an undivided accumulation would sit at roughly four times those norms and clip at essentially every step, making this arm a different optimisation configuration rather than the same recipe at a new width. Clipping sees the accumulated gradient, after the division. The injection schedule is untouched because

the split happens inside a step, after the batch is assembled. Readout runs under `no_grad` with the optimiser state released, which is what makes the 256-candidate exposure scoring fit.

Result.

measure	s=1	s=8	s=20	s=65	massed penalty at s=20
exposure (bits)	$+0.331$	$+0.758$	$+\mathbf{0.769}$	$+0.728$	$+0.439$ $[+0.200, +0.687]$, $p = 9.3 \times 10^{-3}$
depth (nats)	$+0.110$	$+0.275$	$+\mathbf{0.280}$	$+0.247$	$+0.170$ $[+0.050, +0.283]$, $p = 2.7 \times 10^{-2}$

Both resolve, and both survive the panel-covering widening of Appendix E.1, at $[+0.107, +0.770]$ and $[+0.003, +0.337]$; the depth cell survives narrowly and should be read as such. Both are computed under the same convention as the other nine rows of that appendix's table, $1.96\sqrt{\mathrm{SE}^2 + \sigma^2_{\text{between}}}$ on the arm's own standard error. The peak is $s = 20$ on both measures, as in seven of the other nine arms; S and XL are the two whose depth peak is $s = 8$. The interior contrast does *not* resolve, at $+0.042$ $[-0.264, +0.311]$ on exposure and $+0.033$ $[-0.100, +0.163]$ on depth, so this arm supports the massed penalty and not the shape, which is the paper's position everywhere.

Four things it does not establish, stated because it is the arm readers will lean on. First, the effect does not grow with capacity. The peak-to-massed *ratios* here, $2.33\times$ and $2.54\times$, are the largest in the paper, but that is a statement about a small denominator: absolute leakage at $s = 1$ is $+0.331$ bits against XXL's $+1.409$, and the absolute penalty is $+0.439$ against XXL's $+1.160$. Along the whole 3×10^{-4} sequence the absolute penalty *falls*, $+1.484 \to +1.160 \to +0.439$ bits at 85.1M, 302.3M and 1,007.2M, while the ratio rises, $1.75 \to 1.82 \to 2.33$. Which of those matters depends on whether the risk of interest is absolute or relative, and we report both rather than choosing.

Second, the arm is not at matched convergence, and it is the least converged in the paper: mean validation loss 5.961, perplexity 388, against XXL's 5.109 and perplexity 165. One billion parameters trained on 49M tokens is far short of the Chinchilla ratio, at 0.05 tokens per parameter, which is the gap Section 4.8 closes at fixed width; and the validation losses along the 3×10^{-4} sequence are not even monotone (5.233, 5.109, 5.961). So width, convergence and rate remain entangled exactly as Section 4.4 says, and this arm extends the ladder in width without disentangling them.

Third, the count structure is unchanged rather than improved. The per-count massed penalty on exposure is -0.190 at $k = 2$ and $+0.169$ at $k = 4$, then $+0.582$, $+0.414$, $+0.636$ and $+1.020$ at $k = 8$ to 64, with $k = 8$, 32 and 64 resolving and $k = 2$ negative. That is the same high-count pattern Section 4.6 reports at every other width, so scale does not extend the effect downward into the count regime real corpora occupy.

Fourth, one arm at one width is not a scaling law, and its p-values, 9.3×10^{-3} and 2.7×10^{-2} at 12 seeds, are the weakest of any arm that resolves. What the arm removes is the possibility that the effect is confined to models below 302.3M. What it leaves open is how the effect behaves as scale continues, and reachability at production batch size is addressed on the batch axis rather than by width, for the reason the training-length row of Appendix H.3 gives.

F.4 The realistic-pipeline and wide-grid arms

The calibrated attack, on the schedules a pipeline produces. There is an obvious objection to Appendix I: it measures the attack between uniform $s = 1$ and uniform $s = 20$, and no pipeline emits either, so operational evidence on two artificial schedules needs a counterpart on schedules a pipeline emits. The same one-sided offline LiRA of Appendix I is run on all three realistic schedules, on real prose with the heavy-tailed counts, at 48 seeds per arm against 32 dedicated never-injected reference models, with the same 512 non-member passages per run as Appendix I, which is what puts the 1% false-positive point at the 5.12th order statistic rather than between two non-members. Target seeds run 0 to 47 and reference seeds 100 to 131, so unlike Appendix I's ensemble the two ranges are disjoint and no target shares a data order with a reference model.

schedule	runs	TPR at 1% FPR	with the reference ensemble resampled
conditional random placement	48	0.551 [0.518, 0.585]	[0.510, 0.601]
clustered, gap 1	48	0.646 [0.615, 0.677]	[0.607, 0.696]
uniform $s = 20$ (worst)	48	**0.724** [0.691, 0.757]	[0.686, 0.777]

All three contrasts resolve, and the pattern matters more than any single number. Every contrast below is a paired difference over the 48 shared seeds, so it is exactly the difference of the two levels above. *The worst uniform spacing beats the realistic baseline by* $+0.173$ $[+0.143, +0.203]$, surviving resampling the reference ensemble ($[+0.156, +0.236]$), so the operational claim is made against a schedule a pipeline could actually

produce rather than against a second artificial one. *Clustering is worse than random placement by* $+0.095$ $[+0.068, +0.122]$, also surviving reference resampling ($[+0.076, +0.146]$). That result is the one we would least like and it is the more useful: it confirms Section 4.3's refusal to recommend clustering on an *operational* metric rather than on likelihood proxies alone, so the "cluster what you cannot remove" reading is refuted on the attack as well as on leakage and loss. And the contrast that is this arm's analogue of the massed penalty, worst minus clustered, is $+0.078$ $[+0.060, +0.098]$, $p = 1.4 \times 10^{-7}$, and **resolves** both before and after reference resampling ($[+0.056, +0.108]$). So the attack orders the three schedules random $<$ clustered $<$ worst and resolves all three contrasts between them.

Which counts carry it. The heavy-tailed count set includes 2, 3 and 5, so this arm can answer a question the grid arms cannot: whether the effect at counts real corpora actually exhibit is real, or whether a pooled figure is carried by its tail. Restricting the members to $k \leq 5$ and leaving the non-member pool intact, which is required because the 1% threshold is defined on non-members, gives six members per run instead of sixteen.

members counted	random	clustered	worst
all counts, $k \in \{2, \ldots, 55\}$	0.551	0.646	0.724
$k \leq 5$ only	0.212	0.212	0.330
$k \geq 8$ only	0.754	0.906	0.960

The massed penalty is larger at $k \leq 5$ than pooled, not absent: worst minus clustered is $+0.118$ $[+0.080, +0.160]$ over the same 48 seeds, $p = 4.4 \times 10^{-5}$, surviving reference resampling at $[+0.063, +0.174]$, against $+0.054$ $[+0.038, +0.071]$ at $k \geq 8$ and $+0.078$ pooled. Worst minus random is $+0.118$ $[+0.076, +0.163]$ at $k \leq 5$. Two things follow, and they point in opposite directions. The scope limitation that the effect "resolves only from $k \approx 8$" is a statement about the canary grid and its per-cell contrasts, and it does *not* carry over to the attack on real prose, where the low-count cells give the larger contrast. But clustering's loss to random placement does not survive the restriction: at $k \leq 5$ it is 0.000 $[-0.042, +0.045]$, so *on the attack* the refutation of our own advice is carried entirely by $k \geq 8$, where it is $+0.152$ $[+0.117, +0.190]$. Absolute inferability at $k \leq 5$ is also low in every arm, 0.21 to 0.33 against 0.75 to 0.96 at $k \geq 8$, which is the same regime dependence the grid reports. That is not on its own an explanation of the zero, since the same restriction resolves worst minus clustered at $+0.118$; the paragraph below splits leakage by count, where the ordering does survive at $k \leq 5$, at $+0.010$ $[+0.001, +0.019]$ on depth, and locates the zero in the attack's resolution at six members per run rather than in the leakage ordering.

Why the 12-seed interval did not resolve. At 12 seeds and 16 reference models the worst-minus-clustered contrast is $+0.078$ $[-0.016, +0.172]$, which does not resolve. The point estimate is unchanged at 48 seeds; what changed is the interval, for two reasons worth separating. The arm runs 48 seeds and 32 reference models. And the interval is paired. The three schedules are generated from the same seed with the same canary content, the same duplicate counts and the same pre-injection data order, differing only in where the earlier copies land, which we checked holds for all 48 seeds; on the attack they correlate at $r = +0.82$ across seeds, so an unpaired interval on this arm discards most of its precision (Appendix L quantifies what the pairing is worth across every arm). Pairing is what Section 3 specifies everywhere else in this paper and this arm was the exception. At 48 seeds the contrast resolves either way, unpaired at $[+0.031, +0.125]$ and paired at $[+0.060, +0.098]$, so the conclusion does not rest on the choice.

The estimate is stable across the whole seeds-by-ensemble grid, and it is worth printing the grid rather than a diagonal through it, because a diagonal lets two changes offset.

	levels: random / clustered / worst			contrasts		
seeds, reference models	rand.	clus.	worst	w−c	c−r	w−r
12, 16	0.484	0.583	0.662	+0.078	+0.099	+0.177
12, 32	0.521	0.594	0.688	+0.094	+0.073	+0.167
20, 16	0.500	0.616	0.675	+0.059	+0.116	+0.175
20, 32	0.544	0.631	0.709	+0.078	+0.088	+0.166
48, 16	0.499	0.630	0.693	+0.063	+0.132	+0.194
48, 32	**0.551**	**0.646**	**0.724**	**+0.078**	**+0.095**	**+0.173**

Every contrast is the difference of its own two levels, exactly. Across all six cells the massed-penalty analogue ranges $+0.059$ to $+0.094$, so it sits within ± 0.019 of the reported $+0.078$ with no monotone trend in seeds; at sixteen members per run the whole of that range is about a third of one member, which is the resolution of the measure and not a movement in it. Reading only the diagonal of this table, where the ensemble grows with the seed count, would suggest the estimate is identical at 12, 20 and 48 seeds; the honest statement is the ± 0.019 range. What does move systematically is the levels: holding seeds at 48 and going from 16 reference models to 32 raises them by $+0.052$, $+0.016$ and $+0.031$, because a larger ensemble estimates each passage's

out-distribution more sharply. The levels are therefore ensemble-dependent and the contrasts are not, which is the opposite of how they are usually quoted, and it is why we report contrasts.

Absolute rates here are far above Appendix I's 0.101 to 0.193 and *are not comparable with them*: these are 64-token prose passages against 128 alternatives with counts to 55, a much more memorised target than a six-digit canary at $k \leq 64$. The comparison that is licensed is between the three rows above, which share everything but placement.

Against the baseline a practitioner actually runs. Nobody trains with uniform spacing. Three arms use real held-out WikiText-103 passages sharing documents, heavy-tailed counts $\{2, 3, 5, 8, 13, 21, 34, 55\}$ assigned by the same rotation rule as the grid arms, the same pinned last exposure, and 24 runs each, differing only in where the earlier copies sit (Appendix A.3 gives the placement rule).

schedule	mean gap	depth (nats)	exposure (bits)	val. loss
cond. random placem.	613 steps	0.627	1.234	5.1255
clustered at gap 1	1 step	0.721	1.421	5.1276
uniform $s = 20$	20 steps	0.847	1.561	5.1276

Random placement beats uniform $s = 20$ by 0.220 nats $[0.208, 0.232]$ and 0.327 bits $[0.254, 0.401]$. But clustering is *worse* than random placement, by 0.094 nats $[0.082, 0.106]$ and 0.187 bits $[0.107, 0.268]$ at the run level ($p = 1.2 \times 10^{-7\dagger}$, the n=24 Wilcoxon floor, and $p = 2.1 \times 10^{-4}$), and marginally worse on validation loss (paired over the 24 seeds, $+0.0022$ nats; the uniform arm is $+0.0021$, so random placement is best on loss as well as on leakage). In EDR units, using this arm's own β of 0.393 on depth and 0.370 on exposure, random placement is worth 54.8% [44.8, 67.2] and 50.6% [39.3, 63.3] against the worst spacing, and 27.0% [19.3, 37.7] and 32.1% [17.2, 48.8] against clustering, so the refutation is not merely directional.

This leakage arm runs 24 seeds, half the attack arm's 48, so the two contrasts it establishes rest on less precision than the transfer result does. They can be measured again at 48 seeds without new runs, because the attack arm records a per-passage NLL for every run: taking each run's non-member mean as the baseline gives leakage of $+0.036$, $+0.123$ and $+0.252$ nats for random placement, clustering and uniform $s = 20$, so clustering costs $+0.087$ $[+0.078, +0.096]$ nats, $p = 1.4 \times 10^{-14}$, and the worst spacing costs a further $+0.129$ $[+0.123, +0.136]$, $p = 7.1 \times 10^{-15\dagger}$, the $n = 48$ two-sided Wilcoxon floor. The levels differ from the table above because the baseline does, non-members here against a never-injected control set there, but the contrasts agree to 0.007 and 0.003 nats across two arms with different baselines and different seed counts. So the refutation of our own advice now rests on the same 48 seeds as the transfer result.

The low-count null is a property of the attack, not of the leakage ordering. The attack gives clustered minus random as exactly 0.000 at $k \leq 5$. Low inferability does not account for that, because the same restriction resolves worst minus clustered at $+0.118$, so the measure demonstrably separates effects of that size in that subgroup. The per-passage records settle it without new runs, because leakage can be split by count where the attack's 1% threshold cannot be estimated from six members. On *leakage* the ordering at $k \leq 5$ is the expected one on depth, $+0.0101$ $[+0.0011, +0.0189]$, $p = 0.041$, and does not resolve on exposure, $+0.0023$ $[-0.0315, +0.0325]$. So clustering is not neutral against random placement at low count; the attack's zero is its resolution floor at six members per run, and the honest statement is that the sign survives the restriction on one measure and the magnitude does not.

The mean gap is not a single number, and the reconciliation has to be per count. "Clustering leaks least inside the grid" and "clustering loses to random placement" reconcile per count, not through one aggregate mean gap. The aggregate averages over counts whose gaps span a factor of 27, and the account it licenses predicts the *largest* advantage for random placement at low count, where its gaps are widest. The opposite is measured.

k	mean gap, random placem.	clustered − random	worst − clustered
2	1882 steps	-0.004 $[-0.017, +0.010]$	$+0.042$ $[+0.032, +0.052]$
3	1149	$+0.002$ $[-0.011, +0.015]$	$+0.044$ $[+0.031, +0.058]$
5	759	$+0.032$ $[+0.018, +0.046]$	$+0.061$ $[+0.050, +0.073]$
8	476	$+0.097$ $[+0.080, +0.113]$	$+0.078$ $[+0.067, +0.089]$
13	291	$+0.137$ $[+0.122, +0.153]$	$+0.119$ $[+0.104, +0.136]$
21	182	$+\mathbf{0.202}$ $[+0.180, +0.223]$	$+0.148$ $[+0.132, +0.165]$
34	112	$+0.171$ $[+0.146, +0.195]$	$+0.219$ $[+0.202, +0.238]$
55	70	$+0.061$ $[+0.040, +0.082]$	$+\mathbf{0.320}$ $[+0.301, +0.340]$

Depth, nats, 48 seeds; the gap column depends only on the seed and the count, so it is reconstructed exactly rather than measured. Random placement's mean gap falls monotonically with count, from 1882 steps at $k = 2$

to 70 at $k = 55$, while its advantage over clustering is *non-monotone*: near zero at $k = 2$ and 3, largest at $k = 21$, smaller again at $k = 55$.

The crossing-point account is refuted by this table, so we withdraw it and do not replace it with another rule. The account said random placement wins where its gap lands beyond the crossing at $s = 221$. Tested against the column, that is anti-correlated with the data: the three widest gaps (1882, 1149, 759, all beyond 221) give the three *smallest* advantages, averaging $+0.010$, while the three gaps inside the excluded region (182, 112, 70) average $+0.145$, and the maximum sits at a gap of 182 steps, which the rule excludes. Normalising by each count's own dynamic range does not rescue it: the fraction of that range which random placement captures is -0.10, $+0.05$, $+0.52$, $+1.24$, $+1.15$, $+1.36$, $+0.78$, $+0.19$ across the eight counts, still peaking mid-range rather than at the widest gaps.

Two things are worth saying in place of a rule. First, the advantage is bounded by how much there is to redistribute at that count, and that bound rises monotonically with k: worst minus clustered runs $+0.042$ at $k = 2$ to $+0.320$ at $k = 55$. Second, the crossing was measured on the wide grid at *uniform* spacing with many copies, whereas random placement at $k = 2$ is a single earlier copy drawn uniformly from a 3,800-step window, so the two are not the same schedule class and a crossing point measured on one does not transport to the other. Beyond those two constraints the per-count shape is measured and not explained, which is the same position Section 4.7 takes on the descending branch. What the aggregate figure hides is simply that there is no single gap to compare with anything: the per-count gaps span a factor of 27.

The seed enlargement, analysed as held-out. This arm went from 12 to 48 seeds after the 12-seed contrast failed to resolve, which makes a fixed-n interval on all 48 vulnerable to the objection that continuation was result-dependent. Seeds 12 to 47 were never examined before that decision, so they are a clean confirmation at a sample size fixed in advance of seeing them. On those 36 seeds alone, worst minus clustered is $+0.073$ $[+0.052, +0.094]$, $p = 5.5 \times 10^{-6}$; worst minus random $+0.175$ $[+0.139, +0.212]$, $p = 8.4 \times 10^{-7}$; clustered minus random $+0.102$ $[+0.068, +0.137]$, $p = 2.1 \times 10^{-5}$. All three resolve on the held-out seeds alone, at the same signs and within the pooled intervals, so no sequential correction is load-bearing here. For the record the original 12 seeds give $+0.094$, $+0.167$ and $+0.073$ on the same three contrasts, all resolving under pairing; what failed to resolve at 12 seeds was the *unpaired* interval, which is the point the paragraph above makes.

