LAYERWISE FUNCTION ANCHORING:

PRESERVING SUB-MODULE FUNCTIONS ON SAMPLED HIDDEN STATES FOR CONTINUAL DOMAIN ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Fine-tuning a language model on a narrow domain erodes the broad capabilities it arrived with, and can lead to catastrophic forgetting. The classical remedy, rehearsal, mitigates forgetting by replaying pretraining text at every stage. We present Layerwise Function Anchoring (LFA), which preserves capability with a statistic instead. LFA estimates, once, the distribution $p(h)$ of hidden states entering each sub-module of the frozen original when text is passed through: a compact, sequence-free, task-agnostic statistic (under a tenth of the model's size). During adaptation LFA penalizes each sub-module's output drift $\|f_{\mathrm{s}}(h) - f_{\mathrm{t}}(h)\|^2$ independently on point samples $h \sim p(h)$, bypassing resource-intensive self-attention. No exemplars are stored, no text is generated during adaptation, and the statistic can be updated exactly as domains accumulate. Two knobs do the work: λ prevents forgetting by penalizing layerwise functional drift, and μ, an L2-SP penalty on all weights, covers the directions $p(h)$ does not see. Anchoring sub-modules independently risks compounding error through depth; every whole-network measurement below (perplexity, judged answers, rule-scored generation) bounds that risk, though none isolates it, and the cleanest bound is tight: on Qwen3-0.6B adapted to a philosophy corpus under LoRA, unanchored adaptation inflates general-text perplexity by over 400% and LFA holds it to $+1.1\%$. LFA's time overhead falls with model size, $+14.8\%$ per pass at 0.6B, $+6.3\%$ at 1.7B and $+5.5\%$ at 3B, at lower peak memory than Learning without Forgetting (LwF) at 0.6B and 1.7B. Judged by a frontier LLM, LFA matches tuned LwF at each method's best dose in paired per-question tests and is not separable from tuned replay, both strengthened with the same L2-SP penalty. Merging a learned domain's statistics into $p(h)$ supports continual learning with no retained text: after a second domain LFA leads replay and LwF on the new domain's judged quality at all three ranks, and trails replay on old-domain perplexity. After a third domain its old-domain judged quality ends tied with replay, and LFA is ahead of both rehearsal and distillation on rule-scored generation (GSM8K, IFEval), where every perplexity on given text puts replay first. LFA's anchoring statistic need be produced only once, allowing model providers to support continual learning without divulging proprietary training data. When not available, the statistic can be formed using self-generated sequence samples, so LFA can run data-free, with nothing beyond the new domain's own text (one model, one domain). At Qwen3-1.7B (one seed) the continual-learning ordering is the same and every method loses less against the base model. All experiments use LoRA; seed coverage is partial.

1 INTRODUCTION

Adapting a language model to a specialized domain degrades the broad capabilities it arrived with. The remedies that hold up best re-read text: rehearsal retains and replays it at every stage — the original data, a proxy for it, or text the model generates — and distillation runs the old model over it. Both treat preserving like learning, jointly training the model's layers on full sequences. Penalties on the weights alone need no text, and fall short here (§3.4, §7, §8).

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.

We present Layerwise Function Anchoring (LFA), which preserves model layers independently with a sequence-free *statistic* instead, and with it evidence that preserving is a different problem from learning. Learning must train the model jointly on long sequences; preserving only has to keep each trained map doing what it already does. Every map that trains here acts on one position's vector, and the one operation that mixes positions, attention's softmax, has no trainable parameters: sequence dependence lives in the inputs a sub-module receives, not in what it does to a vector, so were each map unchanged on its real inputs, the network would be unchanged on real text (Proposition 1, App. A.1). LFA therefore anchors each sub-module in isolation, on independent vectors: we estimate, once, the distribution $p(h)$ of the hidden states it receives, and during adaptation penalize the adapted sub-module's output drift on samples from it. The anchor holds each map only approximately, so the argument is checked on held-out text (App. A.1). In this work $p(h)$ is fitted on the same public proxy corpus that replay rehearses. The method is *text-free*: it stores, rehearses and generates no text *during adaptation*. It can also run fully *data-free*, in the self-supplied mode of §8, where nothing is supplied beyond the new domain's text (one model, one domain).

What we find is a method that *works*, by a different mechanism; its standing against tuned alternatives is our evidence, not a claim that it is the best option. LFA preserves with no stored, rehearsed or generated text, at the lowest time per pass among the timed methods that preserve; after three domains it is ahead of both rehearsal and distillation on rule-scored generation at two model sizes, with every perplexity on given text putting replay first; and on held-out old-domain text its sub-module functions move 37–313$\times$ less than replay's or LwF+μ's (one seed but for the r=32 replay pair; no predictor of judged quality). All experiments adapt with LoRA adapters (main text r=4, 16, 32); the anchor is defined for any trainable map, and full-weight training is untested. In one domain LFA *matches* tuned LwF+μ (with L2-SP) in paired per-question judged tests, a paired tie at each method's best dose (App. D), and is not separable from tuned replay+μ, an unpaired difference below the minimum detectable effect. After a second domain it leads both on the new domain at all three ranks and trails replay on old-domain perplexity; its old-domain lead is largely answer style, surviving a length control only at r=32 (one of two seeds).

Contributions.

- **A different mechanism** (§3): each sub-module is anchored in isolation on hidden-state samples drawn from a shipped, exactly mergeable estimate of $p(h)$ whose form a pricing identity (Eq. 2) fixes. On held-out old-domain text its functions barely move (§6.1, App. A.1).
- **Evidence that it works, as a match** (§5–§6.1): paired tests against tuned baselines given L2-SP, ties named and "match" defined; after a third, old-domain answering is not separable from entry at r=16† and r=32 (equivalent within the margin at r=32; App. D, G); on rule-scored generation it is ahead of both at two model sizes, with every perplexity on given text reversed (Table 2).
- **An overhead lower at 1.7B and 3B than at 0.6B** (§7, App. I): $+14.8\%$ time per pass at 9.5 GB peak memory against LwF+μ's 15.6 GB (replay's is lower); launch-bound, so fixed per site: $+6.3\%$ at 1.7B and $+5.5\%$ at 3B.
- **L2-SP as a complement** (§7, App. J): this weight penalty, not ours, strengthens replay and a re-tuned LwF (not GeRe), lifting both on rule-scored generation in six of six contrasts; alone it does not preserve.
- **The recipe's inputs can be self-generated** (§8): the model writes both its proxy corpus and its supplement, tying the recipe, $\approx$0.1 behind keeping the frontier supplement (the *data-free mode*; one seed, one model, one domain).

2 Related Work

Regularization-based continual learning. Continual-learning methods form three families (Shi et al., 2024; Wu et al., 2024): replay (rehearsal); regularization of the weights (L2-SP) or of the function (LwF, also called distillation); and architecture (parameter isolation, adapters and PEFT such as LoRA). LFA is function regularization, sequence-free and layerwise-independent rather than joint at the output over sequences, and relies on L2-SP directly as a complementary regularizer (§3, §7). Matching a frozen teacher's intermediate outputs is classical feature distillation (Romero et al., 2014; Jiao et al., 2019; Passalis & Tefas, 2018); Learning-without-Forgetting (Li & Hoiem, 2016) applies it at the output level against forgetting, and L2-SP (Li et al., 2018) is the weight-space pull

we weight by μ. LFA's anchoring loss sits in this lineage and we claim none of these primitives. LwF anchors the teacher's *logits* on the new task's own inputs, constraining the endpoint on whatever those inputs exercise; LFA anchors each *sub-module's* output function on hidden states drawn from a shipped $p(h)$, so it reads no input text and constrains the computation through depth.

Data-free distillation and shipped statistics. Lopes et al. (2017) release per-layer activation statistics as a data substitute, and fitting a parametric activation distribution to resist forgetting is established (Liu et al., 2020; Petit et al., 2022; Goswami et al., 2023). The data-free-distillation family (Nayak et al., 2019; Yin et al., 2019; Haroush et al., 2019; Chen et al., 2019; Micaelli & Storkey, 2019; Ma et al., 2020; 2022; Bai et al., 2023; Smith et al., 2021; Gao et al., 2022; Tong et al., 2025; Krukowski et al., 2026) *synthesizes surrogate inputs* and distills on them, so each sample must be a legitimate input to the whole network; LFA samples hidden states directly into sub-modules of the model being adapted, with no input to reconstruct and no student.

Replay-free adaptation for LLMs. SSR (Huang et al., 2024) (with Yang et al., 2024a;b) also needs no prior data, by having the model *generate* substitute text and rehearsing it; we compare against it in §8. GeRe (Zhang et al., 2025) anchors activation consistency to a frozen reference but replays a buffer of real pretraining text. Anchored Learning (Wang et al., 2026a) anchors the token distribution of an SFT model, not hidden-state functions of the pretrained one. Subspace-constrained LoRA (Xiong & Xie, 2025; Wang et al., 2023; Luo et al., 2025; Wang et al., 2025; Qiao & Mahdavi, 2026), and the intruder-dimension analysis of Shuttleworth et al. (2024) are the nearest LoRA-side work.

A full account of related work is App. O.

3 Method

3.1 The anchoring term, and what it asks of $p(h)$

Let the teacher be a frozen pretrained model with sub-module functions f_{t} (Q/K/V/O, the MLP; lm_head under full-weight training) and the student its copy under fine-tuning, with functions f_{s}. Under LoRA every anchored sub-module is $W_{\text{base}}+BA$, with B and A the trainable low-rank LoRA factors and W_{base} frozen, so the student reads $f_{\mathrm{t}}(h)$ by running its own sub-module with the adapter disabled and no second model is loaded (bit-identical to a separate copy, App. L). With $p(h)$ the distribution of hidden states a sub-module receives on in-distribution text, the preservation term matches student and teacher on samples from it (Fig. 1a):

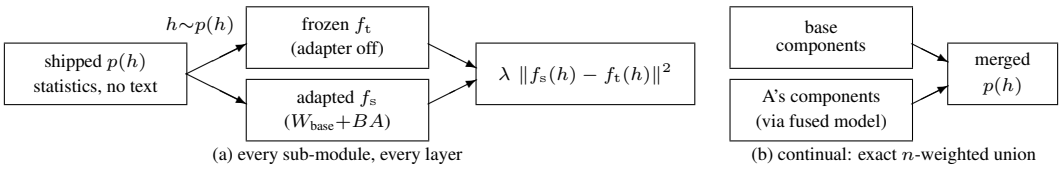
$$\mathcal{L}_{\text{repr}} \;=\; \mathbb{E}_{h\sim p(h)}\Big[\,\|f_{\mathrm{s}}(h)-f_{\mathrm{t}}(h)\|^2\,\Big]. \tag{1}$$

The penalty is finite where Proposition 1 asks for equality, so errors compound, a small change in Q or K can still move a softmax, and inputs drift off $p(h)$; App. A.1 states the gap and measures the part we can (Table 3).

Equation (1) also fixes how faithful the shipped estimate has to be. At a linear sub-module $f_{\mathrm{s}}(h)-f_{\mathrm{t}}(h)=\Delta W h$ with $\Delta W = W_{\mathrm{s}}-W_{\mathrm{t}}$, so

$$\mathcal{L}_{\text{repr}} \;=\; \mathbb{E}_{h\sim p(h)}\,\|\Delta W h\|^2 \;=\; \operatorname{tr}\big(\Delta W\, M\, \Delta W^{\top}\big), \qquad M \;=\; \mathbb{E}_{h\sim p(h)}\big[\,h\,h^{\top}\big], \tag{2}$$

and the loss *and its gradient* depend on $p(h)$ only through its second moment. We call this trace the *price* of an update: directions the model uses heavily are expensive, ones $p(h)$ barely visits are nearly free to adapt, and a mis-specified M *mis-prices* them. At a linear site the term is an EWC-style quadratic (Kirkpatrick et al., 2016) with the Fisher replaced by the input second moment alone (a K-FAC input factor (Martens & Grosse, 2015), identity output factor), and the soft form of null-space projection (OWM, GPM, Adam-NSCL; Zeng et al., 2019; Saha et al., 2021; Wang et al., 2021). What differs: M is fitted once on a proxy corpus and frozen, with no output factor; the MLP input gets a mixture; and domains merge exactly. Two consequences follow. At a linear sub-module a Gaussian carrying the correct *correlated* second moment is exactly sufficient, and it must be correlated: replacing M by its diagonal discards every cross term, a mispricing of 1.6–$4.1\times$ over nine sampled sites by an exact generalized-eigenspectrum comparison (App. B, B.1). That error is *directional*, not a scale the anchor's weight λ (Eq. 3) could absorb: the price is wrong exactly



(a) every sub-module, every layer (b) continual: exact n-weighted union

Figure 1: (a) LFA holds each adapted sub-module to the frozen one on hidden states sampled from a shipped $p(h)$. (b) A learned domain's statistics are merged into $p(h)$; no text is kept.

where the optimizer shops for cheap directions. And two-moment sufficiency stops where linearity does: the MLP's nonlinearity sees the full distribution, so it is the one site that needs a Gaussian mixture.

3.2 The shipped artifact

$p(h)$ is estimated once, offline, at every sub-module's input — by a provider, or, in the one setting we test, by a holder of the weights from text the model itself generates (§8). Equation (2) dictates the form: a 95%-variance PCA basis with correlated covariance at the linear sub-modules, and a $K{=}32$ Gaussian mixture in it at the MLP inputs. The artifact's size is set by the *model*, not the data, $O(\#\text{sub-modules} \times d^2)$: for Qwen3-0.6B, ≈226 MB fp16, or ≈113 MB in blockwise int8, ≈9.5% of the 1.19 GB bf16 checkpoint (cheaper tiers: App. B).

3.3 The objective, the schedule, and knob coupling

$\mathcal{L}_{\text{repr}}$ is blind to directions $p(h)$ does not see, so we add a uniform weight-space pull toward the teacher, $\mathcal{L}_{\text{weight}} = \|W_s - W_t\|^2$ (Li et al., 2018):

$$\mathcal{L} \;=\; \mathcal{L}_{\text{content}} \;+\; \lambda\,\mathcal{L}_{\text{repr}} \;+\; \mu\,\mathcal{L}_{\text{weight}}, \tag{3}$$

with $\mathcal{L}_{\text{content}}$ cross-entropy on the domain data. L2-SP is not specific to our signal: it composes with any content loss, so replay, LwF, GeRe, TM-only and self-generated rehearsal receive it as "$+\mu$", an L2-SP penalty of strength μ (§7). λ follows a per-layer cosine decay that partially compensates for the growth of activation scale with depth — the anchor is an absolute squared error, so effective relative pressure still rises with depth (App. A.3). *Knobs do not port*: LoRA rank is itself an implicit preservation mechanism, so λ is recalibrated with it (App. C.1), and no strength knob, ours or a baseline's, ports across rank, the supervised-answer fraction f of §4, artifact, or optimizer protocol: each is re-tuned, or the port is flagged. L2-SP's own part is measured: it lifts the baselines' rule-scored generation in six of six contrasts (App. J), cuts intruder dimensions at the shipped schedule, $86 \to 7$ (App. A.2), and alone does not preserve (§3.4); our hypothesis, untested, is that the weight pull keeps domain learning integrated with what the model already represents rather than routed through directions the anchor cannot price.

3.4 Evidence that the shape of $p(h)$ does the work

Under one frozen protocol (App. E.1), unanchored adaptation inflates WikiText-2 perplexity by +444% and LFA holds it to +1.1%. At matched $\mu{=}0.05$ and matched content loss, L2-SP (μ alone) forgets +12.9% where $\lambda{+}\mu$ forgets +1.1% (the arms match on rank, corpus, dose and content loss, not on optimizer protocol), so the function anchor and not the weight pull carries preservation. And the *shape* of $p(h)$ carries the effect: replacing a fitted $p(h)$ with an isotropic Gaussian at matched per-site scale ($r{=}16$, the fitted diagonal-moment artifact), at identical domain fit, moves general-text perplexity from −1.9% to +6.8% and lowers judged quality by 0.216 (paired, $n{=}154$, $p{=}1.9{\times}10^{-4}$; robustness judge $p{=}6.1{\times}10^{-4}$; the arms also differ in n, 512 vs. 16, and epochs, 30 vs. 20).

Anchoring also reshapes the update: under one counting protocol, at the shipped schedule, it cuts intruder dimensions (new singular directions; Shuttleworth et al., 2024) from 1,340 to 86 (λ alone) and seven (with uniform $\mu{=}0.05$), a fingerprint that does not predict judged quality among methods that work (App. A.2–A.5).

4 EXPERIMENTAL SETUP

Model, domains, adaptation. Qwen3-0.6B. Domain A is a philosophy-of-mind corpus (77 documents, 5.99 MB); domain B a cybersecurity instruction corpus, distant from A; domain C 100 recent open-access archaeology papers (App. N). Adaptation is LoRA on all seven projections at α/r=2 with frozen embeddings, lr 3×10^{-4} constant, batch 6, seed 42; a checkpoint's *dose* is its epochs trained, *feasible* at held-out domain perplexity $\leq$12 and general-text $\leq$+15%. The corpus carries a **supervised-answer fraction** f: raw text (f=0), or chat-formatted question–answer pairs mixed in to a token fraction f=0.13 and screened against the evaluation questions (written by `claude-sonnet-5` except in §8: the *frontier supplement*); domain perplexity scores such text, so is never compared across f. $p(h)$ is estimated from one 10:1 pretraining:instruction proxy corpus and sampled at n=16 states per site per step (512 in the rank and representation studies, App. N). Baselines: unanchored, L2-SP, real-data replay, LwF, EWC, VeFA, GeRe, "TM-only" (GeRe's activation-matching term alone, on generic text), SSR and, continually, O-LoRA (App. L, E.1, F). Deterministic perplexity (held-out domain; WikiText-2 as *general-text* perplexity) carries every knob-level decision; domain answering has no public benchmark, so an LLM judge scores it.

The two judges, and what a verdict means. Every judged number in this paper is `gpt-5.6-luna@medium` on the frozen protocol of App. M unless the sentence, cell or caption names `claude-sonnet-5`, our robustness instrument; no number from one judge is read against a number from the other, as a level or as a gap. The question sets are *set-A* (154 questions on A), *set-B* and *set-C* (78 each). A verdict is always a paired per-question test (two-sided Wilcoxon signed-rank unless stated, at the 0.05 level; a contrast that does not reach it is a tie), never a difference of levels. Judged means lie near 1 on the 5-point rubric and are reported beside their zero-rate, the share of rubric-dimension scores returned as exactly 0, wherever the source records one: archive-wide 30.6% for the primary judge and 39.3% for the robustness judge, a floor the instrument cannot see past.

Unpaired run-level differences are read against a minimum detectable effect of 0.12 (answer set) and 0.19 (configuration level), both measured on `claude-sonnet-5` (App. E); the same computation gives 0.147 and 0.180 on the primary judge, a reconstruction we take as its resolution. It gives "match" a meaning: a tie is *equivalent within the margin* ($\equiv$ in tables) when two one-sided paired tests place the difference inside $\pm$0.147, and *inconclusive* otherwise; only ties are so classified, a paired separation can be smaller than the margin, and the margin is post hoc, so descriptive. Of the main text's 63 ties, 37 are equivalent and 26 inconclusive; of its 61 separations, seven fail Benjamini–Hochberg correction within this family ($^\flat$; App. D). "Ahead", "behind" and "tie" are the primary judge's paired verdicts; p-values are in tables and captions, † or $*$ marking one the robustness judge reads differently. Generative contrasts are exact McNemar tests paired by item; the 1.7B contrasts are separately corrected replication families (App. H).

Frame, tuning parity, scope. Every cross-arm comparison is checked against a frame record (rank, LoRA α, f, epochs, gradient accumulation (ga), warmup, chunking) and any surviving difference is named. The single-domain search gave LFA 55 configurations against replay's 9, so **no judged ranking of methods is stated in the single-domain frame**: §5 and §8 report pairwise verdicts only. The continual frame gave every baseline its own knob ladders, and there we report pairwise verdicts against LFA with every tie named. Two seeds cover the shipped recipe and its single-domain LwF+μ comparator, the r=32 A→B chain and its canonical-replay (§6) partner, a rank-matched r=4 chain, the r=32 stage-three pair (§6.1) and the three contenders of §8; every other number is one-seed.

5 SINGLE-DOMAIN CALIBRATION: A MATCH, NOT A WIN

The shipped point. LFA ships at LoRA r=32, λ=10^5, μ=0.05, dose 15, on the correlated-covariance + mixture artifact, trained at f=0.13. Across two seeds it reaches held-out domain perplexity 8.76/8.83 while WikiText-2 perplexity moves −10.0%/ − 10.7% (base 18.18), and it judges 0.968/0.974 at zero-rates 26.8%/28.1% (its best feasible dose is 20; Table 13). We read −10% as *no measurable general-text perplexity cost*, not a gain (the anchor can reward proxy-like behaviour, App. C); generative benchmarks fall after stage one (App. F, vi).

Table 1: Continual A→B (seed 42): each method's tuned arm per rank, paired against the LFA arm of its block. ΔA, ΔB: held-out perplexity change over stage two; Wiki: WikiText-2 after stage two (base 18.18). Judged columns: primary judge, zero-rate in brackets where recorded; paired cells give Δ (arm minus LFA), Wilcoxon p, the arm ahead or "tie". Floors: the base model's 0.931 on set-A; on set-B the r=32 chain's pre-B fused model, 1.118 (37.4%, one seed), the only one scored, so indicative at r=4, 16. $^{\equiv}$, $^{\flat}$: §4. $^{\ell}$A tie on the two least length-sensitive dimensions (§6.1). *A tie under the robustness judge (p=0.073). ‡Seed 1337: 1.117, 1.89 (replay 0.901, 1.39); rest one-seed.

	Method (tuned arm)	ΔA↓	ΔB↓	Wiki	set-A	set-B	set-A vs. LFA: Δ, p	set-B vs. LFA: Δ, p
r=4	LFA, stage-2 λ=4×10^4	+3.51	−7.90	16.7	1.034 (21.0%)	1.90 (13.8%)	—	—
	LwF+μ, stage-2 α_{LwF}=8	+0.79	−6.53	18.3	0.997	1.47	−0.036, 0.4355, tie$^{\equiv}$	−0.423, <10^{-4}, LFA
	Replay+μ, canonical	+0.43	−6.78	16.0	0.845 (35.5%)	1.33 (28.7%)	−0.188, 0.0006, LFA$^{\ell}$	−0.567, <10^{-4}, LFA
	Second r=4 block: each method's judged-best arm, paired against LFA*'s judged-best arm*							
	LFA, stage-2 λ=6×10^4 (judged-best)	+3.12	−7.96	16.5	1.062	1.74	—	—
	LwF+μ, stage-2 α_{LwF}=16	+0.25	−5.53	17.7	1.035	1.41	−0.027, 0.444, tie$^{\equiv}$	−0.326, <0.001, LFA
	Replay+μ, A-only buffer, stage-2 μ=0.1	+1.46	−7.70	18.5	0.977 (29.7%)	1.29 (31.3%)	−0.086, 0.136, tie	−0.446, <0.001, LFA
	TM-only+μ (settings carried over)	+1.19	−10.35	17.3	1.057	1.75	−0.005, 0.986, tie$^{\equiv}$	+0.013, 0.694, tie$^{\equiv}$
r=16	LFA, λ=5×10^4	+3.90	−5.03	19.0	1.090	1.72	—	—
	LwF+μ, stage-2 μ=0.1	+2.72	−7.69	28.6	0.981	1.49	−0.109, 0.0496, LFA$^{*\flat\ell}$	−0.233, 0.0034, LFA
	Replay+μ, canonical, stage-2 μ=0.2	+1.16	−3.34	16.4	0.904	1.29	−0.186, 0.0021, LFA$^{\ell}$	−0.433, <10^{-4}, LFA
r=32	LFA, λ=10^5 ‡	+3.55	−4.81	18.5	1.227 (13.8%)	1.81 (16.4%)	—	—
	LwF+μ, μ=0.1	+2.95	−7.75	29.4	0.809	1.44	−0.418, <10^{-4}, LFA	−0.369, 0.0005, LFA
	Replay+μ, canonical, stage-2 μ=0.5	+1.52	−3.75	16.5	0.916	1.41	−0.312, <10^{-4}, LFA	−0.400, 0.0003, LFA$^{\ell}$

The judged result. Each method was tuned on its own preservation–adaptation knob at f=0.13 (Table 14). Against LwF given L2-SP too, LFA **ties** with each of the two LwF+μ arms at its own best feasible dose (p=0.18, 0.13), and **LwF+μ is ahead** at *matched* dose 15, at μ=0.05 (+0.135, p=0.0127) and at μ=0.1 (+0.148, p=0.0168; Table 13, App. E), and again within seed 1337; the rank-matched pairs are ties at r=16 and r=32. The best-dose tie holds within seed 42: against the seed-1337 LwF+μ arm the recipe's best dose is behind. The robustness judge reads three of the ten LwF+μ pairings differently and reverses none (Table 15). Against real-data replay given L2-SP too, the search's best LFA arm is ≈+0.13 higher, an unpaired margin below the 0.19 configuration-level threshold. They are not separable, and no paired primary-judge test exists; against replay without L2-SP that arm is paired-ahead at matched dose (−0.286, p=3.2×10^{-6}; App. E).

Against the never-adapted model. The recipe at dose 15 **ties** the base model on the five-dimension mean (0.968 vs. 0.931, p=0.52; p=0.40 at seed 1337; both $^{\equiv}$), while LwF+μ is ahead of it (p<10^{-4}). The tie is a cancellation: adaptation shortens answers 3.6×, so synthesis falls (−0.188, p=0.0016) as domain accuracy rises (+0.188, p=0.0034). On the two least length-sensitive dimensions (correctness, domain accuracy) the recipe is ahead of base at both seeds under both judges (p≤0.0085).

6 Continual learning with no retained text

Merging a learned domain's statistics into $p(h)$ lets LFA learn a second domain while holding no text of the first.

Mechanism and protocol. After stage one on A (the recipe of §5) the tuner *fuses* the adapter into the base weights and *extends* the artifact: A's hidden states, collected through the fused model while A's corpus is still on hand, are fitted as mixture components and merged by an exact n-weighted union (App. B.3). Stage two on B is §3 again with the merged $p(h)$ and the fused model as teacher. **At stage two nothing of A survives except statistics**: a merge adds ≈4.5 MB, little saving in bytes over A's 6.0 MB corpus, but no text is kept. In Table 1 replay is reported in its *canonical* configuration (every prior corpus in the buffer, L2-SP at both stages); above r=4 only its stage-two μ is re-tuned, its other settings carried over. The arms are not optimizer-step matched (ga=1 for LFA and LwF, 4 for replay); step-matched LFA controls tie the plain chain after B at r=32 (set-A$^{\equiv}$ and set-B, both leads kept under both judges) and after C at every rank (App. F, G).

New domain. On set-B LFA is ahead of the LwF+μ and canonical replay+μ arms of Table 1 at all three ranks; on the two least length-sensitive dimensions five of those six leads hold (the r=16 one over LwF+μ only uncorrected) and the one over replay at r=32 becomes a tie (p=0.078). One LwF arm outside the table ties LFA: the r=16 μ=0.05 mirror chain (LFA's μ at both stages; p=0.087). At

r=32 both five-dimension leads over canonical replay+μ recur at a second seed under both judges (set-A $+0.216$, set-B $+0.503$; all $p<10^{-4}$); on the two least length-sensitive dimensions each seed keeps one of the two under the primary judge (App. F). So anchoring does not cost plasticity.

Old domain. On set-A LFA is ahead of canonical replay+μ at all three ranks and of LwF+μ at r=16 and r=32. The r=16 lead over LwF+μ is a tie under the robustness judge and does not survive the multiplicity correction. At r=4 the result is a **tie band**, a set of arms joined by paired ties with no ordering inside it: LFA ties LwF+μ; replay's *judged-best* r=4 arm ties LFA's judged-best arm, so the lead over replay depends on which arm represents it; and TM-only+μ (settings carried over) ties LFA on both judged axes (App. F).

Perplexity. LFA's fall on B is larger than canonical replay's at every rank; on what came before it has no advantage, trailing canonical replay on ΔA and WikiText-2 at every rank and LwF+μ on ΔA, while LwF+μ's WikiText-2 degrades above r=4 (Table 1).

6.1 After a third domain: old-domain answering against its entry level

Each chain continues to C from its own post-B model with *no knob re-tuned*: LFA merges B's components and passes no text, replay rehearses general+A+B in equal thirds, LwF distills from it.

What each side holds, and perplexity. Replay's general text is load-bearing on *perplexity*: dropping it at stage two raises WikiText-2 after B from 16.66 to 24.16 at r=16 (both μ=0.05). With no text of A, LFA's perplexity on A *improves* over stage three at every rank and its WikiText-2 is flat; replay keeps the lowest prior-domain perplexity at every rank.

The token-wise argument, checked (Table 3). On text the artifact never saw, LFA's anchored functions barely move: relative output drift of the 140 anchored sub-modules on held-out A text, pooled at r=32, reads $0.031 \to 0.037 \to 0.040$ ($\times 10^{-3}$; entry, after B, after C) against LwF+μ's $3.09 \to 6.31 \to 10.12$ and replay's $8.62 \to 10.41 \to 12.44$. On WikiText-2 states their drift is at parity with its value on $p(h)$'s own samples; on held-out domain-A text it is 3.7–4.7 times looser (3.7–7.9 over all ranks), a gap, and still about 10^{-4} of function energy at most, its residue at the attention-output projection. Replay's functions move 275 and 285 times more (entry, after B; 270 and 276 at a second seed, all other ratios being one seed). None of this predicts judged quality (App. A.1).

Judged answering rises, then returns (Table 20). Over stage two LFA's answers lengthen by about half and its judged A *rises* at r=32 (from its fused entry 1.012, $+0.216$, p=3×10^{-5}; a tie under the length control) and at r=16$^\flat$. It then returns: from after B to after C it falls at every rank (-0.147, -0.195, -0.262), and from entry to after C it is a tie at r=16 ($-0.061^\dagger$, a fall under the robustness judge) and r=32 ($-0.047^{\equiv}$) and a small fall at r=4 ($-0.091^\flat$; Table 21). Canonical replay's level and answer length move at neither stage; LwF+μ falls over stage two at r=16 and r=32 and over stage three at every rank.

So the after-B lead of §6 is not a retention margin. Only its r=32 cells survive the length control; about half the r=32 replay gap was there at entry; and at r=16 LwF+μ *entered* ahead of LFA (Table 21), so that lead is a crossing. The stage-three fall survives the length control at r=32 only, where the second seed repeats it. LFA and replay end in a paired **tie at every rank**; the robustness judge reads replay ahead at r=4 and r=16 (†; App. M). Against LwF+μ after C it ties at r=4 and r=16 and is ahead at r=32 (Table 22). On set-B LFA falls by 0.41 to 0.46 at every rank where replay's does not move; after C it leads replay at r=4 only$^\flat$, and not under the length control.

What judged C can see. Rehearsing A's own question–answer pairs inside LFA restores judged A at no rank (App. G). Nor does judged C register C being learned for any method: no chain's judged C separates above the base model's 1.028 (its *recall tier* is unlearnable at these doses; floors: Table 22), and learning shows instead as a 31–48% fall in C perplexity for the three chains (App. M.1).

Behaviour against likelihood (Table 2). After C, LFA is ahead of LwF+μ and of replay+μ on both rule-scored tasks, the baselines tie each other on GSM8K, and LFA itself ends below the base model on both. Every perplexity on given text runs the other way: on GSM8K's gold solutions

Table 2: Generative retention after the third domain (seed 42; correct of 200, sampled decoding, one draw a cell; p: exact McNemar paired by item, the row against LFA, then the arm ahead), with WikiText-2 and the perplexity of GSM8K's gold solutions on the same models. 1.7B: reduced-tuning replication, one seed (App. H); its GSM8K lead over LwF+μ separates on this draw (q=0.022), not quoted as a size. Bold: lowest.

	after C	GSM8K	IFEval	p, GSM8K	p, IFEval	Wiki	GSM8K-PPL
Qwen3-0.6B	base model	101	116	2.9×10^{-5}, base	0.047, base	18.18	4.63
	LFA	65	104	—	—	18.76	3.61
	LwF+μ	32	60	1.3×10^{-5}, LFA	3.2×10^{-8}, LFA	38.72	4.76
	replay+μ	32	79	6.7×10^{-5}, LFA	0.0038, LFA	**17.71**	**3.11**
Qwen3-1.7B	base model	152	136	9.3×10^{-6}, base	0.18, tie	14.96	—
	LFA	121	123	—	—	11.38	—
	LwF+μ	102	89	0.016, LFA	5.5×10^{-6}, LFA	22.54	—
	replay+μ	98	95	0.0044, LFA	2.5×10^{-4}, LFA	**10.81**	—

replay+μ is lowest and LwF+μ highest, and replay is first on WikiText-2, instruction text and every reference-answer column; so the instrument licenses a lead in sampled chat-format task performance and not one in reasoning knowledge. Rehearsal keeps the model close to what it rehearses; both keep factual recall (TriviaQA ties); LFA keeps more format and multi-step reasoning behaviour.

A second model size. At Qwen3-1.7B (knobs re-tuned at that scale; one seed, primary judge only; App. H) the ordering replicates with every loss against the base model smaller: after B and after C LFA is ahead of both baselines on both generative tasks (over LwF+μ on this draw: $q\approx0.02$ after B, and on GSM8K after C), level with base on IFEval, and after C every method is below base on GSM8K. Under the judge after C it is ahead of LwF+μ on all three sets and both axes, ties replay+μ on A and B and is ahead on C; replay+μ is ahead of LwF+μ on set-A, and LFA ties its own entry level on set-A after three stages, as replay+μ does and LwF+μ does not. The instruments disagree on LFA against replay+μ on the retained domains (generation separates, the judge ties) and on the two baselines (generation ties, the judge reads replay ahead on set-A); stated, not resolved.

7 Fairness and cost

What L2-SP adds to other methods. L2-SP (Li et al., 2018) complements other preservation signals. Added to replay it lifts judged quality in one domain under both judges (+0.122, p=0.033, primary) and cuts the old-domain perplexity cost of a second domain from +3.29 to +1.07; it lifts LwF when α_{LwF} is re-tuned with it. Alone it does not preserve (§3), it leaves GeRe unchanged, and its judged effect on replay after a second domain is a tie. LFA carries L2-SP, so against baselines without it part of its margin would be L2-SP's (its standing against the μ=0 arms: App. J). So five baselines carry L2-SP and we compare against the strengthened versions (Table 31).

Cost. Table 30 (App. I) prices one pass at r=4, each overhead against its own control: LFA's +14.8% at 9.54 GB is the lowest among the *timed* methods that preserve (LwF+μ +32.8% at 15.61 GB, canonical replay +31.6% at 8.75 GB; L2-SP is cheaper, +1.3%, and does not preserve; its r=4 A→B arm: ΔA +14.57, WikiText-2 70.6), though not on peak memory or general text. The anchor is launch-bound by measurement (0.51 TFLOP/s against a 21.6–62.2 matmul roofline; 512× the samples cost 1.26× the time): a fixed per-site cost under a step that grows with the model. At 1.7B and 3B it is +6.3% and +5.5% (like-for-like +14.6/6.3/5.4%; LwF+μ +32.9/28.3/28.4%), and LwF+μ is out of memory at micro-batch 12 at 1.7B and 8 at 3B, where LFA fits 16 and 12.

8 Self-supply: the recipe's inputs, and a generating competitor

Two inputs tie the recipe to outside text: the proxy corpus $p(h)$ is fitted on, and the frontier supplement. The model can write both and tie the recipe, at a measured cost, and with every method on the model's own text LFA ties self-generated rehearsal. Verdicts here are pairwise, single-domain, n=154, under the primary judge, which alone scored the both-replaced arm and the head-to-head. We quote paired

t-tests; Wilcoxon agrees on every contrast below but one (App. K). Self-supply was explored last, after the main experiments, so self-supplied inputs through the continual chain are untested.

Each input, then both. At the recipe configuration (r=32, **one seed**), an artifact fitted on 2,500 documents the model wrote from a bare document-start token ties the sample-count-matched real-text artifact in all six comparisons (three λ values, two doses; p=0.079 to 0.81; four of the six equivalent within the margin; Table 35). A supplement written by the base checkpoint ties the frontier-written one at both doses, though at dose 15 the robustness judge reads the frontier one ahead (paired t, p=0.011). With both replaced, the *data-free mode*, the arm ties the recipe at both doses (≡).

What that does and does not show. The fully self-supplied arm is behind the arm that keeps the frontier supplement over the self-generated artifact: by 0.110 at dose 15, where the t-test reads behind (p=0.033) and Wilcoxon reads a tie (p=0.056), and by 0.121 at dose 20 (p=0.023; Wilcoxon $0.039^{\flat}$). The artifact ties say a provider release is a convenience, not a requirement, in this one setting (one seed, r=32, one model, one domain). One boundary is measured: at f=0, where no supplement exists, the self-generated artifact is *behind* the shipped one at dose 20 (−0.138, Table 34).

