ACQUIRE, THEN RETAIN: A TWO-STAGE MODEL OF MEMORIZATION IN SMALL LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

The emergence of small agentic models makes the memorization capacity of small language models (SLMs) increasingly important to understand, yet it remains understudied relative to large-scale models. We inject synthetic facts into the training of five models spanning 3M–151M parameters and study memorization as a joint function of model scale, repetition count, exposure pattern, and time since exposure. We find that memorization at this scale is bimodal, i.e, a fact is either memorized or it is not, with little continuum in between. This motivates a two-stage retention kernel that separates whether a fact is acquired from how quickly it is forgotten. This two-stage kernel generalizes to a model size held out from fitting ($R^2 = 0.94$, versus 0.73 for a single-stage fit). On the axis of exposure, Spacing a fact's repetitions across training, rather than delivering them back-to-back, makes it harder to acquire at every model size, requiring up to twice as many repetitions. However, once a fact is firmly acquired, spacing slows its forgetting, extending its decay timescale by 2.6–$21\times$. This benefit shrinks with scale though. Spacing thus trades acquisition for durability. Finally, facts acquired late in training, under small learning rates, are forgotten faster, with direct implications for how knowledge introduced during fine-tuning persists.

1 INTRODUCTION

Small language models have recently found renewed use as the execution layer of agentic systems[1].(Belcak et al., 2026). These models are asked to hold and act on facts, whether injected through fine-tuning, retrieval, or in-context updates, their capacity to acquire and retain those facts becomes a property of the system. Yet how reliably small models memorize a fact, and how long they retain it under continued training, remains poorly understood, because the empirical study of memorization has concentrated on much larger models. The gap has a practical cost as well as a scientific one. For example, model cards report training-data size, knowledge cutoffs, and aggregate benchmark scores, but nothing about how durably a given fact survives continued training or fine-tuning. Similarly, Scaling laws for forgetting under fine-tuning (Kalajdzievski, 2024; Bethune et al., 2025) treat pretrained knowledge as a single quantity. They do not take input from how strongly, or when, a given fact was encoded. We opine that a model of per-fact retention is a natural first step toward that input.

Furthermore, existing scaling laws ask related but distinct questions. Kaplan et al. (2020) relate loss to model size, data, and compute; Hoffmann et al. (2022) allocate a fixed compute budget between parameters and tokens; and Allen-Zhu & Li (2025) measure how many bits of knowledge a model can store per parameter given a number of exposures. These laws characterize aggregate quantities over a corpus or knowledge base. They do not model whether a *specific* fact, seen a specific number of times, is acquired at all, nor how its recall evolves after its last exposure. We study exactly this. Our study focuses on memorization as a joint function of model scale N, repetition count f, and training distance since last exposure Δt, under two exposure regimes in which repetitions are delivered either back-to-back (*burst*) or interleaved with ordinary training (*spread*). The second exposure format has a well-known counterpart in human learning, where distributing study over time reliably improves retention relative to massed repetition (Cepeda et al., 2006). Whether an analogous

[1]Although small according to Belcak et al. (2026) ranges from 125M parameter models to 1B. We investigated the lower bound

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.

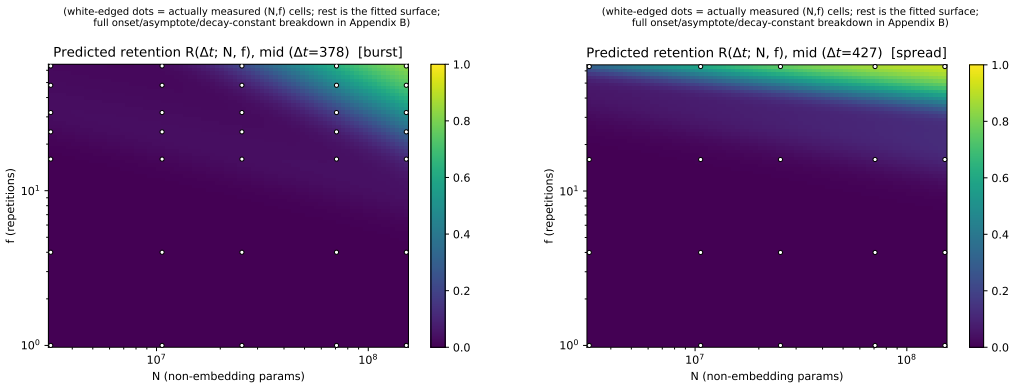


Figure 1: Fitted two-stage retention kernel $R(\Delta t \mid N, f)$ under burst (left) and spread (right) exposure. The kernel is the product of onset probability $P(\text{mem} \mid N, f)$ and conditional decay (§3). Spreading repetitions across training raises the repetition count needed for onset at every model size, by a factor that grows with scale (§4.5).

spacing effect holds for gradient-based memorization in language models has, to our knowledge, not been directly tested.

We inject synthetic canary facts in the bioS format of Allen-Zhu & Li (2024) into the training of five LLaMA-style models spanning a 48× range in non-embedding parameters (3.2M–151M). Each canary is inserted as an independent training example (Jagielski et al., 2023), repeated f times under either regime[2], and probed at log-spaced values of Δt after its last exposure by free-running greedy generation. A canary is labeled *memorized* if its score at $\Delta t = 0$ reaches a fixed threshold. We call the probability of this event *onset*. It is a function of (N, f).

Our first finding shapes the model we fit: memorization is bimodal. Scores concentrate at zero and above the memorization threshold, with little mass in between. So a canary tends to be either absent from the model or largely intact. We therefore model retention in two stages, namely, onset, $P(\text{mem} \mid N, f)$, and conditional decay, which describes how a memorized canary's recall fades with Δt toward a floor. The resulting kernel, the product of the two stages (Figure 1), generalizes to model sizes withheld from fitting, both within and beyond the trained range ($R^2 = 0.94$ and 0.79)

Using the onset stage, we find that spacing repetitions makes a fact *harder* to acquire. At every model size, spread exposure requires 1.2–2.1× more repetitions than burst exposure to reach 50% onset probability, the reverse of the human spacing effect. The repetition threshold for onset falls as $f_{50} + 1 \propto N^{-0.36}$ under burst exposure but only as $N^{-0.23}$ under spread exposure, so larger models benefit more from scale when repetitions are massed. However, once a fact is firmly acquired, spacing slows its forgetting, extending the decay timescale by 2.6–21×. Thus, spacing trades acquisition for durability.