How these rates relate to the primary arm's. They are *this* arm's rates, on its own β, its own count distribution and real prose, so they are a separate measurement and not a re-referencing of the primary arm's 54.4%. The agreement across arms is measure-dependent: on exposure the two land close, 50.6 against 54.4; on depth they do not, 54.8 against 42.2. And "referenced to a shuffle" is ambiguous between two contrasts differing by about a factor of two, 54.8% and 50.6% against the *worst spacing* versus 27.0% and 32.1% against *clustering*, so the phrase has to name which. What survives is the claim the comparison was for: the magnitude of the schedule axis does not depend on referencing it to a schedule nobody runs. The first pair does not hold effective exposure fixed, since the schedules differ in mean gap by a factor of 30, so it is an end-to-end figure for a pipeline choice and not a measurement of the isolated gain effect; the controlled version is the wide grid below. Its mean gap averaged over counts is ten times our widest grid point, but that aggregate is not the quantity to compare with a grid point, and the paragraph above gives them.

The controlled version: extending the grid rather than releasing the control. Because the realistic arm deliberately releases confound (3), we also extended the grid itself into the unsampled region with count still crossed: spacings $\{1, 20, 65, 150, 300, 600\}$, counts $\{2, 4, 8, 16\}$, and a budget giving $s = 600$ room ($T_{\text{last}} = 18{,}000$, 20,000 steps). Across 12 runs per spacing leakage rises to 1.18 to 1.22 times the clustered value near $s = 65$ to 150, then falls to 0.86 on exposure and 0.79 on depth by $s = 600$, crossing back below the clustered value at $s = 221$ $[172, 322]$ on depth and $s = 370$ $[187, 539]$ on exposure. Run-level Friedman is significant on both measures (4.0×10^{-4}, 4.2×10^{-2}), the significant individual contrast is $s = 600$ below massed on depth ($p = 2.1 \times 10^{-2}$), and validation loss is flat across the whole range (0.0013 nats). So spreading rather than clustering rests on a measurement and not only on the uncontrolled arm, with Section 4.7's caveat that this locates where measured leakage returns below the clustered value, not where a gain function does. This arm uses a different grid, count set, canary density and training length from every other, so its s^* is not comparable with theirs (Appendix C.2).

G THE CONVERGENCE LADDER, BATCH SIZE AND TRAINING LENGTH

G.1 THE CONVERGENCE LADDER, PER CELL

Section 4.8's arm at the level of detail the other arms get. Three readings other than retention are excluded here: not the corpus, since each rung has one pre-registered stream; not the spacing grid, since realigning by share leaves none of the top rung's five cells resolving; and not remaining headroom, which is bounded rather than closed, exposure's level falling 2.1× while its best contrast falls from +28.6% to +0.05% of it. The single fitted exponent averages a convex profile: per-doubling slopes on the clustered cell run $+0.10$, $+0.28$, $+0.27$, $+0.49$

and $+0.81$ over $k = 2$ to 64, so one exponent understates the high-count regime. Width is fixed at 25,219,072 non-embedding parameters, the corpus at C4-720M (Appendix A.3), and the token budget moves. So does the retention interval, because T_{last} is pinned at a fraction of T, and that turns out to be what the arm measures.

Steps, and how the pinned copy rounds. A step is 32 sequences of 256 tokens, so 8,192 tokens. $T = \text{round}(\text{tok/par} \times 25{,}219{,}072/8{,}192)$ and $T_{\text{last}} = \text{round}(0.75T)$, to the nearest step. $0.75T$ is not an integer at three of the four rungs, so the pinned last copy sits within half a step of $0.75T$ rather than exactly on it.

What the ladder does not establish is a tokens-per-parameter threshold. The 3.2M arm of Section 4.4 trains at 15.4 tokens per parameter, close to this ladder's top rung, and its massed penalty resolves on both measures. So whatever sets the boundary is not the ratio alone, and we do not offer a value for it.

tok/par	tokens seen	epochs	T	$0.75T$	T_{last}	val. ppl
2	50,438,144	0.0701	6,157	4,617.75	4,618	313
5	126,091,264	0.1751	15,392	11,544	11,544	162
10	252,190,720	0.3503	30,785	23,088.75	23,089	119
20	504,381,440	0.7005	61,570	46,177.5	46,178	100

Cells, seeds, and the run count. The grid is not the same at every rung and the cells are not all the same size, which is why the run totals do not divide by a single seed count. $s = 20$ was a 12-seed pilot at two rungs, and the top rung's two widest cells are 12-seed. Everything else is 48 seeds.

tok/par	s=1	s=20	s=65	s=150	s=300	s=600	runs
2	48	48	48	–	–	–	144
5	48	12	48	48	–	–	156
10	48	–	48	48	–	–	144
20	48	12	48	48	12	12	180
						total	624

Every one of the 624 delivered 504 of 504 injections and none failed. **The comparator** at each rung is the interior cell maximising *depth* among that rung's 48-seed interior cells, read on both measures: $s = 20$, $s = 65$, $s = 150$, $s = 150$. The 12-seed cells are reported below but are not eligible, because using one would set $n = 12$ for a headline.

Every cell, against that rung's own $s = 1$. Spans are the width of the injection window at the largest count, $63s$ steps at $k = 64$, and share is that span as a fraction of T_{last}. Intervals are run-level 95% over seeds paired within cell; win is seeds with a positive paired difference.

				depth (nats)		exposure (bits)	
s	span	share	n	contrast	win	contrast	win
2 **tok/par**, ppl 313; level at s=1 is $+0.3257$ nats and $+1.7261$ bits							
20	1,260	27.3%	48	$+\mathbf{0.0767}$ [$+0.0559$, $+0.0986$]	**42/48**	$+\mathbf{0.4944}$ [$+0.3817$, $+0.6027$]	**42/48**
65	4,095	88.7%	48	-0.0303 [-0.0718, $+0.0069$]	22/48	-0.1262 [-0.2299, -0.0240]	16/48
5 **tok/par**, ppl 162; level at s=1 is $+0.2580$ nats and $+1.2100$ bits							
20	1,260	10.9%	12	$+0.0139$ [-0.0305, $+0.0550$]	8/12	$+0.2152$ [$+0.0748$, $+0.3498$]	10/12
65	4,095	35.5%	48	$+\mathbf{0.0338}$ [$+0.0097$, $+0.0569$]	**35/48**	$+\mathbf{0.2772}$ [$+0.2032$, $+0.3527$]	**42/48**
150	9,450	81.9%	48	-0.0052 [-0.0304, $+0.0192$]	25/48	$+0.1607$ [$+0.0855$, $+0.2385$]	32/48
10 **tok/par**, ppl 119; level at s=1 is $+0.1660$ nats and $+0.9676$ bits							
65	4,095	17.7%	48	$+0.0084$ [-0.0286, $+0.0441$]	24/48	$+0.0879$ [-0.0245, $+0.2027$]	27/48
150	9,450	40.9%	48	$+\mathbf{0.0293}$ [-0.0038, $+0.0634$]	**27/48**	$+\mathbf{0.0988}$ [-0.0262, $+0.2247$]	**26/48**
20 **tok/par**, the Chinchilla-optimal ratio, ppl 100; level at s=1 is $+0.0855$ nats and $+0.8253$ bits							
20	1,260	2.7%	12	-0.0037 [-0.0804, $+0.0814$]	5/12	-0.1267 [-0.4065, $+0.1875$]	3/12
65	4,095	8.9%	48	$+0.0198$ [-0.0267, $+0.0671$]	25/48	$+0.0004$ [-0.1430, $+0.1402$]	27/48
150	9,450	20.5%	48	$+\mathbf{0.0279}$ [-0.0289, $+0.0840$]	**24/48**	$-\mathbf{0.0346}$ [-0.1795, $+0.1073$]	**23/48**
300	18,900	40.9%	12	$+0.0150$ [-0.0570, $+0.0927$]	5/12	-0.1208 [-0.3513, $+0.1447$]	3/12
600	37,800	81.9%	12	$+0.0106$ [-0.0755, $+0.0966$]	6/12	-0.0870 [-0.3520, $+0.1362$]	6/12

Two arguments the body states only as verdicts. Depth's top-rung point estimate is $+32.6\%$ of its own clustered level, so what a coin-flip sign rules out there is a direction and not a magnitude, which is why the one-sided bound below and not the win rate carries the equivalence. Both are properties of the rung as run, at retention 15,392; the matched cell at that budget resolves on exposure. And the ladder's top-rung null rests on a *conservative* estimate: Section 4.7's uncontrolled mechanisms decrease in s and act over the longest gaps anywhere in this paper here, so they push this contrast down, by the same reasoning that makes the positive

result a lower bound. A measured zero is therefore consistent with a real effect masked by stronger suppression, and we run that argument in both directions or neither.

Bold marks each rung's comparator. Permutation p for the comparator, selection absorbed by permuting spacing labels within seed, is 0.0001, 0.011, 0.091 and 0.254 on depth and $<5\times10^{-5}$, $<5\times10^{-5}$, 0.109 and 0.666 on exposure, rising monotonically on both. **Not one of the top rung's five cells resolves on either measure**, from 2.7% to 81.9% of training, which is what Section 4.8 claims; this table is what makes it checkable.

Reading the share-aligned escape off this table. Spacing is in absolute steps, so a fixed s covers a shrinking share of training as the budget grows: $s = 20$ spans 27.3% of T_{last} at the bottom rung and 2.7% at the top. The comparison the escape needs is therefore at matched *share*, and the top rung brackets the bottom rung's winning 27.3% on both sides: $s = 65$ at 8.9%, $s = 150$ at 20.5%, $s = 300$ at 40.9%, with $s = 20$ at 2.7% and $s = 600$ at 81.9% outside. All five are null. The bottom rung's own wide cell, $s = 65$ at 88.7%, is *negative* on both measures, which is the same descending branch Section 4.7 declines to interpret.

The dose test, and what it identifies that the rungs cannot. The identifying variation for retention is not across rungs. It is inside each run. At $s = 1$ every copy of a set lands in consecutive steps, so the injection window is $k - 1$ steps wide: 63 steps at $k = 64$, which is 1.36% of T_{last} at the bottom rung and 0.14% at the top. Retention is therefore constant to within that across the whole $k = 2$ to 64 sweep, within one run. Climbing the ladder multiplies retention ten times. **So the count sweep is a $32\times$ dose manipulation at fixed retention, and the rung is a $10\times$ retention manipulation at fixed dose.**

A retention effect that does not depend on count multiplies every set's level by one factor $r(\text{rung})$, giving $\text{level}(k) = f(k)\,r$. That has two consequences, both testable on runs already in the table above. The count exponent $\mathrm{d}\log\text{level}/\mathrm{d}\log k$ is unchanged across rungs, since $\log r$ enters as an intercept. And the ratio $\text{level}_{20}(k)/\text{level}_2(k)$ is constant in k. Neither holds.

(i) the count exponent, which a count-independent factor cannot move

tok/par	β on depth	β on exposure
2 (published budget)	$+0.494$ [$+0.381$, $+0.704$]	$+0.423$ [$+0.340$, $+0.524$]
5	$+0.246$ [$+0.178$, $+0.337$]	$+0.252$ [$+0.195$, $+0.321$]
10	$+0.099$ [$+0.005$, $+0.215$]	$+0.110$ [$+0.037$, $+0.184$]
20 (Chinchilla ratio)	$+0.027$ [-0.330, $+0.378$]	-0.002 [-0.103, $+0.103$]
20 minus 2	$\mathbf{-0.467}$ [-0.831, -0.140]	$\mathbf{-0.425}$ [-0.548, -0.314]

(ii) $\text{level}_{20}/\text{level}_2$ per count, which such a factor requires to be constant

k	depth	exposure
2	0.671 [0.129, 1.849]	1.030 [0.660, 1.667]
4	0.422 [0.019, 1.013]	0.776 [0.511, 1.145]
8	0.499 [0.084, 1.174]	0.878 [0.582, 1.321]
16	0.259 [0.047, 0.490]	0.428 [0.273, 0.600]
32	0.212 [0.062, 0.367]	0.410 [0.299, 0.535]
64	0.120 [0.013, 0.224]	0.222 [0.162, 0.287]
ratio of k=64 to k=2	**0.178** [0.009, 0.585]	**0.216** [0.127, 0.350]

The exponent falls monotonically over the four rungs on both measures and the fall excludes zero. The ratio falls monotonically in k on both, and its own ratio between $k = 64$ and $k = 2$ excludes one. A single count-independent factor predicts a flat column and a ratio of exactly one, so **a retention effect blind to count is rejected**. That leaves a retention effect that is not blind to count, and the next paragraph measures it directly.

The estimator, stated because the magnitude depends on it. β is the ordinary least squares slope of log of the seed-mean level on $\log k$ over the full sweep $k \in \{2, 4, 8, 16, 32, 64\}$; intervals are 4,000 cluster bootstrap resamples over the 48 seeds, recomputing the means and refitting each time, at the seed of Appendix C.1. Seed-mean rather than per-seed because a per-seed fit needs six positive levels and dropping the seeds without them selects on the outcome; the one rung where resamples go non-positive is the top one, at 223 of 4,000 on depth, which is why that interval is the wide one. A per-seed fit restricted to $k \geq 8$ instead gives 0.705, 0.287, 0.251 and -0.011 on depth. **The monotone collapse to zero replicates under both, and that is what the argument uses**; the magnitude of any single exponent does not carry it. As an internal check, the mean over the six counts of these $s = 1$ levels is $+0.3257$ nats and $+1.7261$ bits at the bottom rung and $+0.0855$ and $+0.8253$ at the top, against the $s = 1$ levels printed in the table above.

What this does not rest on, and what it left. It does not rest on spacing, because it is read entirely at $s = 1$: the spacing-*dependent* forgetting that the ratio-invariance argument below cannot exclude has no spacing here to act on. What it left was a retention effect that is itself count-dependent, acting more strongly on heavily repeated

content than on twice-seen content. We ran the cell that separates that, and it is the one alternative the dose test could not exclude that turns out to be the right one.

The retention-matched top rung, and the withdrawal it forces. The top rung again at $T = 61{,}570$ with $T_{\text{last}} = 60{,}031$, so $T - T_{\text{last}} = 1{,}539$, the bottom rung's retention exactly. The rate is constant at 6×10^{-4} throughout the ladder, so this cell differs from the standard top rung in the retention interval and, necessarily, in the absolute position of the last copy, and in nothing else: same budget, same width, same corpus, same stream, same rate, same spacing, 48 seeds at 900 to 947. **The decision rule was fixed before the cell ran**: if the count exponent returns at matched retention, the collapse up the ladder is retention rather than budget, and the budget attribution is withdrawn.

And then a second matched cell, at a third budget. The cell above leaves one residual, and this is the experiment that tests it: 10 tokens per parameter, $T = 30{,}785$ with $T_{\text{last}} = 29{,}246$, so retention is 1,539 again, 48 seeds at 900 to 947, at $s = 1$ and at the comparator spacing. That puts three budgets $10\times$ apart on the single line that fixed retention permits, with injection at steps 4,618, 29,246 and 60,031. Its rule was fixed the same way.

cell at s=1, $k = 2$ to 64	$T - T_{\text{last}}$	β on depth	β on exposure
2 tok/par, the published budget	1,539	$+0.494$ [$+0.381$, $+0.704$]	$+0.423$ [$+0.340$, $+0.524$]
20 tok/par, as run on the ladder	15,392	$+0.027$ [-0.330, $+0.378$]	-0.002 [-0.103, $+0.103$]
10 tok/par, **retention-matched**	1,539	$+0.402$ [$+0.319$, $+0.520$]	$+0.420$ [$+0.360$, $+0.495$]
20 tok/par, **retention-matched**	1,539	$+0.467$ [$+0.311$, $+0.789$]	$+0.375$ [$+0.281$, $+0.491$]
matched 10 minus the ladder's top rung		$\mathbf{+0.375}$ [$+0.028$, $+0.749$]	$\mathbf{+0.422}$ [$+0.310$, $+0.546$]
matched 20 minus the ladder's top rung		$\mathbf{+0.440}$ [$+0.169$, $+0.781$]	$\mathbf{+0.377}$ [$+0.285$, $+0.485$]
matched 10 minus the bottom rung		-0.092 [-0.272, $+0.042$]	-0.003 [-0.116, $+0.113$]
matched 20 minus the bottom rung		-0.028 [-0.236, $+0.255$]	-0.048 [-0.161, $+0.064$]
matched 10 minus matched 20		-0.065 [-0.364, $+0.111$]	$+0.045$ [-0.075, $+0.167$]

The exponent returns in full. It rises against the ladder's own top rung on both measures, and against the bottom rung it is indistinguishable on both. The matched cell's levels at $k = 2$ to 64 are $+0.755$, $+0.808$, $+0.980$, $+1.181$, $+1.660$ and $+2.908$ bits, against the standard top rung's $+0.862$, $+0.791$, $+0.823$, $+0.736$, $+0.966$ and $+0.774$: the $k = 2$ level is the same to within noise and the $k = 64$ level is nearly four times larger. **So the budget did not remove the accumulation across repetitions. The extra 13,853 steps of forgetting did**, and Section 4.8 says so rather than keeping a reading this cell contradicts.