The self-supplied head-to-head. At r=16, with the self-written supplement as every method's corpus, SSR rehearses 40,000 model-written documents and LFA anchors on the self-generated artifact. **LFA and self-generated rehearsal tie over two seeds** (Table 36). At dose 20 LFA is ahead at seed 42 and ties at seed 1337; at dose 15 they tie at seed 42 ($^{\equiv}$) and *SSR is ahead* at seed 1337 ($^{\flat}$). SSR's dose-20 checkpoints exceed the domain-perplexity bound (14.7/14.4 against 12; LFA's read 10.3/9.8), so dose 15 is its only feasible dose; at each method's best feasible dose (LFA 20, SSR 15) the two tie at both seeds (p=0.49, 0.54; both $^{\equiv}$), LFA with no buffer and no rehearsal text. Against LwF, which holds no buffer either, LFA is ahead at dose 20 at both seeds and ties at dose 15 ($^{\equiv}$ at seed 1337); the μ-only arm is behind at both doses (App. K.4).

Every setting is carried over from the frontier-supplement experiments: one setting per method, not a tuned comparison. The neighbouring LwF and SSR settings we ran leave the verdicts standing, but LFA one step away (λ=5×10^4) is behind SSR at dose 15 (p=0.0072). LFA's dose-20 score costs more retention than SSR's: WikiText-2 drift $+3.7\%/+4.0\%$ against $+0.8\%/+0.9\%$.

9 Limitations

Scale and training method. Two model sizes, LoRA only; the objective applies to full-weight training, untested; the 1.7B replication is one seed, one judge, and reuses 0.6B's equivalence margin. **One chain.** One domain pair in one order, extended once, most arms one-seed (§4), no replay chain dropping an earlier domain's text; the merge grows linearly with domains. **LLM judges.** 17–26% of verdicts are judge-dependent (App. M). **Generation is sampled.** A second draw of the same weights moves a count on this instrument by up to 13 items (0.6B) or 17 (1.7B), and two draws of one checkpoint once separate paired (p=0.002, q=0.029 over fourteen repeats), so a separation with q near 0.02 is a finding of its draw, not a size; and the lead does not extend to multiple choice, which separates the methods once in eighteen contrasts, and is reversed on every perplexity of given text. **The open question.** Whether a signal defined on per-site hidden-state functions can do more for judged answering on a prior domain than keep it near its entry level. Sixteen more: App. P.

10 Conclusion

What a model must not forget can be carried by a statistic instead of text, with no sequence in the penalty: a shipped estimate of $p(h)$, on samples of which each sub-module is held to its frozen self. The evidence that this works is a *match*: level with tuned alternatives in one domain (paired against LwF+μ, unpaired against replay+μ), ahead in paired tests with named ties after a second, behind replay on perplexity retention, and after a third not separable from its own entry level on old-domain judged answering at r=$16^{\dagger}$ and r=32, as replay is not. Its boundary: the old-domain lead after a second domain is largely answer style, not retention, and by the third LFA is tied with replay. LFA preserves generative behaviour with no stored text and matches or beats rehearsal and distillation where produced behaviour is measured, at two sizes; on likelihood of given text, rehearsal is first.

Reproducibility Statement

The supplementary material is an anonymised archive containing: the method code that implements the objective of §3 (the $p(h)$ estimator and sampler, the anchoring and weight terms, the training loop, and the exact mixture merge that extends $p(h)$ to a further domain); the command that trains the shipped operating point of §5; the judge question sets (the two 77-question halves of domain A's 154-question set, the 78-question sets for domains B and C, with each question's source passage only where the source license allows it) and the five-dimension rubric; the evaluation and paired-test scripts; the per-question answer and score files behind every paired test reported here, the judged ones labelled with the judge identity (model and effort) they were produced under; and a one-page README. It contains no corpora, no estimated distribution, no checkpoints and no experiment sweeps, and it is not a reproduction harness for this paper's tables. Training seeds are 42 and, where a result is replicated, 1337; the main-text result tables state their seed coverage. The domain-A corpus is *not* redistributable, and its cleaning pass is unscripted, so a rebuild from the same source pages will be close to but not identical with the corpus used here (App. N). On publication the authors will release the code as a package, the two estimated distributions for Qwen3-0.6B (the correlated-covariance + mixture artifact this paper ships at, $\sim$113 MB int8, and the $\sim$0.5 MB diagonal floor of App. B, a size floor and not a recommended shipping tier) and the complete list of source documents with direct links and the extraction recipe.

AI Use Statement

Generative AI as a research assistant. This work was carried out with Claude Code (Anthropic), an agentic coding assistant running Anthropic's Claude models, at every stage of the project. In the terms of the ICLR policy, we used it to implement methods (the LFA objective, the $p(h)$ estimator and sampler, every baseline, the training and evaluation pipelines); to design and give feedback on experiments and research methodology; to propose and refine hypotheses; to clean and reformat the training corpora; to interpret results and run the statistical analyses; to help develop the paper's conceptual framing; and to help formulate and write out Proposition 1 and its proof. We also used it to create figures, create and edit code, survey, search and format the literature, and draft and edit the text. We did not use it for translation; surveys, interviews and transcription are not applicable to this work. The core idea of LFA — anchoring each sub-module's output function on samples from an estimate of the model's own hidden-state distribution, shipped in place of stored data — originated with the authors, and the authors directed the project throughout: they set its questions and its sequence, accepted or rejected each experiment, hypothesis, interpretation and framing the assistant proposed, and decided which claims the paper makes and at what strength. The working relationship was a close collaboration: the authors were the conceptual lead, involved in every decision and aware of all of the work, and the assistant took direction and carried out the engineering and experimental labour, much as a team of software engineers and graduate students would under the authors' direction. We reviewed all AI-assisted work: code was checked by tests and by re-running it, every reported number was traced to the output file that produced it, each paired test was run from the per-question score files shipped in the supplementary material, and every reference was checked against its source; further LLM instances were used as independent reviewers of code, analyses and text, with their findings checked by the authors against the same sources. We take responsibility for the final content of this work, including text, claims, code and artifacts produced with the aid of generative AI.

LLMs inside the study. Separately, LLMs are part of the method and the instruments, as documented in the body. All judged quality is produced by an LLM judge, `gpt-5.6-luna@medium` as the primary instrument and `claude-sonnet-5` as the robustness instrument, scoring generated answers against source excerpts on a fixed five-dimension rubric (§4, App. M); the judges' resolution, zero-rate, length sensitivity and disagreement rate are measured and reported rather than assumed. Two study inputs are synthetic: the judge question sets, generated from the domain documents by `claude-sonnet-4-6` (domains A and B) and `claude-sonnet-5` (domain C), so that the primary judge wrote none of them (App. M); and the question–answer supplement that makes up the fraction f=0.13 of the training corpus, written by `claude-sonnet-5` except in §8, where the supplement and the artifact's seed corpus are written by the base checkpoint under study, which is the object of that section.

REFERENCES

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory Aware Synapses: Learning what (not) to forget. *arXiv preprint arXiv:1711.09601*, 2017.

Zheyuan Bai, Xinduo Liu, Hailin Hu, Tianyu Guo, Qinghua Zhang, and Yunhe Wang. Data-Free Distillation of Language Model by Text-to-Text Transfer. *arXiv preprint arXiv:2311.01689*, 2023.

Frederik Benzing. Unifying Regularisation Methods for Continual Learning. *arXiv preprint arXiv:2006.06357*, 2020.

Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John P. Cunningham. LoRA Learns Less and Forgets Less. *arXiv preprint arXiv:2405.09673*, 2024.

Hanting Chen, Yunhe Wang, Chang Xu, Zhaohui Yang, Chuanjian Liu, Boxin Shi, Chunjing Xu, Chao Xu, and Qi Tian. Data-Free Learning of Student Networks. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. doi: 10.1109/ICCV.2019.00361.

Dongkyu Cho, Taesup Moon, Rumi Chunara, Kyunghyun Cho, and Sungmin Cha. Forget forgetting: Continual learning in a world of abundant memory. In *International Conference on Learning Representations*, 2026. arXiv:2502.07274.

Jeremy M. Cohen, Simran Kaur, Yuanzhi Li, J. Zico Kolter, and Ameet Talwalkar. Gradient Descent on Neural Networks Typically Occurs at the Edge of Stability. *arXiv preprint arXiv:2103.00065*, 2021.

Gongfan Fang, Jie Song, Xinchao Wang, Chengchao Shen, Xingen Wang, and Mingli Song. Contrastive Model Inversion for Data-Free Knowledge Distillation. *arXiv preprint arXiv:2105.08584*, 2021.

Qiankun Gao, Chen Zhao, Bernard Ghanem, and Jian Zhang. R-DFCIL: Relation-Guided Representation Learning for Data-Free Class Incremental Learning. In *European Conference on Computer Vision (ECCV)*, 2022.

Dipam Goswami, Yuyang Liu, Bartłomiej Twardowski, and Joost van de Weijer. FeCAM: Exploiting the Heterogeneity of Class Distributions in Exemplar-Free Continual Learning. *arXiv preprint arXiv:2309.14062*, 2023.

Hao Gu, Mao-Lin Luo, Zi-Hao Zhou, Han-Chen Zhang, Min-Ling Zhang, and Tong Wei. Spectral Imbalance Causes Forgetting in Low-Rank Continual Adaptation. *arXiv preprint arXiv:2602.00722*, 2026.

Matan Haroush, Itay Hubara, Elad Hoffer, and Daniel Soudry. The Knowledge Within: Methods for Data-Free Model Compression. *arXiv preprint arXiv:1912.01274*, 2019.

Tyler L. Hayes, Kushal Kafle, Robik Shrestha, Manoj Acharya, and Christopher Kanan. REMIND Your Neural Network to Prevent Catastrophic Forgetting. *arXiv preprint arXiv:1910.02509*, 2019.

Jianheng Huang, Leyang Cui, Ante Wang, Chengyi Yang, Xinting Liao, Linfeng Song, Junfeng Yao, and Jinsong Su. Mitigating Catastrophic Forgetting in Large Language Models with Self-Synthesized Rehearsal. *arXiv preprint arXiv:2403.01244*, 2024.

Adam Ibrahim, Benjamin Thérien, Kshitij Gupta, Mats L. Richter, Quentin Anthony, Timothée Lesort, Eugene Belilovsky, and Irina Rish. Simple and Scalable Strategies to Continually Pre-train Large Language Models. *arXiv preprint arXiv:2403.08763*, 2024.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing Models with Task Arithmetic. *International Conference on Learning Representations (ICLR); arXiv preprint arXiv:2212.04089*, 2023.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. TinyBERT: Distilling BERT for Natural Language Understanding. *arXiv preprint arXiv:1909.10351*, 2019.

Hangzhan Jin, Sitao Luan, Tianwei Ni, Sicheng Lyu, Guillaume Rabusseau, Reihaneh Rabbany, Doina Precup, and Mohammad Hamdaqa. RL Fine-Tuning Heals OOD Forgetting in SFT. *arXiv preprint arXiv:2509.12235*, 2025.

Amir Joudaki, Giulia Lanzillotta, Mohammad Samragh, Seyed Iman Mirzadeh, Keivan Alizadeh, Thomas Hofmann, Mehrdad Farajtabar, and Fartash Faghri. Barriers for learning in an evolving world: Mathematical understanding of loss of plasticity. In *International Conference on Learning Representations*, 2026. arXiv:2510.00304.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *arXiv preprint arXiv:1612.00796*, 2016.

Patryk Krukowski, Jacek Tabor, Przemysław Spurek, Marek Śmieja, and Łukasz Struski. Stop Marginalizing My Dreams: Model Inversion via Laplace Kernel for Continual Learning. *arXiv preprint arXiv:2605.11804*, 2026.

Xuhong Li, Yves Grandvalet, and Franck Davoine. Explicit Inductive Bias for Transfer Learning with Convolutional Networks. *arXiv preprint arXiv:1802.01483*, 2018.

Zhizhong Li and Derek Hoiem. Learning without Forgetting. *arXiv preprint arXiv:1606.09282*, 2016.

Xialei Liu, Chenshen Wu, Mikel Menta, Luis Herranz, Bogdan Raducanu, Andrew D. Bagdanov, Shangling Jui, and Joost van de Weijer. Generative Feature Replay For Class-Incremental Learning. *arXiv preprint arXiv:2004.09199*, 2020.

Raphael Gontijo Lopes, Stefano Fenu, and Thad Starner. Data-Free Knowledge Distillation for Deep Neural Networks. *arXiv preprint arXiv:1710.07535*, 2017.

Mao-Lin Luo, Zi-Hao Zhou, Yi-Lin Zhang, Yuanyu Wan, Tong Wei, and Min-Ling Zhang. KeepLoRA: Continual learning with residual gradient adaptation. In *International Conference on Learning Representations*, 2026. arXiv:2601.19659.

Minrui Luo, Fuhang Kuang, Yu Wang, Zirui Liu, and Tianxing He. SC-LoRA: Balancing Efficient Fine-tuning and Knowledge Preservation via Subspace-Constrained LoRA. *arXiv preprint arXiv:2505.23724*, 2025.

Xinyin Ma, Yongliang Shen, Gongfan Fang, Chen Chen, Chenghao Jia, and Weiming Lu. Adversarial Self-Supervised Data-Free Distillation for Text Classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020. doi: 10.18653/v1/2020.emnlp-main.499.

Xinyin Ma, Xinchao Wang, Gongfan Fang, Yongliang Shen, and Weiming Lu. Prompting to Distill: Boosting Data-Free Knowledge Distillation via Reinforced Prompt. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI)*, 2022.

James Martens and Roger Grosse. Optimizing Neural Networks with Kronecker-factored Approximate Curvature. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.

Paul Micaelli and Amos Storkey. Zero-Shot Knowledge Transfer via Adversarial Belief Matching. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Hossein Mobahi and Peter L. Bartlett. Hilbert Operator for Progressive Encoding (HOPE): A Mathematical Framework for Deconstructing Learned Representations in Deep Networks. *arXiv preprint arXiv:2607.21366*, 2026.

Gaurav Kumar Nayak, Konda Reddy Mopuri, Vaisakh Shaj, R. Venkatesh Babu, and Anirban Chakraborty. Zero-Shot Knowledge Distillation in Deep Networks. *arXiv preprint arXiv:1905.08114*, 2019.

Gaurav Kumar Nayak, Konda Reddy Mopuri, Saksham Jain, and Anirban Chakraborty. Mining Data Impressions from Deep Models as Substitute for the Unavailable Training Data. *arXiv preprint arXiv:2101.06069*, 2021.

Nikolaos Passalis and Anastasios Tefas. Learning Deep Representations with Probabilistic Knowledge Transfer. *arXiv preprint arXiv:1803.10837*, 2018.

Lorenzo Pellegrini, Gabriele Graffieti, Vincenzo Lomonaco, and Davide Maltoni. Latent Replay for Real-Time Continual Learning. *arXiv preprint arXiv:1912.01100*, 2019.

Grégoire Petit, Adrian Popescu, Hugo Schindler, David Picard, and Bertrand Delezoide. FeTrIL: Feature Translation for Exemplar-Free Class-Incremental Learning. *arXiv preprint arXiv:2211.13131*, 2022.

Fuli Qiao and Mehrdad Mahdavi. Learn more, but bother less: parameter efficient continual learning. In *Advances in Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=ZxtaNh5UYB.

Fuli Qiao and Mehrdad Mahdavi. Merge before forget: A single LoRA continual learning via continual merging. In *International Conference on Learning Representations*, 2026. arXiv:2512.23017.

Ahmad Rashid, Vasileios Lioutas, Abbas Ghaddar, and Mehdi Rezagholizadeh. Towards Zero-Shot Knowledge Distillation for Natural Language Processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6551–6561, 2021. doi: 10.18653/v1/2021.emnlp-main.526.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for Thin Deep Nets. *arXiv preprint arXiv:1412.6550*, 2014.

Gobinda Saha, Isha Garg, and Kaushik Roy. Gradient Projection Memory for Continual Learning. In *International Conference on Learning Representations (ICLR)*, 2021.

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. Continual Learning of Large Language Models: A Comprehensive Survey. *arXiv preprint arXiv:2404.16789*, 2024.

Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual Learning with Deep Generative Replay. *arXiv preprint arXiv:1705.08690*, 2017.

Reece Shuttleworth, Jacob Andreas, Antonio Torralba, and Pratyusha Sharma. LoRA vs Full Fine-tuning: An Illusion of Equivalence. *arXiv preprint arXiv:2410.21228*, 2024.

James Smith, Yen-Chang Hsu, Jonathan Balloch, Yilin Shen, Hongxia Jin, and Zsolt Kira. Always Be Dreaming: A New Approach for Data-Free Class-Incremental Learning. *arXiv preprint arXiv:2106.09701*, 2021.

Brady Steele. Subspace Geometry Governs Catastrophic Forgetting in Low-Rank Adaptation. *arXiv preprint arXiv:2603.02224*, 2026.

Pengwei Tang, Xiaolin Hu, Yong Liu, Lizhong Ding, Dongjie Zhang, Xing Wu, and Debing Zhang. Put the Space of LoRA Initialization to the Extreme to Preserve Pre-trained Knowledge. *arXiv preprint arXiv:2503.02659*, 2025.

Ruilin Tong, Haodong Lu, Yuhang Liu, and Dong Gong. Model Inversion with Layer-Specific Modeling and Alignment for Data-Free Continual Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Hanqing Wang, Yixia Li, Shuo Wang, Guanhua Chen, and Yun Chen. MiLoRA: Harnessing Minor Singular Components for Parameter-Efficient LLM Finetuning. *arXiv preprint arXiv:2406.09044*, 2024a.

Ke Wang, Nikolaos Dimitriadis, Alessandro Favero, Guillermo Ortiz-Jimenez, Francois Fleuret, and Pascal Frossard. LiNeS: Post-training Layer Scaling Prevents Forgetting and Enhances Model Merging. *arXiv preprint arXiv:2410.17146*, 2024b.

Peng Wang, Minghao Gu, and Qiang Huang. VeFA: Vector-Based Feature Space Adaptation for Robust Model Fine-Tuning. *arXiv preprint arXiv:2510.19155*, 2025.

Shipeng Wang, Xiaorong Li, Jian Sun, and Zongben Xu. Training Networks in Null Space of Feature Covariance for Continual Learning. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); arXiv preprint arXiv:2103.07113*, 2021.

Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuanjing Huang. Orthogonal Subspace Learning for Language Model Continual Learning. *arXiv preprint arXiv:2310.14152*, 2023.

Xinyu Wang, Changzhi Sun, Yuanbin Wu, and Xiaoling Wang. Stabilizing LLM Supervised Fine-Tuning via Explicit Distributional Control. *arXiv preprint arXiv:2605.04468*, 2026a.

Zheng Wang, Wanhao Yu, Li Yang, and Sen Lin. Rethinking continual learning with progressive neural collapse. In *International Conference on Learning Representations*, 2026b. arXiv:2505.24254.

Tongtong Wu, Linhao Luo, Yuan-Fang Li, Shirui Pan, Thuy-Trang Vu, and Gholamreza Haffari. Continual Learning for Large Language Models: A Survey. *arXiv preprint arXiv:2402.01364*, 2024.

Yifeng Xiong and Xiaohui Xie. OPLoRA: Orthogonal Projection LoRA Prevents Catastrophic Forgetting during Parameter-Efficient Fine-Tuning. *arXiv preprint arXiv:2510.13003*, 2025.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-Merging: Resolving Interference When Merging Models. *Advances in Neural Information Processing Systems (NeurIPS); arXiv preprint arXiv:2306.01708*, 2023.

Zhaorui Yang, Tianyu Pang, Haozhe Feng, Han Wang, Wei Chen, Minfeng Zhu, and Qian Liu. Self-Distillation Bridges Distribution Gap in Language Model Fine-Tuning. *arXiv preprint arXiv:2402.13669*, 2024a.

Zitong Yang, Neil Band, Shuangping Li, Emmanuel Candès, and Tatsunori Hashimoto. Synthetic continued pretraining. *arXiv preprint arXiv:2409.07431*, 2024b.

Hongxu Yin, Pavlo Molchanov, Zhizhong Li, Jose M. Alvarez, Arun Mallya, Derek Hoiem, Niraj K. Jha, and Jan Kautz. Dreaming to Distill: Data-free Knowledge Transfer via DeepInversion. *arXiv preprint arXiv:1912.08795*, 2019.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language Models are Super Mario: Absorbing Abilities from Homologous Models as a Free Lunch. *International Conference on Machine Learning (ICML); arXiv preprint arXiv:2311.03099*, 2024.

Lu Yu, Bartłomiej Twardowski, Xialei Liu, Luis Herranz, Kai Wang, Yongmei Cheng, Shangling Jui, and Joost van de Weijer. Semantic Drift Compensation for Class-Incremental Learning. *arXiv preprint arXiv:2004.00440*, 2020.

Guanxiong Zeng, Yang Chen, Bo Cui, and Shan Yu. Continual Learning of Context-dependent Processing in Neural Networks. *Nature Machine Intelligence*, 1(8):364–372, 2019. doi: 10.1038/s42256-019-0080-x.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual Learning Through Synaptic Intelligence. *arXiv preprint arXiv:1703.04200*, 2017.

Yunan Zhang, Shuoran Jiang, Mengchen Zhao, Yuefeng Li, Yang Fan, Xiangping Wu, and Qingcai Chen. GeRe: Towards Efficient Anti-Forgetting in Continual Learning of LLM via General Samples Replay. *arXiv preprint arXiv:2508.04676*, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023.

APPENDIX: ORGANISATION

The appendix follows the contributions of §1. *Mechanism*: what anchoring does to the learned representation (App. A); the statistic, its price and its merge (App. B); the shipped point and its knob coupling (App. C). *Evidence that it works*: what "match" means (App. D), then one, two and three domains (App. E, F, G), then generative retention, the wider suite and the Qwen3-1.7B replication (App. H). *Cost* (App. I), *what L2-SP adds to other methods* (App. J) and *self-supply* (App. K). Reference material closes it: the baselines (App. L), the judge (App. M), the setup in full (App. N), related work (App. O) and limitations (App. P).

A WHAT ANCHORING DOES TO THE LEARNED REPRESENTATION

On hidden states from held-out domain-A text, LFA's 140 anchored sub-module functions move 37 to 313 times less than LwF+μ's or replay's across all eighteen rank-by-stage comparisons (5 to 38 times on the macro site-type average), and 161 to 680 times less on WikiText-2 test states (Table 4). LFA's own drift there is the check on the first half of the argument that independent vectors can anchor a sequence model (§A.1). The ratio is on the quantity LFA minimises and the others are not asked to, it is one seed but for the $r{=}32$ LFA–replay pair, and it does not predict judged quality: after C LFA and replay tie on set-A at every rank (§6.1) across a more than hundred-fold difference in drift. The remaining subsections describe the update's structure under the same caveat, stated here once: far fewer intruder dimensions (§A.2, §A.3), a suppressed destructive footprint with a domain-selective population of modes that unanchored training does not produce (§A.4), and a judged optimum that is interior in (λ, μ) and not at the structurally cleanest setting (§A.5). §A.2–A.5 anchor against the $\approx$0.5 MB diagonal artifact at $r{=}16$, $f{=}0$ (priced in App. B).

A.1 WHY INDEPENDENT VECTORS CAN SUFFICE, AND THE HELD-OUT CHECK

A decoder-only transformer composes two kinds of operation. *Token-wise maps* act on one position's vector at a time: the projections Q, K, V and O, the MLP, the norms, lm_head. The rest have no parameters that train: residual addition, the rotary embedding, and the one operation that mixes positions, the softmax over query–key scores applied to the values. Under LoRA with frozen embeddings, the setting of every LFA run here, the maps that train are exactly the anchored ones: Q, K, V, O and the MLP (as one map) of each layer; the norms (Qwen3's per-head query and key norms included), the embedding and lm_head are frozen and count as fixed operations.

Proposition 1 (the exact-preservation limit)**.** *Let $\mathcal{D}$ be a set of token sequences and, for each trainable map i, let S_i be the set of vectors the teacher presents to it over all positions of all sequences in $\mathcal{D}$. If $f_{\mathrm{s},i}(h) = f_{\mathrm{t},i}(h)$ for every i and every $h \in S_i$, then on every sequence in $\mathcal{D}$ the student computes the teacher's value at every site and position, attention scores and output logits included.*

Proof. Take the trainable maps in a topological order of the computation graph and induct; the base case is the frozen embedding. Suppose every value computed before map i agrees with the teacher's. Then at each position map i receives the teacher's input, which lies in S_i, so its output agrees. Every operation between map i and the next trainable map is fixed, hence the same function in both models applied to values that agree, and its output agrees whether or not it mixes positions. □

The induction is the one behind null-space projection (OWM, GPM, Adam-NSCL; Zeng et al., 2019; Saha et al., 2021; Wang et al., 2021): an update that leaves every layer's output unchanged on the inputs it has seen leaves the network unchanged on them. What a transformer adds is that its only position-mixing operation has no trainable parameters, so per-site marginals suffice and no sequence is needed.

S_i is the support of the true marginal at site i. That marginal carries the sequence dependence implicitly: it is the distribution of vectors contextual processing has already produced, and holding a token-wise map fixed on them needs nothing about the order in which they arose. The samples also break the joint dependence across layers that training forms: each site is anchored in isolation, and in the limit that costs preservation nothing. At a linear site a zero price in Eq. (2) under the true second moment already forces $\Delta W h = 0$ on the span of S_i.

Training does not reach the limit, and we prove nothing about finite λ. The penalty is finite and $p(h)$ is a fitted estimate on a truncated basis, not the marginal, so every map moves a little on S_i; small per-site errors can compound through depth and across positions; the penalty is a squared error on Q and K outputs, not on scores, so a change that is small on $p(h)$ can still move a softmax (item 1 of App. P); and once upstream maps have moved, the student's own inputs (layer 0's Q, K and V apart) are no longer draws from $p(h)$. What remains is a hypothesis with two halves: that maps held near-fixed on samples are near-fixed on *real* inputs the artifact never saw, and that the composition then holds. Table 4 tests the first, on the base model's inputs. For the second we have output-level checks only, but every whole-network instrument in the paper is one: general-text perplexity (§3.4, under a different protocol), old-domain perplexity (Table 20), judged answering (§5–§6.1), rule-scored generation (Table 2) and the judge-free suite (App. H) are all outputs of the composed adapted network, so each bounds the realized cost of compounding. On rule-scored generation LFA is ahead of both baselines after C at both sizes, while judged old-domain answering ties replay+μ and TriviaQA ties (§6.1). None of them isolates compounding from per-site error. The table measures, for the 140 anchored sub-modules (Q, K, V, O and the MLP of 28 layers), the relative drift $\mathbb{E}\,\|f_{\text{model}}(h) - f_{\text{base}}(h)\|^2 \,/\, \mathbb{E}\,\|f_{\text{base}}(h)\|^2$ from the never-adapted base, on hidden states the frozen base produces from held-out domain-A text (the file A's perplexity is scored on, never training text), from the WikiText-2 test split, and on samples from the chain's own $p(h)$, for the three chains of Table 20.

r=32	entry	after B	after C
LFA	0.031	0.037	0.040
LwF+μ	3.09	6.31	10.12
replay+μ	8.62	10.41	12.44

Table 3: Relative output drift of the 140 anchored sub-modules from the base model on held-out domain-A text, $\times 10^{-3}$, pooled (r=32, seed 42; entering stage two, after B, after C). All ranks, WikiText-2 and $p(h)$ samples: Table 4.

Table 4: Relative output drift of the anchored sub-modules from the base model, $\times 10^{-3}$ (seed 42; 512 positions per site and stream, `bf16`; stages: entering stage two / after B / after C). *Pooled*: squared error and base energy summed over the 140 modules before the ratio. *Macro*: the unweighted mean of the four site-type ratios (Q/K/V, O, MLP, and lm_head, frozen under LoRA and ≈0 for every arm; lm_head is excluded from the pooled figure because its energy would swamp it). LwF and replay have no artifact, so their $p(h)$ columns borrow the rank-matched LFA chain's.

		held-out domain-A, pooled			held-out domain-A, macro			WikiText-2 test, pooled			$p(h)$ samples, pooled		
	rank	entry	B	C	entry	B	C	entry	B	C	entry	B	C
LFA	r=4	0.087	0.099	0.103	0.810	0.861	0.880	0.012	0.017	0.021	0.011	0.013	0.017
	r=16	0.047	0.057	0.062	0.468	0.525	0.552	0.009	0.014	0.020	0.008	0.012	0.015
	r=32	0.031	0.037	0.040	0.289	0.325	0.342	0.007	0.010	0.014	0.007	0.009	0.011
LwF+μ	r=4	3.21	6.54	8.67	3.82	6.88	8.61	1.90	4.75	7.14	2.36	5.39	7.61
	r=16	3.44	6.37	9.38	4.43	7.55	10.45	2.87	5.53	8.57	3.35	6.16	8.94
	r=32	3.09	6.31	10.12	4.00	7.66	11.49	2.50	5.38	9.34	2.99	6.10	9.71
replay+μ	r=4	11.84	13.57	16.23	10.28	12.17	14.04	3.24	5.43	11.70	5.63	7.64	11.87
	r=16	8.91	10.61	13.42	10.21	11.75	13.40	4.03	5.76	11.03	5.70	7.45	10.96
	r=32	8.62	10.41	12.44	10.49	11.77	12.87	4.68	6.34	9.78	6.32	8.02	10.18

At r=32 the pooled drift reads $0.031 \to 0.037 \to 0.040$ for LFA against $3.09 \to 6.31 \to 10.12$ (LwF+μ) and $8.62 \to 10.41 \to 12.44$ (replay); on the macro average, where each site type counts once, 0.29–0.34 against 4.0–11.5 and 10.5–12.9. Those ratios give the scale and are not the check: neither baseline is asked to minimise this quantity. What checks the first half is LFA's *absolute* drift on real inputs the artifact never saw, read against its drift on the samples training did see. On WikiText-2 states the two are at parity, 0.007–0.021 against 0.007–0.017 (a ratio of 1.06–1.34 over the nine cells): what the anchor holds on $p(h)$ it holds on the general text $p(h)$ stands for. On held-out domain-A text there is a gap: drift is 3.7–7.9 times its value on samples, which still leaves it at 0.3 to 1.0×10^{-4} of function energy. The residue has a location, 1.20 at the attention-output projection against 0.012 at Q/K/V and 0.157 at the MLP (r=32, after C), and that is the one anchored site whose input position mixing builds directly; LwF+μ and replay move the MLP most (24.1, 24.9). LFA's drift grows by about a quarter from entry to after C at r=32, where LwF+μ's more than triples. The table locates each method's change; it does not rank the methods.

Two further measurements use the same instrument. *A second seed:* at seed 1337 the r=32 chains read $0.03213 \rightarrow 0.03818$ (LFA) and $8.686 \rightarrow 10.53$ (replay) from entry to after B, replay-to-LFA ratios of 270 and 276 against 275 and 285 at seed 42 (ratios are taken before rounding). That pair is the only two-seed comparison: the 37–313× range, and every LwF+μ and after-C cell in it, is one seed. *Dose-response:* in a matched single-domain r=16 ladder (μ=0.05, 10 epochs, gradient accumulation 8, the shipped artifact, final model; one run per rung and no λ=0 rung) held-out domain-A drift reads 0.171, 0.0885, 0.0431 and 0.0194 at λ = 5k, 15k, 45k and 135k: monotone, about half per threefold λ, and in the same order at epoch 5, on WikiText-2 and on the macro average. The gap to the samples is wider there, 15 to 38 times, and widens with λ: tightening the anchor lowers drift on samples faster than on real domain text. *One reading, untested:* LwF+μ's and replay's far larger drift is *joint*, produced under full sequences through the whole network, and may displace sub-module functions along pathways the rest of the network compensates for rather than destroy them, whereas LFA's per-site drift compounds through depth (the argument above), so the same magnitude would not mean the same thing. Whether joint drift costs quality is untested here.

A.2 Anchoring reduces intruder dimensions

An *intruder dimension* (Shuttleworth et al., 2024) is a high-ranking singular vector of a fine-tuned weight matrix that resembles no singular vector of the pre-trained one (maximum cosine below 0.5). Unanchored adaptation produces 1,340 of them and drives general-text perplexity to 93.7 (+427%; Table 5). λ alone (λ=5000) holds perplexity at 20.1 (+13%) with 108, ∼12× fewer, at higher content loss, and μ then trades adaptation for cleanliness monotonically ($108 \rightarrow 36 \rightarrow 12$ as μ goes $0 \rightarrow 0.1 \rightarrow 1.0$). L2-SP alone is far worse than LFA (perplexity 50.2 vs. 18.4): the function anchor, not weight-space L2, carries preservation. Under a ≤13% perplexity budget unanchored LoRA must stop at epoch 2 (content 2.22; still 1% below base there, +18% an epoch later), where LFA trains through epoch 10 to content 1.70.

Table 5: Intruder dimensions under three layer schedules (r=16, diagonal artifact, f=0, 10 epochs). Content: final domain cross-entropy. ΔPPL: general-text perplexity against the training-time subset baseline (17.80 over 200 WikiText-2 documents; the frozen protocol of App. E.1 scores full WikiText-2, 18.18). Intr.: intruders among the top-k=20 directions, summed over sub-modules (L0, L27: layers 0 and 27). Sim: mean maximum teacher/student singular-vector cosine. e: `anchor_end_ratio`. The uniform-μ row's perplexity is the 100-epoch run's epoch-10 checkpoint; the short run of the same configuration reads 18.0 (+1%).

Condition	Content↓	Gen. PPL	ΔPPL	Intr.	L0	L27	Sim
Unanchored (λ=μ=0)	1.017	93.7	+427%	1,340	26	41	0.611
Series 2: e=0.01, μ coupled to the layer schedule							
μ=1.0 only (λ=0)	1.350	50.2	+182%	318	0	51	0.881
LFA (λ only)	1.699	20.1	+13%	108	39	11	0.929
LFA + μ=0.1	1.733	19.8	+11%	36	23	8	0.955
LFA + μ=1.0	1.892	18.4	+3%	12	3	9	0.982
Series 3: e=0.1, μ coupled							
μ=1.0 only (λ=0)	1.395	44.9	+152%	127	1	26	0.920
LFA (λ only)	1.747	19.9	+12%	86	33	0	0.932
LFA + μ=0.1	1.782	19.4	+9%	37	29	0	0.958
LFA + μ=1.0	1.947	18.0	+1%	4	4	0	0.986
Series 4: e=0.1, μ uniform across layers (carried forward)							
LFA + μ=0.05 (uniform)	1.965	17.8	+0%	7	7	0	0.987

A.3 The layer schedule is scale compensation, and L2-SP caps intruder growth

Series 2 leaves residual intruders at layer 27, which is *nominally* under-anchored: at e=0.01 the cosine decay leaves it an effective λ of 4.5 against 342 at layer 0 and 172 at layer 14 while it has the largest weight change ($\|\Delta W\|_F$ 3.34 against 1.67 and 1.64), a nominal pressure $\lambda / \|\Delta W\|_F$ of 1.3 against 204 and 105, ∼150× below layer 0's. That is not the reading we rest on. The anchor is an absolute per-element error and teacher activation scale grows with depth, so effective *relative*

pressure is 17–1,191× *higher* at layer 27 than at layer 0 at the shipped e=0.1: the schedule is scale compensation.[1] Raising e to 0.1 (Series 3) gives layer 27 ∼7× more nominal pressure and eliminates its intruders. Making μ *uniform* across layers (Series 4) is the most effective configuration per unit of μ: μ=0.05 yields 7 intruders (∼191× fewer than unanchored) at similarity 0.987, what coupled μ=1.0 achieves at 20× the weight-space pressure. Layer 0 keeps a QKV intruder population in every setting (3–39; 7 at the carried-forward point), which we take to be structural (the embedding interface). Over 100 epochs at constant learning rate, λ alone slows intruder formation without capping it (105 → 641), while all three μ variants plateau (uniform μ=0.05: 7 → 15, content loss still falling to 1.16), and each schedule fix improves general-text perplexity at 100 epochs (+202% → +121% → +85% → +72%).

A.4 The destructive and non-destructive footprint, measured on real activations

Intruder counts live on the absolute W_{s}. Here we decompose the *change* $\Delta W = W_{\mathrm{s}} - W_{\mathrm{t}}$ into rank-one modes and measure what each injects on *real* activations: $E_{\mathrm{norm}} = \sigma \cdot \mathrm{RMS}_h\langle v, h\rangle$ on general text (the destructive effect) and E_{dom} on domain text. A mode is *non-destructive* (domain-selective) if $E_{\mathrm{norm}}{\approx}0$ while E_{dom} is large; its *entanglement* is the fraction of its input direction inside the base model's 95%-variance activation-PCA subspace, read as the gap Δ_{rand} to a matched random direction's k/d (0.38–0.79 by sub-module). Three predictions hold (Table 6). λ **suppresses the destructive footprint** to a tenth (9.9×; 29× with μ=1.0). **Anchoring creates the non-destructive population:** unanchored training produces *zero* domain-selective modes, anchoring 31–38. **L2-SP re-routes the change through existing structure:** as μ grows, content loss rises (1.75 → 2.36) and the selective modes become constructively entangled (Δ_{rand} −0.27 → +0.16, crossing chance by μ=0.05), which is why the intruder count falls and singular-vector similarity rises while the selective-mode count holds. The baselines each miss one half: replay is structurally close to unanchored fine-tuning yet preserves general-text perplexity best, by rehearsing data and not by restructuring the change; VeFA, L2-SP and EWC damp the change and create no selective modes; TM-only builds 2 against LFA's 37.