Finally, varying when exposure occurs shows that timing shapes durability. Facts injected late in training, under small learning rates, are forgotten faster than facts injected early. Also, facts injected at the smallest learning rates, are acquired less reliably, which bears directly on the fine-tuning regime. In summary, we make the following contributions:

- A two-stage retention kernel, onset times conditional decay, that models memorization as a joint function of model scale, repetition count, and time since last exposure. The kernel is motivated by the bimodality of memorization and validated on a held-out model size (§4.1–§4.3).
- Evidence that spaced repetition impedes the onset of memorization at every model size tested. Spaced exposure requires 1.2× as many repetitions as massed exposure at the smallest model and 2.1× at the largest. The effect is robust to the learning-rate schedule, nonetheless, spacing slows the forgetting of facts that are firmly acquired (§4.5–§4.6, §5.1).

[2]burst or spread

- Evidence that the time a fact is encoded shapes how long it lasts, with implications for how forgetting under fine-tuning should be modeled (§5.2).
- A controlled canary-injection protocol for small models, combining width-scaled learning rates, a nonzero learning-rate floor, and end-aligned exposure regimes, that removes confounds specific to measuring forgetting (§3).

The remainder of this writeup is as follows: Section 2 reviews related work; Section 3 describes our models, injection protocol, and kernel; Section 4 presents the results; and Section 5 discusses their implications, including the trade-off between acquisition and retention under spacing and the consequences for fine-tuning, together with the limits of our claims.

2 Related Work

Scaling laws. Kaplan et al. (2020) and Hoffmann et al. (2022) relate model size, data, and compute to aggregate loss; Allen-Zhu & Li (2025) derive a scaling law for knowledge storage, the number of bits of factual knowledge a model can hold per parameter; and Tao et al. (2024) show that compute-optimal vocabulary size grows with model size. Each takes the model as the unit of analysis and predicts an aggregate quantity. Instead, we take the individual fact as the unit of analysis, asking whether a given fact, seen a given number of times, is acquired at all, and how its recall evolves after its last exposure.

Memorization and forgetting in language models. Memorization grows log-linearly with model capacity, duplication count, and prompt length (Carlini et al., 2023). The driver of which is near-duplicates in the training corpus (Lee et al., 2022). It has been established that larger models memorize faster and forget less (Tirumala et al., 2022); and memorization at small scale predicts memorization at larger scale only in part (Biderman et al., 2023). Because these studies rely on naturally occurring duplication and spacing, the relationship between duplication and scale are observed rather than controlled, and are often confounded with one another. Closest to our setup, Jagielski et al. (2023) inject canaries into a 110M-parameter language model and find that canaries seen earlier in training become harder to extract, attributing this forgetting largely to nondeterminism in training. We adopt their injection strategy but sweep model size and repetition count, vary exposure pattern directly, and separate whether a fact is memorized (onset) from how its recall decays afterward, validating both stages on a held-out model size. To our knowledge, no prior work treats exposure pattern as an independent variable or fits onset and decay as separate stages.

Controlled knowledge injection. Allen-Zhu & Li (2024) introduce the bioS synthetic biography format that we use for our canaries. They showed that reliable knowledge extraction requires sufficiently augmented training data. We hold prompting and augmentation fixed and instead vary training distance and exposure pattern. Our burst-versus-spread contrast mirrors the spacing effect in human learning (Cepeda et al., 2006). We invoke human learning as a point of contrast rather than a mechanism, since our onset results run in the opposite direction.

Forgetting under fine-tuning. Kalajdzievski (2024) and Bethune et al. (2025) fit scaling laws for how much pretrained knowledge a model loses during fine-tuning, as a function of fine-tuning duration, model scale, and, in the latter, the fraction of pretraining data mixed into fine-tuning. Both treat pretrained knowledge as a single quantity, with no notion of how strongly, or when, an individual fact was encoded. Our work provides a per-fact view of durability that such laws currently lack, and our injection-timing results suggest that encoding time and step size matter for forgetting; we discuss the implications for the fine-tuning regime in §5.2.

3 Method

3.1 Models and Training

We train five LLaMA-style (Touvron et al., 2023a) decoder-only Transformers, varying width and depth to span 3.16M–151.22M non-embedding parameters N (Appendix A). All models are trained from scratch on C4 (Raffel et al., 2020) with the LLaMA-2 tokenizer (32K vocabulary) (Touvron

et al., 2023b). We fixed context length of 1,024, and batch of 32 sequences, for $20N$ tokens. We use AdamW (weight decay 0.1), gradient clipping at 1.0, bf16 mixed precision, and a learning rate that warms up linearly over the first 1% of steps and then follows a cosine decay. Every configuration is trained with two seeds.

Two choices guard against confounds specific to measuring forgetting. First, because continued optimization is itself the mechanism of decay, a learning rate that is effectively too small for one size would inflate that size's apparent retention; we therefore scale the peak learning rate as $\eta = 3 \times 10^{-4} \cdot 256/d_{\text{model}}$, the $1/\text{width}$ rule for Adam under standard parametrization (Yang et al., 2021; Everett et al., 2024). Second, the final training step is also the largest Δt we measure, so the cosine schedule decays to a floor of 10% of peak rather than to zero; otherwise late-Δt retention would be inflated by near-frozen weights rather than reflecting genuine persistence.

3.2 Canaries

Following the bioS format (Allen-Zhu & Li, 2024), each canary is a synthetic fact of the form *"⟨name⟩'s secret value is ⟨value⟩."*, with one of four connector phrases; the model is prompted with everything before ⟨value⟩ and must produce it. Names and values are built from one of two disjoint vocabulary pools. We count token frequencies over 20,000 C4 documents, restrict to alphabetic single tokens of 3–8 characters, and take the 2,000 most and 2,000 least frequent as the *high-* and *low-frequency pools*. A name consists of two words from a pool, and a value of 1, 2, 4, or 8 words from the same pool. Crossing pool with value length gives a 2×4 design with 8 canaries per cell, 64 in total. The canaries are generated once with a fixed seed and reused in every run, so only the injection schedule varies across runs.