Is the flat $k = 2$ cell the mechanism or a floor? Partly a floor, and the other measures say so. It is fair to ask whether $+0.8$ bits above control at $k = 2$ is insensitive because retention does not reach a twice-seen sequence, or because exposure is near a floor there that neither cell escapes, and the same runs carry two other reads. They do, and they favour the second. Below, each cell's level above control at $s = 1$, matched minus the ladder's top rung, on the same 48 seeds:

k	exposure, matched − ladder (bits)	per-token top-1, matched − ladder
2	-0.107 [-0.363, $+0.145$]	$\mathbf{+0.026}$ [$+0.017$, $+0.035$]
8	$+0.157$ [-0.113, $+0.451$]	$+0.028$ [$+0.017$, $+0.040$]
32	$+0.694$ [$+0.391$, $+1.000$]	$+0.019$ [$+0.008$, $+0.030$]
64	$+2.135$ [$+1.751$, $+2.544$]	$+0.057$ [$+0.045$, $+0.070$]

Top-1 at $k = 2$ does move with retention where exposure does not, so a twice-seen sequence is affected and the flat exposure cell is partly a measurement floor. We say so rather than keeping the stronger reading. **What the top-1 read does not do is overturn the exponent**: β on top-1 is $+0.219$ [$+0.138$, $+0.329$] at the bottom rung, $+0.120$ [-0.051, $+0.404$] at the ladder's top rung and $+0.167$ [$+0.100$, $+0.241$] at the matched cell, which points the same way as depth and exposure without resolving, its middle interval being wide enough to contain both ends. The quantile behind exposure gives $+0.248$, $+0.033$ and $+0.208$, resolving like exposure. **So the collapse and return of the exponent holds on three of four measures and contradicts none**; what is scoped to exposure and the quantile is the claim that a twice-seen sequence is untouched, and that claim is withdrawn.

What the cell cannot hold, and how the exponent is tested anyway. Retention is $T - T_{\text{last}}$, so token budget, absolute position in training and retention satisfy one identity and *no* experiment fixes two of them while moving the third. Saying this cell moves the retention interval alone would be false, and we do not. What it does is hold the budget and move position by 13,853 steps; the ladder holds position as a fixed fraction of T and moves the budget; and Appendix B.4 holds recency and moves position. Reading the three together:

comparison	what is held	what moves	$\Delta\beta$ on exposure
matched cell vs the ladder's top rung	budget, 20 tok/par	retention 1,539 → 15,392; position 13,853 steps	$+\mathbf{0.377}$ $[+0.285, +0.485]$
matched cell vs the bottom rung	**retention**, 1,539	budget 20 vs 2; position 55,413 steps	-0.048 $[-0.161, +0.064]$
the two matched cells against each other	**retention**, 1,539	budget 10 vs 20; position 30,785 steps	$+0.045$ $[-0.075, +0.167]$
matched 10 vs the bottom rung	**retention**, 1,539	budget 10 vs 2; position 24,628 steps	-0.003 $[-0.116, +0.113]$

Position moves in both comparisons and β moves in only one, the one where retention moves. **And the comparison with four times the position difference is the one with no exponent difference.** That is the wrong way round for an account in which convergence at injection sets β. The matched cell injects into a near-fully-trained model at step 60,031, the bottom rung into a barely-trained one at step 4,618, and their exponents are indistinguishable on both measures. Appendix B.4's position effect, $+0.311$ nats $[0.227, 0.399]$ and $+0.528$ bits $[0.308, 0.722]$ late minus early at fixed recency, is a *level* effect; nothing there moves a count exponent.

What the third budget settles, and what it does not. Fixing retention forces position to $T - 1{,}539$, so every matched cell sits on one line in the (budget, position) plane and no design leaves that line. Two points on it cannot separate a budget effect from a position effect that cancels it. Three can, because one of them lies between the others: a monotone interaction along the line takes different values at 2, 10 and 20 tokens per parameter, and no pairwise difference among the three resolves on either measure. The largest is -0.048 $[-0.161, +0.064]$ bits on exposure and -0.092 $[-0.272, +0.042]$ nats on depth, and the widest of the three intervals spans 0.24 bits and 0.49 nats. **So a monotone position-budget interaction is bounded by those widths rather than left open. What remains** is an interaction non-monotone along the line that happens to agree at exactly these three budgets, and we do not claim against that.

The same matched cell at the comparator spacing, and the massed penalty returns. The exponent above is read at $s = 1$, so it says nothing about the spacing contrast. The matched arm therefore ran two pre-registered cells and not one: $s = 1$ and $s = 150$, the top rung's comparator, 48 seeds each. **There is no selection to absorb at this cell**, unlike the rungs: $s = 150$ is what the pre-registered rule returns on the *ladder's* top-rung runs, and it is applied here unchanged to different runs at a different T_{last}, so the selection and the test use disjoint data. Contrast is comparator minus clustered, averaged over the six count sets within a run and paired across spacings by seed, exactly as the per-cell table above:

cell	$T - T_{\text{last}}$	depth (nats)	exposure (bits)
2 tp, s=20; win 42/48, 42/48	1,539	$+0.0767$ $[+0.056, +0.099]$	$+0.4944$ $[+0.382, +0.603]$
20 tp, s=150; win 24/48, 23/48	15,392	$+0.0279$ $[-0.029, +0.084]$	-0.0346 $[-0.180, +0.107]$
20 tp, s=150, **matched**; win 27/48, **32/48**	1,539	$+0.0285$ $[-0.015, +0.074]$	$+\mathbf{0.2571}$ $[+0.042, +0.472]$
matched minus the ladder's top rung		$+0.0006$ $[-0.074, +0.072]$	$+\mathbf{0.2918}$ $[+0.031, +0.559]$
matched minus the bottom rung		-0.0482 $[-0.097, -0.000]$	-0.2373 $[-0.480, +0.007]$

Permutation p, spacing labels shuffled within seed over 20,000 draws, is 0.021 at the matched cell on exposure against 0.640 at the ladder's own top rung, and $< 10^{-4}$ at the bottom rung. **So on exposure the massed penalty is present at** 20 **tokens per parameter**, and what removed it on the ladder was the retention interval. The matched value is about half the bottom rung's, but the two rungs' comparators are $s = 150$ and $s = 20$, so that gap mixes rung with spacing; it does not resolve either way, and we claim no attenuation with budget, only that the effect survives.

The exposure interval here is within one canary panel, and carrying the between-panel component onto it removes its sign. Appendix E.1 estimates $\sigma_{\text{between}} = 0.1071$ bits on exposure over five panels, and the quadrature this paper applies elsewhere, $1.96\sqrt{\text{SE}^2 + \sigma^2_{\text{between}}}$, gives 0.2995 against this cell's $\text{SE} = 0.1090$: the contrast becomes $[-0.042, +0.557]$ and the retention difference $[-0.045, +0.629]$, and both cover zero. We print that rather than leave it to be recomputed. **We also do not read it as a refutation, for a reason that cuts both ways.** σ_{between} was estimated on five panels that all sit on the *primary* configuration, so carrying it to a different T_{last} and a different budget is an extrapolation rather than an application. If it cannot be transported here, it does not condemn this cell; but then neither can this cell claim to generalise across canary content. What stands on its own terms is the within-panel result, 32 of 48 seeds at a permutation $p = 0.021$ under a rule fixed before the run. What is not shown is that it holds across canary content, because the five-panel replication was never run at matched retention, and that cell is the one experiment that would settle it.

The cross-panel replication, taken two panels at a time. The rule was fixed before the run: rerun the matched cell on an independently generated canary panel at 48 seeds per spacing, and report either way. A third panel has since been run, and the paragraph below reports all three together; this paragraph is the two-panel reading as it

stood, kept because the power account it tests is what the third panel then bears on. On the new panel the contrast is $+0.715$ $[+0.417, +1.009]$ bits and $+0.130$ $[+0.081, +0.182]$ nats, 37 of 48 seeds, permutation $p < 10^{-4}$ on both, against the published panel's $+0.257$ $[+0.042, +0.472]$ and $+0.029$ $[-0.015, +0.074]$. **Both panels are positive and the new one resolves on both measures**, which is consistent with the power account: the depth-to-exposure ratio is 0.111 on the published panel and 0.182 on the new one, and the published panel's depth cell sat below the $+0.064$ nats that 48 seeds can resolve. Two ratios cannot separate depth tracking exposure from this panel simply having a larger effect on both measures, so the account is supported and not established, and the depth dissociation on the published panel stays open. **What two panels cannot give is a variance component.** The between-panel term they imply has one degree of freedom, being a single difference, so no interval built on it is reportable. That is what the third panel changes, and the paragraph below measures the component here rather than transporting it. The magnitudes differ by $+0.458$ between panels, which is consistent with the published panel being the smallest of the five on the primary configuration as well.

The nonparametric version: the penalty against retention, and against position, with nothing fitted to either. Every comparison above reads β, which is fitted inside a separable form Section 4.1 rejects in 7 of 10 arms. The massed penalty is a difference of means and needs no such form, so the same question can be asked of it, and two cells run at the comparator spacing make that possible. Both hold retention at $1{,}539$ and move the last copy: 5 tokens per parameter at $T = 15{,}392$ with $T_{\text{last}} = 13{,}853$, and 10 at $T = 30{,}785$ with $T_{\text{last}} = 29{,}246$, each at $s = 1$ and $s = 150$, 48 seeds at 900 to 947. With the cell above that gives three matched-retention cells whose last copy sits at 13,853, 29,246 and 60,031 steps, a $4.3\times$ span of absolute position at one retention, against three rungs as run at nearly the same positions. **The reading rule was fixed before they ran** and it decides against this paper as readily as for it: if each matched cell sits near the rung at its own position, position sets the penalty and Section 4.8's attribution is withdrawn; if the matched cells disagree, neither axis orders it and both readings go. Read on exposure at $s = 150$, 48 seeds paired by seed, the estimator of Appendix C.3 unchanged.

cell at s=150	$T - T_{\text{last}}$	pos.	t/p	exposure (bits)	win	depth (nats)
matched	1,539	13,853	5	$+\mathbf{0.347}$ $[+0.211, +0.481]$	$\mathbf{37/48}$	$+0.008$ $[-0.016, +0.032]$
matched	1,539	29,246	10	$+\mathbf{0.395}$ $[+0.242, +0.552]$	$\mathbf{38/48}$	$+0.013$ $[-0.024, +0.050]$
matched	1,539	60,031	20	$+\mathbf{0.257}$ $[+0.042, +0.472]$	$\mathbf{32/48}$	$+0.029$ $[-0.015, +0.074]$
as run on the ladder	3,848	11,544	5	$+0.161$ $[+0.086, +0.239]$	$32/48$	-0.005 $[-0.030, +0.019]$
as run on the ladder	7,696	23,089	10	$+0.099$ $[-0.026, +0.225]$	$26/48$	$+0.029$ $[-0.004, +0.063]$
as run on the ladder	15,392	46,178	20	-0.035 $[-0.180, +0.107]$	$23/48$	$+0.028$ $[-0.029, +0.084]$
Spearman over the six cells, exact p over all 720 orderings						
				vs. retention	$-\mathbf{0.941}$, $p = 0.017$	$+0.273$, $p = 0.617$
				vs. absolute position	-0.086, $p = 0.919$	$+0.657$, $p = 0.175$
				vs. token budget	-0.359, $p = 0.533$	$+0.717$, $p = 0.178$

The penalty follows retention and does not follow position. On exposure the three matched cells agree within one another's intervals and each is larger than the rung at its own position, including the pair where the matched cell's last copy sits *later* in training, which is the wrong direction for a position account. Across all six, retention orders the penalty at $\rho = -0.941$ and position does not order it at all; inside the matched set, where position spans $4.3\times$ at one retention, it orders nothing. **The retention difference at fixed budget now replicates at three budgets**, paired by seed on the same stream, width, rate and seed range with only T_{last} moving: $+0.186$ $[+0.034, +0.328]$ at 5 tokens per parameter, $p = 0.020$; $+0.296$ $[+0.103, +0.487]$ at 10, $p = 0.003$; $+0.292$ $[+0.032, +0.552]$ at 20, $p = 0.034$.

And the widening: the component can now be measured where it matters, and it is larger than the one we transported. The objection above is that Appendix E.1's between-panel component covers this cell's interval, and that component is estimated on the primary grid and carried here. The matched cell at 20 tokens per parameter has since been run on three independently drawn canary panels at this exact configuration, canary seeds 777 published, 778 and 779, 48 paired seeds each and 288 runs, differing in the canary seed alone. On exposure they give $+0.257$ $[+0.042, +0.472]$, $+0.715$ $[+0.417, +1.009]$ and $+0.596$ $[+0.439, +0.753]$ bits; on depth $+0.029$ $[-0.015, +0.074]$, $+0.130$ $[+0.081, +0.179]$ and $+0.019$ $[-0.017, +0.055]$ nats. The component measured here is $\sigma_{\text{between}} = 0.196$ bits on exposure, $Q = 8.2$ on 2 df, against the 0.1071 we carried over, so **the transport is anti-conservative by a factor of** 1.8 **at the one configuration where it can be checked**. Widened at the measured component the published cell covers zero, $[-0.183, +0.697]$ bits, and **we withdraw the reading that two of the three matched cells survive widening**: it rested on the transported component, and the other two budgets have no panel replication from which to measure one. What survives is each draw on its own: **all three panels resolve positive on exposure alone**, $+0.257$ $[+0.042, +0.472]$, $+0.715$ $[+0.417, +1.009]$ and $+0.596$ $[+0.439, +0.753]$ bits. **The pooled interval is the part that does not hold.** The DerSimonian-Laird mean over the three panels is $+0.512$ $[+0.255, +0.770]$ bits on exposure, but at $Q = 8.2$ on 2 df this is exactly the regime Appendix C.4 says DL under-covers in, and the HKSJ interval we report beside DL there gives $[-0.067, +1.092]$ and covers zero. So the across-panel exposure evidence is three independent resolutions and not one pooled interval, and we do not quote the pooled figure without this. The published panel is the lowest of the three draws, and on depth the across-panel mean covers zero on

both rules, $+0.059$ $[-0.004, +0.123]$ nats under DL and $[-0.094, +0.208]$ under HKSJ, which is Section 4.8's dissociation again. **Depth stays null on every axis**, with no matched cell resolving and no axis ordering it, so three further cells do not repair the dissociation and we do not report them as if they had. **What this does not do** is isolate retention from position: the three quantities satisfy one identity, fixing retention forces position to $T - 1{,}539$, and every matched cell therefore sits on one line that no design leaves. An interaction non-monotone along that line which happens to agree at these three budgets is still open, exactly as above. What is new is that the ordering argument no longer rests on a fitted exponent.

Depth dissociates on the published panel, and the two retention results do not agree with each other on it. The matched cell reads $+0.0285$ against the ladder's $+0.0279$: matching retention moves depth's spacing contrast by $+0.0006$, nothing, and the matched cell sits resolvably below the bottom rung. Yet depth's *count exponent* does track retention, $+0.494$ to $+0.027$ to $+0.467$. So on depth the exponent collapse is retention and the spacing-contrast decay is not, and we have no account that explains both. Exposure is the pre-specified primary measure for canaries and Section 4.8's claim rests on it, which is the ordering fixed before any of these runs; the split is a real limitation and not a presentational one.

What this does to the equivalence bound above. The one-sided 95% upper bounds of $+0.083$ bits and $+0.076$ nats are properties of *the ladder's* top rung, at retention $15{,}392$. They are not bounds on the effect at 20 tokens per parameter, because the matched cell at that budget resolves on exposure. Read them as bounds at long retention.

What we withdraw, and what we do not. Withdrawn: that the ladder identifies a token budget at which the count exponent vanishes. It identifies no such thing, because the ladder moves retention with budget by construction and the matched cell attributes the effect to retention. Not withdrawn: the exponent's dependence on the retention interval itself, which the matched comparison supports against budget rather than confounding with it, since it holds the budget fixed while position moves with retention. Nor the ladder's per-cell numbers, which are measurements and stand.

Budget and retention are not separated by the ladder's design. T_{last} is pinned at a fixed *fraction* of training, so the retention interval $T - T_{\text{last}}$ is $1{,}539$, $3{,}848$, $7{,}696$ and $15{,}392$ steps up the rungs, where the ten four-spacing arms pin it near $1{,}500$ (Appendix A.1). Pinning the fraction is what keeps the last copy at the same point of the schedule, which is what confound (2) requires; the cost is that the top rung has ten times longer after its last copy than the positive control does. So on the rungs alone the top rung's null is also consistent with forgetting rather than with convergence, which is why the dose test above reads the clustered cell instead.

Two weaker things the rungs settle on their own, the first being the argument that held before the dose test exposed its limit. Any retention kernel that acts equally across spacings multiplies the whole spacing profile by a single factor, so it leaves the contrast *as a fraction of its own level* unchanged. That fraction is $+28.6\%$ at the bottom rung and $+0.05\%$ at the top, so a spacing-blind retention effect cannot produce the collapse on its own: it would have to act differentially in s. That was a constraint on the alternative and not an identification of ours, and **we withdraw it as evidence**: it holds only if forgetting is blind to spacing and to count, the dose test shows it is not blind to count, and the matched cell shows the retention reading is the correct one. We record the argument because it is what the retention reading replaced, and because its limit is instructive.

The equivalence bound, and why it is not a p-value. A coin-flip sign says the direction is undetermined; it does not by itself bound the size. So for the top rung we also report a one-sided 95% upper confidence bound on the contrast, bootstrapped on *the comparator cell this table reports*, so the bound and the point estimate are the same quantity. Against the published penalties of $+0.093$ nats and $+0.733$ bits:

rung	measure	point	one-sided 95% upper	as a fraction of the published penalty
20 tok/par	depth	$+0.0279$	$+0.0756$	81%, so it does *not* exclude an effect that size
20 tok/par	exposure	-0.0346	$+0.0825$	**11%**, so it excludes it
10 tok/par	depth	$+0.0293$	$+0.0579$	62%
10 tok/par	exposure	$+0.0988$	$+0.2069$	28%

So the equivalence result is carried by exposure and not by depth, and Section 4.8 says so. **Each bound is a 95th percentile of the same cell whose interval is tabulated above, so it must sit inside that interval's upper end, and it does**: $+0.0756$ against $+0.0840$ and $+0.0825$ against $+0.1073$ at the top rung. Absorbing the comparator's own selection over both interior cells instead, rather than conditioning on it, widens the top rung to $+0.0777$ and $+0.1238$ and changes no conclusion. Depth's bound is the wide one because its level is small by the top rung, $+0.0855$ nats against the bottom rung's $+0.3257$, so the same absolute noise is a larger share of it.

