Capacity-Gated Forgetting in LoRA Fine-Tuning: Rank, Proximity, and Endogenous Replay in Medical LLMs

Akanksha Narula [1] **Aaditya Sharma** [1] **Dharya Jasuja** [1] **Aditya Dhawan** [1]

Abstract

LoRA fine-tuning on a narrow domain improves target behaviour but is widely reported to erase broad pretrained competence. Almost all such reports use a *single* random seed. We fine-tune `Qwen/Qwen3.5-9B` (the post-trained chat model) in 4-bit QLoRA on MedQA-USMLE and measure per-subject MMLU forgetting across all 57 subjects, then ask whether the result survives re-seeding. **It does not.** At our primary seed (42), the configuration looks textbook: mean forgetting $\bar{f} = 0.352$ at r=16, a clean rank effect ($\chi^2_3 = 166.5$, $p < 10^{-35}$), and an apparent 80% forgetting reduction from a base-model **Endogenous Replay** method ($\bar{f}$: 0.352→0.070). But a matched two-seed replication (seeds $0, 7$) of the replay sweep collapses the entire effect: *every* condition forgets only $\bar{f} \in [-0.022, -0.014]$ (i.e. essentially no net forgetting), and the no-replay→endogenous reduction shrinks from $+0.282$ to $+0.004$/$+0.008$. A faint directional ordering E11 < E10 < E09 < E02 remains detectable (per-seed Friedman $p = 0.017, 0.004$) but at $\sim$0.005 magnitude, far below the 0.04 pre-registered benchmark threshold. The same configuration thus yields 0.35 or $\approx$0 forgetting depending only on the seed. Our contribution is therefore a cautionary, reproducible result: *single-seed LoRA-forgetting measurements — and the mitigation gains derived from them — are not robust*, a problem for a literature that reports them as benchmarks. We retain Capacity-Gated Forgetting (CGF) and Endogenous Replay as seed-42 hypotheses, show exactly where they fail to replicate, and release all data, code, and the matched multi-seed protocol so the phenomenon can be probed further.

[1]Punjab Engineering College, Chandigarh, India. Correspondence to: Akanksha Narula <akanksha.narula8@gmail.com>.

Accepted to the 1st Workshop on Combining Theory and Benchmarks, CTB@ICML 2026, Seoul, South Korea. Copyright 2026 by the author(s).

1. Introduction

QLoRA makes medical adaptation cheap: load a 9B model in 4-bit quantisation, attach low-rank adapters, fine-tune on MedQA-USMLE, and deploy. The cost is forgetting: the fine-tuned model improves on the target domain while losing accuracy on broad evaluations such as MMLU.

The literature gives incompatible explanations. Biderman et al. (2023) emphasise model scale. Luo et al. (2024) argue that forgetting follows semantic proximity to the fine-tuning distribution. Biderman et al. (2024) show that LoRA rank strongly controls the learn–forget trade-off. These claims are hard to compare because they vary base model, corpus, benchmark, and rank range simultaneously.

We fix base model, fine-tuning corpus, and evaluation harness, then vary one axis at a time. The result is a compact battery on `Qwen/Qwen3.5-9B`, `GBaker/MedQA-USMLE-4-options-hf`, and MMLU's 57 subjects. Per-subject evaluation separates mean forgetting from the medical/non-medical proximity gap.

(C1) The central finding: seed-dependence. Under a fixed, matched protocol, the same r=16 MedQA configuration produces $\bar{f} = 0.352$ at seed 42 but $\bar{f} \approx -0.015$ at seeds 0 and 7 — a more than twenty-fold swing in measured forgetting driven by the random seed alone (§4.7). Catastrophic forgetting here is not a stable property of the configuration; it is a high-variance, seed-sensitive event. Because the field almost universally reports single-seed forgetting numbers, this is a direct caution about the reliability of those numbers and of any mitigation gain measured against them.