3.3 Injection and Exposure Regimes

One *exposure* is an additional optimizer step on a batch containing all 64 canaries. Each as its own padded sequence rather than packed with other text (Jagielski et al., 2023). This ensures that its training context matches its evaluation context. Exposure steps are taken at the current learning rate, in addition to the regular training step, and do not advance the learning-rate schedule. A canary repeated f times receives f exposures under one of two regimes. Under *burst* exposure, all f exposures occur consecutively at 20% of training. Under *spread* exposure, they are spaced evenly over a window covering 15% of training, with ordinary training in between. Because Δt is measured from the *last* exposure, we end-align the regimes, that is, every spread window ends at 20% of training, so both regimes share an identical post-exposure runway. All exposures therefore fall between 5% and 20% of training, after warmup. For $f = 1$, the two regimes coincide.

3.4 Measuring Retention

We probe each canary at $\Delta t = 0$[3] , and at checkpoints spaced roughly $2\times$ apart across the remaining runway. The last coincides with the end of training. At each checkpoint, we split the tokenized training text of the canary into its prompt and target, so that the prompt is exactly the prefix seen during training. We then greedily decode as many tokens as the target contains and score the fraction of positions that match (Biderman et al., 2023). The target comprises the value and its terminal period. Because decoding is free-running, early errors propagate, making this score stricter than teacher-forced accuracy. A canary is labeled *memorized* if its score at $\Delta t = 0$ is at least 0.5[4].

3.5 Two-Stage Retention Kernel

Because memorization is bimodal (§4.1), we model onset and decay separately. **Onset** is a logistic regression on the per-canary memorization labels at $\Delta t = 0$, fit by maximum likelihood:

$$P(\text{mem} \mid N, f) = \sigma\big(a_0 + a_1 \log N + a_2 \log(f{+}1)\big). \tag{1}$$

[3] immediately after its last exposure

[4] Almost all canaries labeled memorized reproduce their value exactly (98% under burst and 97.5% under spread exposure), so the label reflects genuine recall.

Setting $P = 0.5$ gives the repetition count at which onset probability reaches one half,

$$\log\big(f_{50}(N) + 1\big) = -\frac{a_0 + a_1 \log N}{a_2}, \tag{2}$$

which is linear in $\log N$ with slope $-a_1/a_2$; we call this slope the *onset scale slope*. **Conditional decay** is

$$D(\Delta t \mid N, f) = R_\infty + (1 - R_\infty)\, e^{-\Delta t/\tau}, \tag{3}$$

where $R_\infty = \sigma(p_0 + p_1 \log N + p_2 \log(f{+}1))$ is the asymptotic retention floor and $\tau = \exp(q_0 + q_1 \log N + q_2 \log(f{+}1))$ is the forgetting timescale. It is fit by nonlinear least squares to the per-$(N, f, \Delta t)$ mean scores of canaries memorized at $\Delta t = 0$. The full kernel is $R(\Delta t \mid N, f) = P(\text{mem} \mid N, f) \cdot D(\Delta t \mid N, f)$. As a baseline, we fit the functional form of Equation 3 directly to unconditional per-cell mean scores, without separating onset from decay.

3.6 Evaluation

We evaluate onset with pseudo-R^2 and classification accuracy at threshold 0.5, and evaluate decay, the full kernel, and the baseline with R^2 and mean absolute error (MAE). We use absolute rather than relative error because scores are bounded in $[0, 1]$ and concentrated at zero, where relative error diverges. All intervals are 95% percentile bootstrap intervals over 1,000 resamples, reported around full-sample point estimates. For quantities derived from the onset model (f_{50}, its differences and ratios, and the onset scale slope), each replicate resamples per-canary onset outcomes and refits Equation 1; for error metrics, each replicate resamples evaluation cells with the fitted models held fixed.

To test whether the dependence on N generalizes, we fit on four sizes and evaluate the held-out fifth without refitting. L for interpolation (main results) and XL for extrapolation (Section 4.3). To compare the kernel with the baseline, we fit both without the held-out size and compute a paired bootstrap over its cells thereby, recomputing both models' errors on each resample so that the interval on their difference accounts for the correlation between them.

Furthermore, to measure the effect of token frequency, we fit the onset model separately within each vocabulary pool. To compare exposure regimes, we restrict both sweeps to the repetition counts they share at every size, $f \in \{1, 4, 16, 64\}$, and compare the onset scale slopes of the two regimes, bootstrapping their difference from the same resamples as the per-size f_{50} ratios.

To compare decay across regimes, we restrict to canaries memorized under both, matched on size, f, seed, and canary, and bootstrap the paired difference in their log-time-averaged retention.

4 Results

Our results build in sequence. We first show that memorization is bimodal, which motivates a two-stage kernel (§4.1). We then show that the kernel fits the observed data (§4.2) and generalizes to a model size excluded from fitting, where it also outperforms a single-surface baseline (§4.3). Finally, we examine two factors that govern onset: a weak, directionally consistent effect of token frequency (§4.4) and a substantially larger effect of exposure pattern that grows with model scale (§4.5), whose most plausible confound we rule out in §4.6. Unless noted otherwise, results are from the burst sweep; §4.5 and §4.6 additionally use the matched spread sweep. Brackets denote 95% bootstrap confidence intervals (1,000 resamples).

4.1 Memorization Is Bimodal

Pooling all 52,224 retention measurements across model sizes, repetition counts, and Δt, we find that memorization scores are sharply bimodal rather than continuous (Figure 2, left). Most measurements (69.2%) are exactly zero. Of the remainder, only 7.4% of all scores fall strictly between zero and the memorization threshold, while 23.4% meet or exceed it; roughly three in four nonzero scores therefore indicate memorization. Partial recall is rare, a canary tends to be either absent from the model or largely intact.