G.2 Does convergence explain the batch-size peak shift?

This subsection states a reading at full strength and then refutes it, so the first two paragraphs are not its conclusion. What it concludes: a fixed interfering-token volume does not explain the peak shift, and at $B = 256$

the measured structure occupies 0.69 of a production batch rather than 0.13. Its canary-share series ends in a candidate driver, not a measured one.

The interfering-token conversion, and the reading it licensed. Appendix H.3's training-length row states the conversion in one sentence and this is the arithmetic. If interfering data rather than steps is the invariant, then at context 256 the fitted s^* of 35.0, 31.2 and 10.7 steps at $B = 16, 32, 64$ correspond to 1.4, 2.6 and 1.8×10^5 interfering tokens. Those three span $1.78\times$, which already exceeds the $1.7\times$ estimator floor established below rather than sitting inside it, so the invariant was never better than marginal even on its own arm. At a production batch of order 2×10^6 tokens per step the same conversion puts the whole measured structure inside about one batch: the peak at 0.13 batches and Section 4.3's crossing points at 0.91 and 1.52. We set that reading out at full strength because the length sweep, which had been our refutation of it, is withdrawn as a convergence effect (Appendix G.3). The four-octave fit below refutes it directly instead, and reverses the consequence: at $B = 256$ the structure occupies 0.69 of a batch, not 0.13, and crosses one batch at $B \approx 420$.

The matched-token experiment, and why the obvious design of it is invalid. The direct way to decouple batch size from convergence is to hold tokens seen fixed: $B = 32$ for 6,000 steps and $B = 64$ for 3,000 steps consume the same 192,000 sequences. Run that way at a constant learning rate, it does not work, and the validation loss says so. The $B = 64$ arm finished at 5.810 against the $B = 32$ arm's 5.126, a gap of 0.684 nats, and showed no spacing peak at any grid point ($p = 1.0$ on exposure). Halving the step budget halves the number of optimiser updates, and at a fixed learning rate that dominates the benefit of a larger batch, so the design trades a coupling through tokens for a stronger one through update count. Its apparent s^* ratio of 13.0 measures that mismatch, and we do not report it as a result.

The linear scaling rule repairs it: doubling the learning rate with the batch brings the $B = 64$ arm to 5.1247 against 5.1262, a gap of 0.0015 nats. Only then are tokens seen and degree of convergence both matched, and only then is the s^* comparison meaningful. On that pair the ratios are 4.5 on depth and 3.6 on exposure for a factor of 2 in batch size, implying γ of 2.16 and 1.85 where the clock predicts 1. We do not claim that exponent: the $B = 32$ exposure peak is clipped at the grid edge, the $B = 64$ depth peak is clipped at the other edge, and the $B = 64$ arm's peak is not significantly above its own massed end ($p = 0.42$ exposure, 1.0 depth). What the arm establishes is the direction under matched convergence, which is what the convergence objection required and what the earlier indirect argument could only suggest.

The wide-grid arm does not obey the token invariant, and this narrows it. The three batch sizes agree that the peak sits at roughly 2×10^5 interfering tokens, but that agreement is measured at a fixed training length of 6,000 steps. The wide-grid arm of Section 4.3 trains for 20,000 steps at the same $B = 32$, where batch size alone predicts $s^* \approx 24$. Measured, its peak is at $s^* = 67.6$ [19, 115] on depth and 88.6 [47, 165] on exposure, and the exposure interval excludes 24. In interfering tokens that is 7.3×10^5, roughly three times the invariant.

We cannot attribute that shift to training length alone, because the wide-grid arm also uses a different canary density and count set. It only bounds the claim: the token invariant holds across a factor of 4 in batch size at fixed length and does not survive a factor of 3.3 in length.

It is tempting to add that a fixed volume of interfering data and a fixed fraction of training fit the batch sweep about equally. On these numbers they do not, and the correct statement is asymmetric. The batch sweep holds length at 6,000 steps, so a length-relative clock predicts s^* *constant in steps* across it under every reading we can construct, fraction of steps, of optimiser updates or of tokens seen; measured, s^* is 35.0, 31.2 and 10.7 and the $B = 64$ interval [9.5, 12.5] excludes the $B = 32$ value of 31.2. Expressed in tokens it fares no better, since it would require s^* in tokens to grow as B, that is 1:2:4, against measured 1.4, 2.6 and 1.8×10^5. So volume fits the batch sweep and fails the wide grid; length-relative does the reverse. Neither single clock fits both arms, and the length sweep below shows why: s^* depends on batch *and* on length, so an arm varying one cannot settle the other and there was never a tie to break, only two exponents to measure separately.

The length sweep, which settles it. Only total length varies: $B = 32$, cpg 6, counts $\{2, 4, 8, 16\}$, $t_{\text{last}} = 0.75T$ so the fraction of training after the last copy is matched, and six spacings at every length before the rightward extension, 12 runs each. The pre-registered reading was written into code before the runs: s^* constant in steps supports the volume clock, s^* proportional to T the length-relative one, anything between is reported as the measured exponent with its interval.

T	runs	argmax	interpolated s^* (exposure)	s^*/T	interfering tokens	val. loss
3,000	12	20	16.1 [10.5, 20.5]	0.54%	1.3×10^5	5.988
6,000	12	65	63.1 [56.5, 100.0]	1.05%	5.2×10^5	5.127
12,000	12	65	67.0 [59.7, 76.4]	0.56%	5.5×10^5	4.595

The fitted exponent is $\delta = +1.03$ [$+0.83, +1.37$], which excludes 0 and does not exclude 1. We read that as the volume clock being out and the length-relative clock surviving. **That reading is withdrawn**: the three cells

differ by 1.393 nats in validation loss, and the matched-convergence arm of Appendix G.3 shows the optimum does not move when length is quadrupled at equal loss, so δ is not identified as a length effect. Everything in the rest of this section describes what the unmatched sweep measured and should be read as such. A clipped peak would bias δ downwards, towards the volume clock, so we extended the longest arm's grid to $s = 400$, the $T = 6{,}000$ arm's edge being $s = 250$, and confirmed its peak is interior; Appendix C.2 gives that test and its two adverse findings, one of which is that the governing Wilcoxon for the $T = 12{,}000$ edge comparison is marginal at $p = 0.052$ where the bootstrap gives 0.010. That the length-relative prediction from the shortest arm, $4 \times 16.1 = 64$, lands on the measured 67 is a coincidence we did not arrange and, given the convergence confound, is not evidence of a length law. The same conclusion follows without any peak estimator from the raw profile of leakage relative to $s = 1$, whose largest entry moves rightwards with length and grows:

T	s=5	s=10	s=20	s=40	s=65	s=100
3,000	0.000	+0.036	**+0.058**	−0.036	−0.077	−0.197
6,000	0.000	−0.004	+0.014	+0.069	**+0.144**	+0.067
12,000	0.000	+0.026	+0.068	+0.093	**+0.273**	+0.164

At $T = 3000$ the wide spacings leak *less* than clustered and at $T = 12000$ they leak considerably more, so the whole curve translates rightwards rather than merely changing height. Depth is not usable in this arm and we say so rather than reporting it as a disagreement: its counts stop at $k = 16$, and Section 4.6 shows the massed penalty on depth resolves only from $k = 16$ upward, so this arm was built without the cells where depth has signal. Its depth s^* is correspondingly non-monotone (26.0, 5.0 clipped, 67.5) with intervals spanning the grid.

The wide-grid comparison, and what it can and cannot establish. The two arms were not designed to check each other, so their agreement is not confirmation. The wide grid trains for 20,000 steps against the batch sweep's 6,000 at the same $B = 32$, a factor of 3.33 in length, and at $\delta = 1.03$ a length-relative clock predicted s^* larger by $3.33^{1.03} = 3.5\times$, that is $31.2 \to 108$ steps. Measured, the wide grid's peak sits at 7.3×10^5 tokens against the $B = 32$ arm's 2.56×10^5, a factor of 2.85, and its exposure s^* is 88.6 against the predicted 108, inside the estimator's $1.7\times$ noise floor. Both comparisons are against the same $B = 32$ baseline.

That agreement is a coincidence between two confounded quantities and is not support. δ is withdrawn (Appendix G.3), so there is no length-relative clock left to do the predicting; and the two arms differ in convergence as well as in length, exactly as the length sweep's cells did. What survives is the measurement, that the wide grid's peak is where it is, and not the inference. Everything above is retained as a record of what the unmatched arms measured.

The second consequence was the reverse of what δ would have licensed, and the batch sweep has since removed it. When δ fell, the interference-volume clock's *invariant* form went from refuted to untested, because the refutation had come from δ; and with δ gone we could not say the worst spacing grows at production lengths, so the pessimistic reading, that the structure sits inside about one production batch and below the resolution at which a pipeline could place duplicates, was live again. Four octaves of batch size now test that invariant form directly and exclude it, on evidence that does not involve δ at all. What the episode leaves is a methodological point rather than an open question: a practical conclusion resting on a single fitted exponent inherits every confound that exponent has, which is why the replacement is an interval over four octaves rather than a second fit.

Batch size and validation loss co-vary by construction in the batch sweep above, so three checks use arms run for other purposes to separate them. *(i)* Validation loss does not order s^* across arms: on depth the lowest-loss arm has $s^* = 40.0$, clipped, while the highest-loss arm has $s^* = 14.3$, and the correlation between validation loss and $\log_2 s^*$ is $+0.22$ on depth but -0.25 on exposure, inconsistent in sign and so not a relation. For $\log_2 B$ it is -0.60 and -0.55, consistent and in the predicted direction. *(ii)* A descriptive joint fit holding validation loss fixed gives $\log_2 s^* = 11.77 - 1.26 \log_2 B - 0.22$ val, that is $\gamma = 1.26$; with few arms and correlated predictors this is descriptive only, and the authority is the four-octave fit below, which does not condition on validation loss. *(iii)* Matched-convergence pairs bound the estimator's noise: two arms with the same batch size, model and validation loss (5.130 versus 5.129) give $s^* = 17.9$ and 29.8 on depth, a floor of about $1.7\times$. The shift in s^* across the first $4\times$ change in batch clears that floor on either estimator, $6.5\times$ by grid argmax on depth and $6.25\times$ interpolated.

The four-octave fit, and what it excludes. Adding $B = 128$ and $B = 256$ takes the lever arm from two octaves to four and replaces the sign with an interval: $\gamma = 0.23$ $[0.05, 0.38]$ on exposure, which excludes both the volume clock's $\gamma = 1$ and 0. Every interval in this appendix resamples *seeds* and carries all of a seed's blocks together, because blocks from one run share a trajectory (Section 3); the block-level version is narrower and we do not report it. On depth two arms clip at a grid end and the remaining three give $\gamma = 0.06$ $[-0.35, 0.65]$, excluding 1 but not 0, so exposure is the measure to read. Three things bound it. The series is *not* monotone, s^* running $35.0, 31.2, 10.7$ then 5.0 at $B = 128$ and 21.0 at $B = 256$, so a power law is a summary and not a form we establish. The $B = 128$ arm at cpg $= 2$ is flat within noise, $1.67, 1.53, 1.32, 1.54, 1.17$ against a standard error near 0.30, so its argmax at the low end is not a located optimum and it is excluded rather than

entered at its edge. And the verdict rests on the widest arm: dropping $B = 256$ gives $\gamma = 0.86$, which *includes* 1 at every replicate count we tried, while dropping any other arm leaves 1 excluded. We quote no bounds for that fit: $B = 16$ clips at the grid edge in a fraction of replicates, and the lower bound moves from -0.32 to $+0.39$ between 4,000 replicates and 80,000, so the interval is not estimable even though the verdict is stable. We report all of this because it cuts against the exclusion, noting also that $B = 256$ is the best-determined arm in the set, 12 seeds with a clean interior peak at $s^* = 21.0$ [16.3, 32.3].

Reachability, in tokens. What practice cares about is not steps but distance in the data stream, $s^*_{\text{tok}} = s^* BL$ with $L = 256$. The volume clock would make this constant; with $\gamma = 0.23$ it grows as $B^{0.77}$: 143k, 256k, 175k, 164k (clipped) and 1,374k tokens at $B = 16$ to 256. The largest arm sits at 0.69 of a 2×10^6-token production batch, and solving $s^*_{\text{tok}} = 2 \times 10^6$ anchored there puts the crossing at $B \approx 420$ [241, 660] sequences, 0.7 octaves out. So at production batch size the structure spans more than one batch rather than hiding inside one, which is the resolution at which a pipeline could place duplicates at all. Both numbers extrapolate, shortly but really, and inherit the non-monotonicity above, so read them as an order of magnitude and not a threshold.

What the widest arm says about the effect itself. These arms were run to move γ, but $B = 256$ is also the best-converged configuration in this paper, so its leakage contrasts bear on the question of whether the massed penalty is an artefact of an undertrained model. Section 4.8 answers it on a ladder, where the penalty survives at 20 tokens per parameter once retention is matched; this arm is a second and independent answer at a single well-converged point. At validation loss 4.507, perplexity 90.6, the lowest of any arm in this paper, all four contrasts resolve at run level, paired over its 12 seeds: against its lowest cell, $s = 5$, the peak at $s = 20$ gives $+0.958$ [$+0.503$, $+1.401$] on exposure ($p = 3.4 \times 10^{-3}$) and $+0.210$ [$+0.108$, $+0.307$] on depth ($p = 4.9 \times 10^{-3}$); against the widest $s = 65$, $+0.694$ [$+0.401$, $+0.964$] ($p = 3.4 \times 10^{-3}$) and $+0.142$ [$+0.050$, $+0.229$] ($p = 1.6 \times 10^{-2}$). It is the only arm in this sweep where all four resolve at once, and the best converged of any, so on this axis the artefact reading is unsupported. **And the massed cell has since been run here, which is the one contrast this arm could not make before.** The published grid was $\{5, 10, 20, 40, 65\}$, with no *massed* cell in Table 1's sense and its lowest at $s = 5$. An $s = 1$ cell has since been run on the same trainer at the same 6,000 steps and the same $\text{cpg} = 2$, 36 seeds against 36 at $s = 20$, and it lands at the same convergence: validation loss 4.5031 against 4.5029, so the contrast is not confounded with how far either cell trained. Paired over the 36 shared seeds the peak beats the massed cell by $+0.376$ [-0.003, $+0.734$] bits on exposure, Wilcoxon $p = 0.038$, 25 of 36 seeds, and by $+0.080$ [$+0.006$, $+0.154$] nats on depth, $p = 0.044$, 24 of 36. **So the massed penalty is positive at the best-converged point in this paper: it resolves on depth, and on exposure the bootstrap lower end sits at -0.003 while the Wilcoxon resolves.** Appendix I's rule is that the Wilcoxon governs where the two disagree; we print both rather than lean on it, and we do not call exposure resolved here. At 0.51 and 0.86 of the primary arm's own penalty the effect is *smaller* at better convergence, not larger, which is the direction the level account predicts. **One reading against us**: $s = 5$ sits below the massed cell on both measures, -0.243 and -0.082, neither resolving, so at this batch size the rise starts later than $s = 5$ and the low end of the curve is not monotone. None of these numbers enters Table 1's tallies, which are the four-spacing grid. Replication at a larger batch does not repair identification, the descending branch still failing its control (Appendix H.1), so what is added is reproducibility at better convergence and not the shape. And $B = 128$ at $\text{cpg} = 2$ peaks at the massed end rather than in the interior, which is the flat arm above and so uninformative rather than contrary; its fraction-restored counterpart at $\text{cpg} = 8$ does peak at $s = 20$, on exposure resolving on both sides and on depth only against the widest point.

Is γ just the falling canary share? It could be: along the $\text{cpg} = 2$ series the canary share of an injection step falls as $1/B$, and the control above names share rather than volume as what tracks s^*, so that series confounds the two and $B^{0.77}$ would assume what the control denies. The share-matched arms test it. Holding share constant with $\text{cpg} = 2, 4, 8$ at $B = 32, 64, 128$ gives $s^* = 31.2, 18.7, 20.3$ on exposure and $\gamma = 0.31$ $[-0.09, 0.65]$, excluding 1 as the main series does and inside its interval, so the exclusion is not an artefact of share. Two caveats: this spans two octaves rather than four, so its interval no longer excludes 0, and on depth it gives $\gamma = 0.45$ $[-0.23, 1.85]$, not excluding 1. Anchored here the crossing is $B \approx 630$ rather than 420. A replicate whose exponent approaches 1 leaves s^*_{tok} almost invariant in B, so its crossing diverges: 30 of 4,000 replicates on exposure and 978 of 4,000 on depth are non-finite and dropped, the surviving percentiles are conditional on that filter, and the depth interval is uninformative. This is why the crossing is quoted as an order of magnitude throughout.

Two facts argue against us. One matched pair, $B = 32$ at validation loss 4.832 and $B = 64$ at 4.837, agrees on s^* (8.9 versus 10.4) despite differing in B, which is what the convergence account predicts; it involves the single arm that fails the massed-penalty contrast on depth, so its peak is the worst-determined in the set, but it is a genuine counterpoint. The canary-fraction control resolves against interference volume. Doubling canaries at $B = 64$ to restore the smaller batch's canary share tracks volume on depth (9.6 versus 10.4, against 29.8 for $B = 32$) and was uninformative on exposure, landing at 18.7 with an interval spanning both candidate values. The same control at $B = 128$, where restoring the $B = 32$ share takes $\text{cpg} = 8$, is informative: it turns an arm with no locatable optimum into a clean interior peak at 20.3 [13.6, 33.3]. Both fraction-restored arms sit above their $\text{cpg} = 2$ counterparts on exposure, by $+0.80$ $\log_2$ units at $B = 64$ $[-0.60, +2.88]$ and $+2.02$ at $B = 128$ $[-1.33, +2.69]$, and both land nearer the $B = 32$ value than their own batch's. But *neither shift*

resolves against zero once seeds are clustered rather than blocks, so the share is a *candidate* driver consistent in direction at two batch sizes, not a measured one. What the share-matched series does establish is narrower and is the point of the next paragraph: the exclusion of a fixed token volume is not an artefact of share. The direct test, holding tokens seen fixed while varying batch, is in Appendix H.3's training-length row.

