When Is Forgetting a Capacity Problem? Dissecting Catastrophic Forgetting under Controlled Capacity Pressure

Anonymous authors
Paper under double-blind review

Abstract

Catastrophic forgetting is often attributed either to exhausted capacity or to optimisation interference despite sufficient capacity. The two explanations imply different remedies, and telling them apart calls for a measured capacity axis, which natural data do not provide: neither the stream's information content nor the model's effective capacity is directly known. To place forgetting on this axis, we build a continual-learning testbed from random facts, rules and noise, with known information content and calibrated model capacity. We define load as learnable information relative to measured capacity and vary it while holding each fact's training exposure fixed, which separates capacity limits from under-training. Three findings emerge in the testbed. First, storage of the old facts fails once they alone exceed capacity. Up to a load of 1 they are still stored almost in full, but as load grows a new linear readout needs more examples to decode them, and the model's own readout decodes them with narrower margins. Second, far below capacity, new facts in the old facts' format collapse old-fact recall within tens of steps, while less than 3% of their information is stored; training without new facts erodes recall much more slowly. During warm-up, collapse comes at a fixed cumulative learning rate, once the parameters have moved by about 1% of their norm; random perturbations of that size leave recall intact. Third, information survives the loss of recall and fades on separate clocks: the model's ranking of the correct answer outlasts exact recall, and brief retraining on a few old facts still restores others (a recoverable trace) once that ranking is at chance. During the collapse, unseen facts also become transiently hard to learn, while old ones stay easy to relearn. Among the consolidation methods we compare at three loads, utility-weighted replay retains the most useful facts and gives the highest future-task utility at each, and, at matched weighting, replay outperforms sleep-phase distillation with less compute. Weighting redirects retention toward useful facts even when all facts fit, but no method stores more than the measured plateau. A pretrained language model (Pythia-160M) fine-tuned on natural-language facts reproduces the collapse, its onset at a fixed cumulative learning rate and the recoverable trace; there, rendering the new facts as statements rather than in the old facts' question–answer format delays the collapse.

1 Introduction

For a learner with finite capacity, continual learning involves deciding how new experience should change what is already known. Some experiences contribute independent facts; others share structure that can be consolidated into more compact representations. The objective is to retain, combine and revise knowledge according to its usefulness for future tasks. Losing detail can be acceptable if useful capabilities are preserved, whereas losing knowledge that will be needed again is costly. Evaluating forgetting therefore requires knowing both what information is lost and what that loss costs in future performance.

Knowing what is lost requires a diagnosis of observed forgetting. Capacity constrains how much information can be retained (Allen-Zhu & Li, 2025; Morris et al., 2025). Interference is a long-standing explanation of forgetting (McCloskey & Cohen, 1989; French, 1999); it can occur even

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.

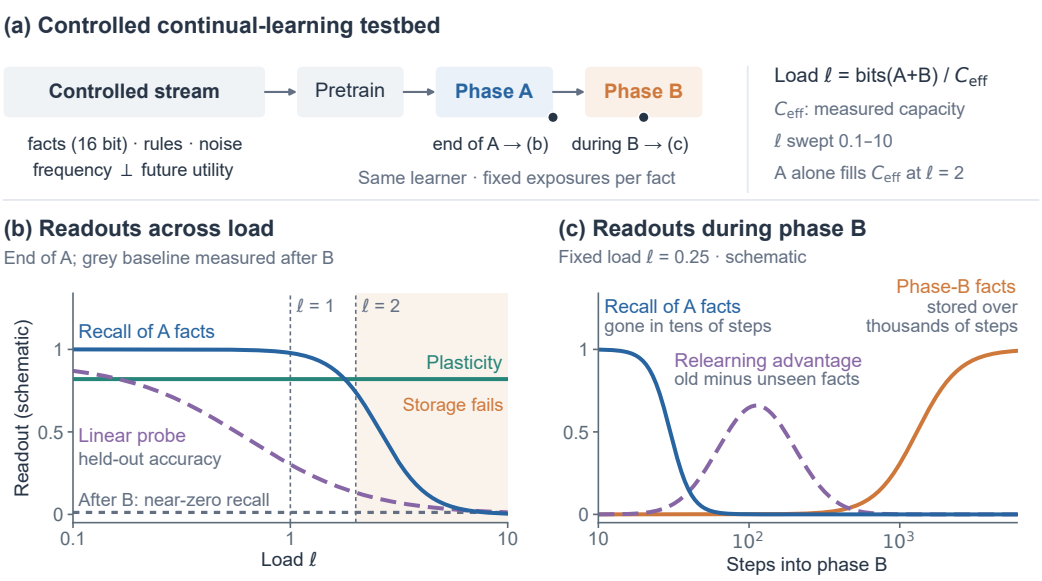

Figure 1: Behavioural forgetting precedes storage failure. (a) The controlled stream and measured load $\ell = \text{bits}(A{+}B)/C_{\text{eff}}$; A alone fills capacity at $\ell = 2$. (b) Schematic end-of-A readouts: accuracy of a probe fitted on 4,200 facts declines from the lowest load, recall stays high until phase A alone exceeds capacity (shaded), and plasticity (unseen-fact learning) stays high throughout; recall after a full phase B (grey) is near zero at every load. (c) At $\ell = 0.25$, recall collapses before much of B is stored; old facts stay easy to relearn while unseen facts are transiently hard to learn. Measurements: Figures 2, 3 and 6.

when a joint solution exists (Evron et al., 2022). In some settings, brief retraining or a change in prompting restores behaviour that appeared lost (Zheng et al., 2025; Kotha et al., 2024; Hu et al., 2025). Failed recall alone therefore does not establish complete information loss. These distinctions matter for intervention: interference calls for protecting knowledge from harmful updates, a surviving trace for restoring access, and a capacity limit for choosing what to retain. To tell which applies, we need to relate the information being learned to the model's effective capacity and examine what survives a loss of recall. In natural data, neither the stream's information content nor effective capacity is directly known, and usefulness can be confounded with frequency.

A testbed for capacity and utility. We build a continual-learning stream that mixes random facts of known bit content with compositional rules and noise, measure effective capacity under a fixed exposure budget, and define load as learnable information relative to that capacity; utility and frequency are assigned independently, so useful retention is separable from repeated exposure (Section 2).

Diagnostics of forgetting. Complementary readouts separate failed recall, recoverable traces and storage failure (Figure 1). Storage fails once the old facts alone exceed the measured capacity, while a new linear readout needs more examples to decode stored values as load grows (Section 3); far below capacity, same-format updates (new facts in the old facts' format) collapse old-fact recall within tens of steps, before much new information is stored, while the model's ranking of the correct answer and a recoverable trace outlast recall (Section 4).

Comparisons under capacity pressure. Comparing replay and distillation under uniform and utility-weighted objectives, we find that utility-weighted replay retains the most useful facts at every load tested, including far below capacity; at matched weighting, replay beats sleep-phase distillation, and no method stores beyond the measured plateau. EWC and gradient projection (OGD, GPM) keep no phase-A facts, whereas displacement-triggered replay keeps more of them the more often it fires, with a knee, at $\ell = 0.25$, near the displacement at which recall collapses (Section 5).

Replication in a pretrained model. Using synthetic facts expressed in natural language, we reproduce in a pretrained language model both recall collapse under same-format updates and recoverable traces, showing that these interference signatures extend beyond the testbed (Section 6).

2 A MEASURED AXIS FOR FORGETTING

World. To measure forgetting in bits, we need a stream with known information content. We construct it from random facts, compositional rules and noise, presented in a pretraining phase followed by phases A and B, which have disjoint facts and rule pairs; forgetting is the loss of phase-A knowledge during B (Figure 1a). Each fact maps a key (a two-token entity and an attribute) to a two-token value. Each value token is drawn independently and uniformly from a vocabulary of 256, and the value is fixed across occurrences. A fact therefore carries exactly 16 bits given its key. Its value cannot be inferred from other facts: the only way to score on it is to store it.

Rules separate storage from abstraction. Each trained rule applies one atomic string transform or an ordered pair of them to inputs from a seeded grammar; held-out pairs and triples test whether rules are stored as instances or learned as transforms (Appendix A). Noise consists of freshly sampled random symbol strings and carries no learnable bits.

Utility and frequency. In natural data, an item's frequency and usefulness are confounded, so whether a method keeps useful facts or merely frequent ones cannot be tested. We separate the two by assigning utility and frequency classes through independent random permutations. High-utility facts (half of each phase) are queried in the future-task evaluation set D_{future}; zero-utility facts are never queried, so forgetting them is benign by construction. Low-frequency facts appear E times per phase, where E is the base exposure count; high-frequency facts appear $4E$ times. U_{future} is the equal-weight mean of seven D_{future} accuracies: high-utility facts and new inputs for the rule compositions of each phase (pretraining, A and B), and held-out composition pairs (Appendix A).

Capacity pressure. To compare forgetting across model sizes (randomly initialised six-layer decoder-only transformers with 5M and 11M parameters; Appendix C), we define load, $\ell = \text{bits}(A{+}B)/C_{\text{eff}}$, as learnable information relative to the measured effective capacity C_{eff}. Learnable bits are $16N_F$ plus a negligible description length for the grammar and transforms, where N_F is the number of distinct facts in A and B. A and B contain equal numbers of facts, so phase A alone has load $\ell_A = \ell/2$: $\ell = 1$ means that A and B together match the measured capacity, and $\ell = 2$ that phase A alone does. Results from the two model sizes align on this axis (Section 3).

We measure C_{eff} as the plateau of stored bits over increasing numbers of facts, under the shared pretraining and fixed-exposure protocol described below. Stored bits are computed from the loss (Allen-Zhu & Li, 2025): $\sum_i \max(0, 16 - \text{nll}_i/\ln 2)$, where nll_i is the total negative log-likelihood, in nats, of the two value tokens of fact i given its key; subtracting this loss in bits from 16 credits partial as well as complete memories. We adopt $C_{\text{eff}} = 6.5$ and 12 Mbit for the 5M and 11M models (observed plateaus 6.1–6.7 and 11.4–12.3 Mbit), or 1.3 and 1.1 bits per parameter. Doubling exposure or training A+B jointly at $\ell = 2.5$ leaves the 5M model's plateau at 6.2–6.6 Mbit, or 1.2–1.3 bits per parameter, within the 1.0–1.4 of a separate facts-only calibration (Table 4).

Controlling learning kinetics. A storage plateau is a capacity limit only if every fact receives enough exposures. Recall stays near zero for many exposures and then jumps, and the exposures needed depend on the number of facts and on model size, non-monotonically for the 11M model; more optimiser steps at smaller batches do not substitute for them (Appendix B). Under a fixed token budget, incomplete learning would thus look like a capacity limit. The sweep instead fixes exposures per fact, batch size and schedule, so tokens and steps grow with the number of facts. Every run starts from the same pretraining phase (10k facts at $E = 400$), after which $E = 60$ in phases A and B gives 97.5–100% end-of-A recall up to $\ell = 1$ (Section 3; Appendix C).

Four readouts of phase-A knowledge. We measure phase-A knowledge at the end of A and track it through phase B with four complementary readouts (Figure 1b,c).

(i) *Retention* measures exact recall (both value tokens correct) on the high-utility half of phase-A facts; relative retention divides it by end-of-A recall. Under a uniform loss, both utility halves are forgotten alike.

(ii) *Recovery* tests whether briefly fine-tuning (200 steps) on 1% of phase-A facts restores recall on the other phase-A facts. *Net recovery* is that recall minus the recall after the stronger of two equal-step controls, fine-tuning on unseen facts (keys never trained) or on noise; a positive value

indicates a *recoverable trace*. *Relearning advantage* is recall on the fine-tuned old facts minus recall on equally many unseen facts trained for the same number of steps. *End-of-A plasticity* is recall on a fixed set of 500 unseen facts after 200 training steps from the end-of-A checkpoint.

(iii) *Probe accuracy* is the accuracy of a linear classifier that predicts the first value token from the frozen residual stream at the key–value separator `[sep]`, trained on 4,200 facts and tested on 1,800 others; shuffled labels serve as a control. The model's final LayerNorm and output head have no bias and already form a linear classifier on this residual (Appendix H); probe accuracy therefore measures how readily such a readout is learned from 4,200 examples under a fixed procedure.

(iv) *Eviction structure* asks which facts are lost. Among facts recalled at the end of A, we relate later forgetting to utility, frequency and the position of their last occurrence in A (recency).

3 Forgetting along the axis

Recall after B is lost at every load; storage fails once phase A alone exceeds capacity. The sweep runs ℓ from 0.12 to 9.9 for the 5M model (Figures 2 and 5; two to ten seeds per load). Up to $\ell = 1$, nearly all phase-A facts are learned (97.5–100% recall, 15.8–16.0 of 16 bits per fact), yet at every load recall falls to near zero after a full phase B (Figure 3a), so end-of-B retention alone cannot distinguish a capacity limit from interference. Storage fails once the phase-A facts alone exceed the measured capacity ($\ell_A > 1$): at $\ell = 2.5$ ($\ell_A = 1.2$) the 500k phase-A facts average 10.9 bits and half of them are learned; at $\ell = 9.9$ the model's 6.1 Mbit are spread over two million facts (3.1 bits per fact), so exact recall is 0.2%.

Before storage fails, a new linear readout needs more examples as load grows, while recall and plasticity stay high. Probe accuracy falls from 0.88 to 0.36 between $\ell = 0.12$ and 1 (s.d. 0.01–0.03) while exact recall stays at 97.5–100%, and an MLP probe fares no better (Table 6). The values stay linearly decodable: trained on all other phase-A facts, the probe comes within 0.02 of the model's own accuracy. At a fixed training size, however, its accuracy still falls with load (0.99 to 0.65 with 10,000 facts; Table 7), while the model's own readout decodes the values with narrower margins. End-of-A plasticity stays flat: after 200 steps on 500 unseen facts, accuracy is 0.75–0.87 across the 5M sweep and 0.94–0.98 for the 11M model.