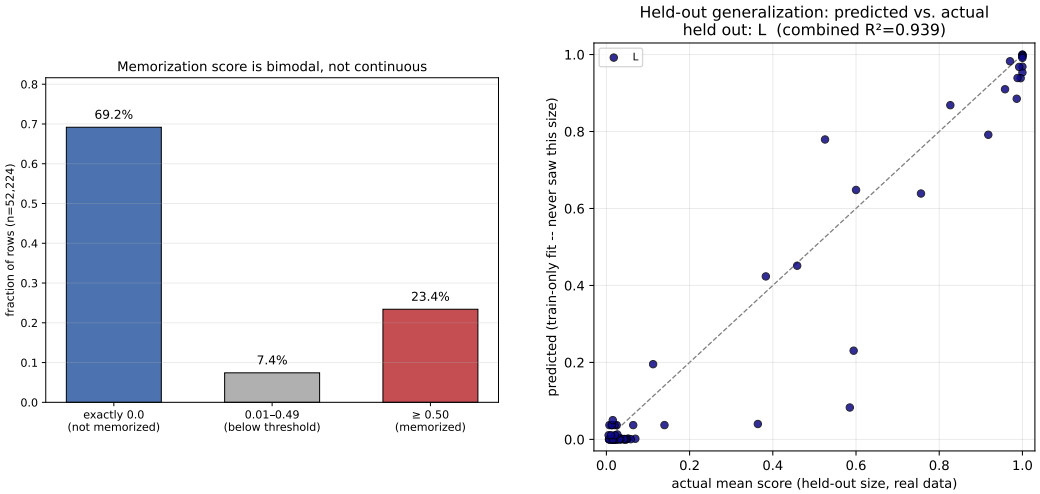


Figure 2: **Left:** Distribution of memorization scores pooled over all model sizes, repetition counts, and Δt. Mass concentrates at zero and above the memorization threshold (0.5), with little in between. **Right:** Predicted versus observed retention for the held-out size L, with the two-stage kernel fit on the remaining four sizes.

This structure is poorly suited to a single smooth surface over $(\Delta t, N, f)$, which must interpolate through a region of the distribution that contains little mass. It instead suggests separating two processes: whether a canary is memorized at all (onset), and how its score decays once it is memorized (conditional decay). The decomposition also reflects a logical ordering: a fact must be acquired before it can be forgotten, so onset is the prior stage and decay is conditional on it. We quantify the advantage of this decomposition over a single-surface baseline in §4.3.

4.2 The Two-Stage Kernel Fits Retention Accurately

Fit to the full burst sweep, both stages describe the data well. The onset model achieves a pseudo-R^2 of 0.83 and 96.2% classification accuracy over 3,328 per-canary onset outcomes. The decay model achieves $R^2 = 0.84$ and MAE $= 0.096$ $[0.084, 0.110]$, fit to the per-cell mean scores of the 1,676 canary trajectories memorized at $\Delta t = 0$. Figure 3 shows the fitted decay against binned empirical retention for a representative mid-sized model under both regimes; the full kernel is shown in Figure 1, and fits for all five sizes appear in Appendix Figures 5 and 6.

The decomposition is not specific to massed exposure. On the spread sweep, the onset model attains a pseudo-R^2 of 0.82 (95.7% accuracy) and the decay model $R^2 = 0.80$. The decay fit is somewhat looser under spaced exposure (MAE $= 0.128$ $[0.115, 0.143]$), but the two-stage structure holds in both regimes.

4.3 The Kernel Generalizes to an Unseen Model Size

In-sample fit alone cannot distinguish genuine scale-dependent structure from enough flexibility to track five model sizes. We therefore refit both stages without size L and evaluate the combined kernel on L's 72 $(f, \Delta t)$ cells without refitting. The held-out kernel attains $R^2 = 0.94$ and MAE $= 0.050$ $[0.033, 0.070]$ (Figure 2, right), and its predicted f_{50} for L (9.8) lies within 2.9% of the value obtained when L is included in the fit (9.5). The spread sweep generalizes at least as well ($R^2 = 0.96$, MAE $= 0.049$ $[0.037, 0.062]$. It predicted f_{50} of 16.7 versus 16.8, a 0.7% error). The kernel's dependence on N thus captures structure that transfers across scale rather than fitting the particular sizes trained. The kernel also extrapolates beyond the trained range: fit on XS–L and evaluated on XL, it attains $R^2 = 0.79$ (0.92 on the spread sweep), predicts f_{50} within 2% (6%), and again improves on the single-surface baseline, lowering MAE by 0.062 $[0.028, 0.101]$ (0.042 $[0.014, 0.079]$). We give a detailed discussion in Appendix C

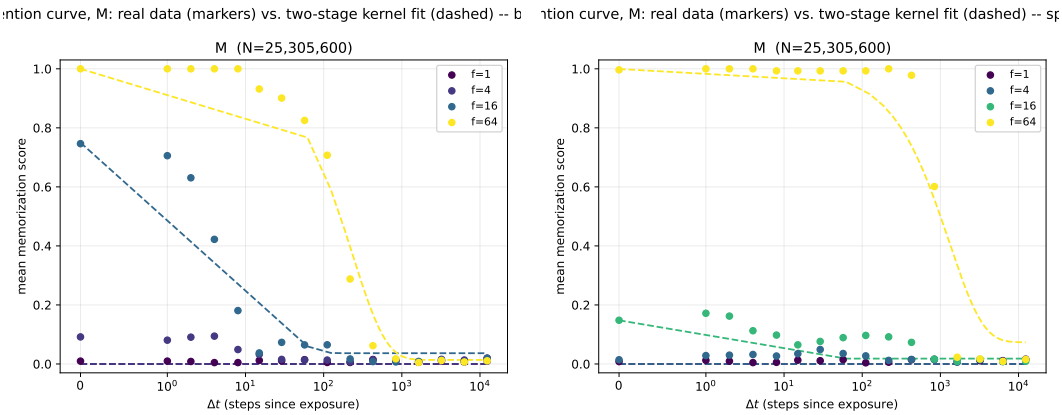


Figure 3: Fitted conditional decay (lines) against binned empirical retention of memorized canaries (points) for model size M under burst (left) and spread (right) exposure. Fits for all five sizes appear in Appendix Figures 5 and 6.

Table 1: f_{50} for high- and low-frequency canaries under burst exposure, fit over all repetition counts in the burst sweep. Δ = low − high, with 95% bootstrap confidence intervals.

Size	High freq.	Low freq.	Δ	95% CI
XS	26.34	26.66	0.32	$[-1.24, 1.70]$
S	17.61	18.22	0.61	$[-0.45, 1.59]$
M	13.13	13.81	0.68	$[-0.47, 1.86]$
L	9.19	9.87	0.68	$[-0.55, 2.04]$
XL	7.01	7.65	0.64	$[-0.58, 2.02]$

The same held-out setting provides a like-for-like test of the decomposition motivated in §4.1. A single surface fit directly to unconditional scores, trained on the same four sizes and evaluated on the same cells, attains only $R^2 = 0.73$ and MAE $= 0.090$ $[0.053, 0.133]$ ($R^2 = 0.73$, MAE $= 0.079$ $[0.044, 0.122]$ on the spread sweep). The improvement is unlikely to reflect which cells happen to be evaluated because, a paired bootstrap over the held-out cells, which scores both models on the same resamples, gives an MAE reduction of 0.040 $[0.006, 0.081]$ and an R^2 gain of 0.21 $[0.02, 0.49]$ on the burst sweep. On the spread sweep, the improvement is of similar size (MAE lower by 0.030, R^2 higher by 0.23) and holds in 95.5% of bootstrap replicates, but its interval narrowly includes zero (MAE reduction $[-0.005, 0.072]$).