G.3 Does the optimum grow with training length, or with convergence?

This appendix withdraws δ, one of the four clocks, on the strength of an arm we ran to test it.

Why the arm was needed. Appendix H.3's training-length row fitted $\delta = 1.03$ from $T \in \{3000, 6000, 12000\}$ at a fixed learning rate. Those cells end at validation losses of 5.988, 5.127 and 4.595, a 1.393-nat spread. Elsewhere we treat a 0.5-nat convergence gap as disqualifying, so a 1.393-nat one cannot be waved past: longer training is also better-converged training, and either could move s^*. The batch arm above is not a matched-convergence version of this sweep, which would be the kind of bad cross-reference that turns an acknowledged confound into an apparent resolution: that appendix's matched-convergence work is the *batch-size* comparison.

Design. $T = 12{,}000$ steps at learning rate $6 \times 10^{-4} \times 0.18 = 1.08 \times 10^{-4}$, the multiplier chosen so that end-of-training loss lands on $T = 3{,}000$'s rather than on $T = 12{,}000$'s. **So this arm is matched to the *least* converged cell of the sweep, validation loss** 5.95 **against the primary arm's** 5.130**, perplexity** 384 **against** 169**.** It is a length test at fixed convergence, not a convergence test, and Section 5's practical claim is not stated for it. T_{last} scales to 9,000 so the pinned last copy sits at the same *fraction* of training, counts $\{2, 4, 8, 16\}$ at cpg 6, batch 32, spacings $\{5, 20, 65, 150\}$, 12 seeds. The cross-arm table below is on the spacings the two arms share, $\{5, 20, 65\}$, because an argmax is not comparable across different grids; the fourth cell is reported separately just after it. Achieved match: 5.9557 against 5.9879, a gap of 0.032 nats, which is 2.3% of the spread the unmatched sweep has.

Result: at matched convergence the optimum does not move, and on exposure it disappears.

arm	val. loss	s=5	s=20	s=65	argmax
exposure, bits relative to $s = 5$					
$T = 3{,}000$, lr 6e-4 (reference)	5.988	0.000	**+0.058**	−0.077	20
$T = 12{,}000$, lr 1.08e-4 (**matched**)	5.956	0.000	−0.001	−0.016	5
$T = 12{,}000$, lr 6e-4 (unmatched)	4.595	0.000	+0.068	**+0.273**	65
depth, nats relative to $s = 5$					
$T = 3{,}000$, lr 6e-4 (reference)	5.988	0.000	**+0.064**	+0.045	20
$T = 12{,}000$, lr 1.08e-4 (**matched**)	5.956	0.000	+0.012	+0.002	20
$T = 12{,}000$, lr 6e-4 (unmatched)	4.595	0.000	−0.003	**+0.043**	65

Quadrupling training length at this arm's fixed convergence does not carry the optimum outwards. On exposure the interior structure is gone *at this arm's level*: the $s = 20$ contrast falls from +0.058 to −0.001 and the profile is monotone decreasing, so the argmax sits at the left edge. On depth the argmax stays at 20 but the penalty shrinks from +0.064 to +0.012, and that shrinkage resolves, −0.052 [−0.078, −0.024] on an unpaired bootstrap over the two arms' runs. The one cell whose optimum is out at $s = 65$ is the *unmatched* $T = 12{,}000$ cell, which is 1.36 nats better converged than the other two.

What we withdraw, and what we do not. δ is withdrawn as a measured clock. It is worth being careful about *why*. The 1.393-nat confound is *sufficient* to account for δ, and the joint-model check above shows the rate lever accounts for the observed movement, so what this arm establishes is that the fixed-rate sweep does not identify a length effect. Calling δ "a convergence exponent" would assert more than two convergence-matched points can support, and we do not assert it. The refutation of the token-invariant reading rested on δ and does not survive with it: at fixed tokens per step, an s^* that does not move with T is what a fixed-token-count description predicts, so this arm does not exclude that description, and the pessimistic consequence Appendix G.2 sets out, that the measured structure sits inside about one production batch and below the resolution at which a pipeline could place duplicates, does not follow from this arm either. The batch arm above excludes that description directly on the batch axis, without δ, which is why nothing here rests on it. We are not claiming the reverse, that s^* *is* a fixed token count: this arm is two convergence-matched points on one measure, and on depth the argmax does not move while on exposure the effect vanishes at this level, which is not the same finding twice.

The $s = 150$ cell. The claim that the matched profile is monotone decreasing would otherwise rest on three of the four points measured, and the fourth is the one that could show a distant peak. It does not. On exposure the full grid relative to $s = 5$ is 0.000, −0.001, −0.016, −0.061, monotone decreasing across all four points with the argmax at the left edge, and the $s = 150$ contrast is the only one on this arm that resolves at all, and it

resolves *downwards*, -0.061 $[-0.111, -0.016]$, $p = 0.034$. On depth it is 0.000, $+0.012$, $+0.002$, $+0.012$, so $s = 20$ and $s = 150$ are *tied* at the maximum to three decimals and calling the argmax 20 picks the lower spacing by convention rather than by measurement; $s = 150$ is itself indistinguishable from clustered, $+0.012$ $[-0.014, +0.040]$. Validation loss at $s = 150$ is 5.945, within 0.011 nats of the cell used for the match. So there is no distant peak on either measure, and the omission cost the arm's own conclusion nothing.

The two-lever objection, resolved with our own exponents. We raise below that the matched cell changes length ($4\times$) and rate ($0.18\times$) at once, and that rate moves s^*. Our own joint form $s^* \propto T^{\delta}\,\mathrm{lr}^{\lambda}$ settles it. Taking λ from Appendix C.2, positive on exposure and flat on depth, where the s^* triple $31.6/31.0/23.8/34.6$ across an $8\times$ rate range gives $\lambda \approx 0$, and starting from the reference argmax of 20:

measure	λ used	predicted at $\delta = 0$	predicted at $\delta = 1.03$	observed argmax
exposure	$+0.54$	7.9	33.0	**5**
depth	≈ 0	20.0	83.4	**20**

Both measures match $\delta \approx 0$ and neither matches $\delta \approx 1$; on depth the prediction is exact. The rate cut alone accounts for the whole observed movement, which is precisely what the two-lever objection needed to establish. On our own exponents the arm's finding is therefore stronger than a narrow negative result.

The arm has two limits of its own. Both matched cells sit at loss ≈ 5.97, far worse-converged than the primary configuration's 5.13, so this comparison is made in a regime where the effect is weak in the first place, and the exposure profile going flat is partly that weakness. And a learning-rate change is itself an intervention: it is the lever Appendix H.3 shows moves s^* with $\lambda = +0.54$, so the matched cell differs from the reference in two things at once, length up $4\times$ and rate down $5.6\times$. A cleaner design would match convergence by width or data rather than by rate, and it is the experiment we would run next on this question. What the arm establishes is that the 1.393-nat confound is sufficient to account for δ, so δ cannot be reported as a length effect; with the joint-model check above, the rate lever accounts for the movement and the length lever for none of it.

The clustered cell, and what this arm can and cannot bear on. The grid above starts at $s = 5$, so every contrast in it is an *interior* one and none of them is the paper's headline clustered-against-spaced contrast. We therefore ran the missing cell, $s = 1$ at 12 seeds, identical in every configured field to the arm's other cells. Against it, at the same 12 seeds:

vs. clustered	depth (nats)	exposure (bits)
$s = 5$	-0.0032 $[-0.0196, +0.0126]$	$+0.0276$ $[-0.0236, +0.0826]$
$s = 20$	$+0.0091$ $[-0.0090, +0.0283]$	$+0.0265$ $[-0.0356, +0.0820]$
$s = 65$	-0.0016 $[-0.0311, +0.0218]$	$+0.0113$ $[-0.0318, +0.0594]$
$s = 150$	$+0.0086$ $[-0.0121, +0.0312]$	-0.0337 $[-0.0710, +0.0038]$

Nothing resolves, on either measure, in any cell. We report that plainly, and then the reason it settles nothing either way: **this arm's leakage level is $10.8\times$ smaller than the published arm's on depth**, $+0.0404$ nats against $+0.436$, because a model at perplexity 384 has barely memorised anything at all. The published $+0.093$ contrast is not merely unresolved here; it is twice this arm's entire clustered level, so it could not appear whatever the schedule did.

Scaled by each arm's own clustered level the depth contrast is 22.5% here against 21.3% in the published arm, which is the same effect in relative terms. *On exposure it is not*: 6.2% against 59.8%. So the two measures disagree about whether anything survived, and with 12 seeds at this level neither is resolvable. **What this arm cannot do is separate "the penalty shrinks when the model is better trained" from "there is almost nothing left to shrink"**, because it holds convergence at the sweep's worst value rather than at its best. This arm does not close the convergence question in either direction. Section 4.8 does, on a ladder built for it: at fixed width and corpus the contrast is a coin flip at 20 tokens per parameter with 15,392 steps after the last copy, and resolves at that same budget with 1,539, so what the rungs moved was retention as much as budget. This appendix's arm remains what it was, a length test at fixed convergence, and is not evidence about convergence either way.

H Mechanism: the two-state model, the exclusion proof and the four clocks

H.1 The two-state model, and whether the descending branch is identified

Why the descending branch cannot be identified here, in full. Section 4.7 states the verdict for space; the conclusion, that the massed penalty is a lower bound and the fall is unclaimed, stays in the body.

Crossing count with spacing cannot identify the fall: an exactly multiplicative $k_{\text{eff}} = kf(s)$ is absorbed as $C(s) = Cf(s)^{\beta}$, invisible in principle. Fitting the decay-weighted count as a nuisance, the best τ is 10^3 to 10^4 steps and absorbs 29 to 170% of the decline. An interior maximum survives in the residual gain in 7 of 10 arm-by-measure combinations but **not in the primary arm on exposure**, its own pre-specified measure, and that tally is confounded with power. The second mechanism is absolute position: at fixed recency, moving one exposure from step 500 to 6,000 moves leakage $+0.311$ nats, which is $5.4\times$ the whole peak-to-widest decline.

These two subjects are in one appendix because they are the same subject, as the last subsection sets out.

H.1.1 Two states, and the condition for an interior optimum

This account has no held-out predictive validity and earns no weight in the paper's argument; it is here because it is the minimal escape from Proposition 1's excluded family, and because its slow timescale turns out to coincide with the nuisance fitted below.

Let a fast trace f and slow trace g decay with $\tau_f < \tau_g$, and at each exposure, after decay, $g \leftarrow g + a_g(1-f)(1-g)$ and $f \leftarrow f + a_f(1-f)$, with readout $f + g - fg$. The gate $(1-f)$ is the mechanism: you consolidate only what you have already partly forgotten, so massed copies arrive while f is saturated and consolidate nothing, while widely spread copies consolidate but g itself decays. An interior optimum needs the timescales to separate, appearing numerically for $\tau_g/\tau_f \gtrsim 10$ and gone by $\lesssim 4$. Fitted on the 324 constant-learning-rate points, $\tau_f = 383$ and $\tau_g = 3946$, a ratio of 10.3, which is *on* that boundary rather than inside it; transferred zero-parameter to held-out schedules it reaches $R^2 > 0$ in only 1 of 5 sets, so on four of five it is worse than predicting the mean. It reproduces the qualitative ordering and nothing more. We claim no novelty for its form, a standard shape in the distributed-practice literature (Appendix K).

H.1.2 Is the descending branch identified against effective exposure?

Confound (3), the decay-weighted copy count, is *partly* controlled by crossing count with spacing, and Appendix B.1 should be read with this subsection. Crossing gives the design power against a *non*-multiplicative nuisance, because $k_{\text{eff}}(k, s) = \sum_{j<k} \phi(js)$ saturates in k at an s-dependent rate and so induces a spacing-dependent count profile. It gives none against a multiplicative one, since $k_{\text{eff}} = kf(s)$ yields $C(s) = Cf(s)^{\beta}$ exactly. Our own rejection of exact separability in 7 of 10 arms (Section 4.1) is the signature such a nuisance produces, and the two facts belong together.

Fitting the nuisance explicitly, with $\phi(d) = e^{-d/\tau}$ and τ profiled, and comparing by held-out error across run folds so the extra parameter cannot flatter itself:

arm	measure	best τ	β	interior in $G(s)$?	% absorbed	held-out MSE	at $\tau = \infty$	improvement
primary	depth	4000	0.520	yes	74	0.013940	0.014325	2.69%
replication	depth	2000	0.820	**no**	120	0.016494	0.017318	4.76%
XL′	depth	2000	0.590	yes	57	0.048944	0.050284	2.67%
natural	depth	4000	0.350	yes	66	0.092700	0.093381	0.73%
wide grid	depth	10^4	0.470	yes	29	0.017905	0.017965	0.33%
primary	exposure	2000	0.830	**no**	168	0.464922	0.480591	3.26%
replication	exposure	1000	1.690	**no**	170	0.250922	0.282491	11.18%
XL′	exposure	4000	0.670	yes	52	0.715744	0.753099	4.96%
natural	exposure	4000	0.380	yes	75	0.427911	0.429340	0.33%
wide grid	exposure	10^4	0.550	yes	43	0.335911	0.335992	0.02%

The nuisance model fits better than the pure-gain model ($\tau = \infty$) in every row, so it is not a straw man, and the best-fitting τ of 1000 to 10^4 steps brackets the independently fitted $\tau_g = 3946$ of Appendix H.1.1. In 7 of 10 rows an interior maximum survives in the residual gain $G(s)$; in 3 it does not, and those include the primary arm on exposure, where the nuisance absorbs 168% of the observed decline, meaning it over-predicts the fall. Reading the same numbers the other way: between a quarter and all of the descending branch is decay-weighted copy count rather than a change in per-exposure gain.

"7 of 10" is a vote count confounded with power, and should not be read as reassurance. The last column is the reason. The nuisance improves held-out error by 1.68% on average in the seven rows where the interior maximum survives and by 6.40% in the three where it dies, a factor of nearly four; three of the survivors improve by under 0.35%, at 0.33%, 0.33% and 0.02%, which is close to no evidence for an ageing term in those arms at all. The two 36-seed arms are the highest-powered in the set and they supply all three failures. So the correct reading is not that the interior optimum usually survives but that it survives where this design has least ability to detect the very term that would remove it. That is the same objection Section 4.4's pooling answers about vote counts, and it applies here against us.

Is the over-absorption kernel mis-specification rather than sufficiency? Two exposure rows have the exponential kernel absorbing more of the decline than there is, which invites the reply that an exponential

is simply the wrong shape. We tested it, since a heavy-tailed kernel is if anything the more standard model of forgetting: a power law $\phi(d) = (1 + d/\tau)^{-p}$ with $p \in \{0.5, 1, 2\}$, fitted over the same τ sweep and the same held-out folds, choosing the winner per row by held-out error alone. The answer goes against us in a small way rather than for us. Selecting the kernel per row, the interior maximum survives in 6 of 10 rows rather than 7, because XL′ on exposure flips from surviving under the exponential ($\tau = 4000$, absorbs 52%, improvement 4.96%) to failing under a power law ($p = 1$, $\tau = 700$, absorbs 117%, improvement 5.79%); and the worst over-absorption gets worse, the replication arm on exposure going from 170% to 215%. No row in which the optimum died under the exponential is rescued by a better kernel. So the exponential was the conservative choice, and we keep the table above as the reported result while recording that the shape-selected version is slightly less favourable to us.

So the interior optimum is not identified in general and we report it as measure- and arm-dependent rather than as a property of memorisation. Second, the crossing point of Section 4.3 inherits this, so it locates where measured leakage returns below the clustered value and not where a gain function crosses.

A third consequence bears on how the practical advice should be read. Everything past the peak, including the crossing point, is a statement about *measured leakage under a nominal duplicate schedule*, and measured leakage at fixed nominal count includes the ageing of the earlier copies. So "spread rather than cluster" is advice about schedules, which is what a data pipeline actually controls, and not about a per-exposure gain. The two come apart for anyone who tries to obtain the same effect by other means: an intervention that reduced per-exposure imprinting without moving the copies apart in time would not inherit these numbers, and neither would one that changed how quickly traces age. That is a real limit on transfer, and it is the price of a result measured on schedules rather than on mechanisms.

What survives untouched is the ascending branch. Confound (3) is monotone decreasing in s at fixed k, so it cannot produce a rise from clustered to intermediate spacing; the measured massed penalty therefore *understates* the underlying gain effect. Confound (3) and the slow-trace decay of Appendix H.1.1 are the same phenomenon, which is why Section 1 crosses it with count instead of removing it: at fixed nominal count the ageing of earlier copies is constitutive of what spacing *is*, and the separable question is whether anything beyond it is needed. Our answer is that the massed penalty needs something beyond it and the descending branch largely does not.

H.2 The single-state exclusion, stated and proved

Proposition 1 (No single state with gap-independent gain)**.** *Let a scalar state $x \in [0, 1]$ update at each exposure as $x \leftarrow \alpha + (1 - \alpha)\phi(d)x$ after a gap d, and decay as $x \leftarrow x\phi(\Delta)$ at readout, with encoding gain $\alpha \in (0, 1]$* ***constant and independent of*** *d and $\phi : [0, \infty) \to (0, 1]$ any non-increasing kernel with $\phi(0) = 1$. At fixed count and last-exposure time the readout is non-increasing in s, so no interior optimum exists.*

With α independent of the gap, the update is affine in x with slope $(1-\alpha)\phi(d) \geq 0$ and intercept α. Composing k exposures at uniform gap s with the terminal decay,

$$x_{\text{final}}(s) \;=\; \phi(\Delta)\,\alpha \sum_{j=0}^{k-1} \big[(1-\alpha)\phi(s)\big]^j .$$

Every term is a non-negative power of $(1 - \alpha)\phi(s)$, which is non-increasing in s because ϕ is; a sum of non-increasing non-negative terms is non-increasing. Hence x_{final}, and so the readout, is maximised as $s \to 0$: massed is optimal and no interior optimum can exist. □

We verified this numerically over 60 configurations, 5 gains × 4 timescales × 3 kernels including heavy-tailed and non-smooth kernels, with zero violations. The step that fails when $\alpha = \alpha(d)$ is the first: the intercept then depends on s, the composition is no longer a geometric sum in a single non-increasing quantity, and the monotonicity argument does not apply.