Two model sizes align on the measured axis. The 11M sweep (two seeds per load) follows the 5M recall and probe curves when plotted against measured load (Figure 2; Appendix J), whereas at equal fact counts its probe is more accurate (0.88, 0.75, 0.53 against 0.75, 0.60, 0.36 at 50k, 100k, 200k facts per phase), so the curves depend on load rather than on the number of facts. Across the sweep, two failures stand apart: recall after B is lost at every load, whereas storage fails once phase A alone exceeds capacity; probe accuracy, meanwhile, declines gradually throughout.

4 Dissecting forgetting below capacity

Below capacity, end-of-B retention hides the dynamics: training on new facts that share the old facts' template and answer vocabulary (*same-format* updates) collapses old-fact recall (Figure 3a) before much new information is stored, while answer ranking and recovery can still reveal surviving information. We fix $\ell = 0.25$ (the 5M model, 50k facts per phase) unless stated otherwise and change one thing at a time; Table 1 summarises the controls.

Both fact sets fit, but small directed updates disrupt recall. Joint A+B training from the shared checkpoint learns both sets in full at $\ell = 0.25$ (Appendix F.1). Under sequential training, recall nevertheless collapses once $\|\theta_t - \theta_A\|_2$ reaches 1.5–2.2, about 1% of $\|\theta_A\|_2 = 143$; we call this distance the *collapse displacement*. Isotropic or per-tensor norm-matched Gaussian perturbations of that size leave recall at 1.00, and random displacements must reach 13–15 before recall drops below 0.9: the trained direction collapses recall where random directions of the same size do not. The collapse displacement shrinks with load (2.2 at $\ell = 0.12$, 0.16 at 4.9), and above capacity random perturbations also become damaging.

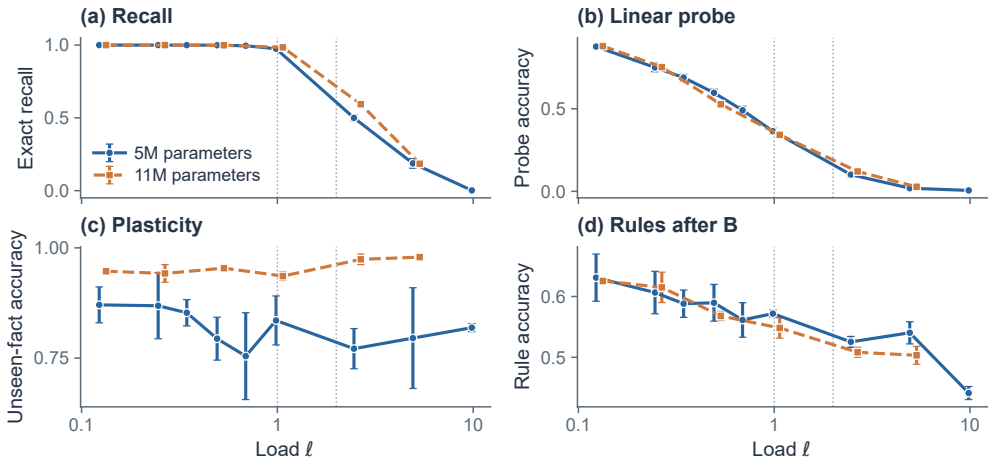

Figure 2: Two model sizes against measured load ℓ (means and standard deviations across seeds). (a) Exact recall of phase-A facts at the end of A. (b) Linear-probe accuracy for the first value token at the end of A (6,000 facts, `[sep]`, best layer). (c) Accuracy on 500 unseen facts after 200 fine-tuning steps from the end-of-A checkpoint. (d) Accuracy on phase-A rule compositions after B. Dotted lines: $\ell = 1$ and $\ell = 2$ ($\ell_A = 1$).

Question	Control	Observation
Do both sets fit?	Joint A+B training from the shared checkpoint.	Both sets are fully recalled at $\ell = 0.25$ (1.6 Mbit).
Does a random displacement of equal size collapse recall?	Gaussian perturbation matched to the trained displacement norm.	Recall remains 1.00 at the collapse displacement.
Does any training collapse recall as fast?	Phase B with rules and noise, but no new facts.	Collapse is about eight times slower and occurs at displacement 22.
Does a lower learning rate prevent collapse?	10^{-4}–10^{-3} learning rates; 100–1,000 warm-up steps.	Lower rates delay collapse, which always comes at a displacement of 1.5–2.2.

Table 1: Controls for recall collapse below capacity in the 5M testbed (Appendix F). Displacement is $\|\theta_t - \theta_A\|_2$; the collapse displacement is its value at collapse, when recall falls below 0.1.

New facts in the shared format drive the collapse. Removing new facts from phase B while retaining rules and noise slows collapse about eightfold. With only 300 distinct new facts, 91–95% of phase-A facts are still recalled after 34 steps, against 6–11% with a full phase B. Sharing 25% or 50% of entities between the phases neither speeds nor slows the collapse (Appendix F.2). Fast collapse thus comes from many distinct same-format facts, even though at collapse they hold only 0.24–0.27 bits each, out of 16 (Appendix F.3); Section 6 varies the format itself.

Collapse comes at a fixed cumulative learning rate. Lowering the peak learning rate from 10^{-3} to 3×10^{-4} and 10^{-4} delays collapse from 34–36 steps to 65–70 and 107–110 (interpolated between checkpoints), all within the $W = 300$-step linear warm-up; holding the cumulative learning rate $\mathrm{lr}\, t^2/(2W)$ at its value at the 10^{-3} collapse predicts 62–66 and 108–114 steps, and a 100-step warm-up shifts collapse the same way. In every case the collapse displacement is 1.5–2.2 (Appendix F.2). A lower learning rate buys time; it does not prevent the collapse.

What survives after recall collapses. At $\ell = 0.25$ (two seeds), recall falls to 6–11% by step 34, yet the correct value still ranks far above chance among the 256 value tokens (median rank 19 for the worse of a fact's two tokens, against 171–184 with values shuffled across facts); the ranking reaches this shuffled baseline around step 114. Brief retraining on 1% of old facts recovers 4–10% of the other old facts beyond the controls between steps 34 and 114, still 2–7% at step 171, when the ranking is already at chance, and at most 1% by step 256 (Appendix F.3). Forgetting is therefore a

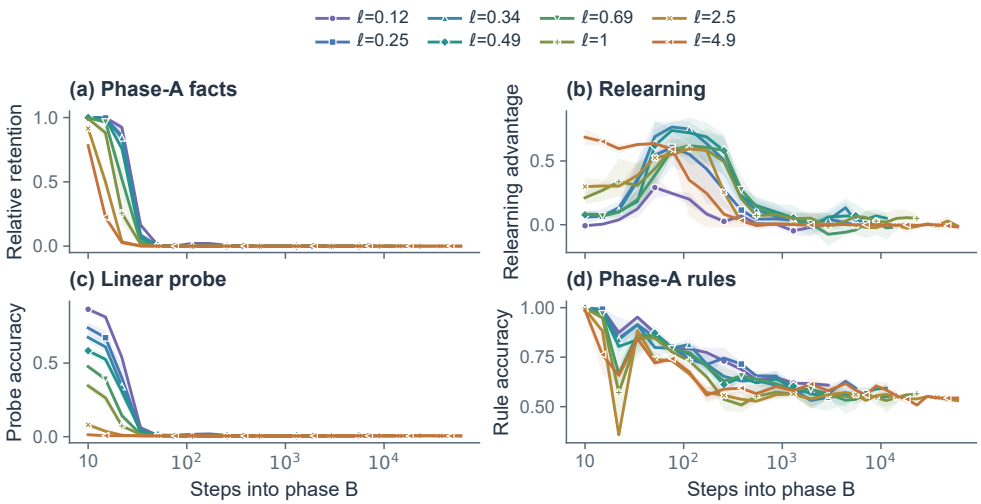

Figure 3: Readouts through phase B at each load ℓ (5M model, naive training; seed means, shading min–max after interpolation in log steps). (a) Phase-A recall divided by its end-of-A value. (b) Relearning advantage after 200 fine-tuning steps: recall on the 1% fine-tuning set of phase-A facts minus recall on equally many unseen facts; the unseen-fact set grows with load and is harder to learn when larger, so levels compare within a load, not across loads (Figure 6). (c) Linear-probe accuracy for the first value token (6,000-fact probe). (d) Phase-A rule-composition accuracy. Panels a, b and d use 21 runs; c uses 26 (2–8 per load).

sequence with separate clocks: exact recall is lost by step 34, most of the ranking signal by step 114 and the recoverable trace by step 256.

New learning is also transiently impaired during the collapse. Splitting the relearning advantage (Figure 3b) into its two raw curves (Appendix Figure 6) shows that at $\ell = 0.25$ old facts stay easy to relearn throughout phase B, whereas unseen facts become hard to learn during the collapse: the 200-step budget teaches 0.82–0.85 of them from the step-10 checkpoint but only 0.10–0.63 from step 51, and learning recovers by about step 600. At this load the relearning advantage therefore rises and falls mainly with this transient loss of plasticity; at $\ell = 1$ and 2.5, by contrast, old facts themselves become hard to relearn, reaching the unseen-fact level by 577 and 384 steps (Table 5).

5 Consolidation under capacity pressure

Under uniform training, low-frequency facts are lost earlier (point-biserial correlation r of forgetting with high frequency -0.18 to -0.30 at the first checkpoint with partial forgetting, $\ell \leq 2.5$), whereas utility, which a uniform loss does not see, plays no role ($|r| \leq 0.02$; Table 5): what survives follows exposure, not usefulness. We therefore ask which methods protect old facts, whether utility weighting redirects retention toward useful ones, and whether any stores beyond C_{eff}.

Operator and arms. To compare distilling from a buffer with replaying it, we use a wake/sleep operator: wake trains on the incoming phase while a buffer keeps 5% of the previous phases' 128-token training windows; sleep then distils two frozen teachers, the pre-wake model on buffer windows and the post-wake model on new windows, into a same-size student for a quarter of the wake steps, optionally weighting each token by its utility (Progress & Compress's distillation step without its EWC penalty; Appendix J). Utilities are oracle labels throughout. The eight arms are naive sequential training; uniform replay, which fills 20% of each batch from the buffer, for 1.0 epoch (the steps of naive training) or 1.42 epochs, bracketing the compute of sleep; uniform and utility-weighted sleep; utility-weighted replay (same buffer and weights, no distillation); and online EWC (Schwarz et al., 2018) with a diagonal Fisher at $\lambda = 0.001$ and 0.01. The arms already act during phase A, so end-of-A recall differs across them (Table 8). All arms have the same parameter count, and compute counts every forward pass, the teachers' included; Table 2 reports $\ell = 0.25$, 1 and 2.5.

	$\ell \approx 0.25$			$\ell \approx 1$			$\ell \approx 2.5$		
Method	Useful A	B (all)	U_{future}	Useful A	B (all)	U_{future}	Useful A	B (all)	U_{future}
Naive	0.00	1.00	0.48	0.00	0.96	0.45	0.00	0.51	0.37
Replay (1.0 epoch)	0.70	1.00	0.87	0.21	0.83	0.74	0.01	0.45	0.61
Replay (1.42 epochs)	0.78	1.00	0.88	0.26	0.87	0.76	0.01	0.46	0.62
Uniform sleep	0.42	0.88	0.79	0.11	0.47	0.67	0.00	0.17	0.58
Utility-weighted sleep	0.60	0.69	0.84	0.31	0.31	0.74	0.03	0.13	0.64
Utility-weighted replay	0.87	0.85	0.91	0.67	0.60	0.88	0.18	0.37	0.75
EWC, $\lambda = 0.001$	0.00	0.95	0.51	0.00	0.57	0.42	0.00	0.00	0.41
EWC, $\lambda = 0.01$	0.00	0.50	0.46	0.00	0.00	0.41	0.00	0.00	0.41

Table 2: End-of-B results by measured load. Useful A: recall on the high-utility half of phase A, not normalised by end-of-A recall, which differs across arms; B (all): recall on all phase-B facts. Seed means; end-of-A recall, seed counts, standard deviations and compute are in Table 8.

Utility-weighted replay performs best; at matched weighting, replay beats sleep. Utility-weighted replay gives the highest useful-A recall (0.87, 0.67, 0.18) and U_{future} (0.91, 0.88, 0.75) at every load (Table 2) with 20–28% less compute than sleep. Replay retains more useful A facts than sleep whether both are uniform or both are weighted, and sleep also learns fewer B facts (Table 2). EWC keeps no A facts, and not only because its penalty also hampers learning A (Table 8): from a common end-of-A checkpoint it loses A as fast as naive training does (Appendix F.4).

Utility weighting redirects retention toward high-utility facts. Uniform sleep and replay forget high- and zero-utility facts alike; weighted sleep keeps almost no zero-utility A facts and raises useful-A recall over uniform sleep (0.60, 0.31, 0.03 against 0.42, 0.11, 0.00). Weighting helps even at $\ell = 0.25$, where both phases fit: with a finite training budget, the weights shape what survives (useful-A recall 0.87 for utility-weighted replay against 0.70 for uniform replay at equal compute). Its B (all) recall trails uniform replay's only on zero-utility B facts (0.71, 0.21, 0.00, against 1.00, 0.99, 0.74 for high-utility ones; Table 8), which D_{future} never queries.

Consolidation keeps rules but does not enlarge the store. Every replay and sleep arm keeps phase-A rule compositions almost entirely (0.96–1.00), even at $\ell = 2.5$ where fact recall is low, whereas naive training ends at 0.53–0.61: a uniform buffer re-presents each training composition hundreds of times but each fact only about three times. These arms also generalise far better to held-out compositions at the end of B (0.33–0.53 against 0.02–0.10 for naive training). No method stores beyond the measured plateau (6.1–6.7 Mbit): at $\ell = 2.5$, naive training and uniform replay hold 5.4–5.7 Mbit by the loss-based measure and joint training 6.6, with different allocations across facts, while utility-weighted replay (4.9) and the sleep arms hold less (Appendix J).