(C2) Capacity-Gated Forgetting (CGF), as a seed-42 hypothesis. The seed-42 rank sweep motivates a capacity ratio $\rho = r_{\mathrm{LoRA}}/d^*(D_{\mathrm{ft}})$ separating a uniform from a proximity-structured forgetting regime, with transition window $r^* \in (16, 64)$, complementary to the subspace-angle account of Steele (2026). We present CGF as a hypothesis generated by single-seed data; its rank sweep (E01–E04) has not yet been re-seeded, so we do not claim it as established (§5).

(C3) Endogenous Replay, and where it breaks. A replay

method whose rehearsal corpus is sampled from the base model, minimising a KL anchoring objective exactly (§6.1). At seed 42 it appears to remove 80% of forgetting; across seeds 0/7 the absolute benefit falls to ~ 0.005 and only a faint ordering survives. We report both the seed-42 mechanism and its non-replication, rather than the headline alone.

(C4) Supporting evidence and infrastructure. A per-subject 57-way forgetting battery on a modern hybrid (Gated DeltaNet/attention) 9B model, an ℓ_2-SP frontier (E12), a K0 target-accuracy check (N=200, all conditions 0.76), and a released, fully specified multi-seed protocol so the result can be reproduced or overturned.

We keep all reported measurements to E01–E13. K1–K4 are planned cross-task/cross-family checks and are not used as evidence here.

2. Related Work

2.0.1. Forgetting, scale, and rank

Biderman et al. (2023) connect large-model behaviour to predictable capacity effects. In the LoRA setting, Biderman et al. (2024) show that LoRA learns less and forgets less than full fine-tuning, and that low-rank perturbations preserve more off-domain behaviour. We build on this rank-centric view but ask when rank produces uniform versus proximity-structured forgetting.

2.0.2. Semantic proximity

Luo et al. (2024) argue that continual fine-tuning harms tasks close to the fine-tuning distribution more than distant tasks. We operationalise proximity by comparing medical and non-medical MMLU subjects and by measuring MiniLM subject-to-corpus similarity (Reimers & Gurevych, 2019).

2.0.3. Subspace geometry and forgetting

Steele (2026) show that forgetting in LoRA is governed by the minimum principal angle $\theta_{\min}$ between task gradient subspaces: $\mathcal{F} = \alpha(1 - \cos^2 \theta_{\min}) + \beta$. At high subspace angles forgetting is approximately rank-invariant; at low angles rank begins to matter. Their framework is validated on synthetic tasks, ViT-LoRA (Split-CIFAR100), and RoBERTa-LoRA (sequential GLUE); the 9B-LLM medical-domain regime is not examined. The CGF capacity ratio ρ operates at a complementary granularity. Concurrently, Ahmad et al. (2026) reach similar conclusions about the role of subspace structure in low-rank PEFT forgetting; our contribution is distinguished by the capacity-ratio framing and the QLoRA medical-domain instantiation.

2.0.4. Standard continual learning methods

Classical regularisation includes Elastic Weight Consolidation (EWC; Kirkpatrick et al. 2017), Learning without Forgetting (LwF; Li & Hoiem 2018), and ℓ_2-SP (Xuhong et al., 2018). We include ℓ_2-SP as an experimental baseline (E12; §7) because Endogenous Replay's KL anchoring objective is geometrically closest to it: both anchor toward θ_0, but Endogenous Replay does so in output (KL) space rather than weight space (§6.1). EWC and LwF remain future baselines (K4).

2.0.5. Replay and pseudorehearsal

LAMOL showed that language models can generate pseudo-examples for lifelong learning (Sun et al., 2020). Most directly related is Huang et al. (2024), who propose Self-Synthesized Rehearsal (SSR): the base LLM generates synthetic instances via in-context learning, refined by the latest checkpoint and selected for diversity before rehearsal, evaluated on NLP task sequences under full-parameter continual fine-tuning. Endogenous Replay differs in three respects: (i) it operates in the QLoRA regime on a 9B medical model rather than full-parameter SFT; (ii) it uses direct base-model sampling without an ICL-prompting or refinement step, reducing compute to a single forward-sampling pass; and (iii) it is grounded in the KL anchoring argument of §6.