H.3 The four clocks, measured

Every candidate here is excluded or withdrawn, so this section supports no claim of its own; it is kept in full because the exclusions are what make Proposition 1's scope narrow rather than vacuous.

The four candidates, and what each predicted. Any model producing an interior optimum does so with a clock. We tested four, three of them pre-registered in code before the runs.

clock	prediction	measured	verdict
accumulated lr	$s^* \propto 1/\text{lr}$	s^* *rises* over 8×; slope excludes -1 and 0	**excluded**
interference volume	falls with B, token-invariant	$\gamma = 0.23$ $[0.05, 0.38]$ over $16 \leq B \leq 256$; 1 excluded	**excluded**
AdamW 2nd moment	$s^* = 1/(1-\beta_2) = 20$ here	500× timescale, argmax only $20 \to 80$	**excluded** (prop.)
one needed at all?	removing it kills the effect	β_2=0 diverges 12/12; SGD will not converge	**open**
training length	$s^* \propto T^\delta$, $\delta > 0$	1.03 at fixed rate; ≈ 0 at matched convergence	**withdrawn**

Whether a second moment is needed at all is open, since no optimiser without one converges here. The grid now reaches $s = 1$ (Appendix H.4), and none of the four sets the optimum. In their place the batch axis leaves a candidate rather than a clock. Restoring the share of a step a duplicate occupies raises the optimum at both batch sizes we can test, $1.7\times$ and $4.1\times$ at fixed volume. Neither shift resolves against zero alone.

Two scopings. Two of the three exclusions are *sign-level*, resting on the direction s^* moves rather than on a fitted timescale; only the third is measured over four octaves, and on a non-monotone series. The remaining exponent rests on one adjacent pair against a $1.7\times$ noise floor, so it too is quoted as a *sign*. And four is the number we tested, not the number available: the retention interval is fixed in every grid arm, forced by the pinned last exposure.

Withdrawing δ reopened whether the structure is reachable at production batch size, and the batch exclusion answers it on the widest arm alone. Since s^* falls more slowly than $1/B$, the optimum in tokens *grows* as $B^{0.77}$: 0.69 of a 2×10^6-token production batch at B=256, crossing one at $B \approx 420$ $[241, 660]$. So the optimum spans more than one batch, on an extrapolation of under one octave.

H.4 Is the optimum AdamW's second-moment timescale?

The coincidence, which is exact, and why it matters. Every arm here trains with AdamW at $\beta = (0.9, 0.95)$, whose second-moment averaging timescale is $1/(1-\beta_2) = 20$ steps; the measured grid argmax is $s^* = 20$ steps. The agreement is exact, and there is a mechanism behind it rather than a numerology: at gaps short relative to the averaging window, consecutive steps carry nearly identical canary gradients, so v accumulates along those directions and the effective per-copy step is damped, while at long gaps v has decayed between exposures and each copy imprints at full size. That predicts the *sign* of the massed penalty and a per-copy gain rising with s and saturating, which is the ascending branch.

The coincidence cannot manufacture the effect, since the canaries are memorised less when clustered whatever the optimiser does. What is at risk is *portability*: if s^* tracks $1/(1-\beta_2)$ then the ascending branch is a property of one optimiser's hyperparameters and no s^* here transfers to a run at a different β_2. The staggered arm of Appendix B.2 cannot address it, because within-canary second-moment accumulation depends only on s, which is the treatment.

The arm, and the massed cell it now reaches. This grid's lowest cell was $s = 5$, so it had no *massed* cell in Table 1's sense, and the mechanism that predicts the sign of the massed penalty was the one mechanism here never tested against the spacing the headline rests on. **That cell has since been run, crossed with β_1, and the account survives its first direct test.** On this sweep's own configuration, not the primary arm's, so counts $\{2, 4, 8, 16, 32\}$ at cpg $= 5$ and 310 injections per run, which is why the base cell reads $+0.602$ and not Section 4.2's $+0.733$. Twelve seeds, $s = 20$ against $s = 1$:

β_1	β_2	exposure (bits)		depth (nats)	
0.9	0.95	$\mathbf{+0.602}$	$[+0.459, +0.749]$	$\mathbf{+0.102}$	$[+0.062, +0.140]$
0.9	0.999	$\mathbf{+0.305}$	$[+0.188, +0.406]$	$+0.045$	$[-0.003, +0.094]$
0	0.95	$\mathbf{+0.625}$	$[+0.420, +0.863]$	$\mathbf{+0.097}$	$[+0.037, +0.155]$
0	0.999	$\mathbf{+0.316}$	$[+0.143, +0.517]$	$+0.042$	$[-0.012, +0.101]$

The first moment is not the mechanism and the second moment is. Removing β_1 entirely leaves the penalty where it was, $+0.602 \to +0.625$ and $+0.305 \to +0.316$ on exposure, while lengthening the second-moment window from 20 steps to 1,000, past the spacing being compared, halves it at either β_1. That is the direction Appendix H.3 predicted before these cells existed, and it is the first test of it at $s = 1$. Exposure resolves in all four cells; depth resolves at $\beta_2 = 0.95$ and not at $\beta_2 = 0.999$, where the halved effect sits under what 12 seeds detect. **Both differences are tested paired, because non-overlapping intervals are not a test of a difference** (Appendix B.2). The four cells share seeds 0 to 11, so each difference is itself paired. The β_2 difference is $+0.298$ $[+0.102, +0.525]$ bits at $\beta_1 = 0.9$, $p = 0.021$, and $+0.309$ $[-0.016, +0.634]$ at $\beta_1 = 0$, the same size and unresolved at 12 seeds. The β_1 difference is -0.022 $[-0.275, +0.230]$ at $\beta_2 = 0.95$ and -0.011 $[-0.201, +0.175]$ at $\beta_2 = 0.999$: null twice, tested rather than asserted.

Convergence is a poor account of the halving, though this is suggestive and not a matched test. Loss runs 5.129, 5.262, 5.308 and 5.539 down the table, and Section 4.4 finds less converged arms leak less in absolute terms, so the objection is fair to raise. It sits badly here because *the larger loss gap is the one that moves nothing*: $\beta_1 = 0$ at $\beta_2 = 0.95$ sits $+0.179$ nats worse than the base cell and changes the penalty by $+4\%$, while $\beta_2 = 0.999$ sits only $+0.133$ nats worse and halves it. **We do not call this identified**, because the two gaps come from different interventions and a level account could depend on *how* the loss degrades and not only how much; the matched test is the $\beta_2 = 0.999$ pair rerun at a rate tuned to the base cell's loss, **and we have since run it**. Normalising each penalty by its own cell's massed level leaves the same split as a ratio of means, 0.467 and 0.640 at $\beta_2 = 0.95$ against 0.264 and 0.296 at $\beta_2 = 0.999$, over massed levels of 1.291, 0.976, 1.153 and 1.066 bits in the table's order, which are the denominators so that these four numbers can be recomputed.

The matched test, and it separates the two measures. The rate was chosen on loss alone and before any contrast was read: lr scaled by 1.25, 1.50 and 2.00 at $\beta_2 = 0.999$, 3 seeds each, giving 5.116, 4.979 and 4.800 against the base cell's 5.129. The first was taken, and at 12 seeds per spacing it lands at 5.105, *below* the base cell, so the matched cell is if anything the better converged of the two. Section 4.4 finds better converged arms leak more in absolute terms, so a convergence account predicts the penalty returns at or above $+0.602$.

cell	loss	exposure (bits)		depth (nats)	
base, $\beta_2 = 0.95$	5.129	$+\mathbf{0.602}$	$[+0.459, +0.749]$	$+\mathbf{0.102}$	$[+0.062, +0.140]$
$\beta_2 = 0.999$, rate untuned	5.262	$+\mathbf{0.305}$	$[+0.188, +0.406]$	$+0.045$	$[-0.003, +0.094]$
$\beta_2 = 0.999$, loss matched	5.105	$+\mathbf{0.272}$	$[+0.118, +0.427]$	$+\mathbf{0.087}$	$[+0.030, +0.157]$
paired, base minus matched		$+\mathbf{0.330}$	$[+0.197, +0.482]$	$+0.014$	$[-0.072, +0.083]$

On exposure the halving is not convergence. It survives matching at $+0.272$, and the paired base-minus-matched difference is $+0.330$ $[+0.197, +0.482]$ over the shared 12 seeds, at $p = 4.9 \times 10^{-4\dagger}$, which is the n=12 floor and so says only that all twelve seeds agreed in sign; the interval is what carries the size. The prediction a convergence account made is refuted at the one place it could be tested, so on exposure the second-moment window is what modulates the penalty. **On depth the apparent halving *was* convergence, and we withdraw that half of the reading.** Matching restores depth to $+0.087$ $[+0.030, +0.157]$ against the base's $+0.102$, a paired difference of $+0.014$ $[-0.072, +0.083]$ that does not resolve: the $+0.045$ at the untuned rate was the less converged optimiser and not the longer window. So the modulation claim holds on exposure, the pre-specified measure for canaries, and not on depth, and we report it at that scope rather than as a single effect on both.

One scoping on the β_1 null: it is a null in absolute units. On the normalised scale the point estimate rises from 0.467 to 0.640 when momentum is removed, which is $+37\%$ and would read as an effect. It does not resolve: the paired normalised difference is -0.347 $[-0.911, +0.051]$ at $\beta_2 = 0.95$ and -0.636 $[-1.872, +0.093]$ at $\beta_2 = 0.999$, neither excluding zero. **Those two figures are on a different statistic from the $+37\%$ they test**, being means of per-run ratios rather than ratios of means, which is why they are larger; we report them because only the per-run form can be paired, and the per-run ratio is unstable here because a run's own massed level can approach zero. So the honest statement is that removing momentum does not *reduce* the penalty on either scale, which is what the mechanism claim needs, and not that momentum provably does nothing. The only probe that removes the second moment outright ($\beta_2 = 0$) diverges at $12/12$ seeds. Six values of β_2 on the learning-rate arms' common grid $\{5, 10, 20, 40, 80, 120\}$, 12 seeds each, counts $\{2, 4, 8, 16, 32\}$ at cpg 5, everything else byte-identical: $\beta_2 \in \{0, 0.5, 0.90, 0.95, 0.99, 0.999\}$, averaging timescales 1, 2, 10, 20, 100 and 1000 steps, and the proportional prediction is $s^* = 1, 2, 10, 20, 100, 1000$. That is a $1000\times$ range as designed, against the $8\times$ range the learning-rate arm managed. As run it is a $500\times$ range over five cells, because the $\beta_2 = 0$ cell diverges at every seed and is reported as a divergence rather than as a measurement.

What is censored, and in which direction. The grid runs $s \in [5, 120]$, so three predictions fall outside it: $\beta_2 = 0.999$ predicts 1000 and is pinned at the right edge, while the two low cells predict $s^* = 2$ and 1 and are pinned at the left. Every one of those censorings pulls the fitted slope $r = \mathrm{d}\log_2 s^* / \mathrm{d}\log_2(1/(1-\beta_2))$ *towards* 0, which is the outcome that excludes the clock, so every censored cell is censored in the direction that suits us. That is why the slope is not the statistic we rest on; the primary statistic is the **grid argmax**, which censoring cannot manufacture (see below).

Two design points. Retuning β_2 is *convergence-neutral*, which this comparison needs and the SGD comparison below cannot have; we check it at the end of training rather than the start, since a 1000-step averaging window is a sixth of the run, and the per-cell losses are in the table below. And $\beta_2 = 0.95$ is re-run here rather than borrowed from the learning-rate arm, because two calibration runs did not reproduce that arm's per-set values to the precision with which it reproduces itself (control NLL 4.806 against 4.664 and 4.682) and we could not localise the difference; every cell therefore comes from one script and the other arm is a cross-check. That cross-check matters: at the same $\beta_2 = 0.95$ it gives argmax 40, $s^* = 36.5$ [26.0, 58.8] and a massed penalty $+0.440$ against the swept cell's 20, 24.4 and $+0.539$. Cross-script variation is therefore as large as everything the $500\times$ sweep produces, which is both the strongest statement that the argmax is insensitive to β_2 and the reason no single cell should be read alone.

Which numbers this reaches, since depth is defined against a control baseline. Both discrepant runs are the 310-injections-per-run configuration on the grid $\{5, \ldots, 120\}$: this sweep's $\beta_2 = 0.95$ cell and the learning-rate arm's $\times 1$ cell. Every row of Table 4, every arm in Figure 2 and every number in Appendix E.1 is a 504-injection arm instead, without exception (Appendix A.3 tabulates the density arm by arm), and Appendix C.5 shows the control baseline is a strong function of injection load, moving 0.353 nats when canaries per set halve. So this is not a baseline any headline depth number is measured against. What the discrepancy does reach is this sweep's per-cell magnitudes, and not merely as a common offset that differencing would cancel: the paired penalty moves $+0.539$ to $+0.440$, so 0.099 bits of it survives. That is why the verdict here rests on the argmax

and not on a magnitude. The direction is not in our favour either. The clock's single hit is $\beta_2 = 0.95$, and that is exactly the cell the second implementation puts at argmax 40, a miss.

Figure 10 and Figure 9 plot the learning-rate and length arms this sweep is compared against.

The rule mapping a predicted timescale to a predicted argmax, since it is a choice. The clock predicts a value of s^* in steps; the grid predicts an argmax. We map the prediction to the nearest grid point in log space, which gives $10 \to 10$, $20 \to 20$, $100 \to 120$, $1000 \to 120$ (capped) and $2 \to 5$ (floored). One of those deserves naming: a timescale of 100 is almost exactly equidistant in log between grid points 80 and 120, and we round to 120, which reports that cell's miss as $3\times$ rather than $2\times$ and so *enlarges* our own margin. Rounding the other way leaves the cell a miss either way and changes no verdict, but in a table otherwise careful about the direction of every bias the convention should not be silent.

The pre-registered reading, fixed before the runs. Let r be the fitted slope of $\log_2 s^*$ on $\log_2$ of the averaging timescale. r indistinguishable from $+1$ means the second-moment clock is the clock and this paper's scope narrows to AdamW at $\beta_2 = 0.95$; r indistinguishable from 0 means a third clock is excluded and the ascending branch is a property of the data and schedule rather than of the optimiser state; anything between is reported as the interval, claiming only the direction. Independently, if the massed penalty appears with no second moment at all then that route is not necessary, whatever r turns out to be, and that conclusion does not depend on the fit.

Result: the clock is excluded, on both measures, and not by way of the censored cell.

β_2	timescale (steps)	val. loss at T	predicted argmax	observed argmax	interpolated s^*	massed penalty (exposure)
0	1	**diverged at all** 12 **seeds**: non-finite validation loss; see below				
0.5	2	5.148	5	**80**	77.7 [31.0, 120]	$+0.273$, $p = 4.3 \times 10^{-2}$
0.90	10	5.158	10	**40**	37.8 [25.2, 65.0]	$+0.393$, $p = 2.4 \times 10^{-3}$
0.95	20	5.129	20	**20**	24.4 [21.6, 61.6]	$+0.539$, $p = 4.9 \times 10^{-4\dagger}$
0.99	100	5.285	120	**40**	38.8 [33.1, 46.5]	$+0.473$, $p = 4.9 \times 10^{-4\dagger}$
0.999	1000	5.265	120	**40**	28.8 [18.0, 43.0]	$+0.290$, $p = 2.4 \times 10^{-3}$

Over the five cells that train, spanning a $500\times$ range in the timescale, the fitted slope is $r = -0.120$ $[-0.244, +0.020]$ on exposure and -0.112 $[-0.216, +0.330]$ on depth. Both **exclude** the proportional clock at $r = +1$, and neither excludes $r = 0$. So s^* does not track $1/(1-\beta_2)$: the location of the optimum is not set by AdamW's second-moment timescale, s^* stays a property of the data and the schedule rather than of one optimiser hyperparameter, and the numbers in this paper are not scoped to $\beta_2 = 0.95$. That was the pre-registered $r \approx 0$ outcome and it is the one we report, *for the proportionality question*.

One thing the interval on r is not is a measure of precision. The bootstrap resamples runs within cells; it does not resample *implementations*, and the cross-script comparison above puts the displacement at fixed β_2 at $\log_2(36.5/24.4) = 0.58$ units against a total slope-driven movement across the sweep of $0.120 \times \log_2(500) = 1.08$ units. A source of variation the interval does not represent is therefore about half of everything the fit describes, so $[-0.244, +0.020]$ locates the sign and the distance from $+1$ rather than bounding small positive r. The exclusion is untouched by this: $r = +1$ would require 8.97 units of movement against the 1.08 observed. The separate necessity question is treated at the end of this appendix, and it is not settled.

The $\beta_2 = 0$ cell does not exist as a measurement, and we report why rather than reporting its numbers. All 12 seeds diverge to a non-finite validation loss. That is mechanical rather than unlucky, and not for the obvious reason: $\beta_2 = 0$ makes $v_t = g_t^2$ exactly, so the step is $m_t/(|g_t| + \varepsilon)$. That is *not* a sign step, because $\beta_1 = 0.9$ leaves the numerator an average over past gradients rather than g_t itself. A smoothed numerator divided by an unsmoothed magnitude explodes whenever $|g_t|$ falls below the recent average, and a fixed 6×10^{-4} rate cannot absorb that. The cell is dangerous in the analysis rather than merely absent: with non-finite logits no alternative is separable from the secret, so *sampled exposure saturates at its ceiling* of $\log_2(2N) = 9.0$ bits in every set at every spacing, and the cell reports a massed penalty of exactly $9.0 - 9.0 = 0$ with its argmax pinned to the first grid point. That zero is the absence of a measurement, not a measured absence of an effect, and its (short timescale, small s^*) coordinate pulls the fitted slope *towards* the proportional clock: including it moves r from -0.120 to $+0.099$ $[+0.019, +0.189]$, which would exclude $r = 0$ and manufacture a weak positive clock out of a cell that never trained. We therefore drop it from every fit and say so here, because a reader recomputing our slope from the table would otherwise get the wrong number for the right reason.