Table 6: Footprint of the change ΔW on real general-text and domain activations (the Series-4 μ ladder and five external baselines; all layers, 10 epochs; 200 general-pretraining documents and all 77 domain documents; entanglement against the base model's 95%-variance activation-PCA subspace, 600k samples). E_{norm}: total destructive footprint, normalized to unanchored=100. Sel.: aggregate $E_{\mathrm{dom}}/E_{\mathrm{norm}}$. #sel: domain-selective modes (Sel>3). ent: their mean entanglement. Δ_{rand}: ent minus each mode's matched random baseline k/d; positive is constructively entangled. Content: final-epoch domain cross-entropy, buffer batches excluded.

Condition	Content↓	E_{norm}	Sel.	#sel	ent	Δ_{rand}
Unanchored	1.02	100.0	1.04	0	—	—
λ (μ=0)	1.75	10.1	1.40	31	0.380	−0.27
λ+μ=0.05	1.97	7.0	1.47	37	0.717	+0.07
λ+μ=0.1	2.06	6.1	1.48	38	0.738	+0.10
λ+μ=1.0	2.36	3.4	1.47	32	0.791	+0.16
Replay (real data)	1.46	44.8	1.19	1	0.749	+0.02
TM-only (GeRe-deriv.)	1.88	29.3	1.30	2	0.798	+0.07
L2-SP (μ only)	1.95	24.7	1.25	0	—	—
EWC (λ_{EWC}=100)	1.99	28.4	1.27	0	—	—
VeFA	2.74	4.1	1.03	0	—	—

A fingerprint of integration style, not of quality. Across the eight continual arms with both a selective-mode count and a judged score the rank correlation is ρ=0.874 (p=0.005), but ρ=0.714 (p=0.11) without the two zero-mode arms, which fail on every axis. Entanglement points the wrong way: replay exceeds LFA in one domain (0.86 vs. 0.76) while their judged contrast does not separate

[1] Qwen3-0.6B, n=32,768 sampled states per (layer, site) in fp32 over the three anchored site types: teacher activation scale $\mathbb{E}_h \|f_{\mathrm{t}}(h)\|^2$ grows 165–11,578× from layer 0 to 27 (min–max across site types) against a schedule decay of 0.103. The magnitudes are model- and corpus-specific; only the sign is load-bearing.

in replay's favour under the primary judge. Within a rank-matched μ ladder on continual replay the count moves $15 \to 1$ while judged quality spans 0.117, and the one pair with a primary-judge paired test (μ=0 vs. 0.1) is a tie (p=0.23). Single-intruder ablation cannot settle it: removing one rank-one direction moves general-text perplexity by $<2\%$. The intruder count behaves the same way under the frozen protocol of App. E.1: across its eleven fine-tuned arms it is rank-correlated with retained accuracy (Spearman $\rho=-0.67$, p=0.02), and adding seed-matched replay, which posts the *highest* retained accuracy (0.442) at 66 intruders, collapses the correlation (n=12: $\rho=-0.43$, p=0.16). Structure tracks preservation (perplexity, retained accuracy) and not judged quality, which is itself the case for anchoring functions and not structure. One checkpoint family, one domain; the entanglement subspace is stable under resampling (its dimension changes by $<1\%$).

A.5 Judged on generated answers, (λ, μ) has an interior optimum

Judged quality rises from 0.62 at μ=0 to 0.68–0.76 at $\mu \in [0.05, 0.1]$ and falls back to 0.58 at μ=1.0 (Tables 7, 8). Under both judges the two interior points tie and the fall separates; the rise separates at μ=0.1 only under the primary judge and at both interior rungs under the robustness judge. So the *structurally cleanest* setting (lowest destructive footprint, near-zero intruders) is generatively *worse* than the interior: optimizing structural proxies overshoots. *Voice capture*, the share of answers written in the corpus author's first person, tracks memorization ($\rho=-0.67$ with content loss; 17% and 14% in the two most-memorized arms against 0% at μ=0.1) and is the strongest correlate of judged quality across the eight fine-tuned arms ($\rho=-0.76$, p=0.028; robustness judge $\rho=-0.71$). Training the same configuration for 100 epochs fits domain text far better (content 1.16 vs. 1.97) and answers no better ($0.68 \to 0.64$, a tie under the primary judge, a drop under `claude-sonnet-5`): perplexity and generative usability decouple. Replay reaches 0.79, the highest level among the fine-tuned arms, and is paired-*below* the base model's 0.93; the best LFA setting ties it (0.76) at $1.5\times$ the unanchored score with no stored text. Both replay verdicts are judge-dependent (under `claude-sonnet-5` replay ties base and is ahead of LFA). Every fine-tuned arm scores below base: on the raw-text corpus adaptation *costs* generative quality, and the dial controls how much (-0.43 unanchored, -0.17 at μ=0.1). The target is use-case-relative: voice capture is a pathology for question answering and the *objective* for persona tuning.

Table 7: Generative evaluation at f=0: mean judged score over the 154 set-A questions (`gpt-5.6-luna@medium`) and voice-capture rate. Content: final-epoch domain cross-entropy (memorization proxy; buffer batches excluded). The two data-using rows cross a protocol boundary: replay runs ga=2, warmup 100 and TM-only ga=8, warmup 100, against the LFA rows' ga=1, warmup 50 (App. L).

Condition	μ	Ep.	Content↓	Judge/5	Voice
Base (teacher)	—	0	—	0.93	0.00
Unanchored	0	10	1.02	0.50	0.17
LFA λ (μ=0)	0	10	1.75	0.62	0.13
LFA λ+μ=0.05	.05	10	1.97	0.68	0.01
LFA λ+μ=0.1	.1	10	2.06	**0.76**	**0.00**
LFA λ+μ=1.0	1.0	10	2.36	0.58	0.03
LFA λ+μ=0.05, 100 ep	.05	100	1.16	0.64	0.14
Replay (real data)	—	10	1.46	0.79	0.01
TM-only anchor (GeRe-deriv.)	—	10	1.88	0.71	0.05

B The statistic: its form, its price, and its exact merge

Equation (2) fixes the statistic's form, and the cheaper forms are measurably mispriced against it. The $\approx$0.5 MB diagonal-moments artifact misprices update directions by 1.6–4.1$\times$ over the nine sampled linear sites, which is worth 4 pp of general-text preservation at the recipe corner, with the same ordering surviving into the continual setting; and with the shape of $p(h)$ removed altogether (an isotropic Gaussian at matched scale) both preservation and judged quality fall at identical domain fit. The diagonal form is the size floor the shipped artifact is priced against, not a recommended shipping

Table 8: Paired tests behind Table 7: two-sided Wilcoxon over the 154 shared questions, second arm minus first. Δ and the first p are `gpt-5.6-luna@medium`; the last column is `claude-sonnet-5` on the same answers (p and verdict only). "Ahead" names the arm that separates at α=0.05.

Contrast (first → second)	Δ	p (verdict), primary	p (verdict), Sonnet
μ=0 → μ=0.05	+0.064	0.21 (tie)	1.1×10^{-4} (0.05 ahead)
μ=0 → μ=0.1	+0.139	0.027 (0.1 ahead)	1.5×10^{-5} (0.1 ahead)
μ=0.05 → μ=0.1	+0.075	0.15 (tie)	0.70 (tie)
μ=0.1 → μ=1.0	−0.170	0.004 (0.1 ahead)	2×10^{-4} (0.1 ahead)
μ=0.05 → μ=1.0	−0.095	0.019 (0.05 ahead)	3×10^{-4} (0.05 ahead)
μ=0.05: 10 → 100 epochs	−0.040	0.30 (tie)	0.016 (10 ep ahead)
base → replay	−0.147	0.006 (base ahead)	0.56 (tie)
LFA μ=0.1 → replay	+0.029	0.46 (tie)	0.003 (replay ahead)
unanchored → LFA μ=0.1	+0.256	3.9×10^{-5} (LFA ahead)	8.1×10^{-8} (LFA ahead)

tier. A learned domain is added to the statistic by an exact n-weighted mixture union, at ≈4.5 MB a round (§B.3).

B.1 How $p(h)$ is estimated, and how its price is measured

$p(h)$ is estimated per layer and per sub-module input level from a 10:1 pretraining:instruction proxy corpus; it is unconditional and task-agnostic. The minimal estimate is the per-level mean and diagonal standard deviation (Gaussian sampling); the shipped default adds a correlated covariance at the linear sites, a 95%-variance PCA basis, and a Gaussian mixture in PCA space at the MLPs. Layer-0 pre-attention states are rendered exactly by $\mathrm{input_layernorm}(\mathrm{embed}(t))$, reconstructed from the teacher at fine-tune time and not shipped. The representation and rank studies (App. A, C.1) were produced on the diagonal floor at r=16; the recipe and the continual chains use the shipped default. The upgrade overturns nothing there: the intruder mechanism and the (λ, μ) account are re-measured at the confirmed artifact and unchanged (App. C), while the footprint split is established at the floor and untested at the confirmed artifact.

The measurement behind the pricing. The 1.6–4.1× figure is the *extreme generalized eigenvalue* of the modelled against the true covariance, the worst-case ratio between a direction's true and quoted price under Eq. (2), read at the nine linear sites the estimator samples (`pre_o` at layers 0, 7, 13, 20, 27 and `pre_qkv` at all of those but layer 0): 3.4–4.1× at five of the nine and 1.6–2.6× at the other four.

Why coverage proxies certify the cheap tier. The diagonal form passes a *geometric* coverage screen, a generator's sample-to-data nearest-neighbour distance over the data's self-distance (1.0 is ideal): it reaches ≈1.34, improved only 1.07× by a PCA basis. That proxy measures where samples land, not whether their variance is *correlated*, and the anchor loss depends on the covariance. Marginal-fit proxies fail the same way: held-out likelihood keeps improving with mixture size K long after K=32 has saturated the downstream result. Artifact fidelity has to be priced against the loss, not against a distance.

Why the zero-byte form of $p(h)$ is unavailable to this architecture. HOPE (Mobahi & Bartlett, 2026) recovers its input distribution from the checkpoint alone: maximum entropy applied to BatchNorm's stored constraints makes the learned affine parameters the pre-activation moments, at zero bytes and with no corpus. The construction is unavailable here for a structural reason. BatchNorm normalizes *per channel*, pinning each channel's pre-activation variance to one before the affine, so the learned γ_k^2 *is* that channel's variance and a layer supplies d constraints. RMSNorm and LayerNorm normalize *across the feature vector*, pinning only the global scale (one constraint per token, not d), so a channel's own variance is never identified and the maximum-entropy solution degenerates toward an isotropic surrogate. The LayerNorm/RMSNorm transformer family therefore lacks the construction, not merely our checkpoint. Measured on our linear-site yardstick, HOPE's completion $\hat{\Sigma} = (\sum_i \lambda_i w_i w_i^\top)^{-1}$ scores SHAPE 0.75–1.01 against 0.62–0.75 for a plain diagonal

and 0.18–0.31 for the correlated covariance we ship: adding weight geometry to the marginals makes the price *worse*, because maximum entropy inflates variance in the unconstrained directions. At the attention-output sites it cannot be formed at all, since fewer neurons read the site than it has dimensions. The artifact is the price of an architecture that does not record its own activation statistics.

B.2 What the floor costs downstream

Table 9 places the shipped point on a $p(h)$-fidelity ladder at matched $f{=}0.13$, so the Domain column is comparable across rows. Against the best point the $\approx$0.5 MB diagonal-moments floor reaches anywhere in the search of App. E (row 1: 9.13 at -1.9%, at $r{=}16$, $\lambda{=}5{\cdot}10^4$, dose 20) it is better on both perplexity axes at both seeds. Rank, λ and dose differ across that pair, so it is an end-to-end operating-point comparison and not an artifact ablation; rows 2–4 hold $(r, \lambda, \text{dose})$ fixed and vary only the artifact, and the paragraphs below price them and the int8 build.

Table 9: The $p(h)$-fidelity ladder at matched $f{=}0.13$ (held-out domain PPL; general-text ΔPPL vs. 18.18). Row 1 is the best point the diagonal-moments floor reaches in the search of §E, and differs from the confirmed rows in rank, λ and dose as well as artifact. Rows 2–4 vary only the artifact at the confirmed point's $(r, \lambda, \text{dose})$. Single-seed except the confirmed arm and its int8 build. Judge: `judge_mean5` under `gpt-5.6-luna@medium` over all 154 questions, used as a *guard*: no paired contrast exists among these rows under either judge and the config-level minimum detectable effect is $\approx$0.19, so no row-to-row judge ordering is read. Zero-rate: the fraction of the 770 rubric scores returned as exactly 0 (archive-wide 30.6%); `--` where not reported. Rows 4 and 8 are absent from the search CSV and are recomputed from their judge sidecars with the aggregator's `judge_mean5`.

Arm	Artifact	dom. PPL↓	gen. ΔPPL↓	Judge (guard)	zero-rate
Diagonal floor ($r{=}16$, $\lambda{=}5{\cdot}10^4$, ep 20)	diagonal	9.13	-1.9%	1.097	—
Diagonal, $r{=}32$, $\lambda{=}10^5$, ep 15	diagonal	8.64	-5.9%	0.980	—
Diag-linear + GMM-MLP, $r{=}32$, ep 15	mixed	8.73	-8.1%	0.940	—
Corr-linear + moments-only MLP ($K{=}1$), $r{=}32$, ep 15	mixed	8.63	-9.4%	1.053	0.239
Confirmed ($r{=}32$, $\lambda{=}10^5$, ep 15), seed 42	corr.+GMM	**8.76**	$\mathbf{-10.0}\%$	0.968	0.268
same, seed 1337	corr.+GMM	8.83	-10.7%	0.974	0.281
same, int8-quantized artifact	corr.+GMM	8.72	-8.9%	0.960	0.281
same, int8 artifact, seed 1337	corr.+GMM	8.84	-9.2%	0.968	0.268

The weighting earns its complexity (ablation). Preservation is carried by the *shape* of $p(h)$, not by the presence of a function anchor. Replacing the fitted $p(h)$ with an isotropic Gaussian at matched per-site scale degrades general-text perplexity from -1.9% to $+6.8\%$, and a unit-scale isotropic Gaussian to $+17.5\%$, at identical domain fit. The ablation runs at row 1's configuration ($r{=}16$, $\lambda{=}5{\cdot}10^4$, dose 20) against row 1's fitted *diagonal-moment* artifact, not the shipped one. Judged quality falls $1.097 \to 0.881 \to 0.681$ under `gpt-5.6-luna@medium`, the last read at dose 10 because the unit-scale arm is infeasible at dose 20. Paired Wilcoxon against the fitted $p(h)$ over the 154 shared questions gives -0.216, $p{=}1.9{\times}10^{-4}$ and -0.418, $p{=}5.4{\times}10^{-11}$, and the two isotropic arms also separate from each other ($p{=}3.5{\times}10^{-4}$). This is the most judge-robust contrast in the section: under `claude-sonnet-5` the same three arms read $0.854 \to 0.681 \to 0.509$ with $p{=}6.1{\times}10^{-4}$ and $p{=}2.7{\times}10^{-11}$.

The floor is measurably mispriced. Matched at the confirmed point's ($r{=}32,\ \lambda{=}10^5$) and swept over $\lambda \in \{5{\cdot}10^4, 10^5\}$, the fully diagonal artifact does not reach the confirmed point: it gives back 4 pp of general-text preservation (-5.9% vs. -10.0%, Table 9). Two controls localize the loss where the pricing argument predicts. Keeping the $K{=}32$ GMM at the MLPs but making the *linear* covariance diagonal costs 2 pp (-8.1%); the complementary cell (correlated linear covariance + moments-only $K{=}1$ MLP) completes the $2{\times}2$ at -9.4%. The marginal contributions *at the recipe corner* are therefore $\approx$1.9 pp (correlated linear covariance) and $\approx$0.6 pp (MLP mixture). The cells do not close additively: the mixture contributes more (2.2 pp) when the linear covariance is diagonal than when it is correlated. But the ordering, linear sub-modules as the dominant fidelity lever, holds on both decomposition paths, as a loss that depends on $p(h)$ only through $\mathrm{tr}(\Delta W\, M\, \Delta W^\top)$ at those sites requires.

Shrink by quantizing, not by reducing fidelity. Truncating the correlated basis' rank is not a viable size lever: it is non-monotone in rank and can underperform the diagonal floor outright, because a low-rank head plus a diagonal residual starves precisely the directions it does not model. Blockwise int8 halves the artifact ($\approx$226 $\rightarrow$ $\approx$113 MB, $\approx 1\%$ covariance reconstruction error) and costs 1.1/1.5 pp of general-text preservation against the same-seed fp16 build (-8.9% vs. -10.0% at seed 42; -9.2% vs. -10.7% at seed 1337) with domain fit unaffected at both seeds (8.72/8.84 vs. 8.76/8.83): an acceptable quantization cost, not a strict tie, and still 3 pp ahead of the fully diagonal artifact.

The gap persists continually, and what the floor is worth. Run through the continual chain at the diagonal artifact's own best rank (r=4), LFA exits stage two at ΔA $+5.71$ with judged A-quality 0.861, against $+4.12/+4.16$ and $0.961/1.071$ for the rank-matched correlated+GMM chain (Table 17). The diagonal row is single-seed and in no paired verdict, so the judged levels are reported and not ordered. The floor is nonetheless a real operating point: it is on the preservation–adaptation frontier in the moderate-preservation regime ($\approx -6\%$ at r=32, -1.9% at its own r=16 point), it preserves far better than weight-space regularization does, and in the baseline comparison the confirmed configuration is not better than its row on every column, trading $\approx$1.1 PPL of domain fit for markedly better preservation (Table 16, ‡).

B.3 Extending the statistic to a learned domain: the exact merge

A learned domain's statistics are merged into $p(h)$ exactly, at $\approx 2\%$ of the payload a round and with no growth in anchoring cost, and the chain's perplexity endpoints approximately commute under a reversed order. The two-domain protocol that uses the merge is in App. F.

The mixture merge. A's hidden states are collected through the fused model ($\approx$40k per site against the base's $\approx$1.54M) and fitted as K_A=8 GMM components per site in the base artifact's whitened per-site PCA basis (whitening keeps EM well-conditioned; components are un-whitened afterwards). They are unioned into the base mixture with n-weighted component weights, the base's scaling by $(1-w)$ and A's by $w = n_A/(n_{\text{base}}+n_A) \approx 0.025$, uniformly at every anchored site, the base density at a linear site being treated as a one-component mixture. The stored per-site mean and basis are kept; A's components carry the shift. The merge is exact for the mixture it represents; the approximations are the per-domain GMM fits, and base components that are one model-generation stale with respect to the fused teacher.

The union neither prunes nor refits, so the component count grows as $K = K_{\text{base}} + R\,K_A$ after R merges, measured 32 $\rightarrow$ 40 at all 84 sampled sites for the single round run here. The per-site PCA bases, 208 of the artifact's 226 MB (92%), are *shared* by the merged components, so a round adds only per-component means, diagonal covariances and weights: 4.5 MB measured, $\approx 2\%$ of the shipped payload, against the 5.99 MB of A's corpus that rehearsal retains. By bytes the contrast is weak; what survives it is that no text is kept in any form and that anchoring cost does not grow: each step draws a fixed n=16 states per level, and the mixture draw is a gather, $O(1)$ in K. The linear parameter growth is unbounded over long chains, and the shared-basis design that would bound it is unbuilt (the limitation is named in §9).

One asymmetry was tested rather than corrected in place: A's states are collected on A's raw prose, not on the question–answer-mixed corpus stage one trains on. Chains that instead collect each domain's components from that stage's own chunked training stream land at after-C judged levels that are paired ties with the shipped chains' (p=0.62 at r=4, p=0.39 at r=32). The fall from after B to after C itself moves (0.147 $\rightarrow$ 0.194 at r=4, 0.262 $\rightarrow$ 0.161 at r=32); what survives is the tie between the two chains, so priced-state coverage is not what §6.1 measures.

Order dependence: the reversed B$\rightarrow$A chain. A single-seed reversed chain (same pipeline with roles swapped, stage two at λ=300k) shows that the *endpoints* approximately commute while the *dynamics* do not. Both orders land at A $\sim$10.7–11.0, B $\sim$6.1–7.0, WikiText $\sim$16.4–16.6, each domain slightly better when trained first, consistent with the merged $p(h)$ being an order-invariant statistic. But the defended domain forgets 13$\times$ less in B$\rightarrow$A (ΔB $+0.15$ vs. ΔA $+1.96$), and the incoming domain pays a larger learning penalty (A -2.3 vs. B -1.0 against its own single-domain fit). The asymmetry tracks manifold occupancy in direction (A exercises 65.5% of the base mixture

mass, B 54.6%), though an 11-point gap is unlikely to be the whole explanation. The judged levels show no order cost on either side. Under the primary judge B reads 1.580 before stage two and 1.577 after (zero-rates 20.3%, 21.3%; n=78), and A learned second reads $1.078/1.094$ against the single-domain recipe's $0.969/0.966$ on the same two rubric variants (n=77); `claude-sonnet-5` reads the same cells $1.341 \rightarrow 1.395$ and $0.839/0.758$ against $0.816/0.758$. These are levels: no verdict file carries a paired test for this chain under either judge.

C THE SHIPPED POINT AND KNOB COUPLING

LFA ships at LoRA r=32, λ=100,000, μ=0.05, dose 15, on the correlated-covariance + Gaussian-mixture artifact of §3 (≈226 MB fp16 / ≈113 MB int8), trained at f=0.13. Across two training seeds it reaches **held-out domain perplexity** $8.76/8.83$ **at general-text ΔPPL** $-10.0\%/-10.7\%$, against the never-adapted base model's WikiText-2 perplexity of 18.18, mean multiple-choice accuracy 0.4278 and judged score 0.93. Two internal tie-arms fix the artifact's remaining knobs (neither tabulated): K=32 saturates the MLP mixture (K=512 ties it downstream, while held-out likelihood keeps improving with K and is itself a failed proxy), and n=16 sampled states per level ties n=512. The λ–rank coupling of §C.1 is why this point exists: r=32 supplies the spare directions that let λ=100,000 bind without choking adaptation. An r=16 ladder with λ *re-tuned* reaches comparable domain fit, but those arms are single-seed and cannot establish that what r=32 buys is only headroom for λ and not a fidelity floor. Where the point sits on the $p(h)$-fidelity ladder is Table 9.

The judge as a guard, and the below-base perplexity. The judge is used here as a *guard* against silent generative failure: judged quality above the non-adapting base (0.93) and no collapse into the over-anchored failure mode of §C.1. Every row of Table 9 passes (0.940–1.097; 0.960–0.974 across the confirmed arm's own seeds and quantizations; under the robustness judge the int8 artifact's second seed reads 0.801 against the seed-42 recipe's 0.788). We claim no judged-quality win, and no row-to-row judged difference there is resolvable: at dose 15 the recipe ($0.968/0.974$) sits ≈0.12–0.13 below the diagonal floor's 1.097, inside the config-level threshold, and at dose 20 (untabulated: domain 9.80 at general-text -5.7%) it reads 1.042 at a zero-rate of 0.222. The guard is also why we read the recipe's -10% as "no measurable cost in general-text perplexity" and not as a capability gain (generative benchmarks do fall after stage one: GSM8K $0.565 \rightarrow 0.415$, IFEval $0.595 \rightarrow 0.545$ on the 2026-07 draw, App. F, finding vi): anchoring optimizes agreement with the teacher on states drawn from proxy-corpus statistics, so, like replay's buffer overlap one level removed, it can partially reward proxy-like behaviour, and §C.1 shows general-text improvement co-occurring with generative collapse in the over-anchored regime.

Why μ=0.05? At the operating point the function anchor and the high-fidelity artifact suppress intruder formation almost unaided: μ=0 leaves only 2 intruder dimensions (vs. 0 at μ=0.05) and *improves* domain fit by ∼0.3 PPL. That cleanliness is a fragile corner. Single-seed checks (Table 10) move one factor at a time away from the operating point, and μ=0 leaks intruder dimensions along all three: longer training ($2 \rightarrow 26$ over 100 epochs), lower anchoring dose ($2 \rightarrow 7$ at λ=$2{\cdot}10^4$), and lower rank (76 at r=16), while μ=0.05 holds the count at 0–1 throughout. Artifact fidelity, rank and dose jointly suppress formation in the favorable corner, and incompletely even there; L2-SP is the factor-independent backstop, a uniform weight-space penalty that caps whatever the other factors leave without needing to locate it. We ship μ=0.05 as this cheap robustness margin rather than as a tuned optimum. Table 10 is that battery: each block fixes everything but μ and moves one factor away from the operating point (single-seed; intruders at top-k=10). Two further checks agree that L2-SP is a factor-independent backstop and not a targeted patch. A no-training pricing diagnostic, which prices each checkpoint's ΔW under the artifact against its price under held-out real activations, confirms that the correlated covariance closes the linear blind spots on the updates training actually makes: a fixed r=16 update priced 43–54% too cheap by a diagonal artifact is priced within 2% by the correlated one (medians over the 21 linear sites probed, for the two r=16 updates; real-against-real floor ≈1%). The same diagnostic shows that L2-SP does *not* reduce a checkpoint's energy in the residual blind spots (the μ=0.05 update sits, if anything, marginally deeper): L2-SP acts by uniform magnitude shrinkage, not by steering direction. And under the mildly stale merged stage-two artifact, μ=0 is slightly worse than μ=0.05 on all three perplexity axes, so the L2-SP effect turns helpful where the anchor degrades.

Table 10: μ-robustness battery. For μ=0.05 the intruder count stays at 0–1 across every departure; for μ=0 it accumulates. Domain PPL shows L2-SP's small adaptation cost at the operating point ($\sim$0.3 PPL; its sign varies off it). The 100-epoch rows are deliberately over-trained. Gen. PPL: WikiText-2 (base 18.18).

Departure from recipe	μ	Intr.	dom. PPL	Gen. PPL
none (r=32, λ=10^5, ep 15)	0	2	8.50	16.8
	0.05	**0**	8.76	16.4
duration (ep 100)	0	26	9.05	34.8
	0.05	**0**	9.21	28.7
dose (λ=$2{\cdot}10^4$)	0	7	7.92	21.2
	0.05	**1**	7.84	18.9
rank (r=16, λ=$5{\cdot}10^3$)	0	76	7.35	23.8
	0.05	**1**	7.66	20.8

C.1 LoRA RANK IS AN IMPLICIT PRESERVATION MECHANISM, AND λ IS RECALIBRATED WITH IT

Rank alone preserves and λ is coupled to rank; at the confirmed artifact, rank is not load-bearing once λ is re-tuned. Rank is a *shared hyperparameter of the training method*, so every LoRA arm of the search of App. E was re-run at r=4 with α/r=2 held fixed (VeFA, a per-channel rescale, has no rank).

(i) Rank alone preserves. Unanchored's general-text degradation at dose 5 falls from $+79.4\%$ to $+18.0\%$ purely by dropping the rank, with *no* preservation mechanism present; the TM-only anchor goes from one feasible dose to four; and **EWC, which has no feasible point at r=16, becomes feasible at r=4** (dose 5, general-text $+6.5\%$ at λ_{EWC}=10^3; judged non-monotone across the swept range: 0.694 at 10^3, 0.790 at 300, 0.724 at the frozen protocol's 10^2, 0.783 re-tuned to 30). Some of the forgetting a reader would attribute to "no anchoring" is attributable to r=16.

(ii) λ is rank-coupled. λ multiplies a loss that constrains the update *subspace*, which LoRA's rank hard-bounds, so shrinking r four-fold makes a λ calibrated at r=16 far more binding. At r=4, λ=50000 reproduces VeFA's failure mode: general-text -8.3% (barely moved) while judged quality collapses on long, fluent, non-degenerate answers. The r=4 ladder $\lambda\in\{5{\cdot}10^3, 2{\cdot}10^4, 5{\cdot}10^4\}$ reads 0.734/0.863/0.721, an interior optimum under both judges (`claude-sonnet-5`: 0.475/0.643/0.476; values recomputed from the judge sidecars with the aggregator's pooled `judge_mean5`, these arms having no row in the search table).

Rank also governs overfitting: from dose 20 to 30 held-out domain perplexity rises at r=16 ($+0.174$; $+0.095$ at a second seed) and *falls* at r=8 (-0.090) and r=4 (-0.314), while $\|\Delta W\|_F$ grows near-identically across ranks. At the second seed EWC is again feasible at r=4 (general-text $+6.5\%/+3.4\%$) and unanchored's forgetting again collapses ($+79.4\% \rightarrow +18.0\%/+13.8\%$ at dose 5, lost by dose 10 at both seeds).

At the confirmed artifact, rank is not load-bearing: λ is. The study above used the diagonal artifact. We also ran the *confirmed* correlated+GMM recipe at r=16 under a λ ladder, every other knob at its default.

	domain PPL $\downarrow$	gen. ΔPPL	judged $\uparrow$
r=16, λ=$2{\cdot}10^4$	8.31	-3.6%	1.047
r=16, λ=$5{\cdot}10^4$	8.71	-9.0%	0.956
r=16, λ=10^5	8.93	-11.4%	0.935
r=32 recipe (two seeds)	8.76/8.83	$-10.0\%/-10.7\%$	0.968/0.974

λ traces a frontier and does not select an optimum, and the frontier passes through the r=32 recipe: interpolating the r=16 ladder to the recipe's -10.0% predicts domain 8.80, against the measured

8.76/8.83. Halving the rank therefore costs nothing here provided λ is re-tuned, and it must be: porting $\lambda{=}10^5$ unchanged lands at the preservation-heavy end. A log-log interpolation between the $r{=}4$ and $r{=}32$ anchors predicts $\sim 5.9{\cdot}10^4$ for $r{=}16$, inside the frontier-matched band ($5{\cdot}10^4$–10^5): a usable starting point for a ladder, not a substitute for one. Judged quality does not resolve the ladder (1.047/0.956/0.935, no zero-rate recorded, a spread below the answer-set threshold): the extreme pair is a paired tie under the primary judge (-0.113, $p{=}0.076$, $n{=}154$), and under `claude-sonnet-5` no paired contrast in the ladder reached significance (largest, the same pair: $p{=}0.10$). The frontier claim is a perplexity claim, and the $r{=}16$ arms are single-seed: a frontier-proximity argument, not a matched-seed equivalence test.

D What "match" means: equivalence and multiplicity for the main text's verdicts

The main text states 124 paired contrasts on the five-dimension judged mean under the primary judge: 61 separate at the 0.05 level and 63 are ties. A failure to reject is not an equivalence, so each tie is also put to two one-sided paired t-tests (TOST, each at 0.05) against a margin of ± 0.147, the primary judge's answer-set minimum detectable effect reconstructed by the recipe that returns the archived 0.12 on the robustness judge (0.180 at the configuration level). **The margin was chosen after the data were seen, so "equivalent within the margin" is descriptive and not confirmatory.** 37 ties are equivalent within it and 26 are inconclusive, neither a difference nor an equivalence (Table 11; 12 are also equivalent within ± 0.114, that judge's largest run-to-run swing on identical weights). The interval is a 95% percentile bootstrap over question ids (10,000 resamples; on set-A both phrasings of a question travel together). TOST at 0.05 corresponds to a 90% interval, so an "equivalent" row may carry a 95% interval that crosses the margin. Multiplicity is controlled within this family of 124 by Benjamini–Hochberg on the Wilcoxon p: 54 of the 61 separations survive and seven do not (Table 12; $^\flat$ in the main text); under the more conservative Holm correction 27 would not. Δ is the second-named arm minus the first.

The length-controlled readings the main text states in words (§5, §6, §6.1) are the same paired Wilcoxon test on the mean of correctness and domain accuracy. They form a separate sub-family of 23 readings (11 separations, 12 ties, two of the ties equivalent within the margin), corrected within itself because each re-reads judgments a family row already reads; one of its separations, LFA's $r{=}16$ set-B lead over LwF+μ, does not survive ($q{=}0.063$), and pooling the two families into one correction changes no row's survival. Three kinds of verdict sit outside both and are reported where they are stated: single-dimension readings (App. E), the with-and-without-L2-SP pairings (Table 31) and the judged-C floor pairings (App. G); the entry contrasts between methods are in Table 21. The rule-scored generative contrasts of Table 2 are exact McNemar tests paired by item (the 0.6B draw's ten tests uncorrected), and the Qwen3-1.7B contrasts are four *separately declared replication families*, each corrected inside itself by Benjamini–Hochberg and never pooled with the family above: A→B judged (24 contrasts on each axis), generative after B ($m{=}24$), generative after C ($m{=}12$) and judged after C (30 contrasts on each axis; App. H).

E One domain: a match under per-method search

Optimizing each method against the judge, under perplexity constraints, puts LFA level with replay and with LwF once each is granted L2-SP, and ahead of neither: the single-domain judged result is a *match*. Match has the meaning of App. D: of the six single-domain ties the main text states, four are equivalent within ± 0.147 and two are inconclusive. Table 14 gives each method's operating point in the search, Table 15 all fifteen paired comparisons with LwF, and §E.1 every baseline under one frozen protocol at $f{=}0$.