3. Experimental Setup

3.0.1. Model and data

The primary battery (E00–E13) fine-tunes `Qwen/Qwen3.5-9B`, the post-trained (chat) checkpoint of the Qwen3.5 family released in March 2026, loaded in 4-bit NF4 with double quantisation and bf16 compute (QLoRA; Dettmers et al. 2023). We emphasise the exact identity because it matters for interpretation: the evaluation applies the model's chat template (§3.0.4), so "forgetting" is measured on the post-trained model, not on the separate `Qwen/Qwen3.5-9B-Base` checkpoint. The fine-tuning corpus D_{ft} is the training split of `GBaker/MedQA-USMLE-4-options-hf` (Jin et al., 2021): $\approx$4000 examples formatted as question–answer pairs with loss on answer tokens only, maximum sequence length 1024.

3.0.2. Hybrid architecture and LoRA coverage

`Qwen/Qwen3.5-9B` uses a hybrid 32-layer stack interleaving Gated DeltaNet (linear-attention) and full-attention layers in a 3:1 pattern: 24 DeltaNet layers and 8 standard full-attention layers. LoRA adapters are attached to the four attention projections (q, k, v, o) of the 8 full-attention layers only; DeltaNet projections and MLP blocks carry no LoRA adapters. The effective rank budget therefore applies to

32 attention-projection matrices, not the full 9B parameter count. Any capacity-ratio interpretation (§5) should be read relative to the LoRA-covered parameter count. Extending LoRA to DeltaNet projections or MLP blocks could shift the transition window and is a natural extension outside the current battery.

3.0.3. Adapter and optimisation

LoRA (Hu et al., 2022) is attached as above with $\alpha = 2r$, dropout 0.05, per-device batch size 4, gradient accumulation 4 (effective batch 16), learning rate 2×10^{-4}, cosine schedule, warmup ratio 0.05, and seed 42. (The replay runs E09–E11 were executed in a separate notebook with effective batch 8 and 50 warmup steps; we flag this minor cross-notebook inconsistency in §10.) The ℓ_2-SP baseline (§7) reuses this configuration and adds a weight-space penalty $\frac{\lambda}{2}\|\Delta\theta\|_2^2$ on LoRA parameters only.

3.0.4. Evaluation

MMLU's 57 subjects (Hendrycks et al., 2021) are evaluated in a 0-shot, chat-template generation protocol: the prompt is formatted with the `Qwen/Qwen3.5-9B` chat template and the model generates an answer letter A–D with greedy decoding (`max_new_tokens=5`, temperature 0). Forgetting for subject s is $f_s = \mathrm{acc}_{\mathrm{base}}(s) - \mathrm{acc}_{\mathrm{ft}}(s)$; positive values mean the fine-tuned model forgot. The seed-42 pre-fine-tuning baseline mean accuracy is $\overline{\mathrm{acc}}_{\mathrm{base}} = 0.776$.

3.0.5. Medical partition

Nine MMLU subjects form the medical group used throughout the analysis: *anatomy*, *clinical_knowledge*, *college_biology*, *college_medicine*, *high_school_biology*, *human_aging*, *medical_genetics*, *professional_medicine*, and *virology*. The remaining 48 subjects are non-medical. (This is the partition actually used to compute every med/non-med number below; an earlier draft listed a slightly different set, which we correct here.)