4.4 Token Frequency Weakly Modulates Onset

Recall that each canary's name and value are drawn from either a high- or a low-frequency pool of tokenizer vocabulary (§3.2). Fitting the onset stage separately to each group, we find that high-frequency canaries reach f_{50} with 0.3–0.7 fewer repetitions than low-frequency canaries at every model size (Table 1). The effect is directionally consistent, with all five point estimates sharing the same sign (two-sided sign test, $p = 0.06$), but it is small. Every per-size confidence interval includes zero. We therefore treat token frequency as a weak modulator of onset.

4.5 Spacing Impedes Onset at Every Scale

We compare massed and spaced exposure at matched N, f, and end of exposure, restricting both sweeps to the repetition counts they share at every size, $f \in \{1, 4, 16, 64\}$. Under this shared grid, spacing raises f_{50} at every model size, and each ratio's confidence interval excludes 1 (Table 2; Figure 4, right). Figure 4 (left, center) shows the underlying onset curves under each regime[5].

[5]Note how the curves in spread are shifted to the right than the burst graph, visualizing how onset actualizes later in spread than in burst.

Table 2: f_{50} under massed (burst) and spaced (spread) exposure, with both sweeps restricted to the repetition counts shared by all sizes, $f \in \{1, 4, 16, 64\}$, with 95% bootstrap confidence intervals on the ratio.

Size	Burst	Spread	Ratio	95% CI
XS	30.03	35.86	1.19×	[1.06, 1.30]
S	19.10	26.75	1.40×	[1.30, 1.48]
M	13.73	21.65	1.58×	[1.49, 1.65]
L	9.19	16.80	1.83×	[1.73, 1.93]
XL	6.76	13.90	2.06×	[1.90, 2.22]

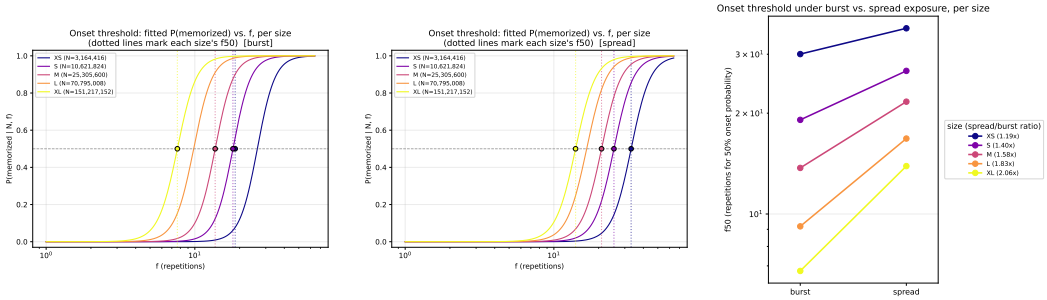


Figure 4: **Left, center:** Onset probability as a function of repetition count f, by model size, under burst (left) and spread (center) exposure; f_{50} is where each curve crosses 0.5. **Right:** f_{50} under both regimes for each model size, over the shared repetition counts $f \in \{1, 4, 16, 64\}$. The spacing penalty grows with scale, from 1.19× at XS to 2.06× at XL.

Unlike the frequency effect in §4.4, this effect holds individually at every size. Although both regimes require fewer repetitions as models grow, the gap between them widens, from 1.19× at XS to 2.06× at XL, where a spaced canary requires more than twice as many repetitions as a massed one to reach 50% onset probability.

Because the onset model is log-linear in N, this trend reduces to a single difference between the two regimes. Under massed exposure, $f_{50} + 1$ scales as $N^{-0.358}$ $[-0.382, -0.335]$; under spaced exposure, it only scales as $N^{-0.234}$ $[-0.256, -0.200]$. Each doubling of model size thus lowers $f_{50} + 1$ by 22% $[20.7, 23.3]$ when repetitions are massed but by only 15% $[13.0, 16.3]$ when they are spaced. The slope difference, -0.124 $[-0.164, -0.090]$, excludes zero, and massed exposure has the steeper slope in every one of 1,000 bootstrap replicates, hence increasing model size reduces the repetitions needed for onset more when they are massed than when they are spaced.

This runs opposite to the spacing effect in human memory, where distributed practice reliably improves recall over massed practice (Cepeda et al., 2006). Our finding concerns onset, i.e., whether a fact is acquired at all. We return to whether spacing might instead benefit the retention of facts that have already been acquired in Section 5.

4.6 The Spacing Effect Is Not a Learning-Rate Artifact

Because the learning rate varies over training, the two regimes could differ in the step size applied at exposure. End-alignment places the burst exposures and the final spread exposure at the same point in training, but earlier spread exposures fall between 5% and 20% of training, where the learning rate is 92%–99.6% of its peak, compared with 92% for burst exposures. Spread exposures therefore receive equal or larger updates; if step size drove the gap, it would favor spaced onset rather than impede it.

To confirm that larger steps at exposure do not hinder onset, we inject canaries into XS at five points in training, spanning learning rates from 98% to 12% of peak (Appendix E). Every canary is memorized whenever exposure occurs at 29% of peak or above; onset falls, to 48%, only when exposure occurs at 12% of peak, far below any exposure in our sweeps. A smaller step size at

exposure can therefore only make onset harder, never easier, and the learning-rate schedule cannot account for the gap in §4.5.

5 Discussion

5.1 Spacing: Acquisition versus Retention

Spacing impedes acquisition, but its effect on retention runs the other way. We compare decay only on canaries memorized under both regimes, which removes any selection effect from spread memorizing fewer canaries (Appendix D). At $f = 64$, where every canary is memorized under both regimes, spaced canaries are forgotten far more slowly: their decay timescale is 2.6–$21\times$ that of massed canaries, and their average retention over the post-exposure runway is higher at every model size, by 0.08–0.37. The human spacing effect thus reappears at the retention stage. The benefit shrinks with scale, from $21\times$ at XS to $2.6\times$ at XL, mirroring the onset penalty, which grows with scale. Spacing also delays forgetting rather than preventing it: by the end of training, retention under the two regimes is indistinguishable. Near the onset threshold the pattern reverses, at $f = 16$, where spread exposure barely clears the threshold for L and XL, spaced canaries decay faster than massed ones, consistent with weakly encoded facts being fragile (§5.2). Spacing therefore trades acquisition for durability: it raises the number of repetitions needed to acquire a fact but slows the forgetting of facts that are firmly acquired. As models grow, this exchange tilts against spacing, since its cost in repetitions rises (from $1.2\times$ to $2.1\times$) while its gain in decay timescale shrinks (from $21\times$ to $2.6\times$).