Why this does not rest on the parts of the test that could flatter it. First, the observed **grid argmax spans** 20 **to** 80, a factor of 4, while the clock's timescale spans $500\times$ across the same cells; and an argmax cannot be pulled down by interpolation clipping, since a true optimum at or beyond $s = 120$ would appear as a monotone increasing profile with its argmax *at* the edge, whereas every cell shows an interior argmax with a fall after it. Second, the clock's prediction is *resolvable* in three of the five trainable cells, since 10, 20 and 100 lie inside the

grid range $[5, 120]$, and it misses in two of the three: by $4\times$ at $\beta_2 = 0.90$, the cell furthest from any edge, and by $3\times$ at $\beta_2 = 0.99$. Restricting to those three leaves $r = +0.060$ on exposure over a $10\times$ range, still far from $+1$. Third, the largest single miss is in the newest cell and it is not a censoring artefact in our favour: at $\beta_2 = 0.5$ the clock predicts the grid's left edge and the observed argmax is 80, a $16\times$ miss, in a cell whose validation loss (5.148) is among the best converged in the arm.

Four things reported against the clean version of this story. The clock's prediction *is* hit at $\beta_2 = 0.95$, where predicted and observed argmax are both 20; we do not read that as support, because 20 is a value the other cells take anyway (on depth the five trainable argmaxes are 80, 20, 20, 40, 20, so $\beta_2 = 0.999$ sits at 20 while the clock predicts 1000). One hit in five says the argmax is insensitive to β_2, not that the clock works at 0.95. Second, the massed penalty resolves at every trainable β_2, which is the result, but its magnitude is not constant: $+0.273$, $+0.393$, $+0.539$, $+0.473$, $+0.290$, a $2.0\times$ spread rising then falling, weakest at $\beta_2 = 0.5$ which only just resolves ($p = 0.043$; 0.077 on depth). We read no trend into that and claim no exact invariance. Third, convergence is close but not exact, validation loss spanning 5.129 to 5.285; that is an order below the length arm's 1.393 nats and it does not order s^*. Fourth, on depth the $\beta_2 = 0.90$ penalty does not resolve ($+0.024$, $p = 0.15$) and the depth slope's interval $[-0.216, +0.330]$ excludes $+1$ without being informative about much else.

Is a second moment *necessary*? Not settled. This is a different question from the one above, and the distinction matters. The sweep establishes that s^* does not *scale with* the averaging timescale. It does not establish that the massed penalty would survive with no averaging at all, because a dependence could be a threshold rather than a proportionality. Our pre-registration treated the two separately, and we owe a separate answer.

We attempted it twice and both attempts failed, for opposite reasons. $\beta_2 = 0$ is the clean form of the control, since $v_t = g_t^2$ removes the averaging entirely while keeping Adam's per-step normalisation, and it diverges at every seed as reported above. SGD with momentum removes the second moment differently and cannot be brought near AdamW's loss, as reported below. So the two designs that remove averaging altogether are respectively unstable and unconverged, and we have no measurement at zero averaging.

What we do have is the $\beta_2 = 0.5$ cell, whose averaging window is 2 steps against the reference configuration's 20. That is a $10\times$ reduction and close to the shortest window that trains at all, and the massed penalty is still present and resolved there ($+0.273$, $p = 0.043$; $+0.084$, $p = 0.077$ on depth), at a validation loss matching the rest of the arm. So the honest statement is that the effect survives shortening the averaging window by an order of magnitude, and that whether it survives removing the window entirely is untested. We do not claim the second moment is unnecessary, and Appendix H.3 states the exclusion as being of the proportional clock specifically.

We also ran SGD with momentum, and it is *supplementary* because it cannot be convergence-matched in this budget. Over five probed rates spanning two decades the best reaches validation loss 7.732 and the loss gets steadily *worse* at higher rates (7.755, 7.914, 8.783, 9.106), so this is not a tuning failure at the top of the range; the best rate is however the *smallest* probed, so a better SGD at some still lower rate is not excluded, which only reinforces treating this arm as supplementary. Over 12 seeds the arm sits at 7.718 against AdamW's 5.118. A 2.6-nat gap is perplexity about 2,250 against 167, larger than any convergence mismatch this paper criticises elsewhere, and at that loss level a null on either spacing condition would be uninformative rather than an answer.

That is exactly what happened, and we report it as uninformative rather than as a result. Exposure sits *below* the never-injected control at every spacing (-0.169, -0.165, -0.164 at $s = 5$, 20, 80), so at this loss level the canaries are not learned above control at all and there is no leakage for spacing to modulate; the best interior contrast is $+0.005$ with $p = 0.47$. A null measured where the quantity being modulated is itself absent says nothing about whether the massed penalty needs a second moment. The $\beta_2 = 0$ cell was meant to be the control that answers this. Since it diverges, neither route is available and the question is open rather than answered.

How much does a run at a fixed seed vary? Three repeats of one cell, same script and seed, differing only in bf16 non-determinism, agree on validation loss to 1.7×10^{-3} nats but spread the *per-set* depth by 0.030, 0.075, 0.164, 0.185 and 0.228 nats across the five counts. The largest is twice the depth massed penalty, which is why Section 3 reports depth over 12 seeds and five counts and why depth contrasts resolve less often than exposure ones (Appendix A.2 gives the two-run comparison).

I MEMBERSHIP INFERENCE, PER SPACING AND PER COUNT

The canary-grid arm, and the two sources of uncertainty in it. On the uniform grid the same attack uses the never-injected control set as the membership counterpart and 16 reference models that are *not* seed-disjoint from the targets, where Section 4.3's arm has 32 that are, so this is the weaker arm and the body reports the other one. Calibrated within-run TPR at 1% FPR is $0.101/0.170/\mathbf{0.193}/0.156$ at $s = 1/8/20/65$. Propagating the ensemble's own uncertainty splits the two contrasts. The massed penalty survives: $+0.092$, $[+0.061, +0.123]$

with the 16 references held fixed and $[+0.030, +0.107]$ once they are resampled. The interior contrast **no longer excludes zero**, at $[-0.001, +0.076]$. Risk concentrates at high count, per cell below.

I.1 THE SAME ATTACK IN A MORE MEMORISED REGIME

This appendix runs the calibrated attack at the primary configuration, the grid arm, where the regime is weak: perplexity about 169 and under 1% probability on the correct digit at $k = 64$. The objection is fair, and the useful answer is our own data rather than an extrapolation. So we re-ran the attack on the high-memorisation arm: a 2M-token subset for the same 6,000 steps, so about 25 epochs, 6-digit secrets, counts to 128, 512 non-member canaries per run and 16 dedicated never-injected reference models. Mean member exposure is 4.74 bits against the primary arm's low single figures.

spacing	runs	TPR at 1% FPR	with the reference ensemble resampled
$s = 1$ (clustered)	12	0.239 [0.206, 0.275]	[0.142, 0.256]
$s = 20$ (worst)	12	**0.392** [0.336, 0.447]	[0.219, 0.394]
massed penalty		**+0.153** [+0.083, +0.219]	[+0.016, +0.217]

Reference resampling does not return an interval centred on the point estimate, here or in the grid arm above: it replaces 16 fixed reference models by a resampled ensemble, and the rows above shift down rather than widen symmetrically, $[+0.083, +0.219]$ becoming $[+0.016, +0.217]$. The grid arm above's $+0.092$ $[+0.030, +0.107]$ is the same asymmetry and not a transcription: 0.092 is exactly $0.193 - 0.101$ from the series above it. The massed penalty on the attack **replicates, and is larger here**: $+0.153$ against the grid arm's $+0.092$, about $1.7\times$, and it survives propagating reference-ensemble uncertainty, which is the test that cost us the interior claim in the grid arm above. Risk concentrates at high count as elsewhere, the per-cell penalty running $+0.014$, $+0.167$, $+0.181$, $+0.278$, $+0.125$ at $k = 8, 16, 32, 64, 128$.

The limits of this arm, one of which matters. It does not answer the regime objection on the axis a reader may care most about. Validation perplexity here is 456, *worse* than the primary arm's 169, because training 25 epochs on a 2M-token subset buys memorisation by overfitting rather than by better modelling. So what we have shown is that the massed penalty strengthens as memorisation rises, not that it survives into a better-converged model; a reader who takes "weak regime" to mean "low model quality" is not answered by this arm, and we do not claim otherwise. Second, the reference-resampled interval only just excludes zero at its lower end, $+0.016$, so this is a resolved result and not a comfortable one. Third, the per-count profile is non-monotone at the top, $k = 128$ giving a smaller penalty than $k = 64$, which is the same compression Appendix E.2 reports at the highest count and which we still do not explain.

Two attacks on the same runs: an uncalibrated loss threshold on the secret NLL, and the one-sided offline LiRA score

$$\lambda(x) = \big[\mu_{\text{out}}(x) - (\text{nll}(x) - \ell)\big]/\sigma_{\text{out}}(x),$$

where μ_{out} and σ_{out} come from 16 reference models trained on the identical corpus with *zero* injections, so every canary is a non-member in every one of them. Each run's overall loss level ℓ is removed before comparing: the reference runs by their own canary mean and the target run by its never-injected control set. That centring is not cosmetic. Measured raw, the per-canary spread across reference models is 0.479 nats against a difficulty spread across canaries of 0.401, so without it the statistic would divide by more noise than the nuisance it is meant to remove.

Both attacks use 24 members and 512 non-members per run over 24 runs per spacing. AUC and TPR are computed *within* a run and then averaged, so the unit of analysis is the run; pooling canaries across runs would repeat the pseudoreplication corrected elsewhere.

attack	s	AUC	TPR at 1% FPR	TPR at 5% FPR
loss threshold	1	0.690 [0.673,0.707]	0.066 [0.047,0.087]	0.172 [0.142,0.201]
	8	0.738 [0.720,0.755]	0.125 [0.094,0.158]	0.253 [0.219,0.288]
	20	**0.789** [0.776,0.802]	**0.172** [0.139,0.207]	**0.325** [0.288,0.361]
	65	0.772 [0.751,0.792]	0.109 [0.082,0.137]	0.307 [0.273,0.340]
offline LiRA	1	0.753 [0.737,0.769]	0.101 [0.076,0.127]	0.278 [0.243,0.314]
	8	0.789 [0.775,0.804]	0.170 [0.141,0.201]	0.349 [0.314,0.385]
	20	**0.818** [0.805,0.832]	**0.193** [0.168,0.219]	**0.405** [0.370,0.439]
	65	0.804 [0.788,0.821]	0.156 [0.134,0.177]	0.347 [0.321,0.372]

I.2 CONSTRUCTION, CALIBRATION AND WHICH STATISTIC GOVERNS

Do the two uid ranges differ intrinsically? Yes, slightly, and in our favour. Members carry uid $= 10000 + 100g + c$ and non-members $90000 + c$, so if the two ranges differ systematically in how hard the secret is to predict given the prefix, the *uncalibrated* threshold is confounded; the calibrated attack is protected because μ_{out} is per canary and absorbs any such difference. The reference models are the clean test, because they inject nothing, so every canary in them is a non-member whatever its uid and any difference is pure difficulty. Over 16 reference models, level-centred, the member range sits at $+0.0641$ nats and the non-member range at -0.0030, a difference of $+0.067$ nats with Mann–Whitney $p = 0.028$: real, and 16.7% of the 0.401-nat difficulty spread across canaries. Its sign is the conservative one. The attack flags *low* loss as membership, and the member range is intrinsically *harder*, so the bias works against the attack rather than inflating it, and the uncalibrated figures are a lower bound for this reason as well as for the one below. It also means the 384-against-8192 imbalance in that comparison is not the explanation, since the effect has the wrong sign to be one.

Calibration raises measured risk, so a loss threshold is a lower bound. At every spacing LiRA beats the loss threshold, and by most at the low-FPR end that matters for privacy: TPR at 1% FPR rises from 0.066 to 0.101 at $s = 1$ and from 0.172 to 0.193 at $s = 20$. A loss threshold alone is the weakest standard attack, and its numbers are a genuine lower bound, now measured rather than asserted.

What 512 non-members buys, and what it does not. The 1% point now falls at the 5.12th order statistic and is estimated rather than interpolated off the end of the sample, which 128 non-members could not do. A 0.1% point is still not expressible: it lies below the first order statistic, and reaching it would need at least 1,000 non-members per run. We report that as a `nan` in the analysis output rather than an interpolated number.

Which statistic governs, where they disagree. It is the **Wilcoxon**: it is the test used for every other contrast in this paper, it is exact at these n, and it assumes nothing about the shape of the paired differences. The bootstrap interval is reported for magnitude, not significance. Both are printed by the analysis so a disagreement is visible rather than resolved silently.

Run-level paired contrasts against the peak at $s = 20$, Wilcoxon p:

attack	statistic	vs. massed	p	vs. widest	p
loss threshold	AUC	$+0.099$	$1.2 \times 10^{-7\dagger}$	$+0.017$	0.049
	TPR at 1% FPR	$+0.106$	4.4×10^{-5}	$+0.063$	0.0067
	TPR at 5% FPR	$+0.153$	2.7×10^{-5}	$+0.017$	0.39
offline LiRA	AUC	$+0.065$	$1.2 \times 10^{-7\dagger}$	$+0.014$	0.043
	TPR at 1% FPR	$+0.092$	1.5×10^{-4}	$+0.037$	0.022
	TPR at 5% FPR	$+0.127$	8.7×10^{-5}	$+0.057$	0.013

Under calibration both interior contrasts resolve on all three statistics; uncalibrated, on two of three. On a smaller arm the uncalibrated count was one of three, so the gain here comes from a stronger attack rather than from a different design. The $n = 24$ Wilcoxon floor is 1.2×10^{-7}, so the AUC-versus-massed rows sit at it† and mean "as small as 24 seeds can show".

Every interval above is conditional on the 16 reference models, and one verdict does not survive dropping that condition. The bootstrap resampled target runs while holding the ensemble fixed, so it propagated sampling error in the targets and none in the calibration. Resampling the reference models with replacement alongside the runs, recomputing μ_{out} and σ_{out} per canary from each resampled ensemble and rescoring every target with the *same* estimator as the table above, widens the *absolute* TPR intervals by $1.73\times$ to $2.55\times$, which is what 16 reference models should be expected to cost.

The contrasts fare better, because the same ensemble scores every spacing so its error is largely common-mode and cancels in a paired difference. That was a prediction and it is only half right: the contrasts widen by $1.22\times$ rather than by 1.7 to $2.6\times$, and one of them still crosses zero.

TPR at 1% FPR	estimate	runs resampled	runs *and* references resampled
vs. massed	$+0.092$	$[+0.061, +0.123]$	$[+0.030, +0.107]$, $1.22\times$ wider
vs. widest	$+0.037$	$[+0.005, +0.068]$	$[\mathbf{-0.001}, +0.076]$, **includes zero**

So the massed penalty on the calibrated attack survives propagating calibration uncertainty and **the interior contrast does not**. We report that as a failure of the operational interior claim, not as a caveat on it: on the reference-fixed statistics both contrasts resolve, and on this treatment the vs.-widest one does not. It is

the third independent reason the interior optimum is reported as measure- and arm-dependent, alongside the decay-weighted-copy-count control (Appendix H.1) and the collapsed pooled interval (Appendix C.4), and all three point the same way.

The recomputation uses the *same* descending-order-statistic interpolation as every other TPR in this paper, so it reproduces the levels above exactly, $0.101/0.170/0.193/0.156$, and the contrast estimates are the grid arm's $+0.092$ and $+0.037$ rather than second versions of them: re-scoring under a different quantile convention moves the levels by up to 0.017, which is enough to leave one estimand with two values, so the convention is fixed once here. And 300 ensemble resamples of 16 models is itself coarse, which can only understate the widening.

TPR at 1% FPR by duplicate count, members of that count only, offline LiRA:

s	k=2	k=4	k=8	k=16	k=32	k=64
1	0.031	0.010	0.062	0.073	0.115	0.312
8	0.042	0.042	0.094	0.062	0.281	0.500
20	0.042	0.031	0.073	0.083	0.312	**0.615**
65	0.052	0.062	0.062	0.115	0.281	0.365

Each cell rests on 4 members per run, so every entry is a multiple of $1/96$ and the per-count rows are coarse; the $k = 64$ row is the one carrying the high-count claim and it is the coarsest place to carry it. At $k \leq 8$ membership is barely inferable at any spacing even under calibration, so the spacing effect on operational risk is a high-count phenomenon, matching Section 4.6. Two limits, stated explicitly. The reference seeds are 0 to 15 and the target seeds 0 to 23, so for sixteen targets one reference model shares its data order; that is a dependence and not a leak, since no reference model saw any canary. It does *not* act on σ_{out}, which is computed across the sixteen reference models alone: what it touches is μ_{out}, by correlating those targets' losses with one of the models entering it. We therefore report it as a dependence without claiming a direction, rather than calling it conservative on the strength of the wrong term. Disjoint seed ranges would be cleaner and are what a further version should use. Section 4.3's arm is built that way, target seeds 0 to 47 against reference seeds 100 to 131, and the massed penalty reproduces there on the same attack, $+0.078$ $[+0.060, +0.098]$ paired over 48 seeds, $p = 1.4 \times 10^{-7}$ (Appendix F.4), so the overlap is this arm's construction rather than the attack's. And the descending branch of the MIA curve is subject to the same identification problem as the descending branch of the leakage curve (Appendix H.1), so the $s = 20$ versus $s = 65$ contrast should not be read as independent evidence about per-exposure gain.