Table 1. Experiment configurations and mean forgetting at seed 42. $\bar{f}$ is averaged over 57 MMLU subjects; $\mathrm{SE}_{\bar{f}} = \hat{\sigma}/\sqrt{57}$; med and non-med are subset means over the partition of §3. Rank sweep: $\chi^2_3 = 166.5$, $p < 10^{-35}$. Replay sweep: $\chi^2_3 = 167.85$, $p < 10^{-35}$; every adjacent pair separates at $p < 10^{-9}$ (paired Wilcoxon). E11 gives an 80% reduction over E02 and ≈50% over E10. *All values are single-seed (seed 42); these magnitudes do not replicate at seeds 0/7 (§4.7).*

ID	r	steps	replay	$\bar{f}$	$\mathrm{SE}_{\bar{f}}$	med	non-med
E01	4	500	none	0.163	0.009	0.159	0.163
E02	16	500	none	0.352	0.013	0.348	0.352
E03	64	500	none	0.540	0.018	0.571	0.535
E04	128	500	none	0.545	0.018	0.577	0.539
E05 (GSM8K)	16	500	none	0.001	0.003	–	–
E05b (Code)	16	500	none	−0.005	0.003	–	–
E06	16	100	none	0.435	0.015	0.435	0.434
E07	16	200	none	0.448	0.015	0.446	0.449
E08	16	1000	none	0.401	0.013	0.412	0.399
E09	16	500	50 real	0.180	0.008	0.182	0.180
E10	16	500	100 real	0.142	0.009	0.152	0.140
E11	**16**	**500**	**100 endogenous**	**0.070**	**0.007**	**0.077**	**0.068**

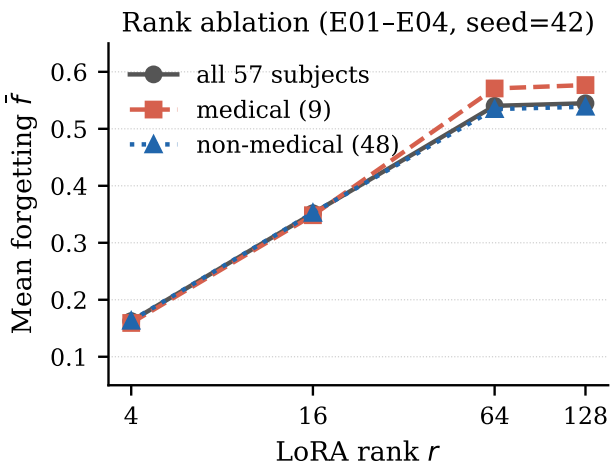

Figure 1. Rank ablation (E01–E04, seed 42). Mean forgetting rises with LoRA rank and plateaus between r=64 and r=128.

4. Results

4.1. Low-rank forgetting is domain-agnostic

At rank 4 and rank 16, medical and non-medical subjects forget within 0.004 of each other. E01 gives 0.159 versus 0.163; E02 gives 0.348 versus 0.352. A paired bootstrap over the 9 medical against the 48 non-medical subjects does not reject equal means. This is inconsistent with a proximity-first account in the low-rank regime.

The domain controls support this interpretation. E05 trains on GSM8K at rank 16 and gives $\bar{f} = 0.001$; E05b trains on CodeAlpaca and gives −0.005. The 0.352 forgetting in E02 is therefore attributable to the MedQA signal, not to the act of fine-tuning alone.

4.2. Rank is the dominant axis

Across E01–E04 (Figure 1), $\bar{f}$ rises from 0.163 at r=4, to 0.352 at r=16, to 0.540 at r=64, and plateaus at 0.545 at r=128. A Friedman test across the 57 subjects gives $\chi^2_3 = 166.5$, $p < 10^{-35}$. The rank ordering is stable in a coarse sense: all 57 of 57 subjects forget more at $r \geq 64$ than at $r \leq 16$. Within the plateau the peak is split — 29 subjects forget most at r=128 and 28 at r=64 — consistent with the near-tie between E03 and E04. The rank-driven span (0.382) is roughly four times the largest step-driven span (0.096, Table 1).