Projection keeps nothing; displacement-triggered replay tracks the collapse. Collapse follows a small directed displacement, so we first block candidate directions: projecting updates off the span of phase-A LM-loss gradients (an OGD variant; OGD itself uses output gradients, Farajtabar et al., 2020) or off each internal linear layer's phase-A input subspace (GPM, Saha et al., 2021; embeddings, LayerNorms and the tied output head stay free) retains no phase-A facts at $\ell = 0.25$ and 1. Collapse also comes at a fixed displacement, so we replay the buffer for five steps each time the parameters have moved a further L2 distance d. At $\ell = 0.25$, bursts at $d = 0.5$, 2, 4, 8 and 16 keep 0.93, 0.78, 0.52, 0.12 and 0.00 of phase A, with a knee near the collapse displacement (1.9); at $\ell = 1$ and 2.5, where that displacement is 0.96 and 0.46, even $d = 0.5$ keeps only 0.54 and 0.16. In Pythia-160M the bursts hold phase A early in phase B only for d below its collapse displacement of 2.3 (Figure 7; Appendix F.4). At similar compute they do no better than uniform replay (Figure 7b).

6 The same dissection in a pretrained language model

The testbed isolates mechanisms, but its facts, model and stream are synthetic. We therefore repeat the interference-regime dissection on a pretrained Pythia-160M (Biderman et al., 2023) with facts of known information content written in natural language. Each of 10,000 synthetic persons has five attributes drawn uniformly from fixed tables and rendered as question–answer lines (43 bits per

person, 434 kbit in total), with a quarter of the persons high-frequency (four times the exposures). Phase A mixes these facts with FineWeb-Edu text (Penedo et al., 2024) at equal token share for 100 exposures; phase B replaces them with the facts of 10,000 new persons (both sets can be learned jointly; Appendix K) or with text alone. The readouts are those of Section 2 (Appendix K). At phase-B learning rate 10^{-4}, the 10,000- and 50,000-person runs have a second seed, which reproduces their recall and end-of-A probe accuracy to within 0.03; other conditions have one seed.

Memorisation kinetics, the collapse and its timing all carry over. Recall of the phase-A facts stays below 0.03 for 24 exposures and reaches 0.999 at 84, with high-frequency facts learned first (Appendix Figure 9a): the plateau-then-jump kinetics of the testbed (Appendix B). In phase B, phase-A recall collapses within tens of steps, at 17, 30 and 55 steps for learning rates 3×10^{-4}, 10^{-4} and 3×10^{-5} (Figure 4a). All three collapses occur during the 100-step linear warm-up at the same cumulative learning rate (4.6×10^{-4}), as in the testbed. Probe accuracy falls with recall, from 0.29 to chance, and an MLP probe does no better; low-frequency facts go first.

The recoverable trace is stronger in a pretrained model. After collapse, fine-tuning on the facts of 1% of the persons for 30 steps restores 0.33–0.37 of the *other* persons' facts, against 0.02–0.03 for the same budget on unseen persons or shuffled answers (Figure 4b); in the testbed, net recovery at $\ell = 0.25$ is 0.04–0.10 (Section 4). The trace is consistent with reports of shallow forgetting in fine-tuned language models (Zheng et al., 2025; Hu et al., 2025), and it has a lifetime: net recovery halves after about 150, 500 and 2,000 steps, roughly in inverse proportion to the learning rate.

Same-format competition accelerates the collapse. With text alone in phase B the facts survive far longer (recall 0.93 after 171 steps, 0.56 after 577, 0.03 after 5,000; Figure 4a), probe accuracy falls in step with recall, and net recovery is still 0.09 at the end. Rendering the same new facts as statements instead of question–answer lines keeps phase-A recall at 0.75 after 34 steps (0.08 as question–answer lines); at matched new-fact storage the statement format keeps more of phase A while the new facts are little stored (0.26 against 0.04 at 0.7 bits per fact), and the two formats converge once the new facts are well stored (Appendix Figure 10). Same-format new facts collapse old-fact recall while holding only about 0.25 bits each, out of 8.7; unrelated text erodes the store about thirty times more slowly (eightfold in the testbed).

More stored facts, later collapse and a stronger trace. With 50,000 persons (250k facts, 2.2 Mbit, about 0.01 bits per parameter), collapse comes later (recall 0.28–0.30 at 34 steps against 0.08), probe accuracy at the end of A is higher (0.60 against 0.29) and the trace is stronger and more persistent (net recovery 0.58–0.67 at its peak, 0.13–0.14 through 5,000 steps, against 0.02). Probe accuracy thus rises with world size here but falls with load in the testbed; the two probes, however, read different targets (attribute classes here, the first value token there; Appendix H).

7 Related work

Interference and recoverability are established themes in continual learning (McCloskey & Cohen, 1989; French, 1999; Mirzadeh et al., 2020; Zheng et al., 2025; Kotha et al., 2024; Hu et al., 2025); we relate their measured dynamics to information load and future-task utility. Marek et al. (2026) jointly study capacity, optimisation and replay in language models, with proxies for capacity. We instead measure known factual information against a protocol-calibrated storage plateau and track ranking, recovery and relearning alongside recall. Synthetic worlds already measure knowledge capacity and distinguish storage from extraction (Allen-Zhu & Li, 2025; 2024; Morris et al., 2025); our non-stationary stream additionally assigns utility independently of frequency. Replay and distillation supply consolidation mechanisms (Schwarz et al., 2018; Buzzega et al., 2020; Arani et al., 2022), which we compare under uniform and utility-weighted objectives. Importance weights shape early training but lose much of their effect as optimisation continues, which L2 regularisation partly offsets (Byrd & Lipton, 2019); consolidation, by contrast, runs on a finite rehearsal budget, within which utility weighting still redirects retention (Section 5). Appendix I expands on these connections.

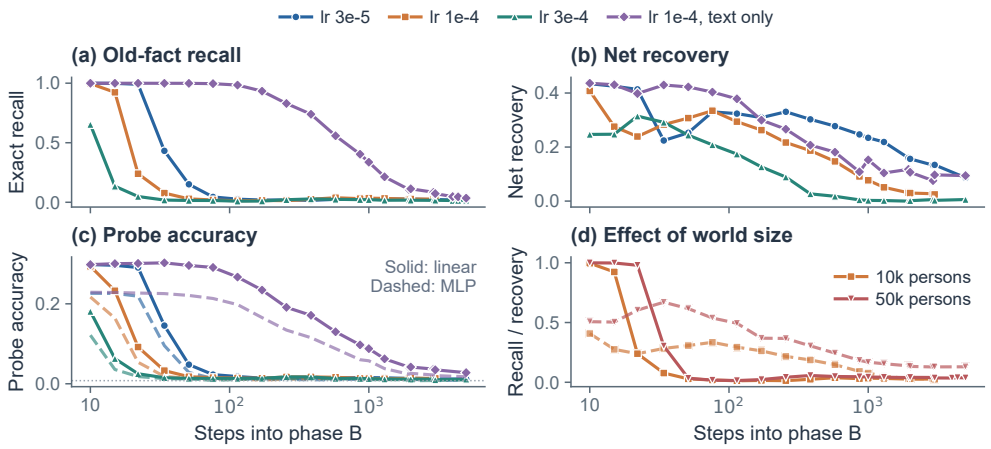

Figure 4: Interference diagnostics in Pythia-160M (one seed shown per condition; 10,000 persons unless stated). (a) Recall of A facts during B. (b) Net recovery on other persons after fine-tuning on 1% of A persons, relative to an equal-step unseen-person control; before recall falls in (a), it mainly reflects recall lost to the control fine-tuning. (c) Attribute-value probe accuracy: linear (solid), MLP (dashed), mean chance (dotted). The shared legend applies to a–c. (d) Recall (solid) and net recovery (dashed) for 10,000 and 50,000 persons at learning rate 10^{-4}.

8 Discussion and Conclusion

Recall after a new phase is lost at every load, whereas the old facts are stored almost in full up to $\ell = 1$ and storage fails once they alone exceed C_{eff}. Where old and new facts jointly fit, same-format updates collapse recall at a small directed displacement, before much new information is stored, and what survives fades on measurable clocks: the model's preference for the correct answer returns to chance before the trace that brief retraining can recover is gone. The timing of consolidation is therefore a concrete design question. Consolidation can act before the collapse, where displacement-triggered replay retains less once bursts are spaced beyond the collapse displacement, or after it, where brief retraining restores part of what recall has lost until the trace fades. Long before it limits storage, load narrows the margins of the model's own readout and raises the number of examples a new linear readout needs; whether the narrower margins set that need remains open.

Consolidation also needs a criterion for what to retain, even when everything fits. Under uniform training, low-frequency facts disappear earlier regardless of their assigned utility; utility weighting redirects retention toward knowledge needed by future tasks, and utility-weighted replay does this best, without any distillation step. These findings motivate treating continual learning as a recurring process of preserving useful knowledge, restoring access while traces remain, and reallocating a fixed storage budget. A fuller consolidation cycle would additionally combine shared structure into more compact representations; developing and testing that cycle remains future work.

Limitations: the world is synthetic and small, and the pretrained-model replication covers only the interference regime (Appendix K); C_{eff} is the storage scale of our training protocol, and effective capacity is non-stationary (Chakraborty & Raghavan, 2025); our two model sizes store similar bits per parameter, so load is confounded with facts per parameter; the testbed controls do not separate competition in the shared template from that in the shared answer vocabulary; we run neither Progress & Compress itself nor parameter-isolation methods; D_{future} is an oracle; we study two phases, not long streams; results at 5M–160M parameters are trends, not predictions for frontier models.

Reproducibility Statement

Appendix A specifies the seeded data generator; Appendices C and D give the training settings, compute and capacity calibration. The diagnostic and consolidation protocols are detailed in Appendices F, H and J, and Appendix K describes the pretrained-model replication. The generator, training loop, probes and experiment configurations will be released.

AI USE STATEMENT

Generative AI tools assisted with literature retrieval and organisation, code implementation and revision, and manuscript polishing. This included code for the synthetic-data generator, training, evaluation and figures, as well as discussions of experimental methodology and the presentation of results. The authors designed the experiments, made all methodological and experimental decisions, and analysed and interpreted the results. They reviewed the AI-assisted material, executed and checked the experimental code, and verified the reported findings. The authors take full responsibility for the final content.

REFERENCES

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.1, knowledge storage and extraction. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 1067–1077, 2024. URL https://proceedings.mlr.press/v235/allen-zhu24a.html. arXiv:2309.14316.

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.3, knowledge capacity scaling laws. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/26d3c9a66836ded8f34a944f2bfe868e-Abstract-Conference.html. arXiv:2404.05405.

Elahe Arani, Fahad Sarfraz, and Bahram Zonooz. Learning fast, learning slow: A general continual learning method based on complementary learning system. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2201.12604. arXiv:2201.12604.

Joshua Barron and Devin White. Too big to think: Capacity, memorization, and generalization in pre-trained transformers. *arXiv preprint arXiv:2506.09099*, 2025. URL https://arxiv.org/abs/2506.09099.

Tina Behnia, Puneesh Deora, and Christos Thrampoulidis. Facts in stats: Impacts of pretraining diversity on language model generalization. *arXiv preprint arXiv:2510.16096*, 2025. URL https://arxiv.org/abs/2510.16096.

Ali Behrouz, Farnoosh Hashemi, Adel Javanmard, and Vahab Mirrokni. Language models need sleep: Learning to self-modify and consolidate memories. *arXiv preprint arXiv:2606.03979*, 2026. URL https://arxiv.org/abs/2606.03979.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430, 2023. URL https://proceedings.mlr.press/v202/biderman23a.html.

Cody Blakeney, Mansheej Paul, Brett W Larsen, Sean Owen, and Jonathan Frankle. Does your data spark joy? performance gains from domain upsampling at the end of training. *arXiv preprint arXiv:2406.03476*, 2024. URL https://arxiv.org/abs/2406.03476.

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. In *Advances in Neural Information Processing Systems*, 2020. URL https://arxiv.org/abs/2004.07211. arXiv:2004.07211.

Jonathon Byrd and Zachary Lipton. What is the effect of importance weighting in deep learning? In *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 872–881, 2019. URL https://proceedings.mlr.press/v97/byrd19a.html. arXiv:1812.03372.

Supriyo Chakraborty and Krishnan Raghavan. On understanding of the dynamics of model capacity in continual learning. *arXiv preprint arXiv:2508.08052*, 2025. URL https://arxiv.org/abs/2508.08052.

Thang Doan, Mehdi Abbana Bennani, Bogdan Mazoure, Guillaume Rabusseau, and Pierre Alquier. A theoretical analysis of catastrophic forgetting through the NTK overlap matrix. In *International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1072–1080, 2021. URL https://proceedings.mlr.press/v130/doan21a.html. arXiv:2010.04003.

Shibhansh Dohare, J Fernando Hernandez-Garcia, Qingfeng Lan, Parash Rahman, A Rupam Mahmood, and Richard S Sutton. Loss of plasticity in deep continual learning. *Nature*, 632:768–774, 2024. URL https://www.nature.com/articles/s41586-024-07711-7.

Itay Evron, Edward Moroshko, Rachel Ward, Nathan Srebro, and Daniel Soudry. How catastrophic can catastrophic forgetting be in linear regression? In *Conference on Learning Theory*, 2022. URL https://arxiv.org/abs/2205.09588. arXiv:2205.09588.

Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. Orthogonal gradient descent for continual learning. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 3762–3773, 2020. URL https://proceedings.mlr.press/v108/farajtabar20a.html.

Robert M French. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135, 1999. URL https://doi.org/10.1016/S1364-6613(99)01294-2.

Daniel Goldfarb, Itay Evron, Nir Weinberger, Daniel Soudry, and Paul Hand. The joint effect of task similarity and overparameterization on catastrophic forgetting: An analytical model. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2401.12617. arXiv:2401.12617.

Yuxian Gu, Li Dong, Hongning Wang, Yaru Hao, Qingxiu Dong, Furu Wei, and Minlie Huang. Data selection via optimal control for language models. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.07064. arXiv:2410.07064.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zheng Leng Thai, Kaihuo Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. MiniCPM: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024. URL https://arxiv.org/abs/2404.06395.

Shengyuan Hu, Yiwei Fu, Zhiwei Steven Wu, and Virginia Smith. Unlearning or obfuscating? jogging the memory of unlearned LLMs via benign relearning. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2406.13356. arXiv:2406.13356.

Jeremias Knoblauch, Hisham Husain, and Tom Diethe. Optimal continual learning has perfect memory and is NP-hard. In *International Conference on Machine Learning*, 2020. URL https://arxiv.org/abs/2006.05188. arXiv:2006.05188.