The search. We maximize the mean judged score *subject to* adaptation (held-out domain perplexity ≤ 12) and preservation (general-text ΔPPL $\leq +15\%$). The constraints are load-bearing: unconstrained, the base model "wins" at 0.931 by not adapting at all (its domain perplexity, 22.2, is infeasible). Scoring is feasibility-first, the judge running only on checkpoints that clear both constraints. Each method is swept on its own preservation–adaptation knob (Table 14's caption). Every row is one

Table 11: Every tie in the declared family of 124 five-dimension contrasts (App. D). Primary judge `gpt-5.6-luna@medium`; Δ: second arm (or later stage) minus first; CI: 95% bootstrap; p: two-sided Wilcoxon; TOST: p of the two one-sided tests at ± 0.147 (post hoc margin). The separations that do not survive correction, and the length-control sub-family, are Table 12. Single-dimension readings, the with-and-without-L2-SP pairings of §7 and the judged-C floor pairings are reported where they are stated (App. E, Table 31, App. G).

	tie	n	Δ	95% CI	p	TOST	within ± 0.147
§5	LFA dose 20 vs. LwF+μ μ=0.05 dose 15	154	+0.061	[−0.029, +0.153]	0.1841	0.0371	equivalent$^{\equiv}$
	LFA dose 20 vs. LwF+μ μ=0.1 dose 15	154	+0.074	[−0.016, +0.168]	0.1306	0.0653	inconclusive
	LFA r=16 vs. LwF+μ μ=0.05 (rank-matched, dose 15)	154	+0.056	[−0.051, +0.165]	0.3124	0.0386	equivalent$^{\equiv}$
	LFA vs. LwF+μ r=32 μ=0.1 (rank-matched, dose 15)	154	+0.054	[−0.062, +0.168]	0.2852	0.0513	inconclusive
	base model vs. LFA recipe, seed 42	154	+0.036	[−0.081, +0.153]	0.5188	0.0182	equivalent$^{\equiv}$
	base model vs. LFA recipe, seed 1337	154	+0.043	[−0.066, +0.154]	0.4024	0.0216	equivalent$^{\equiv}$
§6	r=4: LFA vs. LwF+μ α_{LwF}=8, set-A	154	−0.036	[−0.138, +0.065]	0.4355	0.0123	equivalent$^{\equiv}$
	r=4 judged-best: LFA vs. LwF+μ α_{LwF}=16, set-A	154	−0.027	[−0.118, +0.065]	0.4435	0.0085	equivalent$^{\equiv}$
	r=4 judged-best: LFA vs. replay+μ A-only, set-A	154	−0.086	[−0.187, +0.019]	0.1364	0.1288	inconclusive
	r=4 judged-best: LFA vs. TM-only+μ, set-A	154	−0.005	[−0.121, +0.104]	0.9861	0.0059	equivalent$^{\equiv}$
	r=4 judged-best: LFA vs. TM-only+μ, set-B	78	+0.013	[−0.149, +0.172]	0.6943	0.0493	equivalent$^{\equiv}$
	r=16: LFA vs. LwF+μ μ=0.05 mirror, set-B	78	−0.169	[−0.344, +0.008]	0.0865	0.5982	inconclusive
	r=16 replay: A-only vs. canonical buffer, set-A	154	+0.130	[+0.001, +0.258]	0.0760	0.3953	inconclusive
	r=16 replay: A-only vs. canonical buffer, set-B	78	−0.064	[−0.218, +0.092]	0.5271	0.1522	inconclusive
	r=32: LFA ga=3 step-matched control vs. the plain chain, set-A	154	+0.069	[−0.016, +0.157]	0.1275	0.0432	equivalent$^{\equiv}$
	r=32: LFA ga=3 step-matched control vs. the plain chain, set-B	78	+0.108	[−0.054, +0.274]	0.2843	0.3207	inconclusive
	r=32, seed 1337: canonical replay+μ, entry → after B, set-A	154	−0.070	[−0.190, +0.052]	0.3800	0.1167	inconclusive
§6.1	replay+μ r=4, after B → after C, set-A	154	+0.052	[−0.057, +0.164]	0.4233	0.0615	inconclusive
	replay+μ r=16, after B → after C, set-A	154	+0.036	[−0.068, +0.143]	0.7406	0.0287	equivalent$^{\equiv}$
	replay+μ r=32, after B → after C, set-A	154	−0.004	[−0.113, +0.100]	0.9198	0.0077	equivalent$^{\equiv}$
	replay+μ r=4, after B → after C, set-B	78	−0.061	[−0.221, +0.095]	0.4219	0.1478	inconclusive
	replay+μ r=16, after B → after C, set-B	78	+0.064	[−0.113, +0.238]	0.5064	0.1784	inconclusive
	replay+μ r=32, after B → after C, set-B	78	+0.038	[−0.154, +0.223]	0.4723	0.1351	inconclusive
	LFA r=4, entry → after B, set-A	154	+0.056	[−0.042, +0.152]	0.2285	0.0381	equivalent$^{\equiv}$
	LwF+μ r=4, entry → after B, set-A	154	−0.026	[−0.114, +0.060]	0.4348	0.0064	equivalent$^{\equiv}$
	replay+μ r=4, entry → after B, set-A	154	−0.016	[−0.127, +0.097]	0.5406	0.0147	equivalent$^{\equiv}$
	replay+μ r=16, entry → after B, set-A	154	−0.017	[−0.121, +0.087]	0.5799	0.0088	equivalent$^{\equiv}$
	replay+μ r=32, entry → after B, set-A	154	+0.056	[−0.064, +0.178]	0.2842	0.0582	inconclusive
	LFA r=16, entry → after C, set-A	154	−0.061	[−0.164, +0.043]	0.2597	0.0690	inconclusive
	LFA r=32, entry → after C, set-A	154	−0.047	[−0.174, +0.075]	0.3146	0.0498	equivalent$^{\equiv}$
	replay+μ r=4, entry → after C, set-A	154	+0.036	[−0.096, +0.174]	0.8801	0.0406	equivalent$^{\equiv}$
	replay+μ r=16, entry → after C, set-A	154	+0.019	[−0.092, +0.133]	0.9876	0.0141	equivalent$^{\equiv}$
	replay+μ r=32, entry → after C, set-A	154	+0.052	[−0.074, +0.178]	0.3151	0.0706	inconclusive
	LFA vs. replay+μ after C, r=4, set-A	154	+0.010	[−0.105, +0.133]	0.8239	0.0119	equivalent$^{\equiv}$
	LFA vs. replay+μ after C, r=16, set-A	154	+0.045	[−0.060, +0.152]	0.6738	0.0404	equivalent$^{\equiv}$
	LFA vs. replay+μ after C, r=32, set-A	154	−0.053	[−0.179, +0.073]	0.5626	0.0683	inconclusive
	LFA vs. LwF+μ after C, r=4, set-A	154	−0.030	[−0.140, +0.082]	0.5956	0.0154	equivalent$^{\equiv}$
	LFA vs. LwF+μ after C, r=16, set-A	154	−0.101	[−0.217, +0.010]	0.0711	0.2029	inconclusive
	LFA vs. its ga=3 control after C, r=4, set-A	154	+0.006	[−0.113, +0.125]	0.9366	0.0055	equivalent$^{\equiv}$
	LFA vs. its ga=3 control after C, r=16, set-A	154	+0.031	[−0.081, +0.143]	0.3741	0.0212	equivalent$^{\equiv}$
	LFA vs. its ga=3 control after C, r=32, set-A	154	−0.033	[−0.122, +0.053]	0.5320	0.0096	equivalent$^{\equiv}$
	LFA vs. replay+μ after C, r=16, set-B	78	+0.041	[−0.133, +0.213]	0.4127	0.1210	inconclusive
	LFA vs. replay+μ after C, r=32, set-B	78	+0.059	[−0.100, +0.218]	0.6145	0.1409	inconclusive
	LFA vs. LFA rehearsing A's pairs, after C, r=4, set-A	154	−0.044	[−0.158, +0.066]	0.5200	0.0292	equivalent$^{\equiv}$
	LFA vs. LFA rehearsing A's pairs, after C, r=16, set-A	154	−0.082	[−0.179, +0.018]	0.1445	0.1094	inconclusive
§7	LFA r=4 λ=2×10^4 vs. replay A-only μ=0, set-A	154	−0.078	[−0.187, +0.033]	0.2210	0.1093	inconclusive
§8	LFA vs. SSR, seed 42, dose 15	154	+0.042	[−0.060, +0.147]	0.3276	0.0252	equivalent$^{\equiv}$
	LFA vs. LwF+μ, seed 42, dose 15	154	−0.081	[−0.187, +0.023]	0.1607	0.1098	inconclusive
	LFA vs. SSR, seed 1337, dose 20	154	−0.074	[−0.164, +0.014]	0.1289	0.0642	inconclusive
	LFA vs. LwF+μ, seed 1337, dose 15	154	−0.039	[−0.174, +0.095]	0.4450	0.0400	equivalent$^{\equiv}$
	LFA dose 20 vs. SSR dose 15 (best feasible), seed 42	154	−0.038	[−0.147, +0.073]	0.4219	0.0243	equivalent$^{\equiv}$
	LFA dose 20 vs. SSR dose 15 (best feasible), seed 1337	154	−0.031	[−0.129, +0.058]	0.4738	0.0126	equivalent$^{\equiv}$
	real-text artifact λ=10^5 vs. self-generated, dose 15	154	+0.052	[−0.056, +0.161]	0.3517	0.0447	equivalent$^{\equiv}$
	real-text artifact λ=10^5 vs. self-generated, dose 20	154	+0.044	[−0.054, +0.149]	0.3933	0.0384	equivalent$^{\equiv}$
	real-text artifact λ=5×10^4 vs. self-generated, dose 15	154	+0.106	[−0.014, +0.229]	0.0923	0.2513	inconclusive
	real-text artifact λ=5×10^4 vs. self-generated, dose 20	154	−0.014	[−0.127, +0.096]	0.8531	0.0125	equivalent$^{\equiv}$
	real-text artifact λ=2×10^5 vs. self-generated, dose 15	154	+0.082	[−0.022, +0.188]	0.0835	0.0986	inconclusive
	real-text artifact λ=2×10^5 vs. self-generated, dose 20	154	+0.029	[−0.058, +0.118]	0.7778	0.0105	equivalent$^{\equiv}$
	frontier vs. self-written supplement, dose 15	154	+0.045	[−0.066, +0.158]	0.6326	0.0331	equivalent$^{\equiv}$
	frontier vs. self-written supplement, dose 20	154	+0.021	[−0.071, +0.113]	0.6788	0.0061	equivalent$^{\equiv}$
	recipe vs. fully self-supplied arm, dose 15	154	+0.062	[−0.044, +0.170]	0.2308	0.0439	equivalent$^{\equiv}$
	recipe vs. fully self-supplied arm, dose 20	154	−0.023	[−0.125, +0.078]	0.8633	0.0075	equivalent$^{\equiv}$
	self-gen. artifact + frontier supplement vs. fully self-supplied, dose 15	154	−0.110	[−0.214, −0.005]	0.0562	0.2380	inconclusive

Table 12: Upper block: the seven separations of the declared family that do not survive Benjamini–Hochberg correction over its 124 contrasts (♭ in the main text); q is the adjusted p. Lower blocks: the length-control sub-family (the same paired test on the mean of correctness and domain accuracy; 23 readings, corrected within itself): its twelve ties and the one separation that does not survive. Conventions as in Table 11; the margin was measured on the five-dimension mean, so a TOST verdict on the two-dimension mean is weaker still.

	separation not robust to BH (five-dimension family)	n	Δ	95% CI	p	q	
§6	r=16: LFA vs. LwF+μ, set-A (LFA ahead)	154	−0.109	[−0.197, −0.025]	0.0496	0.1008	
§6.1	LFA r=16, entry → after B, set-A (rises)	154	+0.134	[+0.043, +0.222]	0.0281	0.0612	
	LFA r=4, entry → after C, set-A (falls)	154	−0.091	[−0.187, +0.008]	0.0453	0.0936	
	LFA vs. replay+μ after C, r=4, set-B (LFA ahead)	78	−0.164	[−0.315, −0.010]	0.0254	0.0563	
	LFA vs. LFA rehearsing A's pairs, after C, r=32, set-A (latter behind)	154	−0.164	[−0.287, −0.044]	0.0305	0.0652	
§8	LFA vs. SSR, seed 1337, dose 15 (SSR ahead)	154	+0.126	[+0.014, +0.240]	0.0233	0.0524	
	self-gen. artifact + frontier suppl. vs. fully self-supplied, dose 20 (latter behind)	154	−0.121	[−0.227, −0.014]	0.0394	0.0828	
	tie on the length control	n	Δ	95% CI	p	TOST	within ±0.147
§6	r=32: LFA vs. canonical replay+μ, set-B	78	−0.218	[−0.449, +0.013]	0.0783	0.7275	inconclusive
	r=4: LFA vs. canonical replay+μ, set-A	154	−0.091	[−0.201, +0.013]	0.1239	0.1950	inconclusive
	r=16: LFA vs. LwF+μ, set-A	154	−0.091	[−0.201, +0.016]	0.1422	0.1596	inconclusive
	r=16: LFA vs. canonical replay+μ, set-A	154	−0.036	[−0.156, +0.081]	0.6642	0.0445	equivalent
	r=32, seed 1337: LFA vs. canonical replay+μ, set-A	154	−0.104	[−0.231, +0.020]	0.0700	0.2437	inconclusive
§6.1	LFA r=32, entry → after B, set-A	154	+0.110	[−0.023, +0.240]	0.0851	0.2827	inconclusive
	LFA r=4, after B → after C, set-A	154	−0.045	[−0.162, +0.068]	0.3997	0.0409	equivalent
	LFA r=16, after B → after C, set-A	154	−0.081	[−0.217, +0.052]	0.2309	0.1698	inconclusive
	LFA r=4, entry → after C, set-A	154	−0.127	[−0.260, +0.006]	0.0595	0.3764	inconclusive
	LFA r=16, entry → after C, set-A	154	−0.062	[−0.192, +0.068]	0.5111	0.1170	inconclusive
	LFA r=32, entry → after C, set-A	154	−0.094	[−0.257, +0.065]	0.1974	0.2376	inconclusive
	LFA vs. replay+μ after C, r=4, set-B	78	−0.064	[−0.263, +0.135]	0.4145	0.2079	inconclusive
	separation on the length control not robust to BH	n	Δ	95% CI	p	q	
§6	r=16: LFA vs. LwF+μ, set-B (LFA ahead)	78	−0.186	[−0.365, −0.019]	0.0299	0.0626	

Table 13: Single-domain calibration against LwF+μ (f=0.13, seed 42, n=154), each row paired against the LFA recipe, which judges 0.968 (zero-rate 26.8%) at dose 15 and 1.042 (22.2%) at dose 20, its best feasible. Cells: Δ (LwF+μ minus LFA), the primary judge's verdict and p, then "S", the robustness judge's p and verdict. $^{\equiv}$: §4. LFA is r=32, LwF+μ r=16: not rank-matched.

Arm (dose 15)	Judge (zero-rate)	vs. LFA at its best dose, 20	vs. LFA at matched dose 15
LwF+μ, α_{LwF}=4, μ=0.05	1.103 (20.6%)	+0.061, tie$^{\equiv}$ (p=0.18; S 0.39, tie)	+0.135, LwF+μ ahead (p=0.0127; S 0.0123, LwF+μ)
LwF+μ, α_{LwF}=4, μ=0.1	1.116 (22.3%)	+0.074, tie (p=0.13; S 0.39, tie)	+0.148, LwF+μ ahead (p=0.0168; S 0.44, tie)

run at seed 42 on the $f{=}0.13$ corpus, so domain perplexity is comparable down the table and never against an arm trained at another f.

Judge resolution. Method-level claims rest on paired Wilcoxon tests over the 154 shared questions, which separate differences of $\gtrsim$0.1. Unpaired run-level differences are read against the measured minimum detectable effect, $\approx$0.12 (answer-set) and $\approx$0.19 (configuration-level; irreducible $\sigma_{\text{run}}\approx$0.036 from sampling nondeterminism), and below it they are ties. Both constants were measured on `claude-sonnet-5` and are carried unchanged. The primary judge's own test–retest term is 0.0216 (`claude-sonnet-5`: 0.0350), and its full-pipeline run-to-run term, over 27 arms judged at two independent draws of byte-identical weights, is mean $|\Delta|$ 0.042, max 0.114. Paired sentences quote the mean over the questions a pair shares, which differs from the run-level cell by at most 0.003 (LwF+μ at $\mu{=}0.05$, dose 15: 1.102 run-level, 1.103 paired). Per-arm zero-rates are in Table 14.

Two axes set the level: f and λ. The best feasible judged score of the LFA arms is $\approx$0.76 in the raw-corpus regime of Table 7 and 1.04–1.10 at $f{=}0.13$; the two regimes also differ in rank, artifact and optimizer geometry, and under the search protocol itself the raw corpus reaches 0.893 (the recipe with λ re-tuned to $2{\cdot}10^4$, dose 10) against 0.968–1.042 at $f{=}0.13$. What f mainly buys is a wider (λ, dose) region in which the method is both stable and adapting.

The anchoring strength λ sets how long adaptation can safely run. At fixed $\mu{=}0.05$, $f{=}0.13$ the λ sweep is monotone *in feasibility*: $\lambda{=}1000$ forgets past a light dose (general-text $+105\%$ by epoch 20), $\lambda{=}5000$ is feasible only to epoch 10 ($+39.6\%$ at epoch 20), $\lambda{=}20000$ is feasible at epochs 5, 10 and 20 (its epoch-15 checkpoint was not scored), and $\lambda{=}50000$ at every measured dose through epoch 20 (-1.9%). The best feasible judged score rises with it, from 0.968 ($\lambda{=}5000$, epoch 10) to 1.097 ($\lambda{=}50000$, epoch 20), a rise that separates under the primary judge ($+0.130$, $p{=}0.017$) and not under `claude-sonnet-5` (0.814 and 0.854; $+0.038$, $p{=}0.33$). We claim only what the perplexity constraints support: λ **extends the dose range over which adaptation is safe.** (At the low-dose end the shipped $r{=}32$ recipe is itself infeasible at dose 5, domain PPL 12.32.)

Table 14: Judge-guided search: best *feasible* judged score per method, after giving each method its own primary preservation–adaptation knob sweep (LFA: λ and layer schedule; L2-SP: μ; EWC: λ_{EWC}; replay: buffer fraction; GeRe: weight). **Frame:** $f{=}0.13$ throughout; feasible = domain PPL $\leq$12 *and* gen. ΔPPL $\leq+15\%$; gradient accumulation is set by mechanism and is *not* matched across rows, so no pairing here is optimizer-step matched. **Judge** `gpt-5.6-luna@medium` on the frozen 154-question protocol, /5 rubric. **Zero-rate:** archive-wide this judge scores exactly 0 on 30.6% of judgments, so every mean sits above a floor the instrument cannot see past. **This is the single-domain frame: the table licenses pairwise verdicts and no ordering of methods.**

Method	Judge	zero-rate	Operating point	dom. PPL	gen. Δ
LFA (λ, μ), text-free	1.097	0.227	$\lambda{=}50000$, ep 20	9.1	-1.9%
Replay *(needs data)*	0.815	0.353	$n{=}3$, ep 20	7.3	-19.9%
GeRe *(needs data; authors' code)*[‡]	0.878	0.371	`dy`, ep 20	11.2	-2.0%
its unstable peak	(0.953)	0.300	$w{=}1$, ep 5	7.8	-10.7%
TM-only+μ *(ours; needs data)*	1.043	0.229	$\mu{=}0.2$, final	8.8	$+3.3\%$
L2-SP	0.798	0.379	$\mu{=}0.1$, ep 5	8.0	$+7.2\%$
EWC	—	—	no feasible point at $r{=}16$ (forgets at every λ_{EWC})[†]		
VeFA	—	—	no feasible point (domain PPL >12 always; no rank knob)		
Unanchored	—	—	no feasible point (catastrophic forgetting)		

Notes to Table 14. The LFA row is the $r{=}16$ arm on the $\approx$0.5 MB diagonal-moments artifact (App. B), not the shipped recipe, whose own point under this protocol is 1.042 at dose 20 (Table 31). The replay row is its $n{=}3$ point at the search's matched dose 20; its highest feasible level is 0.847 ($n{=}2$, dose 5), within the 0.12 resolution of the tabulated one. [†]All rows use LoRA $r{=}16$; rank is a *shared* hyperparameter, so every LoRA arm was re-run at $r{=}4$ (App. C.1): EWC becomes feasible there (judge 0.78–0.79 across $\lambda_{\text{EWC}}{\in}\{30, 300\}$) and unanchored borderline feasible at dose 5. The best-over-both-ranks picture changes only for EWC (0.790 at $\lambda_{\text{EWC}}{=}300$) and L2-SP (0.807 at $r{=}4$, dose 5, against 0.798). Under `claude-sonnet-5` each family's best point is re-selected, not re-levelled: EWC at $\lambda_{\text{EWC}}{=}30$ (0.574) and L2-SP at $\mu{=}0.05$ (0.581), neither separable from that judge's replay level (0.665) at its unpaired resolution and both $\approx$0.27–0.28 below its LFA row (0.854). [‡]GeRe is the authors' released implementation, swept over $w \in \{1, 100, \texttt{dy}\}$. `dy` is its only

dose-stable configuration (feasible at all four doses); $w{=}1$ posts the best single point at epoch 5 and then diverges (domain PPL 7.8 $\rightarrow$ 78 by epoch 20), so `dy` is the operating point and the $w{=}1$ peak is in parentheses. Its trainer chunks the corpus once, so GeRe saw the supplement in every epoch and not in 8 of 20 (App. N). TM-only+μ's 1.043 is its `final` draw; the `e20` draw of the same weights reads 0.985.

Against replay. With its buffer fraction as its only knob, replay's tabulated point judges 0.815, below the never-adapted base model's 0.931. Against that $\mu{=}0$ replay rung at the matched dose 20 the search's LFA arm is paired-ahead (replay -0.286, $p{=}3.2{\times}10^{-6}$, $n{=}154$; `claude-sonnet-5`: -0.190, $p{=}3.2{\times}10^{-4}$). **Granting replay L2-SP closes that margin.** The search swept replay's buffer knob but no weight-space penalty, and L2-SP helps rehearsal and LwF, not GeRe (App. J). It lifts replay from 0.815 to 0.963 (the `final` draw; the `e20` draw of the same weights reads 0.933), just above the base model, and the margin to LFA's 1.097 falls to $\sim$+0.13: an unpaired difference below the 0.19 configuration-level threshold, so the two are not separable, and that pairing has no paired primary-judge test. **We therefore state the result as a *match* with replay, not a win over it**; replay reaches this level by rehearsing general-corpus text that LFA replaces with statistics.

L2-SP transfers across preservation signals, though not universally. The same μ ladder cleans replay's spectrum (410 $\rightarrow$ 0 intruders) and raises its update entanglement (0.69 $\rightarrow$ 0.86) with no λ anywhere. On the authors' GeRe it moves the judge by less than the resolution in both our settings, so we scope the claim to rehearsal and, below, LwF. Applied to our TM-only anchor it reaches 1.043 at its best feasible dose, level with the recipe's best (1.042) and with the search's $r{=}16$ LFA row (1.097); neither pairing has a primary-judge paired test, so both are levels, not paired ties. At matched domain fit (8.76 vs. 8.76) TM-only+μ pays +3.3% general-text drift where the recipe holds -10.0%, a cross-configuration read (the recipe is $r{=}32$, ga 1, dose 15; TM-only+μ is $r{=}16$, ga 2, its `final` draw). The search also isolates λ: L2-SP is LFA's $\lambda{=}0$ slice, and at every swept μ it forgets past a light dose, where function anchoring extends the feasible range to the full training budget. At $r{=}16$ none of EWC, VeFA and unanchored is feasible at any knob setting; at $r{=}4$ EWC and, marginally, unanchored become feasible (§C.1). The search still establishes no ranking, since the tuning budgets are not matched (§4).

Against LwF. Table 15 carries the fifteen verdicts; the text adds what it does not show. With α_{LwF} as its only knob, the one LwF configuration feasible past a light dose is $\alpha_{\mathrm{LwF}}{=}16$ (1.008 at dose 15, 1.000 at dose 20); $\alpha_{\mathrm{LwF}}{=}1$ forgets past the +15% bar at every dose and $\alpha_{\mathrm{LwF}}{=}4$ is feasible at dose 5 alone (0.925 at domain PPL 8.21, general-text +4.3%). At dose 15 the recipe reaches domain 8.76 at general-text -10.0% against single-knob LwF's 11.58 at +1.1% (recipe $r{=}32$, LwF $r{=}16$). That ladder gives LwF one knob where LFA has two, so we re-ran it with the identical uniform weight pull at $\alpha_{\mathrm{LwF}}{=}4$, $\mu \in \{0.05, 0.1\}$. L2-SP converts $\alpha_{\mathrm{LwF}}{=}4$ from one feasible dose into three ($\mu{=}0.05$) or four ($\mu{=}0.1$) at a cost of $\sim$0.9–1.9 PPL of domain fit, the same trade it bought replay, and lifts LwF's best feasible judged score from 1.008 to 1.116 (paired, the single-knob best against the $\mu{=}0.05$ arm at dose 15: $+0.093$, $p{=}0.0405$; robustness judge $p{=}0.0001$). At matched dose LwF+μ is ahead of the recipe by $+0.135$ ($\mu{=}0.05$) and $+0.148$ ($\mu{=}0.1$), and again within seed 1337; at best dose each it is ahead by $+0.154$ in the one pairing that *crosses seeds*. **We therefore state the single-domain judged result as a match with LwF and with replay, never a win over either**, and we do not rank the three on judged quality.

What separates LFA from LwF+μ lies on other axes: at matched dose, lower domain perplexity (8.76 vs. 8.97) and general-text drift (-10.0% vs. +8.8%; recipe $r{=}32$ against LwF $r{=}16$), and an anchor evaluated on n sampled states, not a teacher forward over text (App. I); L2-SP zeroes LwF's intruder count as it does ours (0 against 80 for single-knob LwF). Two caveats bind. The judge moves fast with dose (LwF+μ 0.967 $\rightarrow$ 1.102 from dose 10 to 15; the recipe 0.968 $\rightarrow$ 1.042 from dose 15 to 20), so "best feasible dose" selects on a quantity a paired test cannot price; and LwF+μ was given two configurations ($\alpha_{\mathrm{LwF}}{=}4$, $\mu{\in}\{0.05, 0.1\}$) against the recipe's 55.

Against the never-adapted model, by rubric dimension. The recipe's tie with the base model on the five-dimension mean (§5; equivalent within the margin at both seeds, Table 11) is a cancellation. Adaptation shortens the mean answer from 2,196 to 604 characters, and at seed 42 synthesis falls (-0.188, $p{=}0.0016$) as domain accuracy rises ($+0.188$, $p{=}0.0034$), with correctness, completeness and reasoning quality ties ($p{=}0.057$, 0.76, 0.17). On the mean of correctness and domain accuracy

the recipe is ahead of base at both seeds under both judges (primary $+0.156$, p=0.0085 and $+0.179$, p=0.0066; `claude-sonnet-5` p=0.0002 and 0.0025).

Table 15: LFA against LwF in one domain, inside the search of Table 14 (f=0.13; `gpt-5.6-luna@medium`, /5). The two level columns are means over the questions a pair shares, so one checkpoint can read slightly differently against different partners. "Recipe" is the shipped configuration (r=32); "rank-matched" is the r=16 GMM arm (λ=$2{\cdot}10^4$) against LwF at r=16; the last two rows carry LwF+μ=0.1 to r=32. *Best dose*: each method at its own best feasible dose (20 for both LFA arms, 15 for every LwF arm); *matched dose*: both at dose 15. Paired two-sided Wilcoxon, α=0.05. Last column: `claude-sonnet-5` on the same contrast, p and verdict only; five of the fifteen rows change verdict between the judges and none reverses. One run per arm, seed 42 unless marked. *Crosses seeds (recipe seed 42, LwF+μ seed 1337). Zero-rates of these checkpoints: 0.201–0.312.

Comparison	Arms	n	LFA	LwF	paired p (verdict)	Sonnet p (verdict)
LwF as published (μ=0), α=16 at dose 15						
best dose each	recipe	154	1.042	1.009	0.45 (tie)	$1.1{\times}10^{-3}$ (LFA)
best dose each	rank-matched	154	1.151	1.009	0.016 (LFA ahead)	$2.4{\times}10^{-3}$ (LFA)
matched dose 15	recipe	154	0.968	1.009	0.42 (tie)	0.21 (tie)
matched dose 15	rank-matched	154	1.047	1.009	0.52 (tie)	0.62 (tie)
best-domain pt.	recipe vs. α=4@dose 5	154	0.968	0.925	0.48 (tie)	$2.6{\times}10^{-3}$ (LFA)
LwF+μ, α=4, μ=0.05, dose 15						
best dose each	recipe	154	1.042	1.103	0.18 (tie)	0.39 (tie)
best dose each	rank-matched	154	1.151	1.103	0.40 (tie)	0.46 (tie)
matched dose 15	recipe	154	0.968	1.103	0.0127 (LwF+μ)	0.0123 (LwF+μ)
matched dose 15	rank-matched	154	1.047	1.103	0.31 (tie)	0.0005 (LwF+μ)
matched dose 15	recipe, both s1337	154	0.974	1.196	0.0002 (LwF+μ)	$<10^{-4}$ (LwF+μ)
best dose each*	recipe vs. LwF+μ s1337	154	1.042	1.196	0.0031 (LwF+μ)	0.21 (tie)
LwF+μ, α=4, μ=0.1, dose 15						
best dose each	recipe	154	1.042	1.116	0.13 (tie)	0.39 (tie)
matched dose 15	recipe	154	0.968	1.116	0.0168 (LwF+μ)	0.44 (tie)
matched dose 15	r=32 rank-matched	154	0.968	1.022	0.29 (tie)	0.60 (tie)
best dose each	r=32 rank-matched	154	1.042	1.022	0.60 (tie)	0.18 (tie)

E.1 Every baseline under one frozen protocol, at f=0

Table 16 scores every method under the frozen protocol of §4, in one domain at f=0. It establishes three things. Against unanchored adaptation LFA cuts intruders $713 \to 2$ and general-text drift $+444\% \to +1\%$ and improves held-out domain perplexity $17.81 \to 10.33$, at a higher final content loss. The λ term, not the weight-space pull, carries the preservation. And replay on a well-built buffer out-fits LFA on the domain and preserves general-text perplexity best.[2]

The λ contribution. At matched μ=0.05 and matched content loss, L2-SP (μ only) forgets $+12.9\%$ while λ+μ forgets $+1.1\%$. The two arms match on rank, corpus, dose and content loss but not on optimizer protocol (ga 8, warmup 100 against ga 1, warmup 50). L2-SP has *fewer* intruders (0 vs. 2) yet forgets *more*: function-space anchoring, not intruder-count suppression, drives preservation (§A.4).

Replay. Matched replay is the strongest raw preserver: it *improves* general-text perplexity (-18%). A value *below* the base model's signals distributional overlap between the buffer and WikiText-2, so the metric partly rewards replay for training on similar text (App. C flags the analogue for LFA at high anchoring strength).[3] A proxy-fairness audit strengthens this baseline. Rebuilding the buffer to

[2] Counts here use top-k=10, fixed *a priori*, and are therefore lower than the top-k=20 counts of Table 5, whose perplexity baseline also differs (17.80 on a 200-document subset against 18.18 on full WikiText-2).

[3] The overlap is distributional, not literal: a two-layer contamination audit (exact 13-gram containment against the WikiText-2 test split, and article-identity matching against its 61 test articles, which Wikipedia revision churn makes necessary) found zero hits in the $p(h)$ proxy corpus and $\sim$3 marginal documents ($\sim$5 KB) in the 2.4 GB replay buffer.

Table 16: Baseline comparison under the frozen protocol (Qwen3-0.6B, domain A, f=0, 10 epochs; LoRA r16 for every arm except VeFA, whose adapter is its per-channel rescale, and the ‡ row; top-k=10). Content: final-epoch domain cross-entropy, buffer batches excluded; arms differ in gradient accumulation (App. L), so compare at matched Content. General-text PPL: full WikiText-2 (base 18.18). Retain: mean accuracy over 11 multiple-choice benchmarks, one identical command. Domain: held-out chat-formatted Q&A perplexity, not comparable across f. The λ and $\lambda{+}\mu$ rows use the $\approx$0.5 MB diagonal artifact, the GMM row the shipped one. ‡The confirmed configuration (r=32) with λ re-tuned for f=0.

Method	Content↓	Gen. PPL	ΔPPL	Intr.	Retain	Domain
Base (teacher)	—	18.18	0%	0	0.428	22.24
Unanchored	1.02	98.9	+444%	713	0.401	17.81
Replay (uniform buffer)	1.46	14.87	−18%	77	0.420	10.95
Replay (seed-matched buffer)	1.57	15.16	−17%	66	0.442	8.19
TM-only anchor (GeRe-deriv.)	1.88	18.53	+2%	15	0.434	11.67
LwF (α_{LwF}=1, T=2)	1.94	18.50	+1.8%	10	0.429	9.97
L2-SP (μ only)	1.95	20.52	+13%	0	0.427	11.13
EWC (Fisher, λ_{EWC}=100)	1.99	19.53	+7%	5	0.426	10.64
LwF (α_{LwF}=4, T=2)	2.17	17.22	−5.3%	2	0.427	11.69
VeFA (forced integration)	2.74	15.73	−13%	0	0.430	17.20
LFA (λ)	1.75	20.23	+11%	26	0.421	10.82
LFA ($\lambda{+}\mu$=0.05)	1.97	18.38	+1%	2	0.426	10.33
LFA ($\lambda{+}\mu$, r16, GMM artifact, λ=5000)	2.17	16.39	−9.8%	0	0.435	11.13
LFA, confirmed configuration‡	2.03	17.20	−5.4%	0	0.435	11.48

the source proportions of the corpus $p(h)$ is fitted on (including its $\sim$10% chat-formatted instruction share; single-seed), decontaminated at both layers, improves replay's domain fit (10.95 → 8.19), intruder count and retained accuracy, and the below-base perplexity survives the swap, so it is a property of rehearsal and not of one corpus. Its judged domain quality ties the recipe at the unpaired run-level resolution of the primary judge (1.008 vs. 0.966 on the `content_only` rubric variant, rescaled ×5 from the 0-to-1 `domain_quality` convention; that variant is not the `judge_mean5` of Tables 7 and 14 and is never differenced against it). The same swap applied to the continual chain leaves its replay row unchanged on perplexity (ΔA +3.30 vs. +3.29) and on judged A (0.838 vs. 0.883, inside the 0.12 threshold); set-B differs by +0.149 at the level and is in no verdict file, so we read it as untested.

A well-built substitute buffer is thus a strong learner on these axes, and the generative instrument shows where the strength ends (App. F, finding vi): the two replay checkpoints score GSM8K 0.295 (uniform buffer) and 0.250 (seed-matched) and IFEval 0.380 and 0.330, against 0.390–0.415 and 0.520–0.555 for the anchored arms and 0.565/0.595 for the base. Replay's best-in-family multiple-choice retention coexists with the family's *worst* generative retention; rank, epochs and data mix also differ, so rehearsal is the likely but not the isolated driver. Replay also rehearses proxy *text* at tuning time (here a matched 137 MB) and leaves a buffer-dependent 66–77 intruder dimensions against $\lambda{+}\mu$'s 2. What LFA offers against replay in one domain is integration cleanliness and the absence of retained text at matched judged quality, not raw-metric superiority.

The other external baselines. *EWC* improves on L2-SP (+7% vs. +13% general-text at comparable content), so importance weighting helps, but it needs a proxy corpus to estimate the Fisher. The *TM-only anchor* preserves nearly as well as $\lambda{+}\mu$ (+2%) but must replay the real text LFA replaces with statistics; it is a labelled ablation, and the authors' GeRe is evaluated in the search above. *VeFA* has zero intruders *by construction* and improves general-text PPL (−13%) but adapts far less: held-out domain perplexity 17.2 against the base's 22.2 and LFA's 10.3. Forced segregation has no row, because with no previous task to orthogonalize against O-LoRA degenerates to plain LoRA; it is measured continually (App. F, finding vii).

LwF, and the pair that isolates the probe. *LwF* is the strongest external method in the table that holds no text. At matched content its α_{LwF}=1 arm is better than both weight-space baselines on

every functional axis (general-text $+1.8\%$ against L2-SP's $+12.9\%$ and EWC's $+7.4\%$; domain 9.97 against 11.13 and 10.64; retained accuracy 0.429 against 0.427 and 0.426), though not on structure (10 intruders against 0 and 5), and edges $\lambda{+}\mu$ on domain fit (9.97 vs. 10.33) at five times the intruder count. The pair that isolates *what* is anchored and *where it is probed* is matched on rank, artifact, protocol and content: LFA at $r{=}16$ on the shipped artifact ($\lambda{=}5000$, content 2.171) against LwF at $\alpha_{\mathrm{LwF}}{=}4$ (content 2.174). LFA is better on all four reported axes: general-text -9.8% vs. -5.3%, domain 11.13 vs. 11.69, 0 vs. 2 intruders, and retained accuracy 0.435 vs. 0.427, the last inside the $\leq$1-point noise band. One asymmetry is deliberate: the LFA arm carries its shipped $\mu{=}0.05$ and the LwF arm $\mu{=}0$, as LwF is defined; the LwF$+\mu$ arms are judged in the search above, where they reach parity. Above $\alpha_{\mathrm{LwF}}{\approx}4$ LwF stops trading and converges on the teacher (at 256: domain 21.32 against the base's 22.24, general-text -0.8%), preservation by not learning.

The confirmed configuration under the same protocol. The ‡ row is the *configuration* of App. C ($r{=}32$, correlated+GMM artifact) trained on this table's $f{=}0$ corpus at dose 10. Its λ had to be re-tuned ($2{\cdot}10^4$): the 10^5 selected at $f{=}0.13$ over-anchors on raw text, so λ no more ports across f than across rank. The row gives up 1.1 PPL of domain fit (11.48 vs. 10.33) for preservation the weight-space regularizers do not reach while adapting (general-text -5.4%, against $+1\%$ for $\lambda{+}\mu$ and $+7\%$/$+13\%$ for EWC and L2-SP), with zero intruders and, with the $r{=}16$ GMM arm, the highest retained accuracy outside seed-matched replay (0.435 vs. the base's 0.428). The $f{=}0.13$ recipe itself posts zero intruders, retained accuracy $0.433/0.440$ at two seeds and domain perplexity $8.76/8.83$, which this table's Domain column cannot rank: the metric scores chat-formatted Q&A, so an $f{=}0.13$ arm has a train/test format match no row here has. On raw domain prose the $f{=}0$ model scores 6.15 against the $f{=}0.13$ model's 7.89, so the order of the two *inverts* with the evaluation format. Base-level retained accuracy is a *multiple-choice* statement: generatively this checkpoint loses $\sim$15 points of GSM8K (App. F, finding vi), a cost the multiple-choice suite cannot see.

F TWO DOMAINS: THE FULL TABLE, THE SEVEN FINDINGS, AND THE CONTROLS

After a second domain LFA is ahead of tuned LwF$+\mu$ and canonical replay$+\mu$ on the new domain at every rank; on the old domain the $r{=}4$ arms form a tie band and the leads at the larger ranks are read against entry in App. G; on old-domain and general-text perplexity LFA trails canonical replay at every rank. A step-matched control shows that the $r{=}32$ leads are not an optimizer-step artefact. Table 17 is the full table, Table 18 the both-judge paired table behind §6, and Table 19 the $r{=}4$ pairings the findings cite. The cost axis of the same frame is App. I, and how the statistic is extended to a second domain is App. B.3.

Domains and stages. A is the philosophy corpus of §4; B is a cybersecurity instruction-tuning corpus ($\approx$5 MB), the 2,191-document train split of the public `Trendyol/Trendyol-Cybersecurity-Instruction-Tuning-Dataset` at seed 42. Both stages use the confirmed configuration (LoRA $r{=}32$, $\lambda{=}10^5$, $\mu{=}0.05$, dose 15, $n{=}16$ states per level). Stage one anchors to the base correlated+GMM artifact with the base model as teacher; its epoch-15 adapter is fused into the base to form the stage-two teacher. Fusion is algebraically exact up to floating-point rounding. The adapter judges 0.968 (zero-rate 26.8%) and the fused model 1.012 ($n{=}154$ each): the same weights, but two separately sampled and separately scored answer sets, an unpaired $+0.04$ inside the 0.147 answer-set resolution; the $r{=}32$ entry level of §6.1 is the fused model's set. An LFA stage-two ladder or control shares its parent chain's stage-one model and re-runs stage two only; each rank's LFA chain has its own stage-one model, as does each LwF and replay chain of Table 20, and replay, L2-SP and unanchored train theirs under matched budgets.

Arms. The *rank-matched* LFA row re-runs the complete chain at $r{=}4/\alpha{=}8$ with $\lambda{=}2{\times}10^4$ at both stages, everything else identical, fusing at epoch 15. It is replicated in full at a second seed (1337), and its stage-two λ ladder ($\{1,4,6\}{\times}10^4$) re-runs *stage two only* from the seed-42 chain's fused model and merged artifact, the same one-knob-ladder design as replay's μ ladder. *Replay*'s buffer is A's actual stage-one training corpus at the tuned mixing ratio (one buffer micro-batch in four). The *GeRe* rows run the authors' released trainer (App. L) through the same two-stage protocol at w-strategy `dy`, $r{=}4$, effective batch 24 with 6 GeRe rows interpolated per optimizer step; stage two recomputes the thresholds from the fused stage-one teacher over the same generic 1K buffer (no A data at any stage), and the $\mu{=}0.1$ arm adds the L2-SP term of Eq. (3) to the released objective.