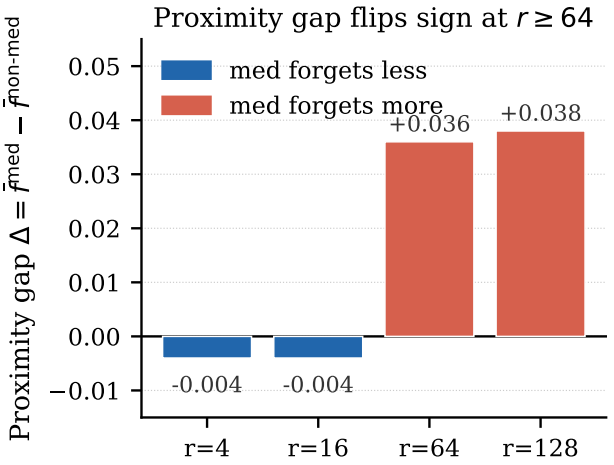

Figure 2. Proximity gap Δ_i across rank conditions. Positive values mean medical subjects forget more than non-medical subjects.

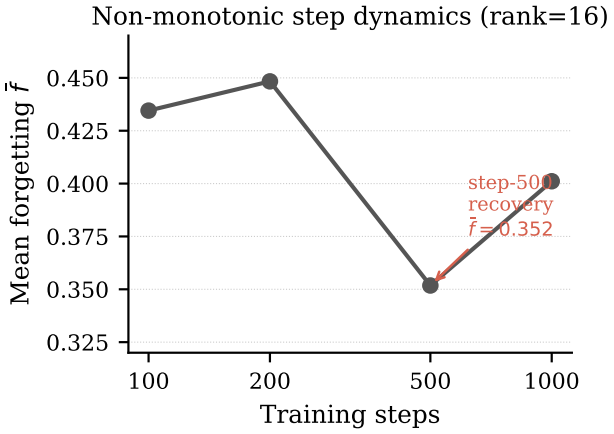

Figure 3. Step dynamics at rank 16. Mean forgetting peaks near step 200, recovers at step 500, and rises again at step 1000.

4.3. A proximity gap appears only above rank 16

Define $\Delta_i = \bar{f}_i^{\text{med}} - \bar{f}_i^{\text{non-med}}$. From Table 1, $\Delta_{E01} = -0.004$, $\Delta_{E02} = -0.004$, $\Delta_{E03} = +0.036$, and $\Delta_{E04} = +0.038$ (Figure 2). The sign flips between r=16 and r=64.

Our data reconcile Luo et al. (2024) and Biderman et al. (2024): Luo's proximity account is correct in the supercritical regime ($r \geq 64$); Biderman's rank-dominance result holds across both regimes. The subspace-geometry framework of Steele (2026) provides a candidate geometric mechanism. CGF packages these accounts under a single capacity-ratio threshold ρ^*. We stress that the gap is small in absolute terms (≤ 0.04) and single-seed; the continuous-distance version of the claim does *not* hold (§4.6).

4.4. Forgetting is non-monotonic in training steps

At fixed rank 16 (Figure 3), forgetting moves from 0.435 at 100 steps, to 0.448 at 200, dips to 0.352 at 500, and rises to 0.401 at 1000. The 500-step dip is robust per-subject relative to the shorter schedules: all 57 of 57 subjects forget less at 500 steps than at either 100 or 200 steps, and 51 of 57 forget less at 500 than at 1000. More steps are therefore not a safe monotone proxy for retention.

Candidate mechanism. The 500-step dip coincides with the trough of the cosine LR schedule. At step 500 the effective LR has reached its minimum, temporarily attenuating gradient magnitude; the adapter weights remain closest to initialisation, producing a transient forgetting minimum. At step 1000 training runs past the intended schedule length at near-zero but nonzero LR, allowing further drift. Because the dip appears in all 57 subjects regardless of medical relevance, a schedule-driven account is more parsimonious than a topic-specific one; we cannot separate the two causally (§10).

4.5. Endogenous Replay outperforms real replay (seed 42)

Fixing rank 16 and 500 steps (Figure 4), no replay gives $\bar{f} = 0.352$ (E02); 50 real examples drop it to 0.180 (E09); 100 real to 0.142 (E10); and 100 endogenous examples to 0.070 (E11). This is an 80% reduction over no replay and ≈50% over matched-budget real replay.