Suhas Kotha, Jacob Mitchell Springer, and Aditi Raghunathan. Understanding catastrophic forgetting in language models via implicit inference. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2309.10105. arXiv:2309.10105.

Jessy Lin, Luke Zettlemoyer, Gargi Ghosh, Wen-Tau Yih, Aram Markosyan, Vincent-Pierre Berges, and Barlas Oğuz. Continual learning via sparse memory finetuning. *arXiv preprint arXiv:2510.15103*, 2025. URL https://arxiv.org/abs/2510.15103.

Sen Lin, Peizhong Ju, Yingbin Liang, and Ness Shroff. Theory on forgetting and generalization of continual learning. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 21078–21100, 2023. URL https://proceedings.mlr.press/v202/lin23f.html. arXiv:2302.05836.

Zhenghao Lin, Zhibin Gou, Yeyun Gong, Xiao Liu, Yelong Shen, Ruochen Xu, Chen Lin, Yujiu Yang, Jian Jiao, Nan Duan, and Weizhu Chen. Not all tokens are what you need for pretraining. In *Advances in Neural Information Processing Systems*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/3322a9a72a1707de14badd5e552ff466-Abstract-Conference.html. arXiv:2404.07965.

Martin Marek, Dongkyu Cho, Shikai Qiu, Rumi Chunara, Pavel Izmailov, and Andrew Gordon Wilson. Forgetting in language models: Capacity, optimization, and self-generated replay. *arXiv preprint arXiv:2605.26097*, 2026. URL https://arxiv.org/abs/2605.26097.

Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation*, 24:109–165, 1989. URL https://doi.org/10.1016/S0079-7421(08)60536-8.

Sören Mindermann, Jan Brauner, Muhammed Razzak, Mrinank Sharma, Andreas Kirsch, Winnie Xu, Benedikt Höltgen, Aidan N Gomez, Adrien Morisot, Sebastian Farquhar, and Yarin Gal. Prioritized training on points that are learnable, worth learning, and not yet learnt. In *International Conference on Machine Learning*, 2022. URL https://arxiv.org/abs/2206.07137. arXiv:2206.07137.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Razvan Pascanu, and Hassan Ghasemzadeh. Understanding the role of training regimes in continual learning. In *Advances in Neural Information Processing Systems*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/518a38cc9a0173d0b2dc088166981cf8-Abstract.html. arXiv:2006.06958.

John X Morris, Chawin Sitawarin, Chuan Guo, Narine Kokhlikyan, G Edward Suh, Alexander M Rush, Kamalika Chaudhuri, and Saeed Mahloujifar. How much do language models memorize? *arXiv preprint arXiv:2505.24832*, 2025. URL https://arxiv.org/abs/2505.24832.

Alex Nguyen and Gautam Reddy. Differential learning kinetics govern the transition from memorization to generalization during in-context learning. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2412.00104. arXiv:2412.00104.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2024. URL https://doi.org/10.52202/079017-0970.

Gobinda Saha, Isha Garg, and Kaushik Roy. Gradient projection memory for continual learning. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2103.09762.

Jonathan Schwarz, Wojciech Czarnecki, Jelena Luketina, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. Progress & compress: A scalable framework for continual learning. In *International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4528–4537, 2018. URL https://proceedings.mlr.press/v80/schwarz18a.html. arXiv:1805.06370.

Idan Shenfeld, Jyothish Pari, and Pulkit Agrawal. RL's razor: Why online reinforcement learning forgets less. *arXiv preprint arXiv:2509.04259*, 2025. URL https://arxiv.org/abs/2509.04259.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning. In *International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/61434. arXiv:2601.19897.

Amelia Sorrenti, Giovanni Bellitto, Federica Proietto Salanitri, Matteo Pennisi, Simone Palazzo, and Concetto Spampinato. Wake-sleep consolidated learning. *arXiv preprint arXiv:2401.08623*, 2024. URL https://arxiv.org/abs/2401.08623.

Timothy Tadros, Giri P Krishnan, Ramyaa Ramyaa, and Maxim Bazhenov. Sleep-like unsupervised replay reduces catastrophic forgetting in artificial neural networks. *Nature Communications*, 13: 7742, 2022. URL https://www.nature.com/articles/s41467-022-34938-7.

Boshi Wang, Xiang Yue, Yu Su, and Huan Sun. Grokked transformers are implicit reasoners: A mechanistic journey to the edge of generalization. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2405.15071. arXiv:2405.15071.

Jiachen T Wang, Prateek Mittal, Dawn Song, and Ruoxi Jia. Data Shapley in one training run. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2406.11011. arXiv:2406.11011.

Meng Wang, Haohan Zhao, Wenzhuo Liu, Lu Yang, Geng Liu, Haiyang Guo, Guo-Sen Xie, Gaofeng Meng, Hongbin Liu, and Fei Zhu. Denser $\neq$ better: Limits of on-policy self-distillation for continual post-training. *arXiv preprint arXiv:2607.01763*, 2026. URL https://arxiv.org/abs/2607.01763.

Zichun Yu, Spandan Das, and Chenyan Xiong. MATES: Model-aware data selection for efficient pretraining with data influence models. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2406.06046. arXiv:2406.06046.

Junhao Zheng, Xidi Cai, Shengjie Qiu, and Qianli Ma. Spurious forgetting in continual learning of language models. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2501.13453. arXiv:2501.13453.

Nicolas Zucchet, Jörg Bornschein, Stephanie Chan, Andrew Lampinen, Razvan Pascanu, and Soham De. How do language models learn facts? dynamics, curricula and hallucinations. In *Conference on Language Modeling*, 2025. URL https://arxiv.org/abs/2503.21676. arXiv:2503.21676.

A WORLD SPECIFICATION

Vocabulary. Eleven special tokens (`pad`, `eos`, `sep`, `rule`, `fact`, `noise`, `dom_r`, `dom_f`, `dom_n`, `q`, `a`), 64 content symbols, 8 transform tags, 256 entity-high and 256 entity-low tokens, 64 attribute tokens and 256 value tokens: 915 tokens, padded to 960 for the embedding.

Facts. Keys $(e_{\text{hi}}, e_{\text{lo}}, a)$ are sampled without replacement from the $256 \times 256 \times 64$ key space, disjoint across phases; values are two i.i.d. uniform tokens, so each fact carries 16 bits given its key. Within each phase, utility (queried / never queried) and frequency class (low / high) are assigned by two independent random permutations, each balanced, so the four quadrants are equally populated and utility and frequency are uncorrelated. A low-frequency fact appears E times per phase and a high-frequency fact $4E$ times; $E = 400$ in the pretraining phase and $E = 60$ in phases A and B.

Rules. A probabilistic context-free grammar with 8 non-terminals, 2–3 productions each of length 2–3, a private set of 12 terminals per non-terminal, non-terminal probability 0.45 and depth at most 5, seeded once per world; inputs are rejection-sampled to length 6–14 (18–22 for length generalisation). The eight atomic transforms are reversal, rotation by 1 and 2, modular addition of 1, 3 and 7, adjacent-pair swap and negation modulo 64. All singletons and 40 of the 56 ordered pairs are training compositions; the training pairs are partitioned into disjoint chunks per phase; the remaining 16 pairs and 16 sampled triples are held out.

Stream. Token shares rules : facts : noise = 0.3 : 0.6 : 0.1; samples are shuffled i.i.d. within a phase and concatenated; a float16 sidecar carries the utility of each token as a prediction target (1 on rule outputs and on the values of queried facts, 0 elsewhere). Training windows of 128 tokens are sampled uniformly from the current phase.

Evaluation. 2000 new instances of each phase's compositions, 2000 held-out pairs, 2000 held-out triples, 2000 long inputs; up to 4000 queried and 4000 unqueried facts per phase (stored bits are extrapolated to the true counts); 500 noise strings. The score U_{future} averages exact accuracy over the high-utility facts and the rule compositions of the pretraining phase, A and B, and the held-out pairs; held-out triples and long inputs are evaluated but not included.

Non-stationary protocol. Phase A trains on rule set R_A and facts F_A; phase B on R_B (disjoint pair compositions, shared singletons) and F_B (disjoint keys). Similarity between phases is parameterised by the fraction of shared entities and shared compositions; the main experiments use zero overlap.

B LEARNING KINETICS

With the 5M-parameter model, batch 128 and learning rate 10^{-3}, 20k facts are memorised at 200 exposures (99.8% recall) and 1000 exposures (100%), but not at 50 or 100 exposures (recall $< 0.1\%$) at any batch size from 16 to 512; at 200 exposures batch 64 fails at learning rate 10^{-3} and succeeds (99.9%) at 5×10^{-4}, so the relevant quantity is the gradient-noise scale $\mathrm{lr}/\mathrm{batch}$. Weight decay 0 and 0.1 behave identically. Larger fact sets lengthen the memorisation plateau: at 200 exposures recall drops from 99.8% (20k facts) to 0% (50k); at 400 exposures it is 94% at 100k and 33% at 200k for the 5M model, and non-monotonic in the number of facts for the 11M model (56% at 100k, 19% at 200k, 94% at 400k; Table 3), because more facts also mean more optimiser steps to exit the plateau (Zucchet et al., 2025). At these budgets the apparent capacity is set by learning kinetics rather than storage. After the pretraining phase, 500k phase-A facts are learned to 50% at 60 exposures (Section 3); without it, 100 exposures give zero recall.

C HYPERPARAMETERS AND COMPUTE

Model: 6 layers, 8 heads, width 256, context 128, no bias, no dropout (5.00M parameters including embeddings; the 11M model has 6 layers, 6 heads and width 384). Optimiser: AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.95$), weight decay 0.1, gradient clipping 1.0, learning rate 10^{-3} with 300 warm-up steps and cosine decay to 10^{-4} restarted in every phase, batch 128 windows of 128 tokens, bf16 autocast, one pass over each phase's stream. Checkpoints at the end of every phase and at steps 10×1.5^n into phases A and B. Recovery test: 1% of phase-A facts (at least 8), 200 steps at learning rate 3×10^{-4}, batch 32. Linear probe: predict the first value token from the residual stream at the `[sep]` position of every block, using 6000 facts (the attribute and entity positions as controls), multinomial logistic regression (300 Adam steps, weight decay 10^{-3}) on a 70/30 split, best layer reported, shuffled-label control, and a two-layer MLP probe (hidden width 512) fitted on the same split. Eviction analysis: 4000 facts. Sleep: 25% of the wake steps, learning rate 3×10^{-4}, temperature 1, half of each batch from a buffer holding 5% of the previous phases' windows, teachers frozen; FLOPs count $6N$ per student token and $4N$ per token for the two teacher forward passes. Replay: 20% of each wake batch from the same buffer, for 1.0 or 1.42 epochs per phase (the latter matches the sleep arms' extra steps). Total compute for the sweeps in this paper: about 35 billion training tokens for the capacity sweep and 27 billion for the consolidation arms, on four RTX 5880 Ada GPUs and one RTX 4090 at about 1.2 million tokens per second each.

D CALIBRATION

Model	Exposures / fact	Facts	Exact recall	Stored (Mbit)
5M	1,000	20,000	1.000	0.32
		50,000	1.000	0.80
		100,000	0.999	1.60
		200,000	0.983	3.18
5M	400	20,000	1.000	0.32
		50,000	0.980	0.79
		100,000	0.940	1.56
		200,000	0.330	2.28
		400,000	0.030	3.32
11M	400	20,000	1.000	—
		50,000	0.998	—
		100,000	0.560	—
		200,000	0.190	—
		400,000	0.940	—

Table 3: Pre-sweep memorisation calibration (facts-only worlds trained from scratch, batch 128, learning rate 10^{-3}), used for the nominal sizing of Appendix J and the kinetics of Appendix B. Dashes denote unreported per-size storage estimates. The largest stored values (5M: 0.64 and 0.67 bits per parameter at 1,000 and 400 exposures; 11M: 6.25 Mbit, 0.57) are lower bounds, not capacity estimates: at 1,000 exposures the largest world is still 98% recalled, and at 400 exposures recall is set by learning kinetics and, for the 11M model, is non-monotonic in the number of facts. Capacity estimates are in Table 4.

Calibration used for the load axis			
Model	Observed plateau (Mbit)	Adopted C_{eff}	Bits / param.
5M	6.7, 6.1	**6.5 Mbit**	1.3
11M	12.3, 11.4	**12 Mbit**	1.1
Protocol checks			
Model	Protocol	Stored (Mbit)	Bits / param.
5M	120 exposures after pretraining, $\ell \approx 2.5$ (two seeds)	6.2–6.4	1.2–1.3
5M	Joint A+B from the shared checkpoint, $\ell \approx 2.5$	6.6	1.3
5M	From scratch, 500k facts, 400 exposures	5.1	1.0
5M	From scratch, 500k facts, 1,000 exposures	7.1	1.4

Table 4: Capacity calibration and protocol checks. Storage is measured by $\sum_i \max(0, 16 - \text{nll}_i/\ln 2)$. The load axis uses the adopted plateau after shared pretraining and 60 exposures per phase; the checks show sensitivity to exposure and training protocol. The nominal 3.2 Mbit sizing reference is described in Appendix J.

Capacity estimates under different protocols. Table 4 collects every estimate of the 5M and 11M models' storage in one place. The sweep's plateau at 60 exposures after the shared pretraining phase is the C_{eff} used for ℓ; the other rows show how far the estimate moves with the exposure budget and the training regime.