Table 17: Continual A→B (philosophy→cybersecurity), every arm. ΔA, ΔB: held-out perplexity change over stage two (A enters at $8.75/8.83$ for LFA, measured post-fusion; $7.58/7.40$ for the A-only replay ablation at r=4; 7.83 for its canonical row). Wiki: WikiText-2 after stage two (base 18.18). Set-A/set-B: `gpt-5.6-luna@medium`, /5, n=154/78; floors as in Table 1. #sel as in Table 6: an undirected fingerprint of integration style, not a quality axis. A-data: whether stage two reads A's text (**req.**). Rows without a rank are r=32 (LFA, A-anchor) or r=4 (GeRe, L2-SP, unanchored); ranks are never compared across. Rows printed as s42/s1337 pairs or labelled by seed are two-seed; every other row, and every #sel outside the first two rows, is single-seed (the canonical r=32 replay row's seed-1337 judged levels: Table 18). Replay's reference row is the **canonical** one (every prior corpus rehearsed, L2-SP at both stages); the A-only and μ=0 rows are ablations. "+A-anchor" and its controls: finding (iv). The rank-matched LFA r=4 row is λ=2×10^4 at both stages; 4×10^4 is the tuned stage-two point. [g]GeRe and TM-only+μ (the hinge at w=1 without the replay term, plus μ=0.1; settings carried over from Table 14's r=16 row) rehearse a fixed *generic* 1K web-text buffer: no A text at any stage, though neither is statistic-only, and both saw the question–answer pairs in every epoch (App. N). [o]O-LoRA enters B at 26.7 against ≈14.5 for the other chains. The joint rows train once on A∪B, so their deltas are undefined. Bold flags values the text discusses, not column optima. Arms are not optimizer-step matched (ga 1 for LFA and LwF, 4 replay, 6 upstream trainers) except [s], the r=32 chain re-run at replay's step count (text below); its set-A 1.158 coincides with the on-manifold control's, a different arm. Paired tests: Table 18.

Method	ΔA↓	ΔB↓	Wiki	set-A↑	set-B↑	#sel	A-data
LFA (merged $p(h)$), s42	+3.55	−4.81	18.45	**1.227**	1.81	**203**	none
LFA (merged $p(h)$), s1337	+3.58	−4.53	18.33	1.117	**1.89**	209	none
step-matched control, ga=3 (s42)[s]	+2.87	−6.17	17.53	1.158	1.70	—	none
LFA **rank-matched** r=4 (s42/s1337)	+4.12/+4.16	−7.7/−7.5	17.2/17.0	0.961/1.071	1.79/1.75	—	none
stage-2 λ ladder $1/\mathbf{4}/6\times10^4$ (s42)	+4.9/+**3.5**/+3.1	−7.3/−7.9/−8.0	17.9/16.7/16.5	1.016/1.034/1.062	1.76/**1.90**/1.74	—	none
LFA + A-anchor (β=0.5)	+2.84	−6.07	17.14	1.140	—	—	none
LFA + A-anchor (β=1)	+2.68	−6.43	16.85	1.191	—	—	none
LFA + A-anchor (β=2)	+2.48	−6.77	16.56	1.192	1.76	—	none
LFA + A-anchor (β=2), s1337	+2.31	−6.55	16.49	1.134	1.76	—	none
LFA + A-anchor, shipped artifact	+2.45	−6.65	16.67	—	1.77	—	none
LFA + A-anchor, merged comps (0-storage)	+2.00	−6.32	16.64	1.227	—	—	none
on-manifold control (β=2)	+2.27	−6.79	16.50	1.158	—	—	none
matched-total λ=300k	+**1.96**	−6.61	16.58	1.160	—	—	none
LwF+μ (holds no prior-domain text; output-distribution anchoring on the current stage's inputs)							
LwF+μ r=4 (α^*=4 both stages; s42/s1337)	+2.18/+2.04	−7.11/−7.04	20.1/19.8	0.986/1.038	1.66/1.71	—	none
tuned stage-2 α=8 (s42; ladder 1–16 in text)	+**0.79**	−6.53	18.3	**0.997**	1.47	—	none
LwF+μ r=16 (μ_{S2}=0.05/0.1)	+3.26/+2.72	−7.58/−7.69	28.9/28.6	0.847/0.981	1.55/1.49	—	none
LwF+μ0.1 r=32 (best-vs-best)	+2.95	−7.75	29.4	0.809	1.44	—	none
LFA r=16 (merged $p(h)$, λ=5×10^4)	+3.90	−5.03	19.0	**1.090**	**1.72**	—	none
Replay r=4 **canonical** (all priors, μ both stages)	+0.43	−6.78	**15.96**	0.845	1.33	—	**req.**
same chain, stage-2 μ=0.2	+**0.41**	−7.37	**15.85**	0.840	1.30	—	**req.**
abl.: A-only buffer, μ=0 (s42/s1337)	+3.29/+2.81	−5.99/−5.46	20.3/21.0	0.883/0.788	1.30/1.28	15	**req.**
abl.: A-only, stage-2 μ (0.05/0.1/0.2)	+1.89/+1.46/+1.07	−7.31/−7.70/−8.06	18.9/18.5/18.2	0.917/0.977/0.860	1.36/1.29/1.27	11/6/1	**req.**
abl.: A+general buffer, μ=0 (s42/s1337)	+0.69/+0.46	−7.27/−6.46	16.8/17.2	0.835/0.948	1.37/1.39	—	**req.**
Replay r=16 canonical, stage-2 μ=0.2	+1.16	−3.34	16.4	0.904	1.29	—	**req.**
abl.: A-only, μ=0 (ported r=4 recipe)	+16.91	+9.87	32.9	0.747	—	—	**req.**
Replay r=32 canonical, stage-2 μ=0.5	+1.52	−3.75	16.5	0.916	1.41	—	**req.**
abl.: A-only, μ=0 (ported r=4 recipe)	+16.96	+15.76	44.5	0.695	—	2	**req.**
GeRe (authors', `dy`, r=4)	+**0.04**	−4.06	17.92	0.808	1.48	—	none[g]
GeRe `dy` +μ=0.1	+0.01	−4.05	17.92	0.783	1.58	—	none[g]
TM-only+μ0.1 (ours; generic buffer, r=4)	+1.19	−**10.35**	17.30	1.057	1.75	—	none[g]
O-LoRA (authors', λ_1=0.5, r=4)	+13.96	−19.52[o]	31.45	0.525	1.59	0	none
Joint A∪B, unanch. r=32	—	—	886.49	0.849	1.29	—	**req.**
Joint A∪B, LFA r=32	—	—	**15.94**	1.060	1.77	—	**req.**
LFA (diagonal $p(h)$, r=4)	+5.71	−8.35	18.53	0.861	—	79	none
L2-SP	+14.57	−8.11	70.57	0.386	—	0	none
Unanchored	+29.07	−6.49	149.04	0.244	—	0	none

Metrics. ΔB is measured against the fused stage-one model's perplexity on B's held-out queries. Set-A is the App. M protocol on A's 154 questions after stage two. Set-B uses 78 questions generated from B's corpus by the same pipeline (generator distinct from the judge) on the same rubric. Its no-learning floor is the fused pre-B model: on the r=32 chain it reads 1.118 (zero-rate 37.4%) under the primary judge and 0.882 (39.7%) under `claude-sonnet-5`, n=78, one seed. That floor is a level with no paired test attached, and no other chain's pre-B model was scored, so for r=4 and r=16 it is indicative. #sel applies the footprint analysis of §A.4 to the stage-two ΔW (change over the fused stage-one model).

The step-matched stage-two control. The tabulated arms are not optimizer-step matched (ga=1 for LFA and LwF, 4 for replay). Re-running the r=32 LFA stage two at ga=3, from the same fused model and merged artifact at seed 42, gives it canonical replay's optimizer-step count exactly (116 after epoch 1 and 1,875 in all, against the plain chain's 347 and 5,608). Under the primary judge the control reads 1.158 on set-A, a tie with the plain chain's 1.227 (+0.069, p=0.13; equivalent within the margin, TOST p=0.043), and is ahead of canonical replay (0.916; p=3.3×10^{-5}) and of LwF+μ (0.809; p=5.3×10^{-11}). On set-B it reads 1.703, a tie with the plain chain's 1.810 (+0.108, p=0.28;

Table 18: The nine continual contrasts of §6 under both judges (seed 42). Primary-judge cells give the paired difference Δ (*second arm minus first*, so negative means the first arm scores higher), the two-sided Wilcoxon p and the verdict at α=0.05 (n=154 on set-A, 78 on set-B), quoted from the verdict file. Sonnet columns: `claude-sonnet-5` on the same answers, p and verdict only (§4). Arms are each method's tuned arm per rank: LFA at stage-two λ=4×10^4 (r=4), λ=5×10^4 (r=16) and the recipe chain (r=32); LwF+μ at α=8 (r=4) and stage-two μ=0.1 (r=16, 32); replay+μ canonical. $*$ marks the two of eighteen cells whose verdict changes between judges, both at r=16. Not optimizer-step matched (ga=1 for LFA and LwF, 4 for replay); the step-matched controls are in the text below (r=32, stage two) and Table 23 (stage three). Levels: Table 17. [§]A tenth row, inside the declared family of App. D: the r=32 LFA and canonical replay+μ chains re-run at seed 1337 (stage-two μ=0.5 as at seed 42). Primary judge: set-A 1.117 against 0.901 (p=9.3×10^{-5}), set-B 1.895 against 1.392 (p=3.4×10^{-5}); `claude-sonnet-5`: +0.197 (p=9.9×10^{-5}) and +0.490 (p=2.8×10^{-6}). Replay's own set-A does not move from entry at that seed either (0.971 → 0.901, p=0.38, a tie, inconclusive; Sonnet p=0.92). On the two least length-sensitive dimensions the primary judge reads the seed-1337 set-A lead as a tie (+0.104, p=0.070) and keeps the set-B lead (+0.353, p=0.013), the mirror image of seed 42 (set-A kept, p=0.0038; set-B a tie, p=0.078); `claude-sonnet-5` keeps both (p=0.0072, 0.0033). LFA's seed-1337 set-A answers are about twice replay's length (1,058 against 535 characters). LwF+μ has no seed-1337 A→B partner, and the after-C chains' seed coverage is unchanged.

Contrast	rank	set-A: Δ, p (verdict)	Sonnet p	set-B: Δ, p (verdict)	Sonnet p
LFA vs. replay+μ	r=4	−0.188, 0.0006 (LFA)	< 10^{-4} (LFA)	−0.567, < 10^{-4} (LFA)	< 10^{-4} (LFA)
	r=16	−0.186, 0.0021 (LFA)	0.0034 (LFA)	−0.433, < 10^{-4} (LFA)	< 10^{-4} (LFA)
	r=32	−0.312, < 10^{-4} (LFA)	< 10^{-4} (LFA)	−0.400, 0.0003 (LFA)	< 10^{-4} (LFA)
	r=32, seed 1337[§]	−0.216, < 10^{-4} (LFA)	< 10^{-4} (LFA)	−0.503, < 10^{-4} (LFA)	< 10^{-4} (LFA)
LFA vs. LwF+μ	r=4	−0.036, 0.44 (tie)	0.11 (tie)	−0.423, < 10^{-4} (LFA)	< 10^{-4} (LFA)
	r=16	−0.109, 0.0496 (LFA)*	0.073 (tie)*	−0.233, 0.0034 (LFA)	0.0003 (LFA)
	r=32	−0.418, < 10^{-4} (LFA)	< 10^{-4} (LFA)	−0.369, 0.0005 (LFA)	< 10^{-4} (LFA)
LwF+μ vs. replay+μ	r=4	−0.152, 0.0036 (LwF)	0.020 (LwF)	−0.144, 0.038 (LwF)	0.014 (LwF)
	r=16	−0.077, 0.12 (tie)	0.13 (tie)	−0.200, 0.013 (LwF)*	0.051 (tie)*
	r=32	+0.106, 0.045 (replay)	0.019 (replay)	−0.031, 0.49 (tie)	0.93 (tie)

inconclusive), and is ahead of replay (1.410; p=0.0029) and of LwF+μ (1.441; p=0.0032). All four leads survive the multiplicity correction of App. D, and `claude-sonnet-5` reads the same six verdicts (p=0.30, 2.2×10^{-6}, 4.9×10^{-11}; 0.84, 8.0×10^{-5}, 5.2×10^{-5}). With a third of the optimizer steps the control also exits lower on all three perplexity axes (A 11.62, WikiText-2 17.53, B 7.39, against the plain chain's 12.30, 18.45, 8.76). It is one rank and one seed; the stage-three controls at all three ranks are in Table 23.

(i) Replay is ahead on prior-domain perplexity; on the old domain's judged quality the methods sit in a tie band at r=4, and where they separate it is against replay's canonical configuration, not its judged-best arm. On ΔA LFA is the worst of the tuned r=4 arms (3.51; LwF+μ 0.79, canonical replay 0.43, GeRe 0.04), partly because it enters stage two better fitted than LwF or GeRe (enter-A 8.72; LwF 10.28; GeRe, at its own footing, 12.40), and on the exit level it is third of the four (12.23; LwF+μ 11.07, replay 8.27, GeRe 12.44). Replay's stage-two μ ladder takes ΔA monotonically down to 1.07 at μ=0.2, the canonical configuration reaches 0.43 at WikiText-2 15.96, and its μ=0.2 sibling's 0.41 at 15.85 is the lowest WikiText-2 in Table 17 and, GeRe aside, the lowest ΔA. **We claim no continual perplexity advantage.** Judged quality is read rank-matched, because adapter rank is a capacity variable. At each method's tuned point LFA (1.034 at stage-two λ=4×10^4), LwF+μ (0.997), TM-only+μ (1.057) and replay's judged-best arm (A-only buffer, stage-two μ=0.1, 0.977) sit within ≈0.98–1.06 on judged A, and every pairing among them that was tested is a tie (Table 19; the robustness judge reads LFA ahead of three of LwF+μ's other arms). The strict rank-matched chain (λ=2×10^4 at both stages; 0.961, second seed 1.071) **ties** the A-only μ=0 replay chain's 0.883 (p=0.22, App. J), at an exit-A perplexity of 12.85 against 8.65 for replay at μ=0.2. Replay's canonical row (0.845), its μ=0 arm (0.883) and GeRe (0.808) sit below the band: LFA is ahead of each, and LwF+μ is ahead of canonical replay (Tables 18 and 31). So LFA's old-domain lead over replay depends on which arm represents replay, and the two readings are consistent because judged A is flat across every r=4 replay configuration run (0.79–0.98) while ΔA spans 3.30 to 0.41. The band is the judge's resolution over tuned points, not a claim that the methods are interchangeable; how the old-domain level reads against the level each chain entered with is App. G. Weight-space regularization alone does not preserve here (L2-SP: ΔA +14.57, WikiText-2 70.6, set-A 0.386), and artifact fidelity matters (ΔA +5.71 for the diagonal-moments artifact at r=4).

(ii) Capacity separates the methods. Rehearsal is capacity-*sensitive*: the configuration tuned at r=4 over-fits when carried to r=16 and r=32 (the two A-only μ=0 *abl.* rows of Table 17) and is recovered only by rehearsing every prior corpus *and* a strong stage-two μ (0.2 at r=16; 0.5 at r=32, where 0.2 over-fits B). Its general text is what carries its perplexity: at r=16 with μ=0.05 at both stages, dropping the general text from the stage-two buffer, the only difference, raises WikiText-2 after B from 16.66 to 24.16 and ΔA from +1.95 to +13.06, while neither judged axis separates the pair (set-A p=0.076, set-B p=0.53; both inconclusive, Table 11). LwF+μ's judged levels fall with rank across one arm per rank (set-B 1.66 → 1.55 → 1.44, set-A 0.986 → 0.847 → 0.809; its r=4 → 16 step is a paired loss on A under the primary judge, Table 19) and its WikiText-2 rises to 28.6–29.4. LFA's set-A level rises with rank (1.034, 1.090, 1.227; no cross-rank pairing was run, and §6.1's reading of that level applies), its set-B holds (1.90, 1.72, 1.81) and its WikiText-2 stays between 16.7 and 19.0.

(iii) The footprint mechanism transfers. LFA's stage-two change is built of domain-selective, entangled modes (#sel 203/209 at r=32), the signature of §A.4. At the rank-matched point #sel is 143 for LFA against 25 for the canonical replay chain and 0 for L2-SP; #sel is undirected and is never read across ranks.

(iv) Raising the stage-two anchoring dose adds a replicated gain on both perplexity axes, and a control cascade shows that *dose*, not allocation, is the mechanism. An additive second anchor (weight β) applies the anchoring loss to sampled A activations at the block sites the *base* mixture explains worst (top half by responsibility entropy), and leaves the merged-artifact term untouched. It improves both perplexity axes monotonically (ΔA +3.55 → +2.48 at β=2, second seed +3.58 → +2.31; ΔB −4.81 → −6.77) with judged A holding (1.134–1.227 across the β=0/0.5/1/2 arms and the second seed). Three controls, the first pre-registered, reproduce the gain without the intended mechanism (rows of Table 17): the same β on the *lowest*-entropy half, where a probe-level slice test finds no targeting signature (slice×selection interaction p=0.34, Mann–Whitney; A is only 6.5–7.7% outside the base artifact's PCA basis by activation variance at five sampled sites, B 7.8–9.3%); sampling only A's *own* merged components, at zero additional storage (retained accuracy 0.434, base level); and a *matched-total* control with no second term at all (λ → 3λ on the

Table 19: Paired contrasts at $r{=}4$ after B that Table 18 does not carry (seed 42; two-sided Wilcoxon, $\alpha{=}0.05$; $n{=}154$ on set-A, 78 on set-B). Each cell is p and the arm that is ahead, or "tie", under the primary judge and then under `claude-sonnet-5` on the same answers. LFA is the tuned stage-two $\lambda{=}4{\times}10^4$ chain unless marked; LwF+μ arms are named by their stage-two knob; "replay best" is replay's judged-best arm (A-only buffer, stage-two $\mu{=}0.1$). [a] LFA's judged-best arm, $\lambda{=}6{\times}10^4$; the robustness judge paired the $\lambda{=}4{\times}10^4$ arm instead. [b] The strict-match $\lambda{=}2{\times}10^4$ chain.

Contrast	set	primary judge	`claude-sonnet-5`
The old-domain tie band			
LFA vs. LwF+μ ($\alpha{=}4$, $\mu{=}0.05$)	A	0.27, tie	$<10^{-4}$, LFA
LFA vs. LwF+μ ($\alpha{=}16$)	A	0.85, tie	0.045, LFA
LFA vs. TM-only+μ	A	0.64, tie	0.97, tie
LFA[a] vs. replay best	A	0.136, tie	0.59, tie
LwF+μ ($\alpha{=}8$) vs. replay best	A	0.68, tie	0.17, tie
TM-only+μ vs. replay best	A	0.21, tie	0.63, tie
LFA[b] vs. replay A-only, $\mu{=}0$	A	0.22, tie	0.71, tie
New-domain placements			
LFA vs. LwF+μ ($\alpha{=}4$, $\mu{=}0.05$)	B	0.0022, LFA	0.015, LFA
LFA vs. LwF+μ ($\alpha{=}4$, $\mu{=}0.1$)	B	0.038, LFA	0.0014, LFA
LFA vs. TM-only+μ	B	0.063, tie	0.22, tie
LwF+μ ($\mu{=}0.05$) vs. replay A-only, $\mu{=}0$	B	0.0005, LwF	0.0006, LwF
LwF+μ ($\mu{=}0.1$) vs. replay best	B	$<10^{-4}$, LwF	0.0003, LwF
LwF+μ ($\mu{=}0.05$) vs. GeRe	B	0.12, tie	0.022, LwF
LwF+μ ($\mu{=}0.05$) vs. O-LoRA	B	0.58, tie	0.86, tie
LwF+μ ($\alpha{=}8$) vs. O-LoRA	B	0.23, tie	0.017, O-LoRA
TM-only+μ vs. replay best	B	$<10^{-4}$, TM-only	$<10^{-4}$, TM-only
TM-only+μ vs. GeRe	B	0.0043, TM-only	0.0003, TM-only
LwF+μ's capacity step, $r{=}4 \to 16$ at $\mu{=}0.05$			
LwF+μ $r{=}4$ vs. $r{=}16$	A	0.0024, $r{=}4$	0.088, tie
LwF+μ $r{=}4$ vs. $r{=}16$	B	0.25, tie	0.036, $r{=}4$

merged mixture, equal total anchor mass to β=2). The dose–response saturates (per-doubling ΔA gains $-0.50/-0.33/-0.17$ across β=4/8/16; no arm above β=16 was run), and no instrument detects displacement at any dose run: retained accuracy stays at base level at β=16 (0.437 vs. 0.428, inside the $\leq$1-point band of the 11-benchmark suite), judged A holds (1.143–1.227 across β=4/8/16), and function-space drift from the teacher over the *base-only* mixture *falls* with dose ($2.7\times10^{-6} \rightarrow 0.9\times10^{-6}$), A's activations overlapping components that carry $\sim$65% of base mixture mass. Apart from β=2 these arms are single-seed. **The operative variable is total anchoring dose**, and the recipe collapses accordingly: continual stage two wants roughly 3$\times$ the single-domain λ, with no second term, no per-domain artifact and no selection machinery.

(v) Judged quality on the *new* domain favors anchoring over rehearsal, and this contrast survives rank-matching. At r=4 LFA reads 1.74–1.90 on set-B across two seeds and the full 6$\times$ stage-two λ range, against replay's 1.30/1.28, the replay μ ladder's 1.27–1.36 and GeRe's 1.48, and every LFA-versus-replay set-B contrast on file keeps its verdict across the two judges. LwF+μ locates the effect in the function-anchoring family: its α_{LwF}=4 arms read 1.66/1.71 at r=4, ahead of the two replay configurations they were paired with, level with GeRe and with O-LoRA (1.59), and behind LFA (Table 19). One data-using arm places *with* LFA: TM-only+μ reads 1.751 on set-B and 1.057 on set-A, ties on both, at a third of LFA's ΔA (1.19) and the lowest exit-B in the table (5.82); it rehearses generic text. **(vi) A generative instrument reveals capability loss that perplexity and multiple-choice retention cannot see, and dose *prevents* it.** On an earlier draw of the generative instrument than Table 2's (the 2026-07 draws; the later draw of the same two states reads 0.550 $\rightarrow$ 0.450 and 0.620 $\rightarrow$ 0.550, Table 25), generative GSM8K falls from the base model's 0.565 to 0.415 after stage one (the diagonal-artifact arm reads the same, 0.410) and to 0.365 after stage two at β=0, while the multiple-choice retention suite stays at base level throughout (anchored arms 0.432–0.437, base 0.428, replay 0.413). Raising the stage-two dose recovers the stage-two loss and more (0.450–0.470 across the ladder, above the stage-one teacher), so the student is not a distillate of its teacher. Stage one's loss can be reduced only up front, and there dose is a trade-off rather than a Pareto move: tripling the stage-one λ recovers about two thirds of it (GSM8K 0.390 $\rightarrow$ 0.490) at a domain perplexity of 9.95 against the recipe's 8.76 (single-seed). A transcript autopsy finds real reasoning degradation under the 3-shot protocol, not a parsing artifact, beside a zero-shot format-compliance drift that IFEval also shows (0.595 $\rightarrow$ 0.545 $\rightarrow$ 0.520, recovering to 0.545–0.555 across the stage-two ladder); TriviaQA is flat.

(vii) Forced segregation, run on the authors' code, does not preserve at this rank. O-LoRA (Wang et al., 2023) with its released λ_1=0.5 achieves the orthogonality it enforces (its Gram penalty falls 4.4 $\rightarrow$ 0.01) and pays ΔA +13.96 for a mid-pack exit-B (7.15): four times LFA's A cost and thirty times replay's. Its judged A, 0.525, is above its own unregularized control (0.244) and below LFA, LwF+μ and replay (each $p<10^{-4}$). Its λ_1 is upstream's, not re-tuned, and the arm is single-seed.

The perplexity/judge dissociation. Perplexity retention and judged quality move *independently*, four ways and twice inside a single method, each readable in Table 17. (1) A 50/50 A+general buffer is better than the A-only buffer on every perplexity column at both seeds while its judged quality stays in replay's band. (2) The seed-matched buffer that strengthens single-domain replay leaves the continual row unchanged on ΔA and judged A (§E.1). (3) Up replay's μ ladder ΔA improves $\sim$3$\times$ while set-A is flat and non-monotone, the perplexity-best rung judging lowest, and set-B is equally flat. (4) Up LFA's own stage-two λ ladder ΔA improves monotonically (4.87 $\rightarrow$ 3.12) while set-B peaks mid-ladder (1.897 at $\lambda=4\times10^4$) and falls at the top (1.738 at 6×10^4), and set-A is non-monotone. The dissociation occurs within family on both sides of the comparison, so no cross-method tuning asymmetry explains it; attribution beyond the partial memorization account (App. L) is open. The continual claims are therefore stated per instrument: after B, replay is ahead on prior-domain perplexity and LFA on the new domain's judged axis, and the old-domain judged axis is read against entry (App. G).

The joint comparator. Training once on A$\cup$B at r=32 is degenerate unanchored (exits A 12.82, B 27.08; WikiText-2 886; judged 0.849 and 1.29 at zero-rates 0.353, 0.272) and, with the recipe's anchor added, a *perplexity* bound: at dose 15 it is below every sequential LFA-family arm on all three perplexity axes at once (A 9.28 vs. the family's best 10.23; B 5.99 vs. 6.56; WikiText-2 15.94 vs. 16.02; a cross-rank read). It judges 1.060 on A and 1.77 on B (zero-rates 0.239, 0.241): the

sequential chain is ahead of it on judged A ($+0.168$, p=0.0021; a tie under the robustness judge, p=0.093) and ties it on judged B (p=0.81; p=0.64). Both joint arms are single-seed and untuned.

Scope, and order dependence. One domain pair; single-seed rows are marked in Table 17. The stage-two preservation budgets are allocated differently by design: replay spends its whole rehearsal share on A's corpus, while LFA's merged anchor is $\sim$97.5% base and $\sim$2.5% A by mixture mass. That favors replay on ΔA and disfavors it on base-capability readouts (WikiText-2, multiple-choice retention). Set-A is the readout where the allocation favors replay most, its full budget defending the domain set-A scores, and neither buffer remedy (the canonical composition, mixed rehearsal) moved replay out of its flat judged-A band (0.845 and 0.835/0.948, in 0.79–0.98); L2-SP at stage one costs replay little on A fit (enter-A 7.83 vs. 7.58 at r=4, 7.10 vs. 7.07 at r=16). Order dependence (a single-seed reversed chain) is in App. B.3.

GeRe and EWC. The authors' GeRe at its dose-stable configuration (`dy`, r=4; protocol above) preserves almost perfectly and barely learns: ΔA $+0.04$, but it *enters* stage two at 12.40 and exits at 12.44. L2-SP moves neither of its judged axes (0.808 $\rightarrow$ 0.783 on A; ties, Table 31), as in the single-domain setting. EWC was not run continually: it has no feasible operating point at r=16, and its r=4 point judges 0.790 (App. E).

G THREE DOMAINS: OLD-DOMAIN ANSWERING AGAINST ITS ENTRY LEVEL

Against the level it entered stage two with, LFA's judged answering on A rises over stage two as its answers lengthen and returns over stage three. From entry to after C it is equivalent within the margin at r=32 (-0.047, p=0.31; TOST p=0.0498), not separable but inconclusive at r=16 (-0.061, p=0.26; TOST 0.069), and a small fall at r=4 that does not survive the multiplicity correction (-0.091, p=0.045; p=0.060, a tie, on the length control). Canonical replay does not move at either stage (from entry to after C equivalent at r=4 and r=16, inconclusive at r=32), and LwF+μ falls from entry at every rank (-0.155, -0.309, -0.317). The robustness judge reads LFA's entry-to-after-C contrast as a fall at r=16 as well as at r=4, and as a tie at r=32 only. Table 21 carries every transition, Tables 22 and 23 the after-C levels and verdicts, and Table 24 the ablation of the third stage's question–answer supplement.

Table 20: Through the third domain (seed 42, primary judge): set-A level entering stage two $\rightarrow$ after B $\rightarrow$ after C, mean answer characters beneath, and A's held-out perplexity. Each of the nine chains enters from its own stage-one model (r=32 LFA: the recipe, fused, re-sampled; its adapter's answers judge 0.968). After C, †: the robustness judge reads replay ahead of LFA; $^{\equiv}$: the arm's tie with LFA is equivalent (§4). The r=4 LwF+μ chain is the α_{LwF}=4, μ=0.05 mirror. p-values: Table 23.

	set-A: entering stage two $\rightarrow$ after B $\rightarrow$ after C			A-PPL, after B $\rightarrow$ after C	
rank	LFA	LwF+μ	replay+μ	LFA	replay+μ
r=4	0.978 $\rightarrow$ 1.034 $\rightarrow$ 0.887 580 / 928 / 586	1.012 $\rightarrow$ 0.986 $\rightarrow$ 0.857$^{\equiv}$ 1607 / 1772 / 847	0.861 $\rightarrow$ 0.845 $\rightarrow$ 0.897$^{\dagger\equiv}$ 547 / 496 / 515	12.23 $\rightarrow$ 11.55	8.27 $\rightarrow$ 8.65
r=16	0.956 $\rightarrow$ 1.090 $\rightarrow$ 0.895 611 / 976 / 576	1.103 $\rightarrow$ 0.981 $\rightarrow$ 0.794 1496 / 1462 / 778	0.921 $\rightarrow$ 0.904 $\rightarrow$ 0.940$^{\dagger\equiv}$ 544 / 525 / 514	12.60 $\rightarrow$ 11.10	8.25 $\rightarrow$ 9.31
r=32	1.012 $\rightarrow$ 1.227 $\rightarrow$ 0.965 641 / 976 / 694	1.022 $\rightarrow$ 0.809 $\rightarrow$ 0.705 1551 / 1155 / 649	0.860 $\rightarrow$ 0.916 $\rightarrow$ 0.912 528 / 528 / 516	12.30 $\rightarrow$ 10.98	8.34 $\rightarrow$ 9.46

The rise, the return, and what each side holds. Over stage two LFA's mean answer on A lengthens by about half (641 $\rightarrow$ 976 characters at r=32) and its judged level rises at r=32 ($+0.216$, p=3×10^{-5}) and, not robustly to the multiplicity correction, at r=16 ($+0.134$, p=0.028); at r=4 it is equivalent to entry. The rise does not survive the length control under either judge (primary $+0.110$, p=0.085 at r=32; `claude-sonnet-5` p=0.71), so we build nothing on it: stage two's instruction-shaped corpus lengthens answers, replay's set-A also rises at r=4 in one seed of its A-only μ=0 arm (0.817 $\rightarrow$ 0.883), the strict-match r=4 LFA chain shows none (0.978 $\rightarrow$ 0.961), and the level returns over a third domain. What looked like judged backward transfer is not a property of the method. LwF's probe is the current stage's inputs, so what it protects is support-conditional (App. O): LwF+μ's judged answering on A falls over stage two at r=16 and r=32 (1.103 $\rightarrow$ 0.981, p=0.0072; 1.022 $\rightarrow$ 0.809, p=6.6×10^{-5}; its r=16 μ=0.05 mirror by -0.256 at the level), both

Table 21: Set-A against its entry level (seed 42, n=154; levels and answer lengths: Table 20). *Entry* is each chain's own stage-one model, scored on the same frozen questions (the fused model's answers; for LFA r=16 and LwF+μ r=16 and r=32, the same weights' adapter answers, sampled separately). Upper block: every stage transition of the three chains, primary judge: paired Δ and two-sided Wilcoxon p on the five-dimension mean, the verdict ($^{\equiv}$, $^{\flat}$: §4) with the TOST p at ± 0.147 for a tie, then the same test on the length control (mean of correctness and domain accuracy). Last column: `claude-sonnet-5` on the same answers, five-dimension mean, p and verdict only. Lower block: each method against LFA (arm minus LFA) at entry, after B, and after B on the length control. The three methods enter from three different stage-one models, so an entry contrast is a stage-one comparison, not a control.

	rank	transition	Δ	p	verdict (TOST p)	LC Δ	LC p (verdict)	Sonnet p (verdict)
LFA	r=4	entry → B	+0.056	0.23	tie$^{\equiv}$ (0.0381)	−0.081	0.29 (tie)	0.0179 (rises)
		B → C	−0.147	0.0030	falls	−0.045	0.40 (tie)	1.2×10^{-7} (falls)
		entry → C	−0.091	0.0453	falls$^{\flat}$	−0.127	0.0595 (tie)	0.0004 (falls)
	r=16	entry → B	+0.134	0.0281	rises$^{\flat}$	+0.019	0.91 (tie)	0.26 (tie)
		B → C	−0.195	0.0009	falls	−0.081	0.23 (tie)	0.0001 (falls)
		entry → C	−0.061	0.26	tie (0.0690)	−0.062	0.51 (tie)	0.0056 (falls)
	r=32	entry → B	+0.216	3.0×10^{-5}	rises	+0.110	0.0851 (tie)	0.0037 (rises)
		B → C	−0.262	3.4×10^{-7}	falls	−0.205	0.0009 (falls)	1.1×10^{-5} (falls)
		entry → C	−0.047	0.31	tie$^{\equiv}$ (0.0498)	−0.094	0.20 (tie)	0.14 (tie)
LwF+μ	r=4	entry → B	−0.026	0.43	tie$^{\equiv}$ (0.0064)	−0.075	0.16 (tie)	1.9×10^{-5} (falls)
		B → C	−0.129	0.0118	falls	−0.016	0.64 (tie)	0.0669 (tie)
		entry → C	−0.155	0.0018	falls	−0.091	0.16 (tie)	1.3×10^{-8} (falls)
	r=16	entry → B	−0.122	0.0072	falls	−0.146	0.0110 (falls)	3.7×10^{-6} (falls)
		B → C	−0.187	0.0004	falls	−0.088	0.18 (tie)	1.2×10^{-6} (falls)
		entry → C	−0.309	2.7×10^{-7}	falls	−0.234	0.0007 (falls)	5.3×10^{-15} (falls)
	r=32	entry → B	−0.213	6.6×10^{-5}	falls	−0.247	3.0×10^{-5} (falls)	2.9×10^{-8} (falls)
		B → C	−0.104	0.0090	falls	+0.003	0.81 (tie)	0.0008 (falls)
		entry → C	−0.317	1.4×10^{-7}	falls	−0.244	0.0007 (falls)	2.2×10^{-12} (falls)
replay+μ	r=4	entry → B	−0.016	0.54	tie$^{\equiv}$ (0.0147)	−0.036	0.55 (tie)	0.41 (tie)
		B → C	+0.052	0.42	tie (0.0615)	+0.078	0.28 (tie)	0.39 (tie)
		entry → C	+0.036	0.88	tie$^{\equiv}$ (0.0406)	+0.042	0.72 (tie)	0.89 (tie)
	r=16	entry → B	−0.017	0.58	tie$^{\equiv}$ (0.0088)	−0.032	0.55 (tie)	0.10 (tie)
		B → C	+0.036	0.74	tie$^{\equiv}$ (0.0287)	+0.036	0.76 (tie)	0.27 (tie)
		entry → C	+0.019	0.99	tie$^{\equiv}$ (0.0141)	+0.003	0.85 (tie)	0.80 (tie)
	r=32	entry → B	+0.056	0.28	tie (0.0582)	+0.075	0.23 (tie)	0.79 (tie)
		B → C	−0.004	0.92	tie$^{\equiv}$ (0.0077)	−0.016	0.96 (tie)	0.13 (tie)
		entry → C	+0.052	0.32	tie (0.0706)	+0.058	0.34 (tie)	0.18 (tie)

rank	arm (minus LFA)	at entry: Δ, p		after B: Δ, p		after B, length control: Δ, p	
r=4	LwF+μ α_{LwF}=8	+0.034	0.63 (tie)	−0.036	0.44 (tie)	−0.055	0.21 (tie)
	canonical replay+μ	−0.117	0.14 (tie)	−0.188	0.0006 (LFA)	−0.091	0.12 (tie)
r=16	LwF+μ	+0.147	0.0080 (LwF+μ)	−0.109	0.0496 (LFA)$^{\flat}$	−0.091	0.14 (tie)
	canonical replay+μ	−0.035	0.55 (tie)	−0.186	0.0021 (LFA)	−0.036	0.66 (tie)
r=32	LwF+μ	+0.010	0.79 (tie)	−0.418	3.8×10^{-14} (LFA)	−0.464	1.3×10^{-12} (LFA)
	canonical replay+μ	−0.152	0.032 (LFA)	−0.312	1.4×10^{-7} (LFA)	−0.201	0.0038 (LFA)

falls surviving the length control (p=0.011, p=3×10^{-5}), and does not move at r=4 (−0.026). So the old-domain leads of §6 are not retention margins. Of the five, the two at r=32 survive the length control and three become ties (lower block of Table 21); about half of the r=32 gap to replay was present at entry (−0.152, p=0.03); and at r=16 LwF+μ *entered* ahead of LFA (+0.147, p=0.008), so that lead is a crossing. Through all of it replay rehearses text and LFA holds none: canonical replay's buffer is 12.7 MB (3,346 rows of general and A text) while learning B and 16.4 MB (5,019 rows; general, A and B in equal thirds) while learning C.