J Claim status: what we withdrew, and what remains limited

This appendix is organised around the paper's claims: the sub-claims the main-text table has no room for, then every reading we withdrew or narrowed, then the limitations in full.

J.1 Every final claim, its status and where it is supported

The authoritative claim-status table is Table 2, in the main text, and this appendix deliberately does not reproduce it: a duplicated registry is a registry that will drift.

Four sub-claims that table has no room for, each graded the same way. *Batch composition is bounded, not excluded, as the source of the massed penalty* (pre-specified; holds on both measures in two configurations, Appendix B.2 Results 2–3). *It is a lower bound rather than a point estimate* (pre-specified; by the monotonicity of the remaining confounds, Section 4.7, and now also by the censoring measured in Appendix E.2). *The separable law $C(s)k^{\beta}$ is an approximation, not a law* (pre-specified; M_1 holds in $10/10$ arms, M_2 is rejected in $7/10$, Section 4.1). *Proposition 1 excludes a gap-independent-gain family* (not applicable; proved, Appendix H.2, and the family it excludes is narrow, since the readout is a geometric sum in a non-increasing quantity, so the proposition is a sanity check on the state class and not a result we lean on).

"Pre-specified" throughout means the reading criterion was written into the run script's header before the runs. Where a row says "post hoc" the arm was run in response to a question about the design rather than to a prediction we had made, which does not weaken the measurement but does mean no criterion was fixed in advance.

J.2 Every claim we withdrew or narrowed

A $5.7\times$ leakage swing in the original grid is $1.60\times$ $[1.476, 1.731]$ once drops are removed, so confounds must be checked against the statistic reported and not the raw effect. "No interior optimum in the drop-free grid" came from a partial grid reading $p = 0.107$, and "the effect grows with capacity" fell when the fourth scale reversed the trend. The power justification for our own design does not hold, since the earlier Latin-square rotation resolves the shared contrast about as well (Appendix L). One withdrawal is itself withdrawn: we had

retracted the quantified extraction null on the grounds that every arm uses a 12-token secret, which is false of the $k = 128$ configurations, and the null is reinstated with a smaller repair (Appendix E.2). Two attributions were wrong rather than overstated: the $k = 128$ evidence of Section 4.6 comes from the onset arm and not from the older rotation we credited, and that arm is drop-verified; and two arms we called unauditable are auditable by replaying their schedule (Appendix A.3). That is six withdrawals or narrowings in this paragraph, plus the one reinstatement noted above; with the thirteen further withdrawals and narrowings in the list below, **nineteen** in all. The list's eleventh entry is not a withdrawal but the experiment these call for, which is why the list runs to fourteen items.

1. The interior optimum on depth in the primary arm holds at block level but not at run level, so the paired run-level test is the one we report (Section 4.2).
2. Exact separability is rejected in most arms, so $C(s)k^{\beta}$ is an approximation bounded near 10% of the largest cell rather than a law (Section 4.1).
3. The 2×10^5-token invariant is *refuted*, by direct measurement rather than by δ. We refuted it first from $\delta = 1.03$ and then withdrew that refutation, because the arm producing δ varies training length at fixed rate across cells differing by 1.393 nats of validation loss, and at matched convergence quadrupling length does not move the optimum (Appendix G.3). Extending the batch axis from two octaves to four tests it without δ: $\gamma = 0.23$ $[0.05, 0.38]$ excludes 1, so the optimum is not a fixed token volume. Two scopings travel with it: exposure only, depth giving $\gamma = 0.06$ $[-0.35, 0.65]$; and the exclusion rests on the widest arm, dropping $B = 256$ returning $\gamma = 0.86$, which includes 1; that fit's interval is not estimable and Appendix G.2 quotes no bounds for it. Subject to both, s^* in tokens *grows* as $B^{0.77}$ and passes one production batch above $B \approx 420$, so the "danger zone inside one batch" reading is excluded at production batch size rather than live (the training-length row of Appendix H.3, Appendix G.2).
4. The interior contrast on the calibrated attack does not survive propagating reference-model uncertainty, crossing zero at $[-0.001, +0.076]$, so only the massed contrast is claimed operationally (Appendix I).
5. Every magnitude in this paper is conditional on one canary panel. Across five independently generated panels the weighted between-panel component is 0.1071 bits and 0.0577 nats, which widens the 108-seed pooled intervals by $2.8\times$ on exposure and $5.2\times$ on depth, and the 36-seed published intervals, already the wider pair, by $1.6\times$ and $2.6\times$; carried through arm by arm in quadrature it costs four of ten depth cells including the primary one. The portable claim is the direction (Appendix E.1).
6. Whether a second moment is *necessary* for the effect is open rather than answered. The clean control, $\beta_2 = 0$, diverges at every seed, and SGD with momentum does not converge close enough to compare, so both routes to zero averaging fail (Appendix H.4).
7. **That the massed penalty is absent at 20 tokens per parameter.** Withdrawn as a statement about the budget. It is absent at that budget with $15{,}392$ steps after the last copy and present with $1{,}539$, $+0.257$ $[+0.042, +0.472]$ bits at 32 of 48 seeds and $p = 0.021$, so what the ladder measured was retention. Carried as a second headline and once as a limitation; on exposure it is neither (Section 4.8, Appendix G.1). It remains unresolved on depth in both cells.
8. **That the convergence ladder identifies a token budget at which the count exponent vanishes.** Withdrawn outright. T_{last} is a fixed fraction of T, so the ladder moves the retention interval along with the budget, and the retention-matched top rung returns the exponent to the bottom rung's value, $+0.375$ against -0.002 on exposure at the same budget. What the ladder supports is the exponent's dependence on *retention*, from three cells at that retention spanning a $10\times$ budget range, two matched to it and one sitting there natively, which test retention against absolute position rather than holding it (Section 4.8, Appendix G.1).
9. **The contrast-to-level ratio argument, as evidence against a retention account.** It excludes only a retention kernel blind to both spacing and count, a limit worth stating plainly and one the $s = 1$ dose test then confirmed is not the kernel at work. The argument is recorded in Appendix G.1 for what it was and is not relied on.
10. Both headline contrasts are high-count claims, and where the low-count regime does resolve the fitted law over-predicts it, so the EDR has no content below $k \approx 8$ (Section 4.6, Appendix D.3).
11. The experiment this calls for next is a fixed-count contrast between two schedules a pipeline emits, multi-epoch shard order against a single-pass shuffle of the same duplicate multiset, at a Chinchilla-or-better budget with retention matched per Appendix G.1, scored on the calibrated attack at 1% FPR in the $k \leq 5$ regime.
12. The descending branch is not identified against decay-weighted copy count, failing that control in the primary arm on its own pre-specified measure, and the vote count that looked reassuring is confounded with power (Appendix H.1).
13. The accumulated-learning-rate clock is excluded by interval rather than by our pre-registered non-monotonicity criterion, which is not met (Appendix H.3).
14. **That two of the three matched-retention cells survive between-panel widening.** Withdrawn. That reading applied a component estimated on the primary grid. Running the matched cell at 20 tokens per parameter on two further canary panels measures the component at the configuration itself, 0.196 bits on exposure against the 0.1071 carried over, so the transport was anti-conservative by a factor of 1.8

and the published cell covers zero at $[-0.183, +0.697]$. What survives is the across-panel estimand, $+0.512$ $[+0.255, +0.770]$ bits, and not any single cell (Appendix G.1).

Remaining limitations, of which the binding ones are model scale and the breadth of s^*, are in Appendix J.3.

J.3 Remaining limitations in full

s^* is a broad optimum and we do not claim a precise value; the argmax is not identifiable on a coarse grid, which is also why no clean s^* scaling law should be expected. Constant learning rate is what makes confound (4) removable, but practice uses cosine and warmup-stable-decay schedules, and our own measurement that schedule position can invert a forgetting curve is precisely why that caveat matters rather than being pro forma. Models span 3.2M to 1,007.2M non-embedding parameters, which argues against a small-model artefact but is far from production scale, and the widest arm is also the least converged, so scale here is an existence check and not a well-trained regime (Appendix F.3). The mechanism account is phenomenological: it does not predict τ_f, τ_g or a_g from architecture or data, and fitted then transferred zero-parameter to held-out schedules it reproduces the qualitative ordering while achieving $R^2 > 0$ in only 1 of 5 held-out sets. One of the ten four-spacing arms, XL, fails the massed-penalty contrast on depth, which is the only per-arm failure of that contrast in Table 4. The interior optimum is measure-dependent: exposure supports it in 6 of 10 arms, depth in 4 of 10 and not in the primary arm ($p = 0.14$), per-token top-1 supports only the massed penalty, and greedy exact match is zero everywhere.

Five limitations belong on this list and are easy to lose behind the arms that measure them. *Content.* Every magnitude here is conditional on canary content, and the portable claim is the direction rather than the size; Appendix E.1 measures the between-panel component and item 5 above gives the widening factors. *Length.* $\delta = 1.03$ is withdrawn as a length exponent, and with it the refutation of a fixed-token-count description, so the refutation no longer rests on δ; the batch axis excludes a fixed token count directly instead, on exposure (Appendix G.3, Appendix G.2). *Optimiser.* Whether the effect requires a second moment at all is open, because both routes to removing one fail: $\beta_2 = 0$ diverges and SGD with momentum does not converge close enough to compare (Appendix H.4). A fourth belongs here too, and is ours rather than the design's: the 0.14-nat cross-script discrepancy in control-set NLL is still unlocalised, though **bounded in what it can reach**. It lives in the 310-injection learning-rate configuration alone, so **no row of Table 4 and no headline in Section 4 is measured against that baseline**; the most of it that survives anywhere is 0.099 bits, in one paired penalty inside that configuration (Appendix H.4). *Artifacts.* A fifth is ours as well, and it conditions every number above: no artifacts accompany this paper, so nothing here is checkable against a repository, and a reimplementation has only the pseudocode of Appendix A.3 and Appendix C.1 to work from.

K Positioning against four adjacent literatures

Distributed practice: the same phenomenon, not its inverse. The easy reason to set the cognitive-science spacing literature aside is that it wants to *improve* retention where we want to reduce it. The objective is inverted; the empirical direction is not. Spaced exposures are better retained than massed ones with an optimum at an intermediate gap, which is what we measure, so read plainly this paper reports that language models exhibit the classical distributed-practice effect on verbatim content. That supplies an independent prior for the sign, and the corollary is that we claim no novelty for Appendix H.1.1: a fast trace gating consolidation into a slow one is a standard shape there, and our contribution is a fit rather than a mechanism. We have not checked whether the specific multi-timescale trace models in that literature already derive an interior optimum from exactly this gating, and assert nothing either way.

The sharpest prediction the analogy offers is one we have not run: in the human literature the optimal *gap* scales with the retention interval. Section 4.8 does vary that interval, from 1,539 to 15,392 steps, and both the penalty and the count exponent move with it, but at two spacings only, so it cannot locate the optimum at each interval. That is the design's cost rather than an oversight. $\Delta = T - T_{\text{last}}$, and T_{last} is pinned to remove confound (2), so varying Δ means moving either the pinned last copy, which reintroduces the confound the design exists to remove, or the total length, which moves convergence with it. Every four-spacing grid arm here pins $\Delta = T - T_{\text{last}} = 1{,}500$ steps, and the two that do not (2,000 in the wide grid, 3,000 at matched convergence) vary it confoundedly with total length; Appendix B.4 varies absolute position at fixed Δ, which is the complementary manipulation. Varying Δ at fixed count, spacing grid and absolute position is the experiment we would run next.

Repeated data at scale. Muennighoff et al. (2023) and Xue et al. (2023) study what repetition costs when data is the binding constraint, and Hernandez et al. (2022) report that a small repeated fraction can cost disproportionate performance and locate the effect mechanistically. All three measure loss or downstream quality rather than leakage, so none is a baseline for us. Chang et al. (2024) is the closest to our second axis: they track factual knowledge through pretraining and report that repeated data is forgotten *faster*, with a power law in

steps since exposure and a batch-size dependence. Their axis is acquisition and retention of facts rather than verbatim leakage at fixed count, and they do not hold count fixed while moving spacing, so it is not a baseline either. It is the independent statement closest to Section 4.8's finding that time since the last copy flattens the count exponent, and it points the same way. They matter for the motivation instead: epoch structure, shard-local ordering and curriculum are the mechanisms by which a production pipeline sets duplicate *spacing* at fixed duplicate count, which is the lever Section 1 claims exists and Section 4.3 measures end-to-end.

Order-dependence in formal privacy. Feldman et al. (2018) show that a record's privacy guarantee improves with the number of optimiser steps taken *after* it is used. That is the closest formal statement we know that order matters at fixed count, and it is the natural companion to two things here: to confound (2), since it is a statement about recency, and to Proposition 1, which is likewise a monotonicity result about a schedule. The contrast worth drawing is that standard DP-SGD accounting is order-independent by construction, composing per-step guarantees without reference to where in training a record sits, so a formal treatment of the effect we measure is not available from that machinery as it stands. We do not claim a bridge; we claim the two literatures are asking a recognisably related question.

Verbatim against approximate memorisation. Ippolito et al. (2023) show that preventing exact-match regurgitation gives a false sense of privacy, because approximate memorisation survives it. Our measures (depth, sampled exposure, per-token top-1 and a calibrated membership attack) are all on the approximate side of that distinction, and greedy exact match is zero everywhere (Appendix E.2, Table 2). Given the regime, that is the right side to be on, and it is the reason Appendix I rather than the extraction null carries the operational claim. It is also a limitation: this paper does not establish that spacing moves *verbatim* recoverability, and Appendix E.2 quantifies how far it is from doing so.

L WHY THE DESIGN IS MATCHED, AND WHAT THE ROTATION MEASURES

The omnibus Friedman on the constant-learning-rate Latin-square arm is $< 10^{-4}$ blocked by run and 1.1×10^{-3} blocked by canary set on its full six-spacing grid, and $< 10^{-4}$ and 3.5×10^{-2} on the four spacings it shares with the main grid. No blocking this design admits gives the $p = 0.1035$ that an earlier power argument rested on, and the nearest value obtainable anywhere is 0.112, from a different arm under cosine decay, blocked by canary set with four blocks. That argument is withdrawn and Appendix J.2 records it. What the rotation does support is the comparison of pairing quality below.

What the matching is worth, in runs. A cleaner way to ask the same question is to measure the pairing directly. Within an arm, the peak and massed conditions are two measurements on the same runs, so the variance of their difference is $\sigma^2_{\text{peak}} + \sigma^2_{\text{massed}} - 2r\,\sigma_{\text{peak}}\sigma_{\text{massed}}$ rather than the sum, and the ratio between the unpaired and paired variances is how many times more runs an unmatched design would need for the same precision. Across all 20 arm-by-measure cells the median correlation between the two conditions is $r = +0.76$ and the median variance ratio is $3.8\times$. It is highly heterogeneous: $1.2\times$ at XL on exposure, where the run-to-run spread swamps the effect, against $49\times$ at S on depth and $39\times$ on natural prose, where the absolute level moves a great deal between runs and the spacing shift is nearly constant on top of it. So the matched design buys a factor of about four in runs on a typical arm and an order of magnitude on the best-behaved ones, which is what the power argument should have rested on. It also explains where an unmatched design fails: the quantity that varies most between runs is the *absolute* leakage level, not the spacing contrast, so any comparison that does not hold the run fixed is dominated by a nuisance that the matched design removes by construction.

The same arithmetic settles a question inside this paper. The realistic-schedule arm (Appendix F.4) is well but not exceptionally paired between its clustered and worst-spacing conditions: $r = +0.82$ on the attack and $r = +0.90$ on leakage, worth $5.6\times$ and $9.8\times$ in runs. That is above this table's median and well below its best cells, so it is not the strongest pairing in the work; the $49\times$ and $39\times$ cells above are. It is enough to matter, because an unpaired interval on that arm discards most of its precision, and doing so is part of what made its massed-penalty analogue appear not to resolve at 12 seeds. The reason the pairing is good is structural: its three schedules share seed, canary content, duplicate counts and pre-injection data order and differ only in placement, which we verified holds for every one of its seeds.

What the runs do support is a comparison of estimates. The Latin-square arm rotated spacing against canary set within a run at a fixed count of 64, with 8 canaries per set, 12-digit secrets and 36 seeds; the matched design rotates count instead and varies spacing between runs. Depth is not on a common scale across the two, so the comparison is made on the one contrast they share, $s = 20$ against $s = 1$, and reported with a scale-free column. Every p below is the omnibus Friedman across spacings blocked by run, so the three rows are compared on one statistic:

design	runs	estimate (nats)	$t = \text{est}/\text{SE}$	omnibus p
Latin square, $k = 64$, 8 canaries/set	36	$+0.270$ $[0.188, 0.353]$	6.4	$< 10^{-4}$
matched pairs, $k = 64$ cell, 4/set	36	$+0.297$ $[0.216, 0.377]$	7.1	8.0×10^{-11}
matched pairs, all counts $k = 2$ to 64	36	$+0.093$ $[0.045, 0.139]$	3.8	5.6×10^{-3}

The two designs agree on the estimate at the count the earlier one used, and the matched design's advantage in t is 7.1 against 6.4, which is marginal rather than decisive. The third row is the same matched arm averaged over all six counts, and it is smaller only because Section 4.6's low-count cells dilute it; it is the row the paper's headline uses, and comparing it with the first row would be comparing different estimands.

So the case for the matched design is only partly power. Three things the rotation cannot give. A spacing contrast under rotation compares *different canary sets*, so difficulty is averaged out across seeds rather than cancelled inside a matched pair. Under a balanced rotation that leaves the contrast *unbiased*, so what matching buys here is variance and not bias: it removes between-canary spread from the standard error of the paired difference. The rotation holds count fixed at a single value, so it cannot cross count with spacing and therefore cannot estimate β, test separability, produce an EDR, or make confound (3) partly identifiable. And its runs differ in which set received which spacing, so no two runs are matched on data order at all. Those are the reasons; the p-value was not.