E DIAGNOSTICS ALONG THE SWEEP

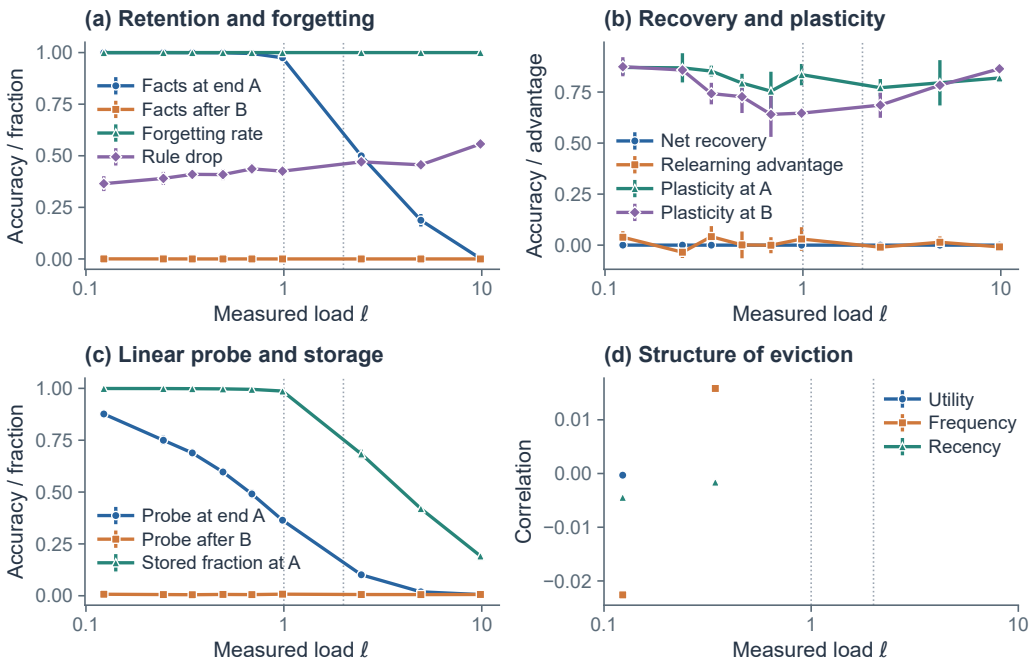

Figure 5: Diagnostics against measured load ℓ (mean and s.d. over seeds, 5M model). (a) Phase-A facts learned at the end of A, kept at the end of B, forgetting rate, and rule drop; (b) net recovery, relearning advantage and plasticity from the A and B checkpoints; (c) linear-probe accuracy for phase-A values at the end of A and B, and stored-bit fraction at the end of A; (d) correlation of forgetting with utility, frequency and recency at the end of B (undefined where everything is forgotten; Table 5 gives the correlations at the first checkpoint with partial forgetting). Dotted lines: $\ell = 1$ and $\ell = 2$.

F MECHANISM AND INTERVENTION CONTROLS

Unless stated otherwise, controls use the 5M model, the worlds of the main sweep and three seeds.

F.1 JOINT TRAINING AND CAPACITY

Joint training. Training i.i.d. on the union of the pretraining phase, A and B with the same total token budget (60 exposures per fact) stores 99% of both fact sets at $\ell = 0.25$ (1.6 Mbit by exact count), 59% at $\ell = 1$ (3.8 Mbit) and 21% at $\ell = 2.5$ (3.3 Mbit), with rules at 1.00. Starting instead from the end of the shared pretraining phase and training on the union of phases A and B with their combined token budget (three seeds) learns both sets in full at $\ell = 0.25$ (1.6 Mbit), 62% of each at $\ell = 1$ (4.0 Mbit by exact count, 5.1 Mbit by loss-based bits) and 18% at $\ell = 2.5$ (2.9 Mbit exact, 6.6 Mbit loss-based); the pretraining facts, absent from this mix, are forgotten. At $\ell = 1$, replay, the best sequential arm, keeps 3.4–3.6 Mbit exact (4.2–4.4 loss-based), 89–95% of the first joint run and 85–90% of the second. At $\ell = 2.5$ both joint runs store fewer complete facts than sequential phase A alone (3.8–4.1 Mbit by exact count) or naive sequential training overall (4.1 Mbit): a million facts learned at once sit longer on the memorisation plateau than half a million, so joint training is a reference with its own kinetics rather than an upper bound, which is why we call C_{eff} the capacity of the protocol. By loss-based bits, naive training, both replay variants and the joint run from the shared checkpoint hold 5.4–6.6 Mbit at $\ell = 2.5$, utility-weighted replay 4.9 and weighted sleep 2.6–3.0 Mbit; the first group lies near the calibration plateau (6.1–6.7 Mbit; Table 4) but does not converge on one value, and its members differ mainly in how those bits are distributed:

joint training spreads them thinly over a million facts (18% complete), whereas sequential training completes about half of the most recent phase.

Exposure budget at the storage transition. Re-running $\ell = 2.5$ with 120 instead of 60 exposures per phase (two seeds) raises phase-A recall from 0.50 to 0.57–0.62 and stored bits from 10.9 to 12.4–12.9 per fact, i.e. from 5.5 to 6.2–6.4 Mbit, which is the plateau the 60-exposure sweep reaches at $\ell = 4.9$ and 9.9 (6.7 and 6.1 Mbit); probe accuracy at the end of A moves from 0.10 to 0.15, and plasticity with the fixed 500-fact set is unchanged (0.73–0.75 against 0.77). Doubling the budget therefore fills the store to its plateau and leaves 40% of the facts unlearned: the transition is a capacity limit under this protocol, not an exposure limit.

Independent capacity calibration. A facts-only world of 500k facts (8.0 Mbit) trained from scratch on the 5M model, outside the sweep, stores 5.1 Mbit at 400 exposures (recall 27%) and 7.1 Mbit at 1,000 exposures (recall 69%), i.e. 1.0–1.4 bits per parameter; the sweep's plateau of 6.1–6.7 Mbit at 60 exposures after the shared pretraining phase lies inside this band. The capacity of this model family is therefore 1.0–1.4 bits per parameter depending on the exposure budget, and the reported loads use the sweep's calibration; changing the protocol shifts the capacity estimate by roughly 20%.

F.2 Collapse controls

Parameter displacement and random perturbation. For every run with log-spaced phase-B checkpoints we compute the L2 displacement $\|\theta_t - \theta_A\|$ from the end-of-A parameters ($\|\theta_A\| = 136$–152) and read off, by log-linear interpolation between checkpoints, the displacement at which phase-A recall passes 0.5 and 0.1. At $\ell = 0.25$: learning rate 10^{-3}, 3×10^{-4}, 10^{-4} (300-step warm-up) give 1.1–1.3 / 1.7–1.9, 0.93 / 1.62 and 0.95 / 1.51 (three seeds for the first); warm-up 100 and 1,000 steps at 10^{-3} give 1.4–1.6 / 2.0–2.2 and 1.0–1.1 / 1.8–2.0; EWC from the same checkpoint ($\lambda = 0.001$) 1.30 / 2.12. The collapse displacement is thus 1.5–2.2 across a tenfold range of learning rates and warm-up lengths, while the step at which it is reached varies from 22 to 115. Along the load sweep (one run per load) the collapse displacement falls: 2.2, 1.9, 1.5, 1.2, 0.96, 0.46, 0.16 at $\ell = 0.12, 0.25, 0.49, 0.69, 0.98, 2.5, 4.9$. With no new facts in phase B (rules and noise only) recall passes 0.5 at a displacement of 12 and 0.1 at 22, after 216 and 338 steps; phase-A recall is still 98% after 51 steps, and probe accuracy falls from 0.76–0.78 at the end of A to 0.31–0.34 at 171 steps and to at most 0.011 by 384. Random perturbation: Gaussian noise added to θ_A with the norm of the displacement at each checkpoint (three seeds, isotropic over all parameters; per-tensor matching gives the same numbers). Below capacity, recall stays at 1.00 up to a displacement of 6 (4% of the norm), is 0.91–0.98 at 9–13 and 0.32–0.64 at 15–22; the collapse displacement, 1.5–2.2, is an order of magnitude smaller. Beyond capacity the store is fragile to noise as well: at $\ell = 2.5$ recall falls from 0.51 to 0.42 at a random displacement of 3.3 and to 0.12 at 8; at $\ell = 4.9$ from 0.18 to 0.13 at 3 and 0.02 at 8.

Warm-up length. Phase B re-run from the end-of-A checkpoints at $\ell = 0.25$ (two seeds) with 100 or 1,000 warm-up steps instead of the default 300, at the same peak learning rate: relative retention falls below 0.1 between 16 and 23 steps with $W = 100$ (0.60–0.73 at 16, 0.01–0.07 at 23), at 34–36 steps with $W = 300$ (two-step resolution, below), and between 52 and 77 steps with $W = 1{,}000$ (0.37–0.40 at 52, ≤ 0.01 at 77). A fixed cumulative learning rate $\mathrm{lr}\, t^2/(2W)$, anchored at the 34–36-step collapse for $W = 300$, predicts 20–21 and 62–66 steps. During warm-up the learning rate is $\mathrm{lr}\, t/W$, so $W = 1{,}000$ at 10^{-3} follows the schedule of $W = 300$ at 3×10^{-4} for 300 steps and collapses at the same checkpoint; the 100-step warm-up is the independent test.

A phase B of 300 distinct facts. Worlds with 10k pretraining facts, N_A phase-A facts and 300 phase-B facts (60 exposures each; 36 steps of phase B), two seeds each at $\ell = 0.25$ ($N_A = 50$k) and $\ell = 2.5$ ($N_A = 500$k). Phase-A recall at the end of A / after the 36 steps of B: 1.00 / 0.89–0.93 at $\ell = 0.25$, 0.51–0.53 / 0.19–0.20 at $\ell = 2.5$; the 300 new facts are learned to 97–100%; phase-A rules stay at 0.82–0.93. Along the timeline (10, 15, 22, 34, 36 steps) recall at $\ell = 0.25$ is 1.00, 1.00, 0.99, 0.91–0.95, 0.89–0.93 and at $\ell = 2.5$ 0.47–0.52, 0.39–0.47, 0.31–0.36, 0.21–0.23, 0.19–0.20. With a full-size phase B the same checkpoints read 0.76–0.95 at 22 steps and 0.06–0.11 at 34 at $\ell = 0.25$, and 0.05 at 34 at $\ell = 2.5$.

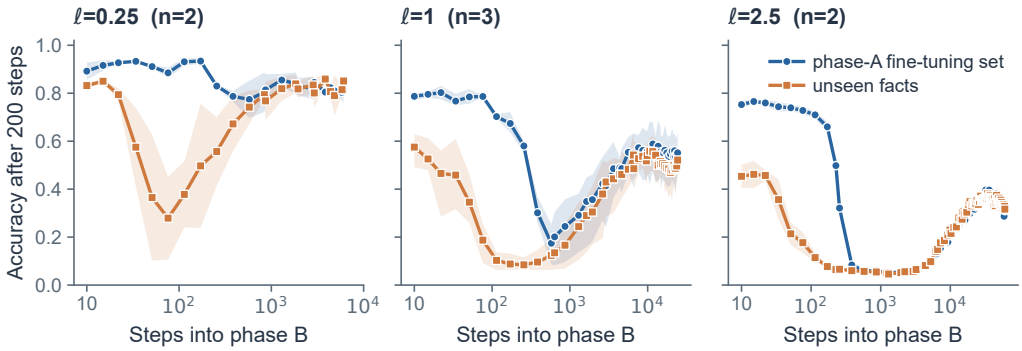

Figure 6: Raw relearning curves behind the relearning advantage: exact match after the 200-step recovery budget on the 1% fine-tuning set of phase-A facts (blue) and on an equal number of unseen facts (orange), from each phase-B checkpoint of the naive runs (mean and range over seeds). The fine-tuning set stays relearnable after exact recall has collapsed (34, 34 and 22 steps at $\ell = 0.25$, 1, 2.5), and the two curves meet at about 600, 600 and 400 steps. Unseen-fact sets are 500, 2,000 and 5,000 facts, so levels compare within a panel, not across panels.

Shared entities between the phases. Worlds at $\ell = 0.25$ in which 25% or 50% of the phase-B facts are about entities of phase A (new attributes; keys stay unique; two seeds each): phase-A recall is 0.84–0.97 at 22 steps, 0.15–0.26 at 34 and 0–0.02 at 51, net recovery 0.03–0.10 between 34 and 114 steps, probe accuracy at the end of A 0.60–0.66 (measured with an earlier 4,000-fact probe protocol, under which disjoint-entity runs give 0.60–0.64) and the peak relearning advantage 0.53–0.85, all within the range of the disjoint-entity runs. Entity overlap between the phases neither speeds nor slows the collapse.

Gradient overlap along phase B. At the end-of-A checkpoint and at each log-spaced phase-B checkpoint of the $\ell = 0.25$ run (seed 0) we compute the mean gradient of the training loss on phase-A windows, g_A, and on phase-B windows, g_B (eight batches of 32 windows each), and the displacement $\Delta = \theta_t - \theta_A$. At θ_A, $\|g_A\| = 0.21$ against $\|g_B\| = 2.19$ and $\cos(g_A, g_B) = -0.01$: phase A sits near a minimum whose gradient is orthogonal to the new phase's. As phase B proceeds the two gradients become anti-aligned, $\cos(g_A, g_B) = -0.20, -0.51, -0.38$ at 10, 15 and 22 steps, while $\|g_A\|$ rises from 0.2 to 4.8 and the phase-A loss from 1.43 to 1.66 (2.46 at 34 steps); after the collapse the cosine returns to ± 0.1. The displacement has no component along the phase-A gradient at θ_A ($\cos(\Delta, g_A(\theta_A)) = 0.00$ at every checkpoint) but acquires one along the current phase-A gradient ($\cos(\Delta, g_A(\theta_t)) = 0.07, 0.21, 0.22, 0.13$ at 10, 15, 22, 34 steps) and is anti-aligned with the current phase-B gradient (-0.34 to -0.18 over the same steps).

Collapse time at two-step resolution. Re-running the first 100 steps of phase B from the end-of-A weights with an evaluation every two steps (fresh optimiser state, as in the learning-rate variants): at $\ell = 0.25$ recall passes one half at step 28 and 0.1 at 34–36 (two seeds); at $\ell = 1$ at 18–20 and 24–26. The log-spaced estimates of Table 5 (34–42 and 22–34) are thus resolved to 34–36 and 24–26.