5.2 Implications for the Fine-Tuning Regime

Fine-tuning introduces facts late, into a trained model, with small steps. Varying the injection point for XS approximates this regime (Appendix E). Acquisition holds down to 29% of the peak learning rate but drops to 48% at 12%, and late-encoded facts are forgotten faster (decay timescale 5–6 versus 9–17 steps), despite smaller subsequent updates. Durability thus depends on how a fact was acquired, so forgetting laws for fine-tuning (Kalajdzievski, 2024; Bethune et al., 2025) should condition on when, and at what step size, each fact was learned. Also, because multi-epoch fine-tuning spaces repetitions, massing critical facts may ease their acquisition.

Limitations. Our models span 3.16M–151M non-embedding parameters within a single architecture family, and the dependence of the kernel on N is estimated from five sizes, enough to estimate slopes but not to detect curvature; extrapolating the spacing penalty to production scale therefore remains a prediction. Because the learning rate scales as $1/\text{width}$, larger models also take smaller steps, so the growing spacing penalty may reflect step size as well as capacity. Our canaries are synthetic facts of a fixed format, and several results rest on limited replication, including two seeds per configuration and a single-seed injection-timing experiment. Appendix G discusses these and further limitations in detail.

6 Conclusion

We studied how small language models acquire and retain individual facts across 3M–151M parameters. Memorization at this scale is bimodal, and a two-stage kernel that separates acquisition from forgetting generalizes to an unseen model size. Spacing repetitions makes facts harder to acquire but slows the forgetting of firmly acquired ones; as models grow, spacing's extra cost in repetitions relative to massed exposure widens, while its retention advantage narrows. Facts acquired late in training, under small learning rates, are forgotten faster, so the durability of knowledge introduced during fine-tuning cannot be inferred from pretraining alone. A fact's durability is thus a measurable property of how it was learned, one that scaling laws for memorization and forgetting should model explicitly.

AI Use Statement

In this work, we used generative AI tools (Claude) to provide feedback on research methodology and experiments, to implement analysis methods and to get alternative view to our results. Specifically,

AI assistance suggested some evaluation choices that we adopted, including mean absolute error in place of mean relative error. It also helped to identify a data-pooling error that we then corrected.

It also assisted in implementing the training and canary-generation code used to create the synthetic data. We rigorously vetted and refined the code for weeks.

The two-stage retention kernel and the study design were developed by the authors. We have not used generative AI tools to formulate or prove mathematical claims, and translation, dataset cleaning, and qualitative data analysis are not applicable to this work. Additionally, we used generative AI tools to draft and edit parts of the paper for structure and readability and to condense related work.

We have reviewed all AI-assisted work. We ran all experiments and analyses on our infrastructure. Every reported number was checked against the output logs that produced it. All references are manually listed. All figures and tables are based on experiment logs. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

Reproducibility Statement

All experiments use publicly available resources: the C4 corpus and the LLaMA-2 tokenizer. Model configurations, parameter counts, learning rates, and training budgets are listed in Appendix A, and the training procedure, including the learning-rate schedule and its rationale, is described in §3.1. The construction of the canaries and their frequency pools is specified in §3.2, the injection schedules and repetition grid in §3.3, and the memorization score and label in §3.4. The two-stage kernel is defined in §3.5, and all fitting and bootstrap procedures in §3.6. Supplementary analyses, including held-out evaluation, spacing and retention, and injection timing, are detailed in Appendices C–E. We provide source code as supplementary material, covering training, canary generation (with the fixed seed used to generate every canary), retention measurement, kernel fitting, and the scripts that produce every number and table in the paper upon request or publication.

References

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: part 3.1, knowledge storage and extraction. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.3, knowledge capacity scaling laws. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=FxNNiUgtfa.

Peter Belcak, Greg Heinrich, Shizhe Diao, Yonggan Fu, Xin Dong, Saurav Muralidharan, Yingyan Celine Lin, and Pavlo Molchanov. Small language models are the future of agentic ai, 2026. URL https://arxiv.org/abs/2506.02153.

Louis Bethune, David Grangier, Dan Busbridge, Eleonora Gualdoni, Marco Cuturi, and Pierre Ablin. Scaling laws for forgetting during finetuning with pretraining data injection. In *Proceedings of the 42nd International Conference on Machine Learning*, ICML'25. JMLR.org, 2025.

Stella Biderman, USVSN Sai Prashanth, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raff. Emergent and predictable memorization in large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=TatRHT_1cK.

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: 354–80, 2006. URL https://api.semanticscholar.org/CorpusID:18831615.

Katie Everett, Lechao Xiao, Mitchell Wortsman, Alexander A. Alemi, Roman Novak, Peter J. Liu, Izzeddin Gur, Jascha Sohl-Dickstein, Leslie Pack Kaelbling, Jaehoon Lee, and Jeffrey Pennington. Scaling exponents across parameterizations and optimizers. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

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, Oriol Vinyals, Jack W. Rae, and Laurent Sifre. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

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, 2023. URL https://arxiv.org/abs/2207.00099.

Damjan Kalajdzievski. Scaling laws for forgetting when fine-tuning large language models, 2024. URL https://arxiv.org/abs/2401.05605.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020. URL https://arxiv.org/abs/2001.08361.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8424–8445, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.577. URL https://aclanthology.org/2022.acl-long.577/.

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. *J. Mach. Learn. Res.*, 21(1), January 2020. ISSN 1532-4435.

Chaofan Tao, Qian Liu, Longxu Dou, Niklas Muennighoff, Zhongwei Wan, Ping Luo, Min Lin, and Ngai Wong. Scaling laws with vocabulary: Larger models deserve larger vocabularies. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=sKCKPr8cRL.

Kushal Tirumala, Aram H. Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. Memorization without overfitting: analyzing the training dynamics of large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023a. URL https://arxiv.org/abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen

Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023b. URL https://arxiv.org/abs/2307.09288.