Table 22: After the third domain, by adapter rank (seed 42; the r=32 LFA chain's stage three was also run at seed 1337). A-PPL: A's held-out perplexity entering and leaving stage three (replay: exit only). Set-A: A's 154 questions after B and after C, `gpt-5.6-luna@medium`. The paired cells compare the plain LFA chain with replay+μ (general+A+B in equal thirds; μ=0.05/0.2/0.5 at r=4/16/32) and with LwF+μ, as *verdict* (p_{Luna} / p_{Sonnet}), n=154. The verdict word is the primary judge's; in the three cells the robustness judge reads differently its verdict follows "S:". Set-C: C's 78 questions, with the chain's own pre-C model in parentheses; the never-adapted base model scores 1.028, and C's recall tier is unlearnable at these doses (a `claude-sonnet-5` measurement, App. P), so no knowledge conclusion is drawn from its mean. The after-B level is itself a stage-two rise; read against entry: Table 21. No knob was re-tuned for C, and every stage-three arm is single-seed except the r=32 pair. Zero-rates are not tabulated (30–41% on set-A after C).

	A-PPL after C		set-A after C			set-C
rank	LFA (after B→C)	replay	LFA (after B→C)	LFA vs. replay+μ	LFA vs. LwF+μ	LFA (pre-C floor)
r=4	12.23 → 11.55	8.65	1.034 → 0.887	0.887 vs. 0.897, tie (0.82 / S: replay ahead, 0.045)	0.887 vs. 0.857, tie (0.60 / 0.58)	0.982 (1.018)
r=16	12.60 → 11.10	9.31	1.090 → 0.895	0.895 vs. 0.940, tie (0.67 / S: replay ahead, 0.047)	0.895 vs. 0.793, tie (0.071 / S: LFA ahead, 0.022)	1.069 (1.031)
r=32	12.30 → 10.98	9.46	1.227 → 0.965	0.965 vs. 0.912, tie (0.56 / 0.82)	0.965 vs. 0.705, ahead (<10^{-4} / <10^{-4})	1.010 (1.010)
seed 1337	12.30 → 10.83	—	1.227 → 0.974	—	0.974 vs. 0.652, ahead (<10^{-4} / <10^{-4})	1.036 (1.010)

Table 23: The paired verdicts behind Table 20 and §6.1, quoted from the primary judge's verdict file (`gpt-5.6-luna@medium`; two-sided Wilcoxon, α=0.05; n=154 on set-A, 78 on set-B). *Over stage three*: a chain after B against itself after C, with the paired Δ printed for the LFA contrasts the text quotes as magnitudes. *After C*: a cross-method contrast at the end of stage three. "S:" gives `claude-sonnet-5` on the same answers where the text reads the two judges differently. The control is the LFA chain re-run at replay's gradient-accumulation geometry (ga=3). Levels: Tables 20 and 22; "falls" is from after B, and the same levels read against entry are Table 21. Every stage-three arm is single-seed except the r=32 LFA/LwF+μ pair.

	set-A over stage three: p (verdict)			set-A after C: p (verdict)		
rank	LFA (Δ, p)	LwF+μ	replay+μ	LFA vs. replay+μ	LFA vs. LwF+μ	control vs. LFA
r=4	−0.147, 0.0030 (falls)	0.0118 (falls)	0.42 (no move)	0.82 (tie; S: replay ahead, 0.045)	0.60 (tie)	0.94 (tie)
r=16	−0.195, 0.0009 (falls)	0.0004 (falls)	0.74 (no move)	0.67 (tie; S: replay ahead, 0.047)	0.071 (tie; S: LFA ahead, 0.022)	0.37 (tie)
r=32	−0.262, <10^{-4} (falls)	0.0090 (falls)	0.92 (no move)	0.56 (tie)	<10^{-4} (LFA ahead)	0.53 (tie)

	set-B over stage three		set-B after C
rank	LFA (Δ, p)	replay+μ	LFA vs. replay+μ
r=4	−0.46, <10^{-4} (falls)	0.42 (no move)	0.025 (LFA ahead)
r=16	−0.41, 0.0001 (falls)	0.51 (no move)	0.41 (tie)
r=32	−0.42, <10^{-4} (falls)	0.47 (no move)	0.61 (tie)

What moves judged A over stage three is the current stage's question–answer supplement, and its size is not the lever. Table 24 ablates it on the LFA r=4 chain. C's pairs *alone* reproduce the full fall on A from its after-B level (a tie with the as-run stage, p=0.66) at one sixth of the optimizer steps; without the prose they also sink judged B (p=0.0001) and push judged C below both floors (p=0.0022 against the chain's own pre-C model, p=0.0037 against the base model). The papers alone collapse answering form instead, giving the longest answers in the study (median 2,545 characters, mean 2,151, largely looping) and the lowest judged-A *level* among the LFA r=4 after-C arms, itself a paired tie with the as-run stage (p=0.11). A quarter or half of the pairs leaves judged A where the full set does (p=0.44, 0.93) and judged C likewise. The supplement is therefore double-edged: it keeps answers judge-readable and it moves the earlier domain's judged level (shortening its answers, 928 → 586 characters as run), and the move is triggered by its presence, not its dose. Read against

Table 24: The third stage's question–answer supplement, ablated: LFA r=4, one seed, every row from the same post-B model at ga=1. Judged columns are `gpt-5.6-luna@medium` levels on the /5 rubric (n=154 on A, 78 on B and C). The post-B model judges 1.897 on B; C's floors are the chain's own pre-C model (1.018) and the base model (1.028). *Pairs*: screened question–answer pairs mixed into C's corpus (pool limit 1,696), with f their epoch-0 token share. *Steps*: stage-three optimizer steps, fewer on the pairs-only row because that corpus is short and the loader's empty epochs are skipped (App. N). p_A: paired p on set-A against the as-run row (every one a tie). *C-PPL*: C's held-out direct-QA perplexity (pre-C floor 18.18); it scores chat-formatted text, is *not comparable across* f, and supports no C-fit conclusion. *Mean chars*: mean answer length on the set-A set.

stage-three corpus	pairs	steps	set-A	p_A	set-B	set-C	C-PPL	mean chars
papers only (f=0)	0	9,279	0.801	0.11	1.400	1.018	13.88	2,151
f=0.03	411	9,279	0.845	0.44	1.349	1.010	11.74	661
f=0.06	847	9,279	0.887	0.93	1.338	0.977	11.57	607
f=0.11 (as run)	1,696	9,279	0.887	—	1.433	0.982	11.48	586
pairs only, no papers	1,696	1,516	0.927	0.66	1.072	0.769	13.29	420

entry (§6.1), the after-B level was itself a stage-two rise that came with longer answers, so the ablation locates what moves the level and does not measure lost content.

Two remedies do not change it. Raising the prior domains' share of the merged mixture five- and tenfold changes nothing the anchor measures: judged A reads 0.887/0.868/1.001 across shares 0.05/0.27/0.50, all ties; including judged B and C the smallest p in the family is 0.0516; perplexities move within 0.15. Adding rehearsal to LFA fails too: a per-step rehearsal stream of A's own pairs lifts judged A no higher than a step-matched control at any rank (at r=16 and r=32 the control is paired-*ahead* of it, p=0.025 and p=0.045, where the robustness judge reads ties), and it costs the un-rehearsed domain; replay's full buffer on top of LFA restores the perplexities without moving judged A (0.944 at r=4, a tie with the control, p=0.41).

Two instrument facts bind these readings. First, the judge tracks answer length within a model (median Spearman per dimension +0.04–0.18 on A, +0.38–0.56 on B, +0.3–0.45 on C), and a training stage changes answer length, so load-bearing contrasts are re-scored on the two least length-sensitive dimensions (correctness + domain accuracy). Under `claude-sonnet-5` every lead and every fall but one survives that control, the r=16 judged-A fall thinning to p=0.056, and several contrasts run against length: at r=4 LFA leads LwF on judged A 0.873 vs. 0.643 (p<0.001) with answers less than half as long (856 vs. 1,834 characters), and at r=32 the contrast is 0.971 vs. 0.581 at matched lengths (under the primary judge the r=4 pairing is a tie, p=0.27, and the r=32 one a separation). Under the primary judge the control changes more (§6.1, App. D): three of the five old-domain leads after B become ties, and the after-C falls at r=4 and r=16 do not survive it, which is why the main text reads old-domain levels against their entry level. The control bounds the confound; it does not remove length from the rubric. Second, C's judge set is not learnable at these doses (recall tier 0.05–0.23 for every scored model including the base, 0.16; `claude-sonnet-5` figures, whose per-tier decomposition has no primary-judge counterpart: synthesis sits below the base's 1.37 for every trained model and analysis straddles the base's 0.79, 23 of 32 stage-three arms above it by at most 0.32), so judged C measures general answering, and C's perplexity is where learning C shows (−31 to −48% for the standard chains against each chain's own pre-C floor). All 30 base-floor pairings in the primary judge's verdict file — the LFA, LwF+μ, replay, GeRe and TM-only chains, with their controls and ablations — are ties (22) or base-ahead (8), so no chain separates above the base model's 1.028. Over stage three the LwF+μ chains move 1.013 → 1.126, 0.777 → 0.915 and 0.718 → 0.887 on judged C at r=4, r=16 and r=32, and the canonical-replay chains 0.710 → 0.854, 0.687 → 0.872 and 0.656 → 0.887. We report the recall tier separately wherever this set is used.

H GENERATIVE RETENTION, THE WIDER SUITE, AND THE QWEN3-1.7B REPLICATION

This appendix carries the evidence behind Table 2 and the two paragraphs of §6.1 that read it: the rule-scored generative instrument at 0.6B with its paired draw and audit, the with-and-without-L2-SP contrasts on that instrument (Table 32, App. J), the rest of the judge-free benchmark suite on the same models, the token-class diagnostic of the likelihood column, and the Qwen3-1.7B *reduced-tuning replication*: its frame, its judged and generative verdicts, its perplexity ladder and its suite. Every 1.7B contrast is **one seed**, every 1.7B judged number is the **primary judge only** (`claude-sonnet-5` was not run on those arms, so no robustness reading exists), and the 1.7B statistics are **separately declared replication families**, corrected inside themselves and never pooled with the family of App. D. No cell in this appendix is read against a judged number from a different instrument, and no level is read across model sizes.

The instrument. GSM8K (3-shot, exact numeric match) and IFEval (0-shot, prompt-level accuracy), 200 scored items a task, decoded by sampling (temperature 0.7, top-p 0.8, top-k 20, presence penalty 1.5). The seed fixes which items are drawn and the few-shot exemplars, not the sampler, so every run of the same weights is a fresh draw; 300 generations are made for 200 scored and a truncated generation is dropped and replaced, so pairing is by item id over the items both arms scored. Paired tests are the two-sided exact (binomial) McNemar on the discordant pair. No judge is involved and no method was tuned on either benchmark. The r=32 chains are those of §6.1; they differ in gradient accumulation (1 for LFA and LwF+μ, 4 for replay), a confound the step-matched control removes after B only.

0.6B, stage by stage (the earlier draw). Table 25 gives correct-of-200 at entry, after B and after C. Unpaired (Fisher exact on the counts, these runs stored counts only), LFA is ahead of canonical replay+μ on GSM8K at every stage and both seeds ($p\leq$0.0029) and ahead of LwF+μ from the second domain on (p=0.0004 after B, 2.8×10^{-6} after C; a tie at entry, p=0.36); on IFEval it is ahead of LwF+μ from the second domain on ($p\leq$0.0002; tie at entry) and of replay+μ at entry and after B at seed 42 (p=0.0019, 0.0004) and at entry at seed 1337, with two pairings not separated (after C, p=0.056; after B at seed 1337, p=0.13). The two baselines separate only at entry, LwF+μ ahead, and tie afterwards. The step-matched control (LFA's stage two at ga=3) reads 81 on GSM8K, the ga=1 chain's own count, and 104 against 111 on IFEval (p=0.55), keeping both leads ($p\leq$0.0048), so the after-B row is not an optimizer-step artefact; entry and after C keep the confound. Against the base model LFA is not separable at entry at seed 42 and is below it on GSM8K after B at both seeds (p=0.0050, 0.0003) and on both tasks after C.

Table 25: Generative retention on the r=32 chains at 0.6B, correct of 200 (proportion in brackets): the earlier draw, from which Table 2's per-item draw of five states is separate (its counts: base 101/116, LFA after C 65/104, LwF+μ after C 32/60, replay+μ after C 32/79, replay+μ entry 58/67). The base model's clean draw is shown; a first, card-contended draw of the same weights read 106/120. Tests on this table are unpaired.

		GSM8K			IFEval		
chain	seed	entry	after B	after C	entry	after B	after C
base Qwen3-0.6B	—	110	—	—	124	—	—
LFA	42	90	81	70	110	111	99
LwF+μ	42	80	47	29	101	73	60
replay+μ, canonical	42	45	43	24	78	75	79
LFA	1337	81	73	—	112	101	—
replay+μ, canonical	1337	52	39	—	79	85	—
LFA step-matched (ga=3)	42	—	81	—	—	104	—

The paired draw and the audit. Table 2's 0.6B rows are a separate draw of five states scored per item; the four pairings the table does not print are LwF+μ against replay+μ after C, a tie on GSM8K (31 vs. 32 of 198, p=1.0) and replay ahead on IFEval (56 vs. 74 of 185, p=0.0051), and the base model against replay+μ at entry, base ahead on both ($p=8.9\times10^{-7}$, 1.0×10^{-8}; only 168 items

pair because that model truncates 53 of 300 IFEval generations). Every unpaired verdict of Table 25 that the draw re-tests survives, and the one after-C pairing that did not separate unpaired (LFA vs. replay+μ on IFEval) separates paired. The GSM8K gap is not an extraction artefact: scored-wrong generations containing the gold number are level across the five models (13, 17, 18, 16, 14 of 200), and in none of those 78 items is the gold number the last number of the generation. Generation style differs and is an observation only: mean tokens over scored GSM8K items are 173 (base), 155 (LFA), 200 (LwF+μ), 131 and 99 (replay+μ at entry and after C); the replay chain adopts the exemplars' «a+b=c»/#### form (148 of 200 after C) and LFA keeps the base model's prose and `\boxed{}` (119 of 200, none with «). Between the two draws the five states move by up to 13 items (0.065), no repeat pair separating (Fisher $p \geq 0.17$); the paired p-values of Table 2 are uncorrected (ten tests).

The rest of the suite at 0.6B (Table 26). The same fifteen model states were scored on deterministic multiple choice (argmax of length-normalised conditional log-likelihood; identical items across models), preference discrimination, sampled short-answer TriviaQA (500 items, no oversampling) and seven perplexities; Table 26 prints the four after-C states. The suite is *ragged*: TruthfulQA, Winogrande, GPQA-diamond, MMLU-Redux and generative NaturalQuestions were dropped after three models (base, LFA at entry and after B, seed 42) had been scored on them and are not tabled; every column below was scored for every model. Tests are Fisher exact on counts (unpaired, although items are shared) with Benjamini–Hochberg inside two families. *Retention screen*, fourteen trained states against base on nine accuracy columns (m=126): eight survive, five *upward* (LFA on ARC-Easy at entry at both seeds, 0.558/0.569; LFA on BoolQ after B at both seeds, 0.689/0.694; the step-matched control on BoolQ, 0.700) and three downward (replay+μ at seed 1337 on BoolQ, 0.595 at entry and 0.567 after B; LwF+μ after C on generative TriviaQA, 0.104 against 0.178). The BoolQ column is disqualified as a method axis: the two seeds of replay+μ differ on it by 0.055 and 0.079, more than any method effect found there. *Method contrast after C*, LFA against each baseline on the nine columns (m=18): one survives, LFA ahead of LwF+μ on ARC-Easy (0.538 vs. 0.482, q=0.0023); nothing against replay+μ (raw p from 0.11 to 0.95), and replay's MMLU 0.325 against the base model's 0.306 does not separate (p=0.11). So no chain shows multiple-choice forgetting and the suite does not rank the methods; generative TriviaQA ties LFA and replay+μ after C (0.142 vs. 0.152). Where the chains move is general text and instruction perplexity, where LwF+μ alone drifts (2.13× and 1.46× base after C, against LFA 1.03× and replay 0.97× on WikiText-2), and the likelihood columns, which put replay+μ first on every one: the gold-solution perplexity of 3.11 against LFA's 3.61 is a 1.16× ratio against a 3.0% two-seed band, and the four short-answer reference columns order the same way by factors (9.45×, 3.47×, 2.55×, 2.76× for replay below LFA). Those four score a bare four-to-seven-token answer under a chat template, a format the base chat model almost never emits (1,539 on MATH-500, 1,999 on BBH) and fine-tuning teaches quickly, so they reward terse answer-shaped continuation at least as much as retained knowledge; that is the limit of the instrument, not a mechanism. Both readings therefore travel together: the generative instrument licenses "LFA retains more sampled chat-format task performance", and every likelihood instrument puts replay+μ ahead on the same subjects.

Table 26: The judge-free suite on the four after-C states at 0.6B (seed 42, 0-shot). Accuracies: MMLU (3,000), HellaSwag (2,000), ARC-Challenge (1,172), ARC-Easy (2,376), PIQA (1,838), SIQA (1,954), BoolQ (3,270; seed-unstable, not a method axis), preference discrimination (500 pairs), generative TriviaQA (500). Perplexities, lower is better: WikiText-2, instruction text (500 Dolly pairs), the gold GSM8K solution (500 items, 126 tokens each), and reference answers of 4–7 tokens on MATH-500, BBH, TriviaQA and NQ. Every column was scored for every model; the five benchmarks dropped from the suite are not tabled.

after C	MMLU	HSwag	ARC-C	ARC-E	PIQA	SIQA	BoolQ	pref.	TriviaQA	Wiki	instr.	GSM8K	MATH	BBH	TQA	NQ
base Qwen3-0.6B	0.306	0.409	0.342	0.504	0.668	0.405	0.649	0.568	0.178	18.18	6.74	4.63	1539	1999	900	378
LFA	0.306	0.441	0.351	0.538	0.657	0.421	0.641	0.540	0.142	18.76	6.45	3.61	93	240	274	96
LwF+μ	0.302	0.423	0.312	0.482	0.646	0.408	0.624	0.536	0.104	38.72	9.83	4.76	206	340	621	295
replay+μ	0.325	0.431	0.326	0.537	0.665	0.418	0.643	0.520	0.152	17.71	6.01	3.11	10	69	108	35

What the gold-solution perplexity is made of (Table 27). Re-scoring the same 63,258 tokens with every token assigned to a class by GSM8K's calculator format, the classes are a property of the gold text: prose is 50.6% of the tokens and 62–72% of the total negative log-likelihood in every run, markup 13–20%, and the computed *result* inside «. . .=X» between 0.18% and 0.57%. Replay's lead

on the aggregate is a lead on prose and on the number copied after ». On the result tokens the order is the generative one, base 1.047 < LFA 1.072 < replay+μ 1.116 < LwF+μ 1.129, every adjacent pair paired-separated over the 495 items carrying such tokens (Wilcoxon on per-item mean NLL; LFA vs. replay+μ p=2.2×10^{-38}, LFA lower on 439 of 495), and on prose and copy tokens the same test runs the other way as strongly (p=3.1×10^{-83}, replay lower on 489 of 500). The result tokens are near ceiling (top-1 accuracy 96.9–98.6%; a 7.8% band on a quantity floored at 1), so the separation is a consistent one, not a large one. This is a diagnostic of one instrument on one dataset with one decomposition and it licenses no generalisation: the likelihood and generative instruments are not in contradiction about the arithmetic, they disagree about the prose and format it is written in.

Table 27: Per-class perplexity of the gold GSM8K solution after C at 0.6B (seed 42; teacher-forced, no sampling). *overall* reproduces the suite's column; *result* is the computed value inside the calculator markup, *copy* the number repeated after it, *prose* everything that is not a number or markup.

after C	overall	result	copy	prose
base Qwen3-0.6B	4.626	**1.047**	3.914	8.822
LFA	3.613	1.072	2.524	5.680
LwF+μ	4.757	1.129	4.517	7.921
replay+μ	3.109	1.116	**1.983**	**4.039**

The Qwen3-1.7B replication: frame. Domains A, B and C are those of §4, at LoRA r=32, α=64, 20 epochs a stage, learning rate 3×10^{-4}, batch 6, sequence 512, warmup 50, frozen embeddings, seed 42, stage one fused at dose 15. $p(h)$ was fitted afresh on the same proxy corpus (1,543,040 samples a site, K=32; 249,412,335 bytes in int8, 6.14% of the `bf16` checkpoint and 10.7% above what linearity in $\sum d^2$ from the 0.6B artifact predicts, the excess a higher 95%-variance rank per site). Knobs were re-tuned on a single-domain perplexity ladder at this size (11 arms, no judge): the 0.6B λ=10^5 sits on the underfitting side, and λ=5×10^4 was chosen; LwF α_{LwF}=4; replay n=1; μ=0.05 in every chain at stage one and in the LFA and LwF+μ chains at stage two (the 0.6B LwF+μ chain runs μ=0.1). Replay's stage two was re-tuned once: the inherited μ=0.05 let its domain-B perplexity rise ($8.48\rightarrow16.09$), and μ=0.5 was chosen on perplexity alone before any judging (ΔB -2.56); the re-tune moved no LFA-against-replay verdict. Nothing was re-tuned for C. Gradient accumulation is 1/1/4 and no step-matched arm exists at this size; the 0.6B feasibility gate was not re-run. Judged verdicts are the paired two-sided Wilcoxon of §4 on the five-dimension mean and on the length control, under the primary judge only; the TOST margin ±0.147 is *carried* from the 0.6B archive at n=154 and applied post hoc, including to the n=78 sets. Four families are corrected separately, as run: A→B judged (24 contrasts, each axis), generative after B (m=24), generative after C (m=12) and judged after C (30 contrasts, each axis).

Judged, after B and after C (Table 28). After B LFA is ahead of both baselines on set-A on both axes and of replay+μ on set-B's five-dimension mean, ties LwF+μ on set-B, and its own judged A *rises* from entry to after B ($1.322\rightarrow1.547$, $+0.225$, p=2.1×10^{-4}) where both baselines are flat or fall. After C it is ahead of LwF+μ on all three sets and both axes, all six surviving correction; ties replay+μ on sets A and B on both axes (no TOST shows equivalence, so "not shown different"); and is ahead of it on set-C on both axes. Replay+μ is ahead of LwF+μ on set-A on both axes and on set-B's length control, and the two tie on set-C, so no three-method ranking closes at this size. Against its own entry model after three stages LFA is a tie on set-A on both axes, equivalent within the carried margin on the five-dimension mean (-0.021, TOST p=0.019), as replay+μ is; LwF+μ's entry model is ahead of its after-C model on both axes (q=0.022, 0.010). Measured from after B rather than from entry, LFA loses on set-A (-0.245, q=0.0024 on the five-dimension mean) and on set-B's five-dimension mean (-0.228, q=0.018): the entry-level tie is the net of a rise and a fall, the 1.7B counterpart of §6.1. The floors, on the five-dimension axis, one seed, one judge: after C LFA is the only method ahead of the never-adapted base model on the new domain's set-C ($+0.290$, q=0.0084; LwF+μ and replay+μ tie it), on set-B it ties the base model on both axes ($+0.067$, p=0.50) while LwF+μ falls below it (length control q=0.020; five-dimension q=0.061, not surviving), and against each chain's *own* pre-C model on set-C not one of the six contrasts survives correction, so nothing here says any chain learned C under the judge; the perplexity evidence that all three did (below) has no judged counterpart. Verdict against verdict on the five-dimension mean, the 0.6B r=32 set-A after-C verdicts replicate exactly (LFA ahead of LwF+μ; LFA and replay+μ a tie; replay+μ ahead

of LwF+μ); LFA ahead of LwF+μ on set-B and of replay+μ on set-C are new at 1.7B, both ties at 0.6B.

Table 28: **Replication, Qwen3-1.7B** (seed 42, `gpt-5.6-luna@medium` only, n=154 on set-A and 78 on sets B and C). Judged levels are five-dimension means with the zero-rate in brackets; *after C* levels for set-A: LFA 1.301 (19.0%), LwF+μ 1.095 (22.2%), replay+μ 1.214 (21.9%), base 1.081 (18.8%); set-B: 1.954 (19.2%), 1.654 (19.0%), 1.838 (18.2%), base 1.887 (15.4%); set-C: 1.456 (30.3%), 1.174 (28.5%), 1.133 (36.4%), base 1.167 (25.6%). Cells: paired Δ (second arm minus first) and Wilcoxon p on the five-dimension mean, the verdict, then the length-control p (verdict). Levels are never read against 0.6B levels; the TOST margin is carried from 0.6B.

contrast	set	after B: Δ, p (verdict)	after B, LC: p	after C: Δ, p (verdict); LC: p
LFA vs. LwF+μ	A	-0.378, 5.4×10^{-9} (LFA)	4.5×10^{-8} (LFA)	-0.206, 4.3×10^{-4} (LFA); 6.8×10^{-6} (LFA)
	B	-0.195, 0.11 (tie)	0.21 (tie)	-0.300, 0.0011 (LFA); 1.6×10^{-4} (LFA)
	C	—	—	-0.282, 0.0040 (LFA); 7.7×10^{-4} (LFA)
LFA vs. replay+μ	A	-0.378, 2.5×10^{-7} (LFA)	4.3×10^{-4} (LFA)	-0.087, 0.15 (tie); 0.57 (tie)
	B	-0.303, 0.0024 (LFA)	0.76 (tie)	-0.115, 0.23 (tie); 0.65 (tie)
	C	—	—	-0.323, 0.0053 (LFA); 0.014 (LFA)
LwF+μ vs. replay+μ	A	$+0.000$, 0.85 (tie$^{\equiv}$)	0.015 (replay)	$+0.119$, 0.018 (replay); 2.0×10^{-4} (replay)
	B	-0.108, 0.32 (tie)	0.23 (tie)	$+0.185$, 0.14 (tie); 0.0028 (replay)
	C	—	—	-0.041, 0.80 (tie); 0.53 (tie)

Generative, at 1.7B (Table 29). The instrument above, scored per item. After B LFA is ahead of replay+μ on both tasks (p=1.2×10^{-6}, 1.4×10^{-5}) and of LwF+μ on both (p=0.015, 0.014; q=0.023, 0.022), a separation on this draw that a second draw reproduces and that is not quoted as a size; at entry it is ahead of replay+μ and ties LwF+μ and the base model. Its own cost against the base model after B is a tie on IFEval (p=0.69; 122 vs. 122 on the second draw) and, on GSM8K, does not survive correction on this draw (p=0.044, q=0.058) but separates on the second (p=4.2×10^{-5}), so "level with the base model" covers IFEval only. The two baselines' after-B ordering (LwF+μ ahead, q=0.014 and 0.022) becomes a tie on both tasks on the second draw, GSM8K with the sign reversed, so no baseline-against-baseline ordering is quoted at this stage. After C the verdicts are those of Table 2, all nine separations surviving inside their family of twelve and all twelve reproducing on the second draw with the same nine surviving; the within-chain step from after B to after C, uncorrected, is a loss on both tasks for LwF+μ (p=3.1×10^{-4}, 0.013), a loss on GSM8K only for LFA (p=0.029; IFEval p=0.56) and flat for replay+μ (p=1.0 on both). Replay's re-tuned and inherited stage-two models tie each other on both tasks (p=0.30, 0.60), so the μ that settles its domain-B perplexity does not move this axis. **Run-to-run spread, measured here.** A second draw of seven states (base, and the three chains after B and after C) moves a count by up to 17 items (0.085; median 3.5), no repeat pair separating unpaired, but one pair separates *paired*: LwF+μ after B on GSM8K, 120 against 102 of 196 shared items (b=25, c=7, p=0.0021; q=0.029 over the fourteen repeat tests). Between 4.6% and 28.6% of items flip between two draws of one checkpoint, so a paired q of about 0.02 is not by itself a demonstration that two arms differ, and no gap of 17 items or fewer is read as a size. The extractor-miss screen is not level across arms here (10 to 27 of 200, the higher-miss arms the lower-scoring ones) and the transcript audit of the 0.6B draw was not run at this size, so the artefact question it settles at 0.6B is open at 1.7B.

Table 29: **Replication, Qwen3-1.7B**: generative retention, correct of 200 (seed 42, one draw a cell, the draw whose paired tests are quoted; the second draw's count follows in brackets where one exists). Replay is the re-tuned stage-two chain; its inherited variant reads 88/99 after B.

	GSM8K			IFEval		
	entry	after B	after C	entry	after B	after C
base Qwen3-1.7B		152 (162)			136 (135)	
LFA	144	137 (136)	121 (123)	141	130 (127)	123 (115)
LwF+μ	137	120 (103)	102 (96)	131	111 (102)	89 (92)
replay+μ	105	98 (110)	98 (94)	100	95 (93)	95 (95)

Perplexity at 1.7B. The A leg reads LFA $9.04 \to 13.84 \to 9.15$ (a $+4.79$ step on B and -4.69 on C, ending $+0.11$ against entry), LwF+μ $12.42 \to 12.47 \to 13.21$ and replay+μ $5.55 \to 6.02 \to 6.36$ (each $+0.80$ cumulative), so the rise-then-return shape of §6.1 has a perplexity counterpart here in which only the anchored chain's A returns; its judged set-A falls from after B to after C while its perplexity on A returns, and the two are read side by side. During C, B's perplexity falls for LFA ($8.25 \to 7.01$) and replay+μ ($5.91 \to 4.31$) and rises for LwF+μ ($5.67 \to 7.46$). C is learned by every chain on perplexity (against each chain's own pre-C model, -15.11, -11.26, -2.52; the never-adapted base reads 91.77), the pre-C column being a per-chain floor and not a ranking (replay's chain, rehearsing chat-formatted text throughout, reads 10.22 on C before any C gradient). General text after C: LFA 11.38 ($0.76\times$ base), replay+μ 10.81 ($0.72\times$), LwF+μ 22.54 ($1.51\times$), the 0.6B direction repeated more mildly. Domain perplexity is never read across f or across sizes.

The suite at 1.7B: format adaptation, not retention. The seven states at entry and after B were scored on the trimmed suite (not ragged: every model, every column). Every trained state is *above* the base model on ARC-Easy by 8 to 16 points, all six surviving correction, and five of six on HellaSwag (LFA after B, $+2.2$, does not separate, p=0.18), replay most; the instruction-tuned base answers these formats below its capacity and any fine-tune on this corpus lifts it, so no multiple-choice rise is read as knowledge gained and no fall as forgetting. Replay's MMLU rise (0.368 against 0.339) is a raw-p statement only (p=0.022, q=0.062), as its 0.6B rise was; the BoolQ column is disqualified by the 0.6B seed spread and this table has one seed. The likelihood columns again put replay+μ first (WikiText-2 after B 10.44, LFA 12.10, base 14.96, LwF+μ 16.74; gold-solution perplexity 2.40, 5.23 for LwF+μ, 6.97, 11.52, a three-way reversal of the generative order); the four short-answer reference columns are degenerate at this size (the base chat model reads 1.15×10^6 on MATH-500) and no ratio on them is quoted. Generative TriviaQA separates nobody: LFA and replay+μ after B both read 0.308 (154 of 500), against base 0.292 and LwF+μ 0.268, which is the column that forbids "every behaviour instrument favours LFA".

I AN OVERHEAD LOWER AT 1.7B AND 3B THAN AT 0.6B: THREE POINTS

LFA's overhead is a fixed per-site cost under a step that grows with the model, so it should fall with scale; we have three points, from two model families: $+14.8\%$ a pass at 0.6B (r=4; $+14.0\%$ and $+14.6\%$ in two further frames), $+6.3\%$ at Qwen3-1.7B and $+5.5\%$ at the one 3B point (timing-only artifact), where LwF+μ runs out of memory at micro-batch 8 and LFA fits 12. On the like-for-like method-footprint rows the series is $+14.6\%$, $+6.3\%$, $+5.4\%$, monotone and falling. Replay's and LwF's overheads are roughly fixed fractions of a step (LwF 32.9%, 28.3% and 28.4% on the like-for-like rows; replay one buffer micro-batch in four, $+31.6\%$, timed at 0.6B only). The cost is launch-bound by measurement (0.51 TFLOP/s against a 21.6–62.2 roofline; $512\times$ the samples cost $1.26\times$ the time). An arithmetic projection to 7 B gives $\approx 0.9\%$, probably low, since the same form under-predicts the 3B point (2.8% against 5.5%).

Against each method's own control. Table 30 carries time and memory per pass at r=4. The overhead is stable across frames: $+14.0\%$ over content-only training at the confirmed operating point (r=32, n=16, batch 6×512 tokens; 0.7010 vs. 0.6147 s/step, built up as $+9.2\%$ to evaluate the anchor, a further $+3.3\%$ to backpropagate it and $+1.1\%$ for L2-SP: successive ratios of four timed arms, not additive shares), $+14.6\%$ in a second campaign (0.7299 vs. 0.6367) and $+14.8\%$ at the continual frame's r=4 (0.6899 vs. 0.6011): three measurements within 0.8 percentage points, with absolute step times differing by about 4%. All are seconds per optimizer *step* on one uncontended RTX 3090, since the loader re-chunks the corpus each epoch and a per-epoch figure is comparable only when the pass is matched. *The int8 artifact times identically to the unquantized one*: in one session the r=4 arm runs at 0.6894 s/step on the chain's 114 MB int8 artifact and 0.6896 on its unquantized one (-0.02%; peak VRAM 9.54 GB for both; content-only 0.5989), the artifact being dequantized once at load. The unquantized *file* is 543 MB because it also carries a 311 MB layer-0 embedding lookup, which is rebuilt from the teacher and never shipped, beside the $\approx$226 MB payload and A's merged components.

Reading the cost scorecard (Table 30). Among the *timed* methods that preserve, LFA has the lowest time per pass and, of the two text-free ones, the lower memory; it is best on no other axis. Each overhead is measured against the method's own control, so the percentages share no denominator

Table 30: **Cost and judged standing in the continual $r{=}4$ frame.** *One pass* is one epoch over the domain corpus and its question–answer supplement, with each method's extra work inside it (buffer micro-batches, teacher forwards, anchor sampling). It is measured on epoch 1, the same $2{,}081$ chunks for every arm under our loader (the interval straddling the epoch boundary is excluded), on one pinned, uncontended RTX 3090 with 20 warm-up then 200 timed steps. Epoch counts are not compared. *Overhead* is against each method's **own control**: content-only at $\mathrm{ga}{=}1$ for the text-free arms, and for replay the identical launcher at $\mathrm{ga}{=}4$ with only the buffer removed. Canonical replay's three cost cells were measured on its stage-two $\mu{=}0.2$ sibling, which differs from it in μ alone; its verdicts are the canonical $\mu{=}0.05$ arm's. †GeRe and O-LoRA run on their authors' trainers at their *own footing* (2,312 domain rows rather than $2{,}081$, GeRe's further $1{,}000$ buffer rows per pass, no gradient checkpointing), so their percentages are not like-for-like and are not differenced against the rest. VRAM is each method's *own* footprint with no redundant teacher copy (under LoRA the student's frozen base is the teacher; App. L), from a second campaign that reproduces the step times within 0.9%. LFA's int8 and unquantized artifacts time identically. Verdicts: primary judge, paired two-sided Wilcoxon ($n{=}154$ on set-A, 78 on set-B), quoted from the verdict files; the $\mu{=}0$ ablation's are those of App. J. Bold flags values the text discusses, not column optima.

Method	A-data at adaptation	one pass (s)	overhead	peak VRAM (GB)	set-A vs. LFA	set-B vs. LFA
LFA stage-2 $\lambda{=}4{\times}10^4$	none; a 113 MB int8 statistic	$\mathbf{239.4}$	$\mathbf{+14.8\%}$	$\mathbf{9.54}$	—	—
LwF+μ stage-2 $\alpha{=}8$	none; the student's own frozen base	277.1	$+32.8\%$	$\mathbf{15.61}$	tie ($p{=}0.4355$)	LFA ahead ($p{<}10^{-4}$)
Replay MIXED+μ (canonical)	12.7 MB (3,346 rows)	276.5	$+31.6\%$	8.75	LFA ahead ($p{=}0.0006$)	LFA ahead ($p{<}10^{-4}$)
abl.: A-only buffer, $\mu{=}0$	7.0 MB (1,673 rows)	244.3	$+17.3\%$	8.55	LFA ahead ($p{=}0.0128$)	LFA ahead ($p{=}7.4{\times}10^{-9}$)
GeRe `dy` (authors')†	none of A; 2.6 MB generic buffer + 1.69 GB activation cache	283.0	$+35.7\%$	10.30	LFA ahead ($p{=}2.7{\times}10^{-5}$)	LFA ahead ($p{=}2.2{\times}10^{-5}$)
O-LoRA (authors')†	none; prior adapter 10.2 MB	244.1	$+17.0\%$	14.21	LFA ahead ($p{<}10^{-4}$)	LFA ahead ($p{=}0.0003$)
L2-SP	none	211.4	$+1.3\%$	8.75	no primary-judge verdict	not run
Content-only (unanchored)	none	208.6	—	8.67	no primary-judge verdict	not run

and are not differenced. The choice of control barely moves them: with ga held at 4 and the buffer removed a pass costs 208.2 s against the ga=1 control's 208.6 s, a tie, and on the ga=1 denominator the two replay rows would read +17.1% and +32.6% in place of +17.3% and +31.6%. Canonical replay is lower than LFA on peak memory and on general-text perplexity, and reads 12.7 MB of A and general text where LFA reads no A text. The two replay rows are different buffers and different μ, not a ladder, and every figure is one model size, rank, card and implementation; why the costs fall where they do, and what is scale-free, follows.