F.3 Readout details

A fixed-size set for plasticity. The recovery test fine-tunes on 1% of the phase-A facts, and the original plasticity measurement learned an equal number of unseen facts with the same 200-step budget, so its set grew from 250 facts at $\ell = 0.12$ to 20,000 at $\ell = 9.9$. Re-running the measurement with 500 unseen facts at every load gives, from the end-of-A checkpoint of the 5M model, 0.87, 0.87, 0.85, 0.79, 0.75, 0.84, 0.77, 0.80, 0.82 at $\ell = 0.12, 0.25, 0.34, 0.49, 0.69, 0.98, 2.5, 4.9, 9.9$ (against 0.89, 0.83, 0.79, 0.72, 0.67, 0.61, 0.34, 0.11, 0.02 with the growing set), and 0.94–0.98 at every load for the 11M model; from the end-of-B checkpoint they are 0.64–0.87 (Figure 5b). End-of-A plasticity is therefore flat across the sweep, and every plasticity number in this paper uses the fixed set.

Raw relearning curves. Figure 6 shows the two accuracies whose difference is the relearning advantage. At $\ell = 0.25$ the fine-tuning set is relearned to 0.86–0.95 from every checkpoint up to 171 steps and to 0.70–0.89 thereafter; unseen facts are learned to 0.82–0.85 from the 10-step checkpoint, to 0.10–0.63 from the 51-step and 0.11–0.45 from the 76-step checkpoint, and to 0.8 again by about 600 steps. At $\ell = 1$ the fine-tuning set holds at 0.75–0.83 through 76 steps, 0.55–0.60 at 256 and 0.09–0.25 at 577, the level of unseen facts (0.06–0.19); at $\ell = 2.5$ it holds at 0.72–0.77 through 76 steps, 0.64–0.68 at 171, 0.30–0.34 at 256 and reaches the unseen-fact level (0.05–0.10) at 384. Unseen facts at both loads are learned to 0.4–0.6 from the 10-step checkpoint and to 0.06–0.13 from the 114–384-step checkpoints; both curves then rise together over thousands of steps as the network's ability to take up same-format facts returns (to 0.4–0.55 at $\ell = 1$, 0.3–0.4 at 2.5). At a matched checkpoint the relearning advantage is evidence of a trace; its time course depends on both curves, and at $\ell = 0.25$ mainly on the recovery of unseen-fact learning.

Rank of the correct answer at the collapse. For 2,000 phase-A facts and every phase-B checkpoint of one naive run per load we compute, under teacher forcing and in fp32, the rank of the correct value among the 256 value tokens, taking the worse of a fact's two value tokens, and the rank of the first value token over the whole vocabulary. Chance is not the uniform median: for the worse-of-two statistic it is about 181 under independence, and the model's output is not uniform, so we also report the same statistic with the target values shuffled across facts, an empirical baseline at each checkpoint (171–184 for the worse-of-two statistic among values, 121–140 for the first token over the vocabulary once phase B is under way). At $\ell = 0.25$: 22 steps, exact 0.76, median rank 1; 34 steps, exact 0.06, median 19 (first token 12), 91% of the facts within the top 100; 51 steps, median 92 (70); 76 steps, 125 (93); 114 steps, 168 against a shuffled 175, i.e. chance, and stays there except for a marginal residual ordering between 256 and 1,300 steps (medians 133–157 against 177–181). At $\ell = 1$: 22 steps, exact 0.16, median 6; 34 steps, exact 0.01, median 37; 51 steps, 114; 76 steps, 154; 114–171 steps, 167–176 against 181–182. At $\ell = 2.5$ (exact 0.54 at the end of A): 15 steps, exact 0.29, median 4; 22 steps, exact 0.02, median 37; 34 steps, 114; 51 steps, 158; 76–114 steps, 172–175 against 179–180. The ordering of the values thus outlives exact recall by two to three checkpoints, some 40–80 steps, and reaches the shuffled baseline at 114–171 steps at $\ell \leq 1$ and 76–114 steps at $\ell = 2.5$, while net recovery at $\ell = 0.25$ is still 0.02–0.07 at 171 steps and at most 0.01 at 256. The same pass gives the phase-B facts' loss-based bits: at the collapse of phase A (34 steps) they hold 0.24–0.27 bits per fact (out of 16) at every load and 0.1% exact match, and they pass 1 bit only after about 250 steps at $\ell = 0.25$ and 900 at $\ell = 1$. The fp32 pass also reproduces the bf16-autocast recall of the training-time evaluation at the end of A to within sampling error (1.000, 0.975 and 0.539 against 1.000, 0.976 and 0.527 on 2,000 facts), so the testbed's evaluations do not share the precision problem of Pythia (Appendix K).

F.4 Consolidation controls

EWC from a common checkpoint. Online EWC started from the naive end-of-A checkpoint at $\ell = 0.25$, with the Fisher estimated on phase A (100 batches, normalised to mean one) and $\lambda \in \{0.001, 0.01, 0.1\}$: phase-A recall is 0.96 at 23 steps, 0.15–0.16 at 35 and ≤ 0.02 at 52 for every λ, indistinguishable from naive training; phase-B facts learned are 98%, 68% and 0%, and rules kept 0.77, 0.82, 0.84.

Utility-weighted replay. Replay with the per-token utility weights of the sleep arm and no distillation (Table 2): high-utility phase-A facts kept 0.87, 0.67, 0.18, zero-utility facts kept 0.05, 0.00, 0.00, phase-B facts learned 1.00/0.71, 0.99/0.21, 0.74/0.00 (high/zero utility), U_{future} 0.91, 0.88, 0.75, total store by exact count 1.1, 3.0, 3.7 Mbit, increasingly concentrated in high-utility facts.

Diagnosis-guided remedies (Section 5). *Projection.* OGD-style: at the end of each finished phase, 64 gradients of the training loss on that phase's windows are orthonormalised into a basis (128 directions after phase A), and every AdamW update in phase B is projected off it (update projection, so the parameter displacement stays orthogonal to the basis). GPM-style: at the end of each finished phase the inputs of each internal linear layer on 32 batches of that phase are collected and the leading right-singular directions carrying 90% or 97% of their energy kept (per layer, 14–130 and 35–340 directions out of 256 or 1,024); every weight update is projected so that it acts only on the orthogonal complement (embeddings, layer norms and the shared output head are not projected).

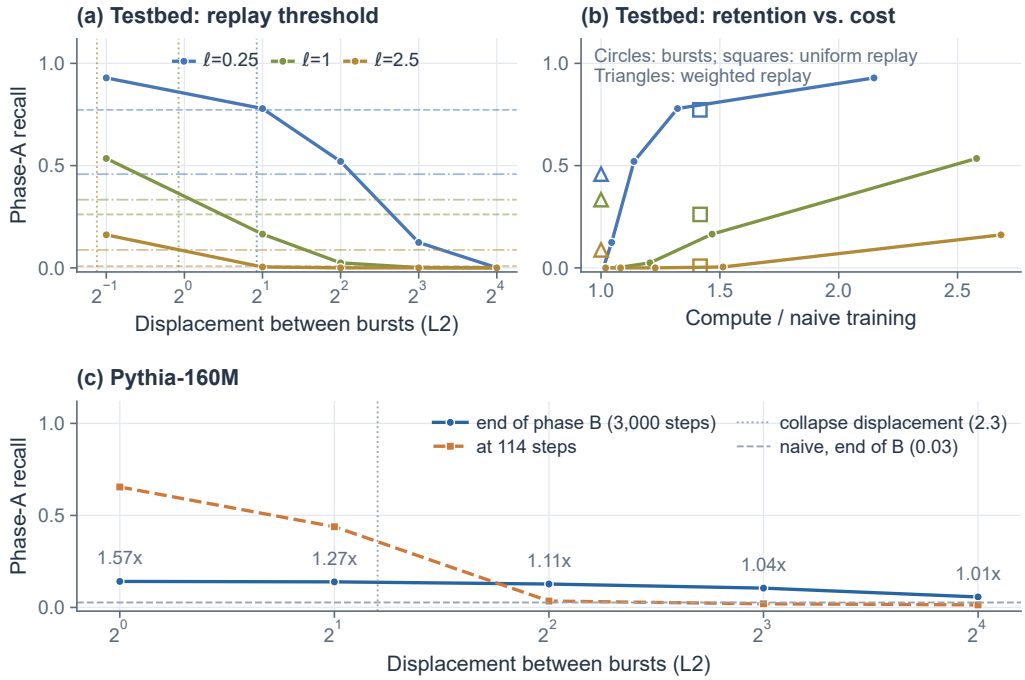

Figure 7: Displacement-triggered replay. (a) Testbed: phase-A facts kept at the end of B against the displacement allowed between bursts, per load; dotted lines mark the collapse displacement, dashed and dash-dot lines uniform and utility-weighted replay. (b) The same against compute relative to naive training. (c) Pythia-160M: phase A held at 114 steps and at the end of phase B, labels give the compute ratio.

Results at the end of phase B, three (OGD) or two (GPM) seeds: OGD keeps 0.000 of phase A at $\ell = 0.25$ and 1 with phase B learned to 1.00 and 0.96; GPM at 90% keeps 0.000 of phase A with phase B at 0.99, 0.55 and 0.19 ($\ell = 0.25$, 1, 2.5), at 97% keeps 0.000 with phase B at 0.66, 0.17 and 0.02: the more of the activation subspace is protected, the less of phase B is learned, and phase A is never kept. At $\ell = 2.5$ the $d = 0.5$ bursts keep 0.16 of phase A (two seeds) with phase B learned to 0.27 instead of 0.52, at 2.7× the compute; OGD keeps 0.000 there as well. *Displacement-triggered replay.* A buffer of 5% of the previous phases' windows is replayed for five steps whenever the L2 displacement since the last burst reaches d; the extra steps are counted. Phase-A facts kept and compute relative to naive training, $d = 0.5, 2, 4, 8, 16$: $\ell = 0.25$ (three seeds) 0.93, 0.78, 0.52, 0.12, 0.00 at 2.15, 1.32, 1.14, 1.05, 1.01×; $\ell = 1$ (two to three seeds) 0.54, 0.17, 0.03, 0.00, 0.00 at 2.58, 1.47, 1.21, 1.07, 1.01×; $\ell = 2.5$ (one to two seeds) 0.16, 0.00, 0.00, 0.00, 0.00 at 2.7, 1.5, 1.2, 1.1, 1.0×, so beyond capacity the bursts reallocate the store without enlarging it. With a 1% buffer at $d = 0.5$: 0.44 and 0.30 at $\ell = 0.25$ and 1. Uniform replay (20% of every batch, 1.42 epochs) keeps 0.77 and 0.26 at 1.42×. In Pythia-160M (10k persons, 5% of the persons as buffer, phase B of 3,000 steps at learning rate 10^{-4}, collapse displacement 2.3): for $d = 1, 2, 4, 8, 16$ the first burst fires at 22, 34, 51, 76 and 114 steps, phase A is at 0.65, 0.44, 0.03, 0.02, 0.01 after 114 steps and at 0.14, 0.14, 0.13, 0.11, 0.06 at the end of phase B (naive: 0.03), at 1.57, 1.27, 1.11, 1.04, 1.01× the compute.

Fate of the 10k pretraining facts. The shared pretraining phase stores its 10k facts in full (400 exposures). Under naive training their recall collapses during phase A as phase-A recall does during B: 0% at the end of A at every ℓ and both model sizes, and 0% at the end of B. EWC at $\lambda \leq 0.01$ leaves them at 0%; $\lambda = 10$ and 100 keep 10–17% and $\lambda = 1000$ all of them, while learning nothing new. Because the buffer samples all previous phases, the consolidation arms keep them better than they keep phase-A facts, which had 60 exposures: 1.42-epoch replay 0.99 / 0.84 / 0.53 at $\ell = 0.25$ / 1 / 2.5 (one-epoch replay 0.99 / 0.80 / 0.55), uniform sleep 0.92 / 0.79 / 0.58, utility-weighted sleep 0.97 / 0.92 / 0.82 of the high-utility half against 0.29 / 0.03 / 0.00 of the zero-utility half, utility-

weighted replay 1.00 / 1.00 / 0.90 against 0.66 / 0.36 / 0.01. Joint training on all three phases keeps 1.00 / 0.98 / 0.84; joint training on A and B from the shared checkpoint, in which they do not appear, forgets them.

G READOUT THRESHOLD TIMES AND EVICTION STRUCTURE

		Recall thresholds		Relearning advantage			Eviction r	
Load ℓ	n	$t_{1/2}$	$t_{0.1}$	Peak	t_{peak}	$t_{<0.1}$	Freq.	Util.
0.123	2	34	51	0.30	64	214	−0.18	0.01
0.246	2	34	42	0.64	64	480	−0.20	0.00
0.345	3	34	34	0.78	89	673	−0.23	0.00
0.492	4	34	38	0.77	89	724	−0.27	0.00
0.689	3	30	34	0.65	120	787	−0.30	−0.01
0.985	3	22	34	0.63	139	577	−0.27	−0.02
2.46	2	18	22	0.60	142	384	−0.28	0.00
4.92	2	15	22	0.69	62	346	−0.06	−0.05

Table 5: Readout threshold times in phase-B steps and eviction structure (5M model; seed means). $t_{1/2}$ and $t_{0.1}$ are the first sampled steps with relative retention below 0.5 and 0.1. $t_{<0.1}$ is the first step after the relearning peak with advantage below 0.1 (it may rise again later). Times are means of per-run crossings on log-spaced checkpoints (10, 15, 22, 34, 51, 76, . . .), not interpolated crossing times. Relearning times refer to the relearning advantage, which also depends on unseen-fact learning (Figure 6). Eviction r: point-biserial correlation between being forgotten and belonging to the high-frequency or the high-utility class, among the facts of a 4,000-fact phase-A sample that are recalled at the end of A, at the first checkpoint with relative retention between 0.05 and 0.95 (Figure 5d shows the end of B). Load labels are rounded.

H LINEAR AND NON-LINEAR PROBES

What the fitted probe measures. The testbed's final LayerNorm and output head have no bias. For a last-block residual h, centring matrix $C = I - \mathbf{1}\mathbf{1}^\top/m$ (m the residual width) and learned LayerNorm scale γ, the logits are $W\,\mathrm{diag}(\gamma)Ch/\sigma(h)$, with $\sigma(h) > 0$. The common positive denominator does not change the argmax. Thus $W\,\mathrm{diag}(\gamma)C$ is a linear classifier on the same residual whose value-token accuracy is at least the model's exact fact recall, so an accurate linear readout exists wherever recall is high. The fitted probe measures how readily such a readout is learned from 4,200 facts under a fixed procedure; below we vary the number of training facts, the optimisation, the regularisation and the features. Pythia's probe instead predicts attribute-value categories, which need not correspond to individual output tokens.