Greg Yang, Edward J. Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tensor programs v: tuning large neural networks via zero-shot hyperparameter transfer. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

A Model Configurations

Table 3: Model configurations. All models are LLaMA-style decoder-only Transformers with RoPE, RMSNorm, SwiGLU feed-forward layers, standard multi-head attention (head dimension 64), and untied input and output embeddings. N counts all parameters except the input and output embeddings. Each model is trained on $20N$ tokens with a batch of 32 sequences of 1,024 tokens; the peak learning rate is $3 \times 10^{-4} \cdot 256/d_{\text{model}}$.

Size	d_{model}	Layers	Heads	FFN dim.	N	Peak LR	Tokens	Steps
XS	256	4	4	688	3.16M	3.0×10^{-4}	63.3M	1,931
S	384	6	6	1024	10.62M	2.0×10^{-4}	212.4M	6,483
M	512	8	8	1376	25.31M	1.5×10^{-4}	506.1M	15,445
L	768	10	12	2048	70.80M	1.0×10^{-4}	1.42B	43,209
XL	1024	12	16	2736	151.22M	7.5×10^{-5}	3.02B	92,295

B Decay Fits for All Model Sizes

Figures 5 and 6 show the fitted conditional decay against binned empirical retention of memorized canaries for all five model sizes, under burst and spread exposure respectively. The full kernel surfaces for both regimes appear in Figure 1.

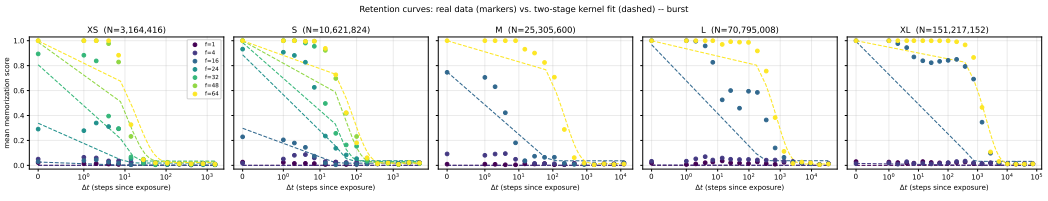


Figure 5: Fitted conditional decay against binned empirical retention of memorized canaries under burst exposure, for all five model sizes.

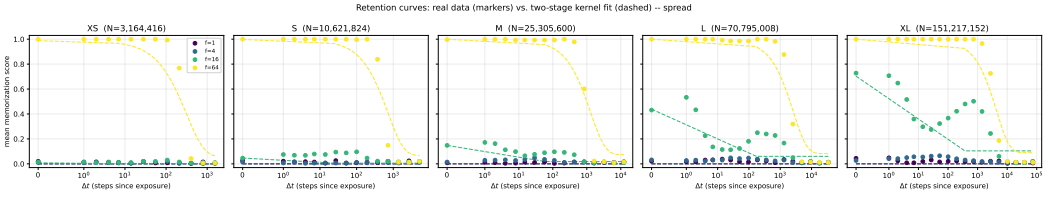


Figure 6: Fitted conditional decay against binned empirical retention of memorized canaries under spread exposure, for all five model sizes.

C Held-Out Generalization: Interpolation and Extrapolation

To test whether the dependence of the kernel on N reflects structure that transfers across scale, we hold out one model size at a time. All fits, including the onset model, the decay model, and the single-surface baseline, use the remaining four sizes, and the held-out size is evaluated without refitting. Holding out L, which lies between M and XL, tests interpolation; holding out XL, which lies about $2.1\times$ beyond the largest training size, tests extrapolation. Table 4 reports both tests for both sweeps. The onset stage is scored on per-canary outcomes at $\Delta t = 0$, the decay stage on per-cell mean scores of memorized canaries, and the combined kernel and the baseline on unconditional per-cell mean scores. Kernel-versus-baseline differences are computed with a paired bootstrap over held-out cells.

Table 4: Held-out evaluation. Each column fits every model on the four remaining sizes and evaluates the held-out size without refitting. Reductions and gains are paired kernel-versus-baseline differences on the same cells, with 95% bootstrap confidence intervals. f_{50} is compared with the value from a fit that includes the held-out size.

	Interpolation (L held out)		Extrapolation (XL held out)	
	Burst	Spread	Burst	Spread
Held-out cells	72	72	76	76
Onset pseudo-R^2 / acc.	0.97 / 99.8%	0.72 / 89.6%	0.96 / 99.8%	0.71 / 92.0%
Decay-only R^2	0.71	0.71	0.54	0.45
Kernel R^2 / MAE	0.94 / 0.050	0.96 / 0.049	0.79 / 0.077	0.92 / 0.070
Baseline R^2 / MAE	0.73 / 0.090	0.73 / 0.079	0.65 / 0.139	0.71 / 0.112
MAE reduction	0.040 [0.006, 0.081]	0.030 [−0.005, 0.072]	0.062 [0.028, 0.101]	0.042 [0.014, 0.079]
R^2 gain	0.21 [0.02, 0.49]	0.23 [−0.00, 0.60]	0.14 [−0.02, 0.38]	0.21 [0.03, 0.50]
f_{50} predicted / fitted	9.8 / 9.5	16.7 / 16.8	7.2 / 7.3	14.7 / 13.9
f_{50} error	2.9%	0.7%	1.9%	5.8%

The combined kernel remains accurate under both tests, with R^2 between 0.79 and 0.96, and predicts the onset threshold of the held-out size within 6%. It improves on the single-surface baseline in all four settings, significantly so on MAE in three of them; the exception, interpolation on the spread sweep, favors the kernel in 95.5% of bootstrap replicates. The two stages transfer differently. Onset generalizes well in both tests, whereas the decay stage alone loses accuracy under extrapolation (R^2 of 0.54 and 0.45 on XL, against 0.71 on L). The combined kernel stays accurate because onset accounts for most of the variation in unconditional retention: whether a canary is memorized at all matters more than how quickly a memorized canary fades. Extending the decay stage to larger scales is therefore the more pressing target for future work.