One larger size. On `Qwen2.5-3B-Instruct` (36 layers, d=2048, 3.09 B parameters) at r=32, sequence 512 and a matched effective batch of 8 on one 24 GB card (452 optimizer steps a pass for every arm), LFA costs +5.5% over content-only (2.3940 vs. 2.2688 s/step) and LwF+μ +28.2% (2.9078). The fit boundary is a result of its own: probed at micro-batch 8, 12 and 16, LwF+μ raises `torch.OutOfMemoryError` at each and fits only at 4, where LFA fits 12 and content-only 16 (the largest probed), so LwF reaches the effective batch only through ga=2 and pays the throughput. Every floor at micro-batch 1 sits far inside the card (8.17, 13.05, 11.30 GB; peaks as run 16.54, 17.69, 17.13 GB), so the difference is activation growth: the teacher's vocabulary logits, a term independent of d and of who holds the weights. No offloaded number was constructed for the arm that did not fit, no 3B model was scored on any domain, and the $p(h)$ used is *timing-only* (20,000 samples a site against the recipe's 1,543,040; per-step cost is set by n and the site list). The overhead falls with size, but not as fast as the simple form derived below says: carried from 0.6B, $1/(d^2T)$ predicts about 2.8% where 5.5% was measured, and two points that also differ in T (effective batch 8 against 6) do not establish an exponent. Carrying the same n and T to 7 B (d=4096) puts the +14.0% near 0.9% (1.1% at the d=3584 of Qwen2.5-7B): an arithmetic projection, not a measurement, and probably a low one, since the same form under-predicts the 3B point. LwF's overhead is one teacher forward per micro-batch and so a roughly fixed fraction of the step (32.8% at 0.6B, 28.2% at 3B).

A second size, and the first scale point where all three share a micro-batch. On Qwen3-1.7B (28 layers, d=2048, 2.03 B parameters) at r=32, sequence 512, one RTX 3090 at a 290 W power cap, 20 warm-up then 100 timed steps, every arm at its method footprint (no teacher copy) and the same matched effective batch of 8, all three methods fit micro-batch 8 and take the same 452 optimizer steps a pass: content-only 1.4480 s/step (654.5 s a pass, 13.98 GB peak), LFA 1.5391 (695.7 s, 15.41 GB, +6.3%) and LwF+μ 1.8581 (839.9 s, 23.51 GB, +28.3%). Probed at micro-batch 8, 12 and 16, content-only and LFA fit all three (LFA peaking at 20.75 GB at 16) and LwF+μ raises `torch.OutOfMemoryError` at 12 and 16, so the boundary result of the 3B point reproduces one size down. Because the three arms share a micro-batch, this is the first *scale point* at which the memory comparison is confound-free (the 0.6B r=4 frame of Table 30 already shared micro-batch 6; at 3B LwF was forced to 4): the anchor costs +1.44 GB over content-only on `nvidia-smi` (+1.61 GB allocated) against LwF's +9.53 GB (+10.84 allocated), a factor of about 6.6 (deltas from byte counts, not rounded cells), and at a matched micro-batch of 8 the anchor's memory cost does not scale with the model (1.44 GB at 1.7B, 1.15 GB at 3B). The artifact this campaign read is the shipped int8 statistic of the 1.7B chains of App. H, 249,412,335 bytes; no training run read the int8 file itself (the chains trained on its `fp16` twin and their own merges). The 1.7B-to-3B pair holds d and T fixed and would predict a flat overhead; it fell by 0.9 points on one run per arm, recorded and not read as evidence about the exponent. Replay was not timed at this size; one seed, one session, one run per arm, and no timing artifact records power at any size, so absolute step times are never read across campaigns.

The arithmetic floor, and a launch-bound implementation. The anchor evaluates the anchored sub-modules on n sampled vectors while the content loss evaluates the whole model on T tokens, and the shared parameter count cancels, so its arithmetic is $\Theta(n/T)$ of a step *independently of model size*. Here n/T=16/3072=0.52%: the anchor's forward pass over teacher and student is 28.2 GFLOP against the content step's 11.0 TFLOP, 0.26% of the step's arithmetic (from the model's shape: 440 M anchored projection parameters, two FLOPs per parameter per sample). The distance from 0.26% of the arithmetic to +9.2% of the time is launch overhead, and it is measured. On a quiet RTX 3090 the anchor's forward pass (85 sampler calls over 28 layers) takes 55.6, 54.5, 56.9 and 70.1 ms at n=1, 16, 64 and 512: 512× more samples cost 1.26× the time. At n=16 it achieves 0.51 TFLOP/s against a dense-matmul roofline, measured in the same process on the same card, of 21.6 (fp32) and 62.2 (`bf16`) TFLOP/s: each product is $[16, 1024] \times [1024, \cdot]$, far too small to occupy the device.

Fixed per-site work yields to engineering. Caching the sampler's immutable per-site tensors on the device takes sampling from 46.7 to 10.9 ms and the whole anchor forward from 94.2 to 55.5 ms, and computing the L2-SP penalty from the LoRA factors, $\|BA\|_F^2 = \mathrm{tr}(B^\top B\, AA^\top)$, in place of the materialised weight difference takes that term from 39.2 to 4.1 ms (each timed before and after in one process). The fixed overhead grows with depth alone while the content step grows with depth $\times$ width$^2\times$ tokens, so the relative overhead should fall as $\approx 1/(d^2T)$.

Why it is cheap. Each step draws n states per anchored site and pushes them through the frozen and the adapted projections and MLP: work $O(n \times$ anchored parameters$)$, independent of sequence length and corpus size, that never enters attention (the o_proj anchor *samples* the concatenated head outputs). *No teacher copy exists under LoRA*: W_{base} is frozen, so the student produces $f_{\text{t}}(h)$ by running its own sub-module with the adapter disabled, exactly and not approximately, and equally so for LwF (verified bit-identically, App. L); the memory figures of Table 30 and of the 3B point above are each method's own footprint measured that way. A teacher proportional to the model (1.11 GB at 0.6B, 5.75 GB at 3B, about 14 GB at 7 B in `bf16`) is therefore not a cost of the LoRA case; what remains is the anchor's activations, 0.9 GB at 0.6B and 1.15 GB at 3B. The 6.1 GB gap to LwF+μ is LwF's full teacher forward, which materialises logits over a 152k-token vocabulary for every sequence at every step, where the anchor evaluates sub-modules on 16 sampled vectors. Rehearsal's cost is a share of the gradient budget (25–75% of micro-batches) spent on full passes over real sequences, nominally +33% to +300%. The upper end holds almost exactly (three micro-batches in four: +303.2% a pass); the lower end is loose, +17.3% (A-only buffer) to +31.6% (mixed) at one in four, because a buffer row is shorter than a domain row (133 and 213 mean tokens against 431). The *share* sets the cost, not the buffer's size: tripling the buffer to 5,019 rows at the third stage leaves it at +31.9%. Under *full-weight* training a teacher is needed; precomputing its outputs on a *fixed* anchor set is possible, but the shipped code samples fresh anchors every step and whether fixed anchors preserve the effect is unmeasured.

The one footprint that grows with the model is the artifact, measured at three sizes with one identical timing-only fit: 96.8, 271.4 and 577.5 MB `int8` at 0.6B, 3B and 7B. A 20,000-sample fit understates a recipe-scale one by a measured 15% at 0.6B (112.5 MB shipped) and by more at the larger widths, so the two larger figures are floors. Size grows *linearly* in $\sum_{\text{sites}} d^2$ (fitted exponent 0.984), because the 95%-variance rank ratio k/d stays flat (0.50, 0.56, 0.50, d-weighted). The one recipe-scale artifact above 0.6B, the 1.7B one (1,543,040 samples a site; $\sum d^2$ exactly 2.000$\times$ the 0.6B recipe artifact's), is 249,412,335 bytes int8, 10.7% above the linear prediction and 11.9% above the fitted exponent's, the whole excess a higher d-weighted k/d (0.662 against 0.590); it is 6.14% of its `bf16` checkpoint against 9.46% at 0.6B, one point from a different family than the fitted series, so no exponent is re-fitted. Extrapolated, with nothing built at that size, a 70B artifact is 8–13 GB `int8`, 6–10% of its `bf16` checkpoint.

J WHAT L2-SP ADDS TO OTHER METHODS

L2-SP (Li et al., 2018), the weight pull whose strength is μ, is not ours, and it complements other preservation signals. Added to replay in one domain it lifts judged quality at matched dose under both judges (+0.122, p=0.033 primary; p=0.034 robustness). Over a second domain it cuts replay's old-domain perplexity cost from +3.29 to +1.07 up its stage-two ladder, while its judged effect there is a tie (p=0.23). It lifts LwF when α_{LwF} is re-tuned with it (+0.093, p=0.041), and self-generated rehearsal is infeasible without it (general-text +30.8% at μ=0, Table 33). It does not preserve alone (§3), and it leaves the authors' GeRe unchanged in both frames (ties, p=0.066 and 0.72). Because LFA carries L2-SP, five baselines carry it too, and the main text compares against those strengthened versions, naming an arm without it where it quotes one; Table 31 gives each baseline with and without L2-SP, and LFA's verdict against the arm without it.

L2-SP is a penalty on the adapter's own weights: it composes with any content loss and touches nothing that makes a baseline the method it is. Five baselines received it: replay, LwF, GeRe, TM-only and self-generated rehearsal (EWC, VeFA and O-LoRA were not run with it). We swept it wherever we ran a ladder, re-tuned per protocol: single-domain replay $\mu\in\{0, 0.01, 0.05, 0.1, 0.2\}$; the continual replay chain $\mu\in\{0, 0.05, 0.1, 0.2\}$; LwF $\mu\in\{0.05, 0.1, 0.2\}$ across its α_{LwF} ladder, of which the single-domain search used the α_{LwF}=4 corner; self-generated rehearsal $\mu\in\{0, 0.05, 0.1, 0.2\}$

(App. K). The authors' GeRe and TM-only received the single value μ=0.1. For LFA itself the judged optimum in μ is interior (§A.5), and L2-SP acts by uniform shrinkage and not by steering direction (App. C).

Table 31: Each baseline at μ=0 and at its μ-tuned best. Judge: primary-judge level (`gpt-5.6-luna@medium`), zero-rate in % in brackets where recorded; run-level cells, within 0.003 of the paired rendering. "μ effect": paired p of the μ arm against the μ=0 arm at matched dose, primary then robustness judge, with the verdict they share. "LFA vs." and "Sonnet": LFA's paired p and verdict against the μ=0 arm under the primary judge and under `claude-sonnet-5` on the same answers (two-sided Wilcoxon, α=0.05). †Re-tunes α_{LwF} alongside μ (16→4): two knobs against one, not a μ ablation. *Verdict changes between the judges; neither reverses. Frame: a recipe row differs from a baseline row on rank, LoRA α and ga, and the GeRe rows also on epochs, warmup, `freeze_embed` and the released trainer's ≈3× supplement exposure; continually ga is 1 for LFA, 4 for replay and 6 for GeRe.

Standard (μ=0)	Judge	With μ, re-tuned	Judge	μ effect	LFA vs.	Sonnet
Single domain (f=0.13), n=154; LFA = the recipe at dose 20, judging 1.042						
Replay n=3, ep 20	0.815 (35.3)	replay+μ=0.2, ep 20	0.933	0.033/0.034 (μ ahead)	0.0004 (LFA)	1.3×10^{-4} (LFA)
Replay n=2, ep 5	0.847 (33.9)	*no frame-clean μ arm at this n*			0.0007 (LFA)	2.7×10^{-7} (LFA)
LwF α=16, ep 15†	1.008 (24.5)	LwF α=4+μ=0.05, ep 15	1.102	$0.041/7.7\times10^{-5}$ (μ ahead)	0.45 (tie)*	0.0011 (LFA)*
LwF α=4, ep 5	0.925 (31.2)	*μ counterpart: the row above, at dose 15*			0.047 (LFA)	1.9×10^{-5} (LFA)
GeRe dy, ep 20	0.878 (37.1)	GeRe dy+μ=0.1, ep 20	0.771	0.066/0.42 (tie)	0.0013 (LFA)	1.3×10^{-7} (LFA)
GeRe w=1, ep 5	0.953 (30.0)	*no w=1+μ arm was run*			0.16 (tie)	0.21 (tie)
Continual A→B, r=4, set-A (n=154); LFA = tuned chain, $\lambda=4\times10^4$, 1.034						
Replay r=4, A-only	0.883 (30.9)	replay+μ_{S2}=0.1	0.977	0.23/0.10 (tie)	0.0128 (LFA)*	0.10 (tie)*
its perplexity up the stage-two ladder	μ_{S2}=0/0.05/0.1/0.2: ΔA +3.29/+1.89/+1.46/+1.07; WikiText-2 20.3/18.9/18.5/18.2; set-A 0.883/0.917/0.977/0.860					
GeRe dy, r=4	0.808 (37.8)	GeRe dy+μ=0.1	0.783	0.72/0.25 (tie)	2.7×10^{-5} (LFA)	8.3×10^{-9} (LFA)
LwF, any rank	*no μ-free LwF continual chain was run*: every such chain on file carries $\mu\geq$0.05					

What L2-SP did to each baseline. In one domain L2-SP lifts replay at matched dose (paired +0.122) and lifts LwF when α_{LwF} is re-tuned with it (+0.093); on the authors' GeRe it is a tie (−0.109). In the continual frame it is a tie for replay (+0.093) and for GeRe on both judged axes, while replay's perplexity responds strongly: up its stage-two ladder μ=0/0.05/0.1/0.2, ΔA falls +3.29/+1.89/+1.46/+1.07 and WikiText-2 20.3/18.9/18.5/18.2 as set-A reads 0.883/0.917/0.977/0.860 (Table 17). The other available draws agree in verdict: replay's μ=0 dose-20 rung against the μ=0.2 `final` draw (p=0.049/0.036, μ arm ahead), and GeRe dy against its μ arm at dose 15 (p=0.71/0.80, tie). Replay's mixed-buffer continual chain is left out of the table because its μ pair does not share a stage one and so is not a stage-two μ ladder; that contrast is also a tie under both judges (p=0.89/0.93).

LFA against the μ=0 baselines. In one domain the recipe at its best feasible dose is ahead of four of the six μ=0 arms and ties the other two, single-knob LwF at α=16 and GeRe's unstable w=1 peak. At the *shipped* dose 15 it ties five of the six ($p\geq$0.065), the exception being the μ=0 replay n=3 rung (p=0.022), and only the rank-matched r=16 arm at its judge-maximum is ahead of all six (p between 0.016 and 6.9×10^{-7}). None of this is a judged ranking: the tuning budgets are unmatched (55 LFA configurations against nine for replay and fewer for LwF), whichever configuration of a baseline is scored. In the continual frame the tuned r=4 chain is ahead of every μ=0 comparator on both judged axes under the primary judge, and under the robustness judge on all but its set-A pairing with A-only replay, which that judge reads as a tie (p=0.10). Against replay's mixed-buffer chain at μ=0 it is ahead on both axes under both judges ($p\leq5.3\times10^{-5}$). The strict rank-matched $\lambda=2\times10^4$ arm is ahead of GeRe on both axes (p=0.0042 on set-A, 1.2×10^{-4} on set-B) and of A-only replay on set-B ($p=5.2\times10^{-6}$), and on set-A it stays a **tie** with that replay chain (p=0.22; robustness judge p=0.71). This is the matched-rank tie band of finding (i) of App. F: it survives the removal of L2-SP and is written as a tie. One cell is empty: **no μ-free LwF continual chain was run**, so LwF's μ=0 comparison rests on the single-domain rows and the continual LwF comparison of record is the μ-tuned one of Table 18.

On the generative instrument, L2-SP helps both baselines (Table 32). On the judge-free instrument of App. H, six of six with-and-without-L2-SP contrasts separate in μ's favour and none against: in one domain at r=16, LwF at fixed α_{LwF}=4 (a μ ablation with one knob differing) and replay at fixed buffer, and continually at r=32 the L2-SP-free ported replay chain against the canonical replay+μ chain at entry and after B. So the matched baselines of Table 2 are the *stronger* versions on

the axis LFA leads them on, the conservative direction. Two limits: the r=32 rows compare across draws (spread up to 13 items; the entry pairing survives against the canonical entry model's second draw too, p=1.9×10^{-5} and 0.0069), and the after-B pair differs in stage-two buffer composition as well as in μ and the L2-SP-free chain fails the continual feasibility gate at r=32, so only the entry pair is a one-knob contrast there. This measures μ's effect on the baselines, not why they were given it, which is the judged evidence above.

Table 32: With and without L2-SP on the generative instrument at 0.6B (correct of 200, seed 42, one draw a cell; Fisher exact on the counts, two-sided, within frame). Single domain: r=16, f=0.13; LwF at α_{LwF}=4, dose 15; replay at n=3, dose 20. Continual: the r=32 chains, the L2-SP-free one at μ=0 at both stages against Table 25's canonical counts.

frame	pair	GSM8K		IFEval	
		μ=0	with μ	μ=0	with μ
single domain	LwF, μ=0.05	32	79 (p=2.1×10^{-7})	82	106 (0.021)
single domain	replay, μ=0.2	39	64 (0.0059)	61	96 (0.0005)
continual, entry	replay, canonical	23	45 (0.0049)	42	78 (0.0001)
continual, after B	replay, canonical	15	43 (9.8×10^{-5})	39	75 (9.7×10^{-5})

Limits. Six of the 37 distinct contrasts behind this section change verdict between the two judges, five in the single-domain frame and one on continual set-A (the $*$ of the continual block); none reverses direction. Where LFA is paired-behind a baseline in one domain, that baseline carries L2-SP: LwF+μ (Table 15), and self-generated rehearsal at dose 15, at one of the two seeds and one λ rung away from the head-to-head arm (§8). The μ=0 counterpart of the matched-dose LwF pairing, single-knob α_{LwF}=16 at dose 15, is a tie (p=0.42).

K SELF-SUPPLY: THE SELF-GENERATED INPUTS, AND THE RESULTS IN FULL

§8 consumes three bodies of text the model wrote for itself. This appendix says how each was produced, what it looks like, and how the self-generated rehearsal baseline was placed at its operating rung. It is descriptive: nothing here is an argument about why generated text helps, or about what it preserves. The arm with both inputs self-written is the *data-free mode* of §1: the statistic and the supplement are both written by the model, and nothing beyond the new domain's own text is read. Its scope is one setting (r=32, one seed, one domain): with the supplement present the self-generated statistic ties the count-matched real one in six comparisons, the fully self-supplied arm ties the recipe at both doses, and with no supplement the self-generated statistic is behind at dose 20.

K.1 HOW THE TEXT WAS SAMPLED

The artifact corpus (2,500 documents). Free-running continuation from the base checkpoint, prompted with nothing but the document-start token (<|endoftext|>), which is not kept in the output. Sampling is at temperature 1.0 and top-p 1.0 with *every* truncation knob set explicitly: top-k disabled (k=0), min-p 0, repetition penalty 1.0. The explicitness matters for reproduction: this checkpoint's shipped generation configuration sets top-k=20, which a call passing only temperature and top-p inherits silently, and passing a single `eos_token_id` overrides the configuration's stop list and drops the document-boundary token, which we leave in the stop set. There is no burn-in. Length is capped at 2,048 new tokens. Filters are effectively off (minimum 1 character; a repeat-ratio ceiling of 1.0 on 1 − (distinct 8-grams/total 8-grams)), and 9 documents were dropped. Seed 42. The artifact is then fitted by the shipped builder with the shipped settings (PCA at 0.95, diagonal GMM with K=32, embedding lookup) at 600,000 samples per site.

The rehearsal buffer (40,000 instances). The SSR baseline (Huang et al., 2024) rehearses 36,400 free-running documents, sampled as above but capped at 512 new tokens and filtered (minimum 200 characters, repeat-ratio ceiling 0.3), plus 3,600 prompt/response instances written in-context from K=2 demonstrations, the SSR authors' own convention: demonstrations are drawn from a pool of 200, prompts are sampled at temperature 0.9/top-p 0.6 and responses at 0.7/0.8, generated at 2× oversampling and selected near the centroids of 20 K-means clusters. The demonstration pool is *real* instruction pairs, so the buffer is model-written but not demonstration-free; the artifact corpus sees no demonstrations of any kind. The buffer was screened against the WikiText-2 test set for contamination (0 removed). Seed 42.

The self-written supplement ($2{,}658$ pairs). The question–answer supplement is produced by the base checkpoint from the same domain passages and the same pre-baked template as the frontier-written one, and mixed into the domain corpus at the same $f{=}0.13$. Only the writer changes.

K.2 What the generated text looks like

The supplement's pairs have a mean length of 372 characters as standalone documents. Its *answers* run about 206 characters and one sentence against the frontier-written supplement's 435 and two; that difference is a property of the writer, and we report it rather than engineer it away. Scored under the base model as standalone documents, mean per-token perplexity is 8.24 for the free-running prose and 3.76 for the model-written instances ($1{,}000$ documents per source at 512 tokens, seed 0), and 59.7 for the self-written supplement against 52.2 for the frontier-written one ($2{,}658$ documents each). The supplement figures carry a confound the archive cannot remove: each pair was written conditioned on a passage the scorer does not see, so short pairs read as high-perplexity by construction. We draw no conclusion from any of these numbers.

K.3 Where the self-generated rehearsal baseline sits

The SSR baseline enters §8 at $\mu{=}0.05$, a rung ported from the frontier-supplement frame. Table 33 is the ladder that selects it, and the selection is on the judge-free axis: at $\mu{=}0$ the arm is infeasible, drifting $+30.8\%$ on general-text perplexity, and μ sets that axis monotonically across the ladder. The four rungs' judged levels sit within 0.15 of one another at dose 15 with overlapping zero-rates. Paired, the $\mu{=}0$ rung is behind $\mu{=}0.05$ and $\mu{=}0.1$ at dose 15 (paired t $p{=}0.012$, 0.022; Wilcoxon agrees), and every other pairing of the four rungs at either dose is a tie under both tests (paired t $p{\geq}0.13$).

Table 33: The self-generated rehearsal baseline's μ ladder, which places it at $\mu{=}0.05$. Frame: No-Anchor, LoRA $r{=}16/\alpha{=}32$, gradient accumulation 4 with three rehearsal micro-batches per domain batch, lr $3{\cdot}10^{-4}$, batch 6, seed 42, 20 epochs, `teacher-mode separate`, corpus $f{=}0.13$ with the *frontier-written* supplement; only μ differs between rows. Judge `gpt-5.6-luna@medium`, $n{=}154$, zero-rate in brackets. General-text Δ is WikiText-2 perplexity against the base model at dose 15; the feasibility bound of §E is $+15\%$. Judged levels are the verdict file's per-question pooled means over the questions a pairing shares; they differ from the search CSV's run-level cell by at most 0.002.

μ	Judge e15	Judge e20	dom. PPL e15	gen. Δ e15
0.2	1.126 (23.5%)	1.121 (19.9%)	10.92	-6.2%
0.1	1.197 (21.0%)	1.127 (23.9%)	10.25	-3.9%
0.05	1.201 (20.8%)	1.119 (22.1%)	9.58	$+0.1\%$
0	1.053 (26.1%)	1.132 (21.8%)	7.97	$+30.8\%$ (infeasible)

K.4 The substitutions and the head-to-head in full

Table 34 gives the recipe-input substitutions in full (each external input replaced, then both, at the confirmed recipe, $r{=}32$, one seed), with the paired verdict against the recipe row at the same dose; Table 35 gives the artifact-provenance block against the *sample-count-matched* real-text artifact; and Table 36 the $r{=}16$ head-to-head, every method on the model's own supplement, including the neighbouring ported rungs, GeRe, the μ-only arm and the unanchored control, each paired against the LFA row of the same seed. All verdicts are paired per-question tests under `gpt-5.6-luna@medium` over the same 154 questions in the single-domain frame, which was not tuning-matched: the tables license pairwise verdicts and no ordering of methods. No difference in them mixes the pooled and the run-level rendering of a judged level.

Three readings. The one measured boundary is the $f{=}0$ block of Table 34: with no supplement to carry answering, the self-generated artifact is paired-*behind* the shipped one at dose 20 (-0.138, $p{=}0.0046$), so its dose-15 lead over the shipped artifact does not carry to that frame. The two artifact blocks use different reference arms: Table 34's recipe row is the shipped artifact, fitted on the real proxy corpus at $1{,}543{,}040$ samples per site, while Table 35 pairs the self-generated artifact against a real-corpus artifact refitted by the same builder at its $600{,}000$ samples per site; we draw no conclusion

from the difference. And Table 36's SSR rows are feasible at dose 15 only, so the head-to-head at each method's best feasible dose (LFA at 20, SSR at 15) is the pairing §8 quotes: a tie at both seeds, -0.038 (paired t p=0.49, Wilcoxon 0.42) and -0.031 (0.54, 0.47).

Frame details. Both tables train at lr $3{\cdot}10^{-4}$ and batch 6. In the head-to-head LFA anchors on the self-generated artifact, SSR rehearses the model-written buffer, GeRe uses a 1,000-document self-generated buffer with thresholds computed on the same text, and LwF, μ-only and unanchored consume no preservation data. Each method sits at its best rung under the primary judge in the frontier-supplement frame, carried across supplement provenance and, for LFA, across the artifact as well: one rung per method, not a tuned comparison. EWC and a tuned L2-SP rung are absent because they were infeasible at every frontier rung (general-text $+100\%$ to $+1{,}900\%$) and never judged; the μ-only row is L2-SP's one scored rung.

Table 34: Self-generated inputs to the recipe: the confirmed recipe of §C (LoRA r=32/α=64, λ=10^5, μ=0.05, gradient accumulation 1, seed 42; doses are checkpoints of one 20-epoch run) with only the named input changed. Corpus f=0.13 except the two f=0 rows. Judge `gpt-5.6-luna@medium`, n=154, zero-rate in brackets; levels are the verdict file's per-question pooled means, within 0.002 of the search CSV's run-level cells. IFEval: 200 prompts, seed 42, dose 15, base model 0.580; `--` where unscored. Paired columns: per-question difference against the recipe row at the same dose (paired-t p in parentheses), positive meaning the row scores above the recipe; the second f=0 row is paired against the first. †Domain perplexity scores frontier-written held-out question–answer text, so it favours the frontier-supplement rows and is never comparable across f (§4): printed for feasibility only.

				dom. PPL†	gen. Δ	IFEval	paired vs. the recipe	
Artifact	Supplement	Judge e15	Judge e20	e15	Δ e15	e15	e15	e20
real corpus	frontier	0.968 (26.8%)	1.042 (22.2%)	8.76	−10.0%	0.550	—	—
self-gen	frontier	**1.140** (22.6%)	**1.139** (21.6%)	8.51	−8.9%	0.585	ahead +0.173 (0.0019)	tie +0.097 (0.075)
real corpus	self-written	1.013 (29.2%)	1.062 (24.0%)	9.78	−10.3%	0.575	tie +0.045 (0.41)	tie +0.021 (0.68)
self-gen	self-written	1.030 (26.1%)	1.018 (23.5%)	10.06	−8.9%	**0.600**	tie +0.062 (0.21)	tie −0.023 (0.64)
f=0: no supplement exists, so only the anchor can preserve answering; the second row is paired against the first								
real corpus	none	0.825 (31.9%)	0.947 (26.4%)	11.96	−8.1%	0.555	(f=0 floor)	
self-gen	none	0.877 (29.7%)	0.809 (31.7%)	12.02	−6.9%	—	tie +0.052 (0.26)	behind −0.138 (0.0046)

Table 35: Artifact provenance at matched sample count: the self-generated artifact against a real-corpus artifact refitted by the same builder with the same settings at the same 600,000 samples per site, so the two differ only in the corpus they are fitted on (App. K). The shipped artifact of Table 34's recipe row is fitted at 1,543,040 samples per site and is *not* count-matched. Frame: the recipe configuration (LoRA r=32/α=64, μ=0.05, f=0.13, gradient accumulation 1, seed 42, **one seed**; doses are checkpoints of one 20-epoch run), three λ rungs by two doses. Judge `gpt-5.6-luna@medium`, n=154, zero-rate in brackets, levels as in Table 34. Δ is the per-question difference, self-generated minus count-matched. Pairwise verdicts only.

λ	dose	count-matched real	self-generated	Δ	paired t p	Wilcoxon p	verdict
$5{\cdot}10^4$	15	0.975 (28.4%)	1.082 (25.1%)	+0.106	0.079	0.092	tie
$5{\cdot}10^4$	20	1.156 (21.2%)	1.142 (21.7%)	−0.014	0.81	0.85	tie
10^5	15	1.088 (27.7%)	1.140 (22.6%)	+0.052	0.35	0.35	tie
10^5	20	1.095 (23.5%)	1.139 (21.6%)	+0.044	0.45	0.39	tie
$2{\cdot}10^5$	15	0.922 (30.0%)	1.004 (27.7%)	+0.082	0.106	0.084	tie
$2{\cdot}10^5$	20	1.066 (24.0%)	1.095 (24.0%)	+0.029	0.57	0.78	tie

L BASELINES: IMPLEMENTATION AND TUNING

Table 37 specifies every baseline, and the paragraphs after it say what each was given so that it enters at its best. The frozen-protocol arms (§E.1) share Qwen3-0.6B, the domain-A corpus (5.99 MB, 77 documents), 10 epochs, batch 6, sequence length 512, constant learning rate $3{\times}10^{-4}$, AdamW (weight decay 0.01) and seed 42. The LFA, unanchored and μ-only arms use LoRA r=16/α=32 on all seven projections (q/k/v/o/gate/up/down), gradient accumulation 1, warmup 50. The externally run baselines use warmup 100 and differ in accumulation: replay 2 (one buffer micro-batch per optimizer

Table 36: The self-supplied head-to-head: every row trains on the model's own question–answer supplement (f=0.13), with self-generated preservation data wherever the method consumes any. Frame: LoRA r=16/α=32, `teacher-mode separate`, 20 epochs, seed 42 except the seed-1337 block; gradient accumulation set by mechanism (4 SSR, 2 GeRe, 1 otherwise). Judge `gpt-5.6-luna@medium`, n=154 (153 in three cells: LwF α=16 at both doses, μ-only at dose 15), zero-rate in brackets; levels as in Table 34, as are IFEval and the domain-perplexity caveat. Every setting is carried over from the frontier-supplement frame rather than re-tuned; *nb.* marks the neighbouring rungs run to test that. Paired columns: per-question difference against the LFA λ=2·10^4 row *of the same seed* (paired-t p in parentheses), positive meaning the row scores above LFA; both LFA rows sit at μ=0.05. The feasibility bound of App. E is domain PPL $\leq$12, and §marks the SSR rows' dose-20 cells as *infeasible*: their domain perplexity at that dose reads 14.72, 14.81 and 14.41, against 10.26, 10.80 and 9.76 for the LFA rows, so dose 15 is the SSR rows' only feasible dose. The LwF α=16 row is over the bound at both doses (14.44, 14.71). $^{\equiv}$, $^{\flat}$ (§4; $^{\flat}$ on the Wilcoxon p) are marked on the two-seed contender rows only. **Pairwise verdicts only: the single-domain frame licenses no ordering of methods.**

			dom. PPL	gen. Δ	IFEval	paired vs. LFA λ=2·10^4	
Method	Judge e15	Judge e20	e15	e15 / e20	e15	e15	e20
LFA λ=2·10^4	1.143 (23.2%)	**1.222** (18.2%)	9.53	−3.4 / +3.7	0.505	—	—
LFA λ=5·10^4 *(nb.)*	1.026 (30.3%)	1.106 (21.8%)	10.00	−8.0 / −4.0	0.555	behind −0.117 (0.033)	behind −0.116 (0.026)
SSR μ=0.05	**1.184** (19.4%)	1.069§ (24.2%)	11.89	−3.1 / +0.8	0.570	tie$^{\equiv}$ +0.042 (0.44)	behind −0.153 (0.0027)
SSR μ=0.1 *(nb.)*	1.183 (20.6%)	1.049§ (23.9%)	11.97	−6.4 / −2.6	0.585	tie +0.040 (0.45)	behind −0.173 (0.0014)
LwF α=4+μ0.1	1.062 (25.6%)	1.055 (21.6%)	10.43	+0.7 / +4.8	0.575	tie −0.081 (0.14)	behind −0.168 (0.0020)
LwF α=1+μ0.1 *(nb.)*	1.061 (24.5%)	1.043 (25.2%)	8.72	+47.7 / +91.6	0.445	tie −0.082 (0.13)	behind −0.179 (0.0019)
LwF α=16+μ0.1 *(nb.)*	0.996 (25.0%)	0.941 (27.7%)	14.44	−4.4 / −4.4	0.615	behind −0.142 (0.0049)	behind −0.275 (< 0.0001)
LwF α=4+μ0.05 *(nb.)*	1.055 (24.4%)	1.008 (21.4%)	9.99	+6.4 / +13.7	0.555	tie −0.088 (0.083)	behind −0.214 (0.0001)
GeRe+μ0.1	1.012 (28.4%)	1.113 (22.3%)	8.62	+8.8 / +18.4	0.550	behind −0.131 (0.018)	behind −0.109 (0.0499)
μ only (μ=0.4)	0.961 (30.6%)	0.927 (33.5%)	8.86	+40.3 / +97.3	0.360	behind −0.184 (0.0031)	behind −0.295 (< 0.0001)
unanchored	0.738 (44.2%)	0.810 (39.7%)	10.40	+1,225 / +2,405	0.215	behind −0.405 (< 0.0001)	behind −0.412 (< 0.0001)
seed 1337*: the three contenders re-run end to end; the paired columns are against the seed-*1337 LFA *row*							
LFA λ=2·10^4	1.096 (25.8%)	**1.253** (15.3%)	9.60	−1.9 / +4.0	0.540	—	—
SSR μ=0.05	**1.222** (17.5%)	1.179§ (18.6%)	10.92	−1.9 / +0.9	0.605	ahead$^{\flat}$ +0.126 (0.027)	tie −0.074 (0.12)
LwF α=4+μ0.1	1.057 (25.2%)	1.032 (24.4%)	10.33	−0.5 / +4.9	0.585	tie$^{\equiv}$ −0.039 (0.53)	behind −0.221 (< 0.0001)

step); L2-SP, EWC, LwF, VeFA and TM-only 8 (TM-only: one buffer micro-batch per 8). Arms therefore differ in optimizer-step count at equal data budget and are read at matched final content loss, which every table reports. Retained accuracy averages 11 multiple-choice benchmarks (MMLU, MMLU-Redux, HellaSwag, ARC-Challenge, ARC-Easy, TruthfulQA, PIQA, WinoGrande, SIQA, BoolQ, GPQA-Diamond); held-out domain perplexity uses 77 domain question–answer probes.

How each baseline was given its best chance. Five baselines receive LFA's second knob (replay, LwF, GeRe, TM-only, self-generated rehearsal; App. J). Replay's rehearsal share is swept per setting and its buffer rebuilt to the corpus $p(h)$ is fitted on, which makes it stronger. LwF is a transcription of its authors' layer, checked by permanent tests. GeRe and O-LoRA run on their authors' trainers, and where a defect control moves GeRe's judged level we disclose it. The details, with every number, are the notes below.

Replay: share. Standard guidance is 1–10% (Ibrahim et al. (2024) sweep 0/1/5/10/50%; 5% captures nearly all the benefit), but 20–100 epochs over a 5.99 MB corpus is a far higher update count per unit of data. In one domain, rehearsing the 2.4 GB general corpus, general-text drift improves monotonically over 50/67/75% (−12.8/−17.1/−19.9% at a matched epoch 20; −16.5% at the 50% arm's own best dose, epoch 5) while judged quality does not order with it (0.825 at that epoch-5 point against 0.709/0.828/0.815 at epoch 20, a span of 0.119; the 75% arm's zero-rate is 0.353); we report 75%. The share does not port: continual stage two rehearses A's own training corpus (1,673 documents), where 75% drives held-out A perplexity to 390.0 (from 7.79 at entry) while perplexity *on the buffer* falls to 1.35, memorization and not forgetting. The response is U-shaped and every continual table uses the optimum, 25%.

Replay: buffer. The *uniform* general buffer is a 927,398-document sample of RedPajama-1T. The *seed-matched* one (41,766 entries, 119.5 M characters) is rebuilt to the character shares of the corpus $p(h)$ is fitted on, sized so no document is reused, and screened against the WikiText-2 test split by a 13-gram scan (one document removed; the uniform buffer scans at 2 of 927,398, neither a test article). Composition matters as much as share: rehearsing A as raw prose (77 documents, no pairs) when A is scored on question–answer text gives ΔA 22.04, worse than L2-SP's 14.57, so every replay number here rehearses A in the format it is scored in. The "A+general" rows of Table 17 blend A's corpus 50/50 by entry with the seed-matched buffer (25/75 and 75/25 are also exposed): at two seeds that arm betters the A-only buffer on every continual perplexity column, reads

Table 37: The baselines. *Knob*: the preservation–adaptation knob each was swept on; a knob is re-tuned per protocol or its port is flagged where it occurs (§3). μ is the strength of the L2-SP pull of Eq. (3), granted to the five baselines marked $+\mu$ (App. J). SSR: App. K.