The within-condition ordering E11 < E10 < E09 < E02 is robust under paired Wilcoxon signed-rank tests on per-subject forgetting (n=57): every adjacent pair separates at $p < 10^{-9}$ (E09 vs. E02: $p = 7.5 \times 10^{-11}$; E10 vs. E09: $p = 1.1 \times 10^{-9}$; E11 vs. E10: $p = 5.1 \times 10^{-11}$). The replay-sweep Friedman test gives $\chi^2_3 = 167.85$, $p < 10^{-35}$. *These statistics establish a within-seed ordering only. At seeds 0 and 7 the effect collapses to ∼0.005 (§4.7); the impressive seed-42 separation is not a robust property of the method.*

4.6. Honest negative: no continuous distance gradient

MiniLM subject-to-corpus similarity is essentially uncorrelated with forgetting at every rank: Pearson $r = +0.065$ (p=0.63) at r=4, -0.036 (p=0.79) at r=16, $+0.036$ (p=0.79) at r=64, and $+0.045$ (p=0.74) at r=128. We therefore observe a binary medical/non-medical partition at high rank, not a reliable continuous distance gradient at embedding resolution.

4.7. The headline does not replicate across seeds

We re-ran the replay sweep E02/E09/E10/E11 on the chat model `Qwen/Qwen3.5-9B` at seeds 0 and 7 under the matched protocol of §3.0.4 (4-bit NF4, r=16, α=32, 500 steps, chat-template evaluation with `enable_thinking=False`; an earlier replication mistakenly used the `-Base` checkpoint and a non-chat scorer

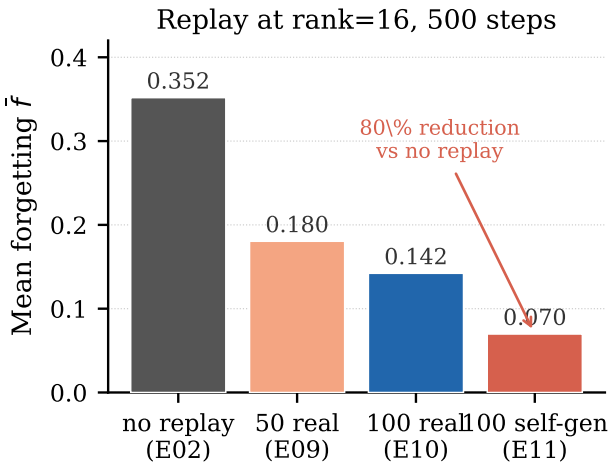

Figure 4. Replay at rank 16, 500 steps, *seed 42*. 100 endogenous examples achieve an 80% forgetting reduction over no replay. This effect does not replicate at seeds 0/7 (Figure 5).

Table 2. Replay sweep across seeds (mean per-subject forgetting $\bar{f}$, n=57). Seed 42 is the paper battery; seeds 0/7 are the matched chat-model replication. The configuration yields $+0.352$ or ≈ -0.015 forgetting depending only on the seed, and the no-replay→endogenous reduction (last column) falls from $+0.28$ to ~ 0.005. Per-condition means are rounded to 3 dp; the E02−E11 column is computed from unrounded means (seed 7: $-0.0135 - (-0.0211) = 0.0075$).

Seed	E02	E09	E10	E11	E02−E11
42 (paper)	+0.352	+0.180	+0.142	+0.070	+0.282
0	−0.016	−0.022	−0.017	−0.020	+0.004
7	−0.014	−0.015	−0.016	−0.021	+0.008

and is discarded). The seed-42 magnitudes do not survive (Table 2, Figure 5).

Forgetting itself collapses. At seeds 0 and 7, *every* condition — including no-replay E02 — shows near-zero net forgetting, $\bar{f} \in [-0.022, -0.014]$, versus $+0.352$ for E02 at seed 42. Whatever drives the large seed-42 drop, it does not recur at two other seeds of the same configuration.