D Spacing and Retention After Onset

We compare conditional decay between regimes on canaries memorized under both, matched on model size, f, seed, and canary; $f = 1$ is excluded because the regimes coincide. Each trajectory is summarized by its mean score over post-exposure checkpoints, which, given their logarithmic spacing, is a log-time average of retention; we bootstrap the paired spread − burst difference over matched trajectories. The decay timescale τ is fit to each regime's mean retention curve. Of the 803 canaries memorized under spread exposure, 798 are also memorized under burst exposure, and unmatched comparisons give nearly identical results. Retention at the end of training does not differ between regimes in any configuration.

E Learning Rate and Injection Timing

We train XS with burst exposure at $f = 64$ (one seed), varying only the point in training at which canaries are injected, and compare the resulting onset and decay with the kernel fit to the main sweep, in which injection occurs at 20% of training. Table 6 reports the results.

Onset is unaffected down to 29% of the peak learning rate and degrades only at 12%, so a larger step at exposure never hinders memorization. Because $f = 64$ lies well above the onset threshold

Table 5: Decay of canaries memorized under both regimes. τ is the decay timescale in training steps. Δ is the spread $-$ burst difference in log-time-averaged retention, with 95% paired bootstrap confidence intervals.

Size	f	n	τ_{burst}	τ_{spread}	τ ratio	Δ [95% CI]
XS	64	128	17.9	375.1	21.0×	$+0.37$ [$+0.36$, $+0.38$]
S	64	128	69.8	759.0	10.9×	$+0.26$ [$+0.25$, $+0.27$]
M	64	128	216.2	1325.4	6.1×	$+0.19$ [$+0.17$, $+0.21$]
L	64	128	850.0	3086.1	3.6×	$+0.12$ [$+0.10$, $+0.13$]
XL	64	128	2031.9	5286.0	2.6×	$+0.08$ [$+0.07$, $+0.09$]
L	16	52	191.3	3.6	–	-0.12 [-0.19, -0.06]
XL	16	93	1353.9	6.0	–	-0.16 [-0.22, -0.10]

Table 6: Onset and decay as a function of injection point (XS, $f = 64$, burst). The learning rate is given as a fraction of its peak. τ is the decay timescale, in training steps, of an exponential fit to each injection point's memorized canaries alone. The residual is the mean difference between observed retention and the prediction of the kernel fit to the main sweep.

Injection	LR at exposure	Memorized	Memorized canaries	τ	Kernel residual
10%	98.2%	100%	64	9.0	-0.10
30%	82.3%	100%	64	12.3	-0.08
50%	55.7%	100%	64	16.9	-0.04
70%	28.9%	100%	64	5.7	-0.18
90%	12.2%	48%	31	4.9	-0.25

for XS, this test establishes the direction of the effect rather than its size near f_{50}. Later injection also coincides with a more-trained model and a shorter post-exposure runway, so the learning rate is not the only variable that changes. Decay, in contrast, depends on when exposure occurs: the timescale varies more than threefold across injection points, is shortest for the two latest injections, and the kernel fit to exposures at 20% of training overpredicts retention at every injection point, most strongly for late ones.

F Transfer to a Sparse Mixture-of-Experts Architecture

As a preliminary test of whether the kernel transfers across architectures, we apply the kernel fit to the LLaMA burst sweep, without refitting, to a sparse mixture-of-experts model (four SwiGLU experts per layer, top-2 routing) at size S, trained with burst exposure at $f \in \{1, 4, 16\}$ (one seed). Because only two of the four experts process each token, we evaluate the kernel with two notions of model size: total non-embedding parameters (31.9M) and active parameters per token (17.7M). With total parameters, the kernel predicts retention with $R^2 = 0.94$ (MAE 0.044) and predicts the memorized fraction at $f = 16$ as 0.86, against 0.89 observed. With active parameters, accuracy drops to $R^2 = 0.70$ and the predicted memorized fraction to 0.60. Memorization thus appears to track the parameters a model stores rather than those it activates per token. The model's fitted onset threshold ($f_{50} = 10.3$) lies below even the total-parameter prediction (12.5) and well below that of the dense model of matching width and depth (17.9), suggesting that sparse experts are particularly efficient at storing isolated facts. These results rest on a single size and seed and should be read as a direction for future work.

G Extended Limitations

Scale. Our models span 3.16M–151M non-embedding parameters, far below production scale. The dependence of the kernel on N is estimated from five sizes, four of which constrain each held-out fit; this suffices to estimate a slope but not to detect curvature in the assumed log-linear form. Holding out the largest model tests extrapolation beyond the trained range only by a factor of about two in N; there, the combined kernel remains accurate, but the decay stage alone extrapolates less

well than onset (Appendix C). Taken at face value, the fitted onset slopes predict a spacing penalty of roughly $2.4\times$ at 1B parameters and $3.1\times$ at 7B, a prediction that requires larger models to test.

Scale and step size. Because the peak learning rate scales as $1/\text{width}$, larger models also take smaller steps. The growing spacing penalty may therefore reflect step size as well as capacity; separating the two requires training a given size at another size's learning rate.

Functional form. The logistic onset model and the exponential-to-floor decay are assumed rather than derived. They fit well in-sample and on a held-out size, but other forms, such as decay that is linear in $\log \Delta t$, cannot be ruled out with the range of Δt available at small scale, where runways span only a few thousand steps for the smallest models.

Design choices. Each result is conditioned on fixed design choices: exposure ends at 20% of training in the main sweeps, the spread window covers 15% of training, and all exposures are full-batch steps on the 64 canaries together. The size of the spacing penalty may depend on the window width and on how canaries are mixed with other data, which we do not vary.

Architecture and data. All models share one architecture family (LLaMA-style) and one pre-training corpus (C4). Our canaries are synthetic facts of a fixed format, injected as isolated sequences; naturally occurring facts vary in form, context, and co-occurrence with related text.

Measurement. We measure memorization by free-running greedy decoding from the exact training prefix and label a canary memorized at a score of 0.5. This label is conservative with respect to format: 98% of canaries labeled memorized reproduce their value exactly. Other probes, such as sampling, paraphrased prompts, or likelihood-based extraction, may reveal partial knowledge that greedy decoding misses, and could shift onset thresholds.

Statistical power. Each configuration is trained with two seeds, one spread configuration (XL, $f = 1$) completed a single seed, and the injection-timing experiment uses a single size, repetition count, and seed. Held-out cells within a model size share one trained model and are therefore correlated, so bootstrap intervals over cells are likely somewhat narrow. The design has limited power for small effects: per-size intervals on the token-frequency effect all include zero, and on the spread sweep the kernel's improvement over the single-surface baseline does not reach significance at the 95% level.