Method	What is run	Knob and sweep	Selected	Source
Replay $+\mu$	Plain cross-entropy on buffer micro-batches of real text	rehearsal share $50/67/75\%$ (one domain), $\{25, 50, 75\}\%$ against dose (continual); μ	75% (one domain); 25% (continual)	ours
LwF (Li & Hoiem, 2016) $+\mu$	$\alpha_{\mathrm{LwF}}\,\mathrm{KL}(p_{\mathrm{t}} \,\|\, p_{\mathrm{s}})$, $p{=}\mathrm{softmax}(z/T)$, on the adaptation batches; mean over non-padding tokens, $T{=}2$ (the authors' default)	α_{LwF} in powers of four, $\{1/64, \ldots, 256\}$ at $f{=}0$ and $\{1, 4, 16, 64, 256\}$ at $f{=}0.13$; μ	1 and 4 ($f{=}0$); 16, or 4 with μ ($f{=}0.13$)	ours, transcribing the authors' layer
EWC (Kirkpatrick et al., 2016)	$\lambda_{\mathrm{EWC}} \sum_i F_i(\Delta W)_i^2$ on the adapter-inclusive weight change; diagonal empirical Fisher over the 10:1 proxy corpus (500 batches of 4)	λ_{EWC} over 10^{-3}, 10^{-2}, 10^{-1}, 10^{2}, 10^{3}	10^{2}	ours
L2-SP (Li et al., 2018)	The μ-only configuration, uniform layer weights	μ	0.05 (frozen protocol)	ours
VeFA (Wang et al., 2025)	A trainable per-input-channel rescale k on each of the seven projections (initialised 0; 286,720 parameters, 0.05% of the model), base frozen, plain cross-entropy; merged as $W_{\mathrm{t}}\,\mathrm{diag}(1{+}k)$; no $\|k\|_\infty$ clamp	none	—	ours
TM-only $+\mu$	GeRe's threshold-margin loss on last-layer states against the frozen teacher, thresholds $\tau_k^{\pm}{=}m_k{\pm}\sigma_k$ from a 1,000-sample generic buffer; no replay cross-entropy	margin weight 0.5–2; μ	labelled ablation, not GeRe	ours
GeRe (Zhang et al., 2025) $+\mu$	The authors' `GeReTrainer`, unmodified but for an in-memory buffer loader (removed API); `bf16`, its default	$w \in \{1, 100, \mathtt{dy}\}$; μ	`dy`	authors' code, pinned commit
O-LoRA (Wang et al., 2023)	The authors' `UIETrainer.training_step`, unmodified, over our adapter; continual only, $r{=}4$	$\lambda_1{=}0.5$, $\lambda_2{=}0$ as released; not re-tuned	as released	authors' code (MIT), pinned, checksummed

WikiText-2 *below* the base model's 18.18, and leaves both judges inside replay's band, so the judged contrasts of §6 are not buffer-composition effects.

LwF. The authors publish no language-model implementation, so the formulation of record is their MatConvNet layer (`vl_nnsoftmaxdiff.m`), whose forward and backward expressions are checked against ours as permanent tests (two deviations and three transposition choices are tabulated with the vendored source). We take KL where they take cross-entropy (the gradient is identical, to $5.6{\times}10^{-17}$ in double precision), reduce by mean (under their sum the effective α_{LwF} of a causal LM would depend on document length), and apply no T^2 rescaling, as their backward confirms (it carries $1/T$), so a value quoted from a T^2 implementation does not port. One structural difference is unavoidable: the authors distil on the *original* head and train a new task head, whereas a causal LM's single tied `lm_head` puts both losses on the same output units in opposing directions, so our α_{LwF} trades adaptation against preservation directly.

GeRe. Batch interpolation is enabled at 3 buffer rows per 12-row step (6 per 24 continually): 25%, replay's share, above the demo default's realized 19.8%. The release combines its two losses per micro-batch, so a nominal $w{=}100$ realizes 17–75 (75 at our geometry). It chunks the corpus once, so GeRe and the continual TM-only$+\mu$ chain, which runs on the same trainer, saw the supplement in every epoch, $\approx 3\times$ the exposure of the arms they are compared with (App. N). And its loss is numerically fragile: the teacher-activation cache does not bitwise-match batched training, and in `bf16` the margin saturates at arithmetic resolution. Controls that remove these (teacher recomputed in-graph, domain perplexity 12.40 against 12.41; fp32 throughout, 12.85) stay within 0.44 of the released arm at $w{=}100$ and do not explain its under-adaptation. They do move the judge: the fp32 control reads 0.877 against the released arm's 0.732 under the primary judge, an unpaired $+0.145$ above the 0.12 answer-set threshold and paired-tested under neither judge (`claude-sonnet-5`: $+0.072$). We report the released behaviour and include a defect report with the code.

O-LoRA. We follow the source where it differs from the paper: an L_1 penalty on the Gram matrix between the previous and current A factors, a *dual* adapter whose forward sums both, and the standard-order scripts' $\lambda_1{=}0.5$, $\lambda_2{=}0$ (one long-sequence script sweeps both). Seven runtime shims touch no loss arithmetic, and two smoke tests gate every arm (the orthogonality term is exactly 0 on a first task; the dual forward at step 0 equals the stage-one adapter's output). Its stage one is the unanchored chain's $r{=}4$ adapter, which makes that chain the matched no-regularizer control. Our frame penalises all 196 factor pairs (seven projections $\times$ 28 layers) where the released runs penalise two projections per layer; we do not assert the direction of that difference, and a λ_1 down-ladder is unrun.

EWC, TM-only. At $\lambda_{\text{EWC}} \leq 0.1$ EWC's penalty is numerically negligible and its mild preservation comes from the reduced optimizer-step count alone; 10^2 is selected (content 1.99, general-text $+7\%$), and 10^3 preserves slightly more (18.99) at content 2.05. TM-only is weaker than released GeRe in three ways (margin weight 0.5–2 against 100 or dynamic; the reduction averages over all elements, not violating ones; no replay cross-entropy) and differs in two that do not weaken it (teacher states recomputed in-batch; padding not masked); its constraint gradient is $\sim 0.5\times$ the content gradient against the release's 33–66$\times$. Its buffer size is not load-bearing: drawing from the full replay corpus scores nearly identically (general-text 18.62 vs. 18.53; domain 11.53 vs. 11.67; 17 vs. 15 intruders).

The teacher under LoRA. No teacher copy is needed: W_{base} is frozen, so the student reads $f_{\text{t}}(h)$ by running its own sub-module with the adapter disabled, for LFA and for LwF alike. We verified it at 0.6B by training bit-identically without the copy (all 392 adapter tensors `torch.equal` over 50 optimizer steps, per-step loss terms identical), and every memory figure in App. I is a method's own footprint measured that way.

Our LwF baseline follows the released implementation of Zhizhong Li and Derek Hoiem (`https://github.com/lizhitwo/LearningWithoutForgetting`), vendored unmodified under its BSD-style license, which conditions redistribution on acknowledging the authors in accompanying materials; we acknowledge them here.

M LLM-JUDGE DETAILS

Protocol. The question sets are 154 fixed probes on domain A (77 "attributed" + 77 "content-only" phrasings) and 78 each on B and C, generated from the domain documents and held fixed across all arms and all judge instruments. A's set (2026-03) and B's (2026-07) were written by `claude-sonnet-4-6` and C's (2026-09) by `claude-sonnet-5`. Answers are greedy-decoded locally from each checkpoint with identical generation settings. Judging is one call per question: each call scores {correctness, completeness, domain_accuracy, reasoning_quality, synthesis} $\in [0, 5]$ against the question's source excerpt, with a written justification (Zheng et al., 2023), and Judge/5 in Table 7 averages the five dimensions over all 154 questions. Three checks preceded its use: its written rationales are reference-grounded and specific (e.g. flagging a factual inversion of the reference's persistence claim); length bias on set-A is weak (median Spearman ρ=0.12 across the eight fine-tuned arms of Table 7, range -0.01 to $+0.16$, `claude-sonnet-5`; it is not weak on every set, App. G); and the two 77-question phrasings agree closely per arm. Every answer set is scored by `gpt-5.6-luna@medium` and, under the identical rubric, by `claude-sonnet-5`. Judge identity is the pair (model, effort): the same model at a different effort is a different instrument. A domain's question set is shared by every arm, so who wrote it affects the absolute level of a judged score and not the within-domain cross-arm contrasts we draw.

Voice capture is a regex detection of first-person-authorial markers ("I have argued", "in my view", "as I argued", self-citation), validated against hand-read examples; assistant-voice first person ("I think the answer is…") does not match. An example of the pathology, from the $\lambda{+}\mu$=0.05 arm, where it is rare at 1.3% (it reaches 14–17% in the most-memorized arms): *"I have argued in* The Conscious Mind *(Chalmers, 1996) that the third-person model of consciousness is not the only possible model… "*, fluent domain voice with fabricated content, judged 1/25 with the rationale "fails to explain the actual argument… incoherent, self-contradictory, and repetitive." The same protocol under `claude-sonnet-4-6` is directionally consistent but compressed (all arms $\approx 1/5$).

Two judges, and why this one is primary. We chose `gpt-5.6-luna@medium` on two counts. *Independence*: the `claude-sonnet` family wrote all three question sets and the domain question–answer supplement the models train on, and `claude-sonnet-5` itself wrote C's set, so on C it has no generate-and-grade separation; `gpt-5.6-luna@medium` appears nowhere else in the pipeline. *Stability*: re-scoring identical frozen answers, its mean absolute test–retest difference is 0.0216 over 28 answer sets against `claude-sonnet-5`'s 0.0350 over 15, and it scores exactly 0 on 30.6% of judgments (n=383,050) against 39.3% (n=379,220), so it sees past the rubric's floor more often.

We report the second judge rather than discard it, and the agreement between the two is itself evidence. Of the 116 single-domain and two-domain contrasts testable under both, 86 (74%) preserve their verdict and 30 flip (19 separations becoming ties, 11 ties becoming separations); the most stable section is new-domain judged quality after stage two, at 34 of 43 (79%) with zero sign reversals; the

three-domain section preserves 258 of 309 (83%), its 51 flips concentrated on the after-C old-domain contrasts. Across all 425 dual-judge contrasts **not one separation under one judge becomes an opposite-direction separation under the other**: the eight cases in which a separation reverses sign do so by becoming a tie. All 621 judged cells that moved when we changed judge moved *up* (mean $+0.238$), and on identical frozen answers the primary judge scores 1.19 to 1.63 times the second across a 28-set campaign. A judged number is therefore never read against a number from the other judge, as a level or as a gap; only verdicts and ratios are compared.

Reading conventions. Cross-arm contrasts are paired two-sided Wilcoxon tests over the questions the two arms share (154 on A, 78 on B and C, less any a run drops at parsing), quoted with the means over that shared set. Unpaired run-level differences are read against a measured minimum detectable effect (App. E); a new domain's judged score is read against two floors, the chain's own pre-stage fused model and the base model (§6.1); and, because the judge tracks answer length within a model, load-bearing cross-stage contrasts are re-scored on the two least length-sensitive dimensions (§6.1, App. G).

What the second judge changes. *Robust under both:* the new-domain judged advantage of §6 (34 of 43 set-B verdicts preserved, zero sign reversals, every LFA-versus-replay set-B contrast preserved), the matched-dose LwF+μ result at μ=0.05, LFA's judged-A lead over LwF+μ at r=32 after three domains, every LFA judged-A fall across the third stage, and the six verdicts of the step-matched stage-two control (App. F). *Judge-dependent:* the single-domain best-dose lead over single-knob LwF (present under the second judge, a tie under the primary except rank-matched at r=16); the matched-dose comparison at LwF's μ=0.1 rung (a tie under the second judge, against us under the primary); the rank-matched matched-dose comparison (against us under the second judge, a tie under the primary); replay's judged-A advantage after the third domain and the optimizer-step explanation built on it (both only under the second judge; §M.1); LFA's judged-A lead over LwF+μ at r=16 after three domains; and the reading of LFA's old-domain level against entry at r=16 (from entry to after C a tie under the primary judge and a fall under the second, -0.132, p=0.0056, which also reads the r=4 fall more firmly, -0.155, p=0.0004, and registers the stage-two rise at r=4 and r=32 where the primary registers it at r=16 and r=32; Table 21). *Instrument gaps:* the per-tier decomposition of set-C exists only under the second judge; no base-model set-B floor exists under either judge, so set-B levels are read against the r=32 chain's pre-B fused model (1.118); and C's question set was written by the second judge itself, so under that judge C lacks generate-and-grade separation, which all three sets have under the primary judge. `claude-opus-4-8` (a sidecar on five continual rows) and `claude-sonnet-4-6` (historically) are labelled where they appear and carry no claim.

M.1 What the study says about evaluating continual adaptation

Three findings from the continual experiments are about the *instruments*, not about LFA.

(1) Perplexity retention and judged answering dissociate, four ways and twice inside a single method (App. F); through a third domain the arm with the best prior-domain perplexity of its rank judges in the text-free band at every rank. *Rule:* report both instruments, and state each claim per instrument.

(2) A stage's question–answer supplement moves every earlier domain's judged level, and its size is not the lever. A third stage trained on the new domain's pairs alone reproduces the full fall on the first domain at one sixth of the optimizer steps and sinks the second, never-supplemented one further; a quarter of the pairs reproduces it; removing them collapses answering form instead; and no remedy we tried lifts the level above a step-matched control (the ablation above). *Rule:* treat a supervised-answer supplement as a stage-level intervention, ablate its presence and not its fraction, and read an earlier domain's judged level against the level it entered the stage with, beside its answer length.

(3) Judged comparisons across methods can be confounded by optimizer-step geometry, and the control is cheap. Rehearsal methods interleave buffer micro-batches, so at a fixed epoch count they take a fraction of the optimizer steps a text-free method takes (here 3,103 against 9,279). Re-running each rank's text-free chain at replay's gradient-accumulation geometry, the confound is *null on judged A at every rank* under the primary judge: the plain chain and the step-matched control both tie replay (largest departure p=0.56 for the plain chain, p=0.64 for the control), and the control differs from

the plain chain by no more than $p{=}0.37$. Under the robustness judge the same three configurations read as a replay lead at $r{=}4$ and $r{=}16$ that the control removed, but the control-versus-plain contrast was itself non-significant ($p{=}0.084$), so the step effect was inferred from two contrasts against a third arm and never measured. *Rule:* run the text-free arm at the rehearsal arm's step geometry, *and test the control against the un-matched arm directly*.

One further check travels with these: a loader that re-chunks from a random offset each epoch can silently drop short documents, so a mixed corpus's nominal composition is an epoch-0 quantity. Ours carried its question–answer files in 8 of 20 epochs, every knob was selected under that schedule, and changing it moved the dose optimum by five epochs (App. N).

N Experimental setup, in full

This appendix gives what §4 condenses: corpus provenance, the loader's realized supplement schedule and what it does to dose selection, and the anchoring, metric and series details.

Model, domain, adaptation. Qwen3-0.6B (28 layers, hidden size 1024). The domain is a philosophy-of-mind corpus ("chalmers"): 77 documents, 5.99 MB of text extracted from author-hosted copies of David Chalmers' published papers, retrieved in February 2026 from https://consc.net/papers-menu/ and, for the *Constructing the World* excursuses, https://consc.net/constructing-the-world/. **The most recent document in the corpus is dated 2020**, so the domain predates the base model and cannot contain material the model could not have seen.[4] Adaptation is LoRA[5] on all seven projections at $\alpha/r{=}2$ with frozen embeddings, learning rate 3×10^{-4} constant, batch size 6, 50 warmup steps, cosine anchor schedule, seed 42, 10 epochs unless noted. Rank and λ are properties of an operating point rather than global defaults, because the two are coupled (§C.1); where a table sweeps neither, the mechanism-study setting $r{=}16$, $\lambda_{\text{qkv}}{=}\lambda_{\text{mlp}}{=}5000$ applies. The domain corpus is used in two forms, distinguished by a **supervised-answer fraction** f: the raw extracted text ($f{=}0$), and a QA-mixed variant in which chat-formatted domain question–answer pairs are mixed into that same text to a target *token* fraction $f{=}0.13$, screened for contamination against the evaluation questions. The Series tables and the frozen-protocol comparison (§E.1) use $f{=}0$; the judge-guided search and everything selected against the judge (the rank study and the confirmed operating point included) use $f{=}0.13$. What f buys is measured in §E.

The realized supplement schedule. One loader detail bears on reproduction. It is a defect, found after most training runs were complete; re-running every arm under a corrected loader was not feasible, so every arm is reported under the realized schedule and the exposure it implies. The corpus is re-chunked each epoch from a random offset, and a document shorter than the offset contributes no chunk that epoch, so the short question–answer files are present in 8 of 20 epochs at the recorded seed. One of the eight carries the supplement only fractionally, so the realized exposure is $\approx$7 full-epoch equivalents at a mean share of 0.35: f is the epoch-0 share (≈ 0.045 averaged over training), and each pair is seen $\approx$7 times rather than 20. The schedule is identical for every arm trained through our loader. The exceptions run on the authors' released GeRe trainer (the faithful-GeRe arms and, in the continual setting, our TM-only$+\mu$ chain), which materialises a corpus once at its epoch-0 chunking: those arms saw the supplement in every epoch at the epoch-0 share, $\approx 3\times$ the realized exposure of the arms they are compared with (App. L). O-LoRA is unaffected: its stage one is a plain-LoRA adapter trained through our loader, and its own stage trains on B, which carries no supplement. Every

[4] Eighty sources were retrieved (59 PDF, 21 HTML) and converted to markdown by our extraction script (marker-pdf for PDFs, BeautifulSoup+markdownify for HTML), with a 1000-character minimum, then cleaned by hand to remove running headers/footers and passages that extracted garbled. Three retrieved documents were excluded and are named here for completeness: *Actuality and Knowability*, *Revisability and Conceptual Change in "Two Dogmas of Empiricism"*, and *What Do Philosophers Believe?* (the last co-authored, unlike the rest of the corpus). On publication the authors will release the complete list of source documents with direct links, together with the extraction recipe, rather than the texts themselves, which remain under their publishers' copyright; all 77 links were re-verified in August 2026. Because the cleaning pass is unscripted, a rebuild from the same pages will be close to but not identical with the corpus used here.

[5] The reference implementation uses an internal abbreviation throughout (`lra_*` modules and `outputs/lra/` run directories); it names the objective described here.

operating point and every baseline knob in this paper was selected under that schedule, and App. N reports what changes when the supplement is kept in every epoch.

Epoch selection under the realized schedule. Dose and λ are coupled to the *realized* supplement exposure as they are to f. A single re-run that keeps every short document in every epoch, everything else identical, reaches the same perplexity point five epochs earlier (domain perplexity 8.75 at -12.0% general-text drift at epoch 10, against the shipped 8.76 at -10.0% at epoch 15) and then over-fits the pairs, domain perplexity rising to 10.50 at epoch 15 and 16.77 at epoch 20. Its judged score is level with the shipped run at epoch 15 (0.959 vs. 0.968; shipped zero-rate 26.8%), and the shipped epoch-20 maximum does not exist under that schedule (0.913 vs. 1.042; 22.2%). This pairing is in no verdict file under either judge; the paired p-values on record are `claude-sonnet-5`'s, on its own levels (0.739 vs. 0.788, p=0.27; 0.690 vs. 0.877, p=0.001). The numbers in this paper are specific to the recorded schedule. For the uniform-μ=0.05 100-epoch run of App. A.3, the best checkpoint under a given preservation budget is: $< 5\%$ PPL $\rightarrow$ epoch 10 (content 1.965, 7 intruders); $< 10\% \rightarrow$ epoch 13; $< 20\% \rightarrow$ epoch 29; $< 50\% \rightarrow$ epoch 63; general-text perplexity is non-monotone in later epochs while content loss keeps decreasing, a preservation plateau and not continued degradation.

The learning-rate-decay confound. Decaying the learning rate to zero (cosine decay over the full run) severely harms preservation: a 100-epoch run reached content loss 0.447 but general-text perplexity 128.1 ($+620\%$), against content 1.108 and perplexity 32.9 ($+85\%$) for its constant-LR twin. Near-zero LR appears to remove the implicit regularization that keeps optimization off sharp minima (Cohen et al., 2021). We use a constant LR throughout and flag the LR schedule as a confound for anti-forgetting studies.

Anchoring specifics. Both $p(h)$ artifacts are estimated from the same 10:1 pretraining:instruction proxy corpus (1.54M states per level for the correlated+GMM artifact, 1.2M for the diagonal-moments one). The default is the correlated+GMM artifact at its own operating point (r=32, λ=100,000, n=16 sampled states per level), used by the recipe and the continual chains. The representation study and the rank study (App. A, C.1) run at the diagonal-moments artifact: diagonal-Gaussian sampling from the per-level moments, with the exact embedding lookup at layer 0, n=512 states per anchored level per step, anchoring Q/K/V/O and the MLP of every layer. Runs freeze embeddings, layernorms and lm_head, so two further anchors of the general (full-weight) objective are *inactive* and omitted from Eq. (3): $\lambda_{\text{embed}}\mathcal{L}_{\text{embed}}$ on the embedding/first-layernorm interface (weight: λ's layer-0 value) and the lm_head anchor.

Metrics, and why these instruments. We report the domain *content loss* (training cross-entropy on the domain corpus; lower is more adaptation) and two preservation signals: *general-text perplexity* on WikiText-2, and the *intruder-dimension count* via the detector of Shuttleworth et al. (2024) (Algorithm 1, cosine threshold 0.5; top-k=20 within the Series tables and a frozen top-k=10 for the baseline comparison), with the mean maximum cosine similarity between student and teacher singular vectors. Deterministic perplexity carries every knob-level decision (λ, μ, rank, dose) because it is cheap and resolves differences of a fraction of a percent. The multiple-choice suite cannot separate those arms: retained accuracy sits at base level, inside a $\leq$1-point band, for nearly every anchored arm, so it serves as a guard and a table column. Generative benchmarks do see what multiple choice misses (App. F, finding (vi)), but at hundreds of generations per checkpoint they are not affordable across a fifty-configuration search, so they appear as a finding and not a screen.

Design and baselines. The three series of App. A.2 are a chain, not a grid: each changes one layer schedule to repair a defect the previous one exposed, and Series 4 is carried forward. A Series 1 pilot at λ=10,000 runs without L2-SP, falls outside the two-parameter objective, and is not reported. "Unanchored" is the same recipe with λ=μ=0 (a separately tuned LoRA baseline is unrun); "μ only" sets λ=0. The *frozen protocol* of App. E.1 is a metric set fixed *a priori* and applied identically to every arm: tier-1 perplexity benchmarks, full WikiText-2, the intruder detector at top-k=10, and held-out domain perplexity. Its *training* frame is not uniform: gradient accumulation is 1, 2 or 8 and warmup 50 or 100 depending on the arm's own tuned protocol, and VeFA is not a LoRA arm, so within-group comparisons (LFA against LwF, against L2-SP/EWC) are frame-clean and cross-group rows are not (App. L).

Code and artifacts. The Reproducibility Statement describes the supplementary archive and what follows on publication. One detail is added here: the diagonal floor of App. B is $\sim$0.5 MB, or $\sim$1 MB with the optional token frequencies of §3.

O EXTENDED RELATED WORK

§2 places LFA against three families. This appendix gives each in more detail, then the nearest precedents for the premise, the sampling and the constraint.

Feature and function distillation. $\mathcal{L}_{\text{repr}}$ descends from FitNets (Romero et al., 2014), TinyBERT (Jiao et al., 2019) and probabilistic knowledge transfer (Passalis & Tefas, 2018), from LwF (Li & Hoiem, 2016) at the output level, and from L2-SP (Li et al., 2018) in weight space. The family's standing weakness is that a function-space constraint is only as good as the input distribution it is evaluated on, which has meant a representative set of prompts, and therefore data; $p(h)$ *is* that distribution, estimated once and shipped as statistics. LwF's probe is the current task's own inputs, so what it buys is *support-conditional*: it protects what those inputs exercise and has no handle on a domain that has left the input stream. Read against the level each chain entered stage two with, LwF+μ's judged answering on the *previous* domain falls over the next stage at $r{=}16$ and $r{=}32$, both falls surviving the length control, where LFA's does not fall (Table 21); over a third domain both methods' judged answering on the first falls (Table 23). LwF's cost is retention and general-text fit rather than plasticity: at $r{=}16$ and $r{=}32$ it fits the new domain *better* than LFA in perplexity after one round (Table 1).

Second-moment regularizers and null-space projection. At a linear site the anchor is $\mathrm{tr}(\Delta W M \Delta W^\top)$ (Eq. 2). EWC (Kirkpatrick et al., 2016) penalizes a quadratic in the weight change weighted by a diagonal Fisher, and K-FAC (Martens & Grosse, 2015) factors a layer's Fisher into an input second moment and an output-gradient factor; LFA keeps the input factor alone (identity output factor), so it depends on no loss or label and can be fitted once on a proxy corpus, frozen and shipped. OWM (Zeng et al., 2019), GPM (Saha et al., 2021) and Adam-NSCL (Wang et al., 2021) established that an update orthogonal to the span of previously seen features does not change the response to prior inputs. Our destructive/non-destructive split is that statement applied to intruder dimensions: a new direction is destructive precisely when it carries mass under $p(h)$. They *enforce* the principle by projection, so used directions cannot be adjusted and the free budget depletes over a task sequence; LFA *prices* it, so the optimizer may still move in a used direction when $\mathcal{L}_{\text{content}}$ makes that worth the cost. Their covariance comes from previous-task data held at adaptation time, and we *measure* the split on real activations rather than assuming it (§A.4). SI (Zenke et al., 2017) and MAS (Aljundi et al., 2017) weight parameter changes by other importance estimates, unified by Benzing (2020) as (root-)Fisher weighting; LFA's weight pull is L2-SP and complementary.

Data-free distillation and shipped activation statistics. Lopes et al. (2017) release per-layer activation means and covariances as a data substitute. The family that followed optimizes inputs until the teacher reproduces sampled targets or stored BatchNorm statistics (Nayak et al., 2019; 2021; Yin et al., 2019; Haroush et al., 2019; Fang et al., 2021), or trains a generator (Chen et al., 2019; Micaelli & Storkey, 2019); for text, AS-DFD (Ma et al., 2020) generates pseudo-*embeddings* because tokens are discrete, PromptDFD (Ma et al., 2022) and DFKD-T^3 (Bai et al., 2023) let a pretrained language model write the transfer set, and Rashid et al. (2021) perturb out-of-domain text. The continual descendants (Smith et al., 2021; Gao et al., 2022; Tong et al., 2025; Krukowski et al., 2026) invert the frozen previous model per task and rehearse on the result. Every member *synthesizes surrogate inputs* and distills on them, so each sample must be a legitimate input to the whole network. For a generative LLM the input-space branch has no runnable form (no class head to invert, no gradient through a discrete token), and the branch in which a language model writes the corpus is the self-generated rehearsal below. LFA samples hidden states straight into sub-modules of the model being adapted: no input to reconstruct, no forward pass for a synthetic sample to survive, and no student. The inversion line's finding that feature *covariance*, not diagonal statistics, improves the synthesized set matches the one we reach for a different use of the statistic (App. B).

Replay-free adaptation for LLMs. On *claim* the closest neighbour is SSR (Huang et al., 2024) (with SDFT (Yang et al., 2024a) and synthetic continued pretraining (Yang et al., 2024b)): it needs no prior training data, has the model *generate* substitute text and trains on it, so its preservation signal is token-space; LFA generates and stores no text *during adaptation*. On *mechanism* the closest is GeRe (Zhang et al., 2025), which anchors activation consistency to a frozen reference but replays a buffer of *real* pretraining text and anchors whole-forward-pass states, not per-sub-module functions.

Anchored Learning (Wang et al., 2026a) shares our vocabulary but not our mechanism: having trained an ordinary SFT model, it re-adapts the base toward it in bounded steps, distilling toward an interpolation between the current model and that SFT model, so stability is a property of the step size toward the *adapted* model and not a pull toward the pretrained one. We compare to GeRe and to SSR empirically (§8); the Anchored Learning comparison is unrun (App. P).

Intruder dimensions and subspace-constrained LoRA. Shuttleworth et al. (2024) introduced intruder dimensions and showed LoRA and full fine-tuning differ spectrally at matched task performance; Biderman et al. (2024) report that LoRA "learns less and forgets less", Steele (2026) relate forgetting to the principal angle between task subspaces, and Jin et al. (2025); Gu et al. (2026) to rotation of dominant singular directions. OPLoRA (Xiong & Xie, 2025), O-LoRA (Wang et al., 2023), SC-LoRA (Luo et al., 2025), LoRA-Null (Tang et al., 2025) and MiLoRA (Wang et al., 2024a) constrain updates into, or initialize them within, a subspace deemed safe. The constraint is on the *update*, and what it precludes is encoding new knowledge by adjusting an existing direction, which a distribution-weighted penalty prices and does not prohibit. We run O-LoRA on its authors' code (App. F, finding vii). VeFA (Wang et al., 2025) is the closest peer on framing: it uses intruder dimensions, works in feature space, and rescales channels so the effective update stays inside the pretrained column space. It forbids new directions by construction, where LFA penalizes only *functional* change on $p(h)$, which permits non-destructive novelty and catches in-subspace distortion.

Recent continual LoRA, and the turn to plasticity. The most recent continual-LoRA methods stay in this family. KeepLoRA (Luo et al., 2026) projects the gradient away from the pretrained weights' principal subspace and from the accumulated dominant feature directions of earlier tasks, then initializes the adapter from what remains (CLIP and LLaVA); Qiao & Mahdavi (2024) project gradients to keep task subspaces orthogonal, and their SLAO (Qiao & Mahdavi, 2026) initializes each task's adapter orthogonally to the last and merges it into a single LoRA with a $1/\sqrt{i}$ scale (Llama and Qwen, 3–13B, scored by task accuracy on classification and SuperNI streams). All three constrain or initialize the *update*; none measures what the pretrained model could already do, which is what $\mathcal{L}_{\text{repr}}$ prices, and we ran none of them: O-LoRA is the family's one measured member here. A second line argues that once memory is plentiful forgetting is the lesser problem: Cho et al. (2026) find plain replay with a large buffer level with specialized methods, the difficulty moving to plasticity, and add parameter resets and weight averaging to a replay learner; Joudaki et al. (2026) trace loss of plasticity to frozen and cloned units. Our results agree that tuned replay is strong (§6); the setting differs, in that here the earlier text is unavailable, not merely costly, and our three-domain chains are too short to test plasticity over a long horizon (App. P). Class-incremental methods that shape the classifier's geometry, such as progressive neural collapse (Wang et al., 2026b), have no counterpart in open-ended generation.

Function-space capacity, feature replay, and merging. Concurrent with this work, HOPE (Mobahi & Bartlett, 2026) scores each neuron by a function-space norm under an estimated input distribution, the object λ prices, in order to *delete* it where $\mathcal{L}_{\text{repr}}$ scores its *perturbation* in order to *preserve* it; its continual arm, DEFT, thresholds the score into a frozen core and a plastic slack ($\lambda{=}\mu{=}\infty$ on the core, 0 on the slack) where §3 keeps two continuous knobs (why its zero-byte input distribution is unavailable here: App. B.1). Fitting a distribution over activations and sampling it to resist forgetting is established: generative feature replay (Liu et al., 2020) samples a *class-conditional* density over penultimate features into the classifier head, deep generative replay (Shin et al., 2017) does the same over *inputs*, latent replay (Pellegrini et al., 2019) and REMIND (Hayes et al., 2019) retain *real* past activations, FeTrIL (Petit et al., 2022) and SDC (Yu et al., 2020) keep class means, and FeCAM (Goswami et al., 2023) fits per-class Gaussians as a Mahalanobis classifier, finding, as we do for a different use (App. B), that covariance beats an isotropic treatment. In each the stored object is indexed by a class label, lives at one representation level, and is consumed by a classification loss; LFA ships an unconditional $p(h)$ at the input of *every* sub-module and consumes it with a teacher–student *function* discrepancy. Post-hoc merging (Ilharco et al., 2023; Yadav et al., 2023; Yu et al., 2024; Wang et al., 2024b) composes a fine-tuned checkpoint with its base after training, so it composes with LFA and does not substitute for it: interpolation cannot recover a direction the update already destroyed. We do not evaluate merging.

P EXTENDED LIMITATIONS

The sixteen items that §9 points here for.

1. **Attention scores are not anchored directly.** The mixing has no trainable parameters, so it is preserved exactly when Q and K are (Proposition 1); but $\mathcal{L}_{\text{repr}}$ prices squared error on their outputs over $p(h)$, not scores, so a change small there, or one off $p(h)$, can alter score mixing unseen by $\mathcal{L}_{\text{repr}}$. Only L2-SP would see it, and we have not looked for one.
2. **The destructive/non-destructive split is validated on one checkpoint family** (§A.4); other domains, scales and full-weight training are unrun.
3. **Training-method agnosticism is a design property.** Full-weight replication on Qwen3-0.6B is unrun.
4. **"Text-free" is a statement about *adaptation* time, not about how the statistic was built.** $p(h)$ is fitted once on a general proxy corpus, never on the adaptation domain, and is a data-derived artifact that someone must supply; "data-free" is reserved for the self-supplied mode, one setting (App. K); the zero-byte alternative is unavailable to this architecture (App. B.1).
5. **The artifact is evaluated on two models of one family and two domains.** Broader reuse is untested; no inversion-attack check has been run, a question the correlated form makes *more* acute, so we make no privacy claim; and a provider-side estimate over the true training distribution is an affordance, not a measured result.
6. **External baselines are reimplementations, except GeRe and O-LoRA** (App. L). Of the forced-segregation family only O-LoRA, of which OPLoRA (Xiong & Xie, 2025) is a projection-based refinement, is run, at one seed and upstream's λ_1; a separately tuned LoRA baseline is unrun.
7. **One family is unrepresented, and the closest rival is one rank deep.** Post-hoc merging is not compared (App. O); self-generated rehearsal is scored at one rank, one domain, settings carried over, with pairwise verdicts that do not all run one way (§8), and we have not priced a generation pass and a buffer against an artifact; Anchored Learning (Wang et al., 2026a), which releases no code and needs a preliminary SFT model and a second resident copy, is unrun.
8. **Self-supply is one seed, one domain, one model, and is not "LFA needs no data".** Only the $r{=}16$ self-supplied arm has a second seed, which replicates its level ($p{=}0.38$, $p{=}0.55$) but not its position against self-generated rehearsal; the measured cost is about 0.1 of judged quality at $r{=}32$; two pairings read differently under the robustness judge and are labelled where they appear; another size, domain or writer is untested (App. K.4).
9. **The accuracy signal is coarse.** At 0.6B the spread among preservation methods is ≤ 1 point of retained accuracy (App. E.1), so perplexity carries the preservation signal; other scales are untested.
10. **Two judges, 154 questions, one domain each.** The instruments, their resolution and their agreement are in App. M and App. E; what the second judge changes is classified in App. M. Both are large proprietary models; an open-weights judge and human agreement are untested.
11. **The judge tracks answer length, and training stages change answer length.** The control bounds the confound and does not remove length from the rubric; a rubric that scored brevity would rank some arms differently (App. G).
12. **One of the three question sets measures general answering, not domain knowledge.** Set-C's recall tier cannot separate any scored model from the base, and its per-tier decomposition exists only under the robustness judge (App. G); a longer dose, a larger rank and a different question generator are untested.
13. **Third-stage arms are single-seed with one exception, and nothing was re-tuned for the third domain**, so stage three measures how a *stage-two* operating point transfers, not best achievable behaviour on C; its borderline contrasts are paired over questions within one seed.
14. **The cost axis is three model sizes, one card, one implementation, one run per arm.** The percentages fix no scaling exponent, at 0.6B they are dominated by a launch-bound term that belongs to our implementation (App. I), replay was timed at 0.6B only, and full-weight LFA would carry a teacher copy or a precomputed store, neither measured.

15. **The Qwen3-1.7B replication is reduced-tuning.** One seed; knobs chosen on a perplexity ladder rather than a judged search, with replay's stage-two μ re-tuned once on perplexity; μ matched at 0.05 between LFA and LwF+μ where the 0.6B chains are not; the primary judge only, so no robustness reading exists; an equivalence margin carried from the 0.6B archive and applied post hoc, including to the 78-question sets; gradient accumulation unmatched with no step-matched control; nothing re-tuned for C; and the 0.6B feasibility gate not re-run (App. H).
16. **The generative instrument is sampled, and its lead is one reading of two.** A second draw of the same weights moves a count by up to 13 items at 0.6B and 17 at 1.7B, two draws of one checkpoint once separate under the paired test, and between 4.6% and 28.6% of items flip between draws; the after-C LFA lead over LwF+μ and replay+μ on GSM8K and IFEval does not extend to deterministic multiple choice, ties on short-answer recall, and is reversed on every perplexity of given text, so it licenses "more sampled chat-format task performance" and not "more reasoning knowledge" (App. H).