		End of A		End of B			
Load ℓ	n	Linear	MLP	Linear	Shuffled	MLP	Shuffled
0.123	3	0.876	0.733	0.007	0.005	0.007	0.004
0.246	3	0.750	0.593	0.006	0.005	0.006	0.006
0.345	8	0.689	0.529	0.005	0.006	0.005	0.006
0.492	8	0.597	0.444	0.006	0.005	0.006	0.006
0.689	9	0.491	0.364	0.005	0.006	0.005	0.006
0.985	3	0.364	0.265	0.008	0.006	0.007	0.005
2.46	3	0.101	0.083	0.006	0.006	0.007	0.005
4.92	3	0.019	0.017	0.006	0.006	0.006	0.006
9.85	2	0.006	0.007	0.006	0.007	0.006	0.006

Table 6: Held-out accuracy of linear and two-layer MLP probes (hidden width 512), reading the first value token at `[sep]` in the 5M model. Each probe uses 6,000 facts split 70/30 and reports its best layer; shuffled-label controls use the same selection rule. Values are seed means (sample s.d. ≤ 0.032); theoretical chance is $1/256 \approx 0.0039$. The MLP probe is less accurate than the linear one wherever either is above chance, and end-of-B accuracies match the shuffled controls.

The number of training facts limits the fitted probe. We refit the last-block probe on 15 end-of-A checkpoints of the 5M sweep, with a fixed test set of 1,800 phase-A facts per run and nested training sets drawn from the other phase-A facts (Table 7). With 4,200 training facts the decline is reproduced on this fixed split (0.88 to 0.38 between $\ell = 0.12$ and 0.98, against 0.36 with the 70/30 split of Table 6), and train accuracy is 1.00 at every load, so the probe's errors are errors of generalisation, not of fit. Optimisation and features do not explain the decline: at 4,200 facts, seed-mean accuracy changes by at most 0.01 with 3,000 instead of 300 optimiser steps, with unstandardised features, or with features normalised by the final LayerNorm but not rescaled by γ, and removing weight decay lowers it by 0.02–0.12. More training facts close the gap to the model's own readout: with all other phase-A facts the probe comes within 0.02 of the model's accuracy at every load. The number of facts needed grows with load, however: with 10,000, 20,000 or 40,000 training facts accuracy still falls with load. Over the same range the model's own readout loses margin. Between $\ell = 0.12$ and 0.98 the median logit margin of the correct value over the runner-up falls from 8.4 to 5.4, the median geometric margin, a distance in the LayerNorm-normalised space where features have norm 16, from 1.63 to 0.74 and its 10th percentile from 1.15 to 0.31, and the share of the standardised residual's variance explained by the means of the 256 first-value classes from 13% to 6%. These trends are consistent with values that crowd together as load grows, so that a new readout needs more examples to place its boundaries; as correlations across loads, they do not identify which property sets that need.

Probe sensitivity to layer and position. Figure 8 reports the fitted probes at every layer. Their highest accuracy is at the last layer (0.88 to 0.10 across loads, at most 0.03 at the preceding layer); the MLP has a similar profile. At the attribute and entity positions the probes perform near chance. A probe thus reads the value only where the model is about to emit it.

I Extended related work

Consolidation and theory. Progress & Compress (Schwarz et al., 2018) distils an active column into a fixed-size knowledge base under online EWC; dark experience replay (Buzzega et al., 2020), complementary-learning-system methods (Arani et al., 2022) and sleep-inspired consolidation (Tadros et al., 2022; Sorrenti et al., 2024) vary what is replayed and when; for language models, self-distillation from an earlier state forgets far less than supervised fine-tuning (Shenfeld et al., 2025; 2026), dense self-distillation can collapse in continual post-training (Wang et al., 2026), Behrouz et al. (2026) distil into a grown network during a sleep phase, and sparse memory finetuning (Lin et al., 2025) routes updates to memory slots. Our comparison adds a measured load axis and independently assigned utility; our operator (Section 5) is the distillation step of Progress & Compress without its EWC penalty, plus a buffer and utility weights, used as an instrument. On the theory side, optimal continual learning under a capacity constraint is NP-hard (Knoblauch et al.,

			Probe, by training facts					Model's margin		
Load ℓ	n	Model	4.2k	10k	20k	40k	All	Logit	Geom.	Geom. p10
0.123	3	1.00	0.88	0.99	1.00	—	1.00	8.41	1.63	1.15
0.246	3	1.00	0.75	0.95	0.99	1.00	1.00	7.75	1.35	0.90
0.345	2	1.00	0.69	0.92	0.98	1.00	1.00	7.43	1.22	0.79
0.492	2	1.00	0.61	0.88	0.96	0.99	1.00	7.01	1.08	0.68
0.689	1	1.00	0.48	0.78	0.93	0.97	0.99	6.30	0.93	0.52
0.985	3	0.98	0.38	0.65	0.83	0.92	0.97	5.38	0.74	0.31
2.46	1	0.60	0.12	0.21	0.32	0.45	0.58	0.98	0.17	−0.98

Table 7: Linear-probe accuracy against the number of training facts, and margins of the model's own readout (5M model, end of A, last-block `[sep]` residual; seed means). Each run has a fixed test set of 1,800 phase-A facts; training sets are nested subsets of its other phase-A facts, and "All" uses every one of them (23k at $\ell = 0.123$ to 498k at $\ell = 2.46$); —: fewer than 40k facts available. Model: accuracy of the model's own readout for the first value token on the test facts. Train accuracy with 4.2k facts is 1.00 at every load. Margins, on the test facts: median logit margin of the correct value over the runner-up (Logit), and the median and 10th percentile of that margin divided by the norm of the difference between the two rows of $W\,\mathrm{diag}(\gamma)C$ (Geom.), a distance in the LayerNorm-normalised feature space.

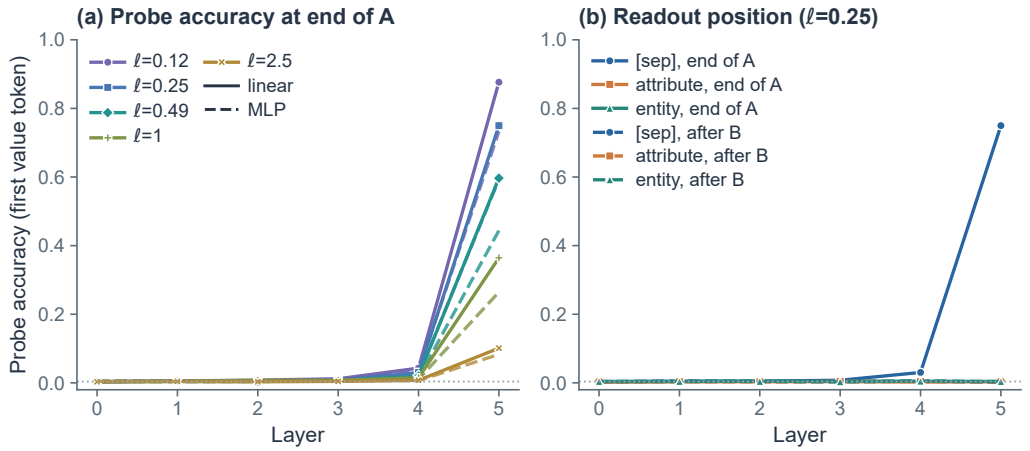

Figure 8: (a) Linear (solid) and MLP (dashed) probe accuracy for the first value token against layer at the end of phase A, 5M model, mean over seeds, one colour per load. (b) At $\ell = 0.25$: the `[sep]`, attribute and entity positions at the end of A and after phase B. Dotted line: chance.

2020); in over-parameterised linear regression the joint solution exists and forgetting is set by task geometry and order (Evron et al., 2022; Lin et al., 2023; Goldfarb et al., 2024), in the NTK regime by kernel overlap (Doan et al., 2021); loss of plasticity (Dohare et al., 2024) can masquerade as overwriting. Marek et al. (2026) study how capacity, optimisation and replay jointly affect forgetting, using model size and pretraining conditions as capacity proxies; our controlled fact stream adds an explicit information measure and complementary readouts; Nguyen & Reddy (2025) show that the memorisation-to-generalisation transition can be driven by learning kinetics rather than capacity, which we control with a fixed exposure budget.

Utility signals and synthetic testbeds. Reference-model excess loss selects tokens worth learning (Lin et al., 2024; Mindermann et al., 2022), influence-style scores estimate marginal contribution to a target (Yu et al., 2024; Gu et al., 2025; Wang et al., 2025), and late upsampling of high-utility data is standard (Hu et al., 2024; Blakeney et al., 2024); Byrd & Lipton (2019) find that importance weighting shapes early training but loses much of its effect with further optimisation, and that regularisation modulates this decay; consolidation, by contrast, works with a finite rehearsal budget on a changing stream. Controlled synthetic worlds have measured knowledge capacity (Allen-Zhu & Li,

2025; Morris et al., 2025), separated storage from extraction (Allen-Zhu & Li, 2024), and studied facts against statistics (Behnia et al., 2025), facts against rules (Wang et al., 2024; Barron & White, 2025) and spurious forgetting (Zheng et al., 2025). Our testbed combines independently controlled frequency and utility with a capacity sweep and non-stationary phases.

J Extended results and protocol details

Table 8: Complete consolidation results underlying Table 2. Values are mean ± sample s.d. over n seeds. A initial is recall at the end of A; all other recall and rule scores are measured at the end of B. All, useful and zero refer to all facts, the high-utility half and the zero-utility half of the indicated phase. Final recall uses the full indicated population, not only facts learned in A. PFLOPs include student and teacher computation. Load labels are rounded; U_{future} is defined in Section 2.

$\ell \approx 0.25$ **— Old knowledge**						
Method	n	A initial	A (all)	A (useful)	A (zero)	A rules
Naive	3	1.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.61±0.04
Replay (1.0 epoch)	6	1.00±0.00	0.70±0.01	0.70±0.01	0.69±0.01	0.99±0.00
Replay (1.42 epochs)	6	1.00±0.00	0.77±0.02	0.78±0.01	0.77±0.02	0.99±0.00
Uniform sleep	3	0.96±0.00	0.41±0.05	0.42±0.04	0.41±0.05	0.96±0.02
Utility-weighted sleep	6	0.82±0.00	0.30±0.01	0.60±0.03	0.00±0.00	0.98±0.01
Utility-weighted replay	6	0.93±0.01	0.46±0.01	0.87±0.01	0.05±0.01	1.00±0.00
EWC, $\lambda = 0.001$	3	0.99±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.71±0.04
EWC, $\lambda = 0.01$	3	0.72±0.05	0.00±0.00	0.00±0.00	0.00±0.00	0.77±0.01
$\ell \approx 0.25$ **— New knowledge and compute**						
Method	n	PFLOPs	B (all)	B (useful)	B (zero)	U_{future}
Naive	3	9.9	1.00±0.00	1.00±0.00	1.00±0.00	0.48±0.01
Replay (1.0 epoch)	6	9.9	1.00±0.00	1.00±0.00	1.00±0.00	0.87±0.01
Replay (1.42 epochs)	6	14.1	1.00±0.00	1.00±0.00	1.00±0.00	0.88±0.01
Uniform sleep	3	12.4	0.88±0.02	0.87±0.02	0.88±0.02	0.79±0.02
Utility-weighted sleep	6	12.4	0.69±0.01	0.94±0.01	0.43±0.02	0.84±0.01
Utility-weighted replay	6	9.9	0.85±0.01	1.00±0.00	0.71±0.02	0.91±0.00
EWC, $\lambda = 0.001$	3	10.1	0.95±0.01	0.95±0.01	0.96±0.02	0.51±0.01
EWC, $\lambda = 0.01$	3	10.1	0.50±0.05	0.50±0.06	0.50±0.05	0.46±0.02
$\ell \approx 1$ **— Old knowledge**						
Method	n	A initial	A (all)	A (useful)	A (zero)	A rules
Naive	3	0.98±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.57±0.01
Replay (1.0 epoch)	6	0.95±0.01	0.21±0.01	0.21±0.01	0.21±0.02	0.99±0.00
Replay (1.42 epochs)	6	0.97±0.01	0.26±0.02	0.26±0.02	0.26±0.02	0.99±0.00
Uniform sleep	3	0.64±0.01	0.12±0.01	0.11±0.01	0.12±0.01	0.98±0.00
Utility-weighted sleep	6	0.41±0.01	0.15±0.01	0.31±0.02	0.00±0.00	0.99±0.00
Utility-weighted replay	6	0.73±0.01	0.33±0.01	0.67±0.02	0.00±0.00	1.00±0.00
EWC, $\lambda = 0.001$	3	0.74±0.04	0.00±0.00	0.00±0.00	0.00±0.00	0.62±0.06
EWC, $\lambda = 0.01$	3	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.79±0.06
$\ell \approx 1$ **— New knowledge and compute**						
Method	n	PFLOPs	B (all)	B (useful)	B (zero)	U_{future}
Naive	3	27.8	0.96±0.02	0.96±0.02	0.96±0.02	0.45±0.01
Replay (1.0 epoch)	6	27.8	0.83±0.02	0.83±0.02	0.83±0.02	0.74±0.01
Replay (1.42 epochs)	6	39.4	0.87±0.01	0.87±0.01	0.88±0.01	0.76±0.00
Uniform sleep	3	37.8	0.47±0.01	0.47±0.01	0.47±0.01	0.67±0.01
Utility-weighted sleep	6	37.8	0.31±0.01	0.60±0.02	0.02±0.01	0.74±0.01
Utility-weighted replay	6	27.8	0.60±0.02	0.99±0.00	0.21±0.03	0.88±0.01
EWC, $\lambda = 0.001$	3	28.0	0.57±0.07	0.57±0.07	0.57±0.07	0.42±0.03
EWC, $\lambda = 0.01$	3	28.0	0.00±0.00	0.00±0.00	0.00±0.00	0.41±0.01

Continued on next page

Table 8 (continued)