The replay benefit collapses with it. Because there is almost nothing to forget, the no-replay→endogenous reduction falls from $+0.282$ (seed 42) to $+0.0042$ (seed 0) and $+0.0075$ (seed 7). A faint directional ordering E11 < E10 < E09 < E02 is still statistically detectable (per-seed Friedman $\chi^2_3 = 10.2/13.1$, $p = 0.017/0.004$; Wilcoxon E11 vs. E02 $p = 0.04/<10^{-3}$), but its magnitude (~ 0.005) is two orders below the seed-42 effect and below the 0.04 benchmark threshold our analysis code preregistered. The strict per-subject ordering holds for only 4–5% of subjects.

Seed or protocol? One caveat keeps this from being a pure seed effect: the seed-42 *E02* number (0.352) was produced by the rank-sweep configuration (`nbA`: effective batch 16, cosine schedule, sequence length 1024), whereas the seed-0/7 runs use the replay configuration (effective batch 8, linear schedule, sequence length 256). The cross-seed contrast therefore conflates seed with training configuration. This does not rescue the headline — it means the reported 0.352 is sensitive to *both* seed and protocol, neither of which was varied when it was presented as a benchmark. Disentangling the two (seed 42 under the replay config; seeds 0/7 under the rank config) is the immediate next step; the protocol is released in Appendix D.

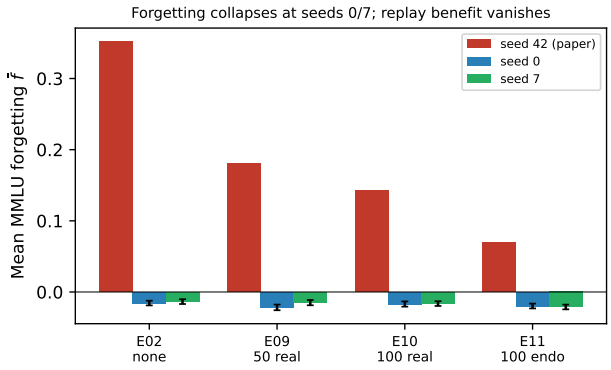

Figure 5. Replay sweep at r=16, 500 steps. The large seed-42 forgetting and its monotone replay reduction (red) do not appear at seeds 0 and 7 (blue/green), where all conditions sit near zero. The mitigation gain is a property of the seed, not the method.

Format caveat. Forgetting is scored through the chat template, so part of the seed-42 drop may be degradation of instruction-following *format* rather than knowledge loss (Open Problem 3); the near-zero seed-0/7 result is consistent with little genuine knowledge loss at those seeds.

5. Capacity-Gated Forgetting

Let θ_0 denote the base parameters and $\Delta\theta$ the LoRA-induced perturbation after fine-tuning on D_{ft}. We define the capacity ratio

$$\rho = \frac{r_{\text{LoRA}}}{d^*(D_{\text{ft}})}. \tag{1}$$

The intrinsic dimension $d^*(D_{\text{ft}})$ is the minimum random-subspace dimension required to solve the task at near-full performance (Aghajanyan et al., 2021). Equation (1) is a hypothesis variable: we do not measure d^* for MedQA-USMLE, and the CGF claim does not rest on identifying the exact value of ρ^*. The empirical content is the two-regime structure observed in Δ_i across the seed-42 rank sweep, with transition window $r^* \in (16, 64)$ for `Qwen/Qwen3.5-9B` under LoRA covering the 32 full-attention projection matrices (§3).

Sub-critical ($\rho < \rho^*$). The adapter rank is insufficient to span the intrinsic task subspace; optimisation cannot

Table 3. CGF regime classification by rank condition (seed 42). Schematic ρ values require measuring $d^*(D_{\text{ft}})$ (Open Problem 1).