$\ell \approx 2.5$ **— Old knowledge**

Method	n	A initial	A (all)	A (useful)	A (zero)	A rules
Naive	3	0.50±0.03	0.00±0.00	0.00±0.00	0.00±0.00	0.53±0.01
Replay (1.0 epoch)	3	0.49±0.01	0.01±0.00	0.01±0.00	0.01±0.00	1.00±0.00
Replay (1.42 epochs)	3	0.51±0.01	0.01±0.00	0.01±0.00	0.01±0.00	1.00±0.00
Uniform sleep	3	0.23±0.02	0.00±0.00	0.00±0.00	0.00±0.00	0.98±0.01
Utility-weighted sleep	3	0.16±0.02	0.02±0.00	0.03±0.01	0.00±0.00	0.99±0.00
Utility-weighted replay	3	0.46±0.01	0.09±0.01	0.18±0.01	0.00±0.00	1.00±0.00
EWC, $\lambda = 0.001$	3	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.80±0.01
EWC, $\lambda = 0.01$	3	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.81±0.03

$\ell \approx 2.5$ **— New knowledge and compute**

Method	n	PFLOPs	B (all)	B (useful)	B (zero)	U_{future}
Naive	3	63.6	0.51±0.01	0.51±0.02	0.51±0.01	0.37±0.00
Replay (1.0 epoch)	3	63.6	0.45±0.01	0.45±0.00	0.45±0.01	0.61±0.02
Replay (1.42 epochs)	3	90.1	0.46±0.01	0.46±0.02	0.47±0.01	0.62±0.00
Uniform sleep	3	88.5	0.17±0.00	0.18±0.00	0.17±0.01	0.58±0.00
Utility-weighted sleep	3	88.5	0.13±0.00	0.27±0.01	0.00±0.00	0.64±0.00
Utility-weighted replay	3	63.6	0.37±0.00	0.74±0.01	0.00±0.00	0.75±0.01
EWC, $\lambda = 0.001$	3	63.8	0.00±0.00	0.00±0.00	0.00±0.00	0.41±0.01
EWC, $\lambda = 0.01$	3	63.8	0.00±0.00	0.00±0.00	0.00±0.00	0.41±0.01

Nominal sizing and measured load. Before the sweep, we set world sizes using $k =$ learnable bits$/C_{\text{ref}}$, with $C_{\text{ref}} = 3.2$ Mbit for the 5M model and 6.4 Mbit for the 11M model. The 3.2 Mbit reference came from the pre-sweep calibration with 200k facts and 1,000 exposures: 98.3% recall, or 0.64 bits per parameter for a nominal 5M parameters (Table 3). The larger model's worlds were sized using a nominal 10M-parameter reference, giving $0.64 \times 10\text{M} = 6.4$ Mbit, although the model is reported as 11M. Consequently, $\ell = kC_{\text{ref}}/C_{\text{eff}}$, giving $\ell \approx 0.49k$ for the 5M and $\ell \approx 0.53k$ for the 11M model. The measured load itself uses the actual number of facts and the measured capacity. The measured plateaus behind C_{eff} are 6.7 and 6.1 Mbit at $k = 10, 20$ for the 5M model (we use 6.5) and 12.3 and 11.4 Mbit at $k = 5, 10$ for the 11M model (we use 12).

Operator and arm details. Frequency and utility are decorrelated by construction, so no frequency proxy could recover the oracle utility. One-epoch replay uses 20–28% fewer FLOPs than the sleep arms and 1.42-epoch replay 2–13% more (Table 8). Online EWC uses a diagonal empirical Fisher normalised to mean one; $\lambda \geq 0.1$ blocks learning altogether, and at $\lambda = 0.001$ phase-B facts are learned to 57% and 0% at $\ell \geq 1$, with rules kept at 0.62–0.80 against 0.53–0.61 under naive training. Three seeds per arm and load, six for the replay, utility-weighted replay and utility-weighted sleep rows at $\ell = 0.25$ and 1.

Fact stores by exact count. Counting complete facts at 16 bits each, the store at $\ell = 1$ is 3.1 Mbit for naive training, 3.4–3.6 for replay and 1.5–1.9 for sleep, against 4.0 Mbit for joint i.i.d. training from the shared checkpoint (weighted sleep keeps its high-utility part at 1.49 Mbit against 1.54 for naive training); at $\ell = 2.5$ it is 4.1 Mbit for naive training, 3.6–3.8 for replay, 1.4 for uniform sleep and 1.2 for weighted sleep; by loss-based bits, naive 5.7, replay 5.4–5.6, utility-weighted replay 4.9, weighted sleep 2.6–3.0 and joint training 6.6 Mbit. This count treats a fact as stored or not, unlike the loss-based bits used for C_{eff}, which also credit partial memories.

Per-load values behind Section 3. Linear-probe accuracy at the end of A is 0.88, 0.75, 0.60, 0.36, 0.10, 0.02 at $\ell = 0.12, 0.25, 0.49, 0.98, 2.5, 4.9$ for the 5M model and 0.88, 0.75, 0.53, 0.34, 0.12, 0.03 at $\ell = 0.13, 0.27, 0.53, 1.1, 2.7, 5.3$ for the 11M model; rule retention after B stays at 0.44–0.63 across the sweep; at $\ell = 4.9$ ($\ell_A = 2.5$) one million facts hold 6.7 bits each on average and 19% of them are learned; per-load plasticity values are in Appendix F.3.

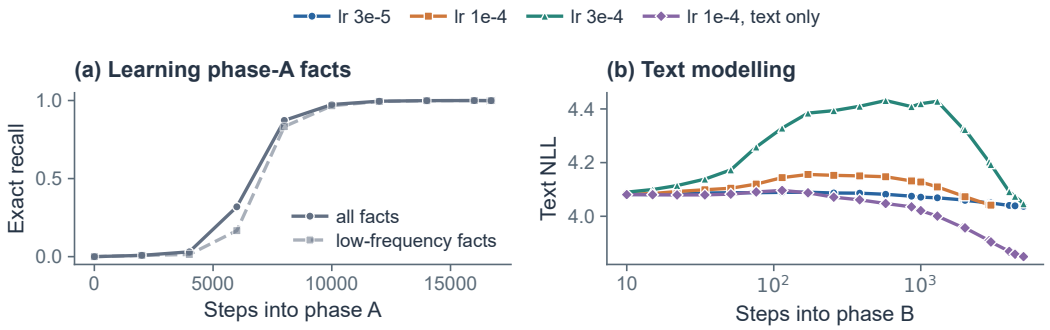

Figure 9: Training context for Figure 4 (Pythia-160M, 10,000 persons, one seed shown per condition). (a) Recall of all phase-A facts and of the low-frequency subset during phase A (100 exposures = 16,710 steps). (b) Held-out text NLL during phase B for the three learning rates and text-only control (shared legend; lower is better).

K PRETRAINED-MODEL REPLICATION

Model: Pythia-160M (EleutherAI/pythia-160m), 12 layers, width 768, GPT-NeoX tokenizer. Facts: five attributes per person, values drawn uniformly from tables of 64 years $\times$ 12 months $\times$ 28 days (14.4 bits), 256 cities (8), 128 universities (7), 64 fields (6) and 256 employers (8); names are unique first–last pairs, and the phase-A and phase-B populations share no name. Each fact is one line, e.g. "Q: In which city was Mateo Patterson born?\nA: Wellington", about 16 tokens; a quarter of the persons are high-frequency (four times the exposures). Stream: batches of 32 sequences of 512 tokens, half from the fact stream (facts concatenated in a fresh random order each pass) and half from random windows of a 300M-token FineWeb-Edu sample. Phase A: 100 exposures (16,710 steps and 274M tokens for 10,000 persons; 84,138 steps for 50,000), AdamW, learning rate 3×10^{-4} with 100 warm-up steps and cosine decay, weight decay 0.1. Phase B: 3,000–5,000 steps at the learning rates given in the text, checkpoints at the same log-spaced steps as in the testbed. Recall is teacher-forced exact match on the answer tokens of a sample of 5,000 training facts; stored bits per fact are $\max(0, b_i - \mathrm{nll}_i / \ln 2)$ with b_i the attribute's bits. Probes read the residual stream at the final prompt token and predict the value index of one attribute at a time (6,000 facts, 70/30 split, best layer); the recovery test fine-tunes on all facts of 1% of the persons for 30 steps at learning rate 10^{-4} and reports exact match on the facts of the other persons; net recovery subtracts an equal-step control on unseen persons, the counterpart of the testbed's unseen-fact control (Section 2), and a shuffled-answer control is reported alongside in place of the noise control. Numerics: Pythia's forward pass is inaccurate under bf16 autocast (text NLL 3.24 in fp32 against 3.75 in bf16 for the pretrained model), so training uses fp16 autocast with loss scaling and every evaluation runs in fp32. Re-checked on 16 held-out batches of 16×512 tokens: fp32 3.31, fp16 autocast 3.32, bf16 autocast 3.81 (4.17 with the eager attention kernel), weights cast to bf16 3.81; after phase A the gap is gone (3.950 in fp32 against 3.951 in bf16), so the loss is a property of the released, fp16-trained weights and not of our evaluation, and none of the reported numbers is affected. Phase A costs the model 0.8 nats of text NLL (3.31 to 4.08), of which little is recovered in phase B.

Joint learning of both fact sets in Pythia-160M. Trained from the pretrained weights on the union of the two 10,000-person worlds (100k facts, 868 kbit) with the protocol of phase A, the model recalls 99.1% of each set after 60 exposures (98–100% of the high-frequency and 99% of the low-frequency facts), with text NLL 4.19 against 4.08 when one world is learned alone. Both sets fit at once, so the sequential collapse of Section 6 is not a shortage of room.

Format-only control. Phase B of the 10k-person Pythia-160M run re-run from its end-of-A checkpoint with the same 10,000 new persons and the same facts rendered as statements ("Jacob Rios was born in the city of" $\rightarrow$ "Beijing") instead of question–answer lines, at learning rate 10^{-4}: phase-A recall 0.97, 0.84, 0.75, 0.67, 0.65, 0.58, 0.57, 0.56, 0.59 at 15, 22, 34, 51, 76, 114, 171, 256 and 384 steps (same-format: 0.92, 0.24, 0.08, 0.03; text only: 1.00 through 76 steps and 0.22 after 1,297), net recovery 0.24–0.34 throughout, linear-probe accuracy from 0.29 to 0.16 over the same steps, and the new facts learned to 0.22, 0.42 and 0.53 after 1,946, 3,000 and 5,000 steps.

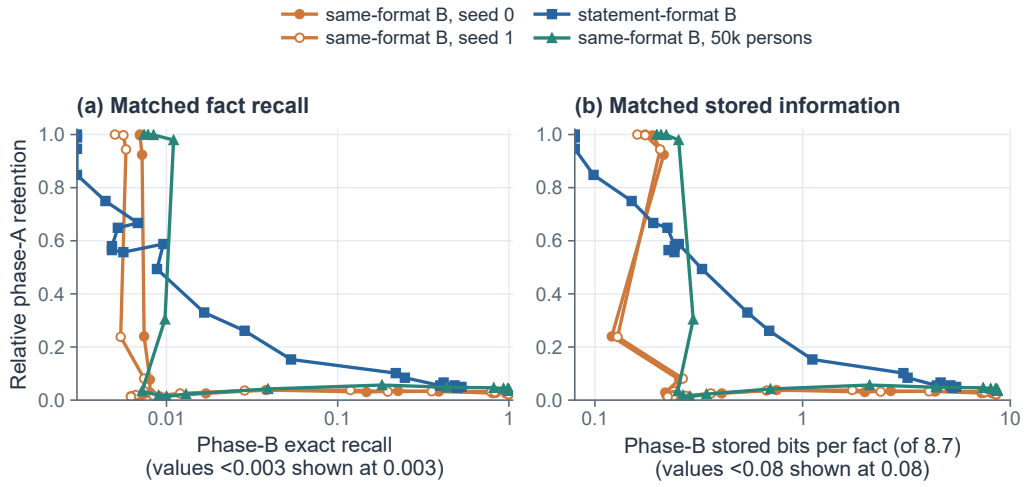

Figure 10: Pythia-160M: relative phase-A retention against phase-B facts recalled (a) and loss-based bits stored per phase-B fact (b), one point per log-spaced checkpoint. Same-format (question–answer) new facts collapse phase-A recall while holding 0.15–0.25 bits per fact (out of 8.7); with the statement format phase A keeps most of its recall until the new facts hold about 0.3 bits and then declines as they are stored.

Formats compared at matched phase-B storage. Figure 10 plots phase-A retention against what phase B has stored, exact match and loss-based bits, at every log-spaced checkpoint of the Pythia runs. With same-format new facts phase A falls to a quarter while they hold 0.15–0.25 bits per fact (out of 8.7; 0.7% exact); with the statement format phase A stays above 0.55 until the new facts hold about 0.3 bits and then declines as they are stored: at 0.7 bits per fact the statement format keeps 0.26 of phase A against 0.04 for the same format, at exact match 0.22 it keeps 0.10 against 0.03, and at 4.4 bits 0.055 against 0.03. The format therefore changes both the clock and the trade-off while the new facts are little stored, and the two converge once phase B is well stored; the 50k-person world follows the same-format curve with a later drop.

Why the capacity regime is not reached in a pretrained model. We tried to push a real model past its capacity with Pythia-14M (about 17 Mbit, the information in about 400k persons, if the testbed's 1.1–1.3 bits per parameter applied to all its parameters, most of which are embeddings). The attempt fails for kinetic, not capacity, reasons: at 100 exposures the model stores 63% of the facts of 20k persons (0.9 Mbit, with the low-frequency facts at 52%), 4% of the facts of 80k persons (3.5 Mbit) and under 1% at 200k–800k persons halfway through phase A, even when warm-started from the 20k-person checkpoint and trained at a higher learning rate. The memorisation plateau lengthens with the number of facts faster than the exposure budget we can afford, exactly the confound the testbed removes with its shared pretraining phase; a storage failure measured there would be exposure-limited, so we do not report it. The 20k-person run itself reproduces the interference regime once more (collapse at 22–34 steps, net recovery 0.32, probe accuracy to chance in step with recall).