	$r = 4$	$r = 16$	$r = 64$	$r = 128$
Regime	sub-crit.	sub-crit.	super-crit.	super-crit.
Forgetting	uniform	uniform	structured	structured
Δ_i	-0.004	-0.004	$+0.036$	$+0.038$

Algorithm 1 Endogenous Replay (SSRA)

Require: base model θ_0; anchor prompts $\{p_i\}_{i=1}^N$; task corpus D_{ft}
Ensure: LoRA adapter trained on $D_{\text{ft}} \cup \mathcal{R}$
1: $\mathcal{R} \leftarrow \emptyset$
2: **for** $i = 1, \ldots, N$ **do**
3: sample $y_i \sim p_{\theta_0}(\cdot \mid p_i)$
4: $\mathcal{R} \leftarrow \mathcal{R} \cup \{(p_i, y_i)\}$
5: **end for**
6: train the LoRA adapter on $D_{\text{ft}} \cup \mathcal{R}$

selectively concentrate on task-relevant directions, and gradient pressure spreads roughly uniformly across pretrained representations. Forgetting is high in aggregate but structurally indifferent to topic proximity — consistent with the high-subspace-angle, rank-approximate-invariance result of Steele (2026) and with E01–E02.

Super-critical ($\rho \geq \rho^*$). The adapter has capacity slack beyond the task subspace; gradient descent allocates this slack to directions correlated with the task distribution, and forgetting inherits that selectivity, concentrating on pretrained representations in the same topic cluster as D_{ft} — consistent with Luo et al. (2024), the low-subspace-angle regime of Steele (2026), and E03–E04. The structuring is observable as a binary partition (medical vs. non-medical), not a continuous gradient at MiniLM resolution (§4.6).

Because the proximity gap is small and single-seed, CGF is offered as a falsifiable hypothesis, not an established law. Informal geometric and Bayesian arguments are given in Appendix B; the decisive tests are K1–K2 (§9).

6. Endogenous Replay

Endogenous Replay (Self-Synthesized Rehearsal Anchoring; SSRA — distinct from the SSR of Huang et al. 2024, which uses ICL-based generation and LLM refinement in a full-parameter SFT setting) samples a rehearsal set $\mathcal{R}$ from the base model θ_0 on anchor prompts $\{p_i\}_{i=1}^N$ and trains on $D_{\text{ft}} \cup \mathcal{R}$.

6.1. KL anchoring

For an anchor distribution $\mathcal{P}_{\text{anchor}}$, the ideal retention objective constrains the fine-tuned model to stay close to the base model:

$$\mathcal{L}_{\text{anchor}}(\theta) = \mathbb{E}_{x \sim \mathcal{P}_{\text{anchor}}} \mathrm{KL}(p_{\theta_0}(\cdot \mid x) \,\|\, p_\theta(\cdot \mid x)) \,. \quad (2)$$

Sampling $y \sim p_{\theta_0}(\cdot \mid x)$ turns the cross-entropy term into an unbiased Monte Carlo estimator of Equation (2), up to the θ-independent entropy of the base model. Around θ_0,

$$\mathcal{L}_{\text{anchor}}(\theta_0 + \Delta\theta) = \tfrac{1}{2} \Delta\theta^\top F_{\text{anchor}} \Delta\theta + O(\|\Delta\theta\|^3), \quad (3)$$

where F_{anchor} is the Fisher matrix under $\mathcal{P}_{\text{anchor}}$. Equations (2)–(3) explain why endogenous samples act as a sample-based anisotropic Fisher penalty.

Relationship to EWC and ℓ_2-SP. EWC (Kirkpatrick et al., 2017) imposes a Fisher-weighted penalty centred at the previous task's optimum; ℓ_2-SP (Xuhong et al., 2018) regularises toward the pretrained initialisation with a uniform ℓ_2 ball. Endogenous Replay sits between them: it anchors at θ_0 (like ℓ_2-SP) but with an anisotropic Fisher penalty determined by the anchor distribution (like EWC), in output (KL) space rather than weight space. In the QLoRA regime this matters: θ_0 is the 4-bit quantised anchor, and an ℓ_2 penalty in weight space is distorted by the dequantisation grid, which the output-space objective avoids.

Table 4. K0 target-accuracy check on the chat model `Qwen/Qwen3.5-9B` (r=16, 500 steps, N_{test}=200). All three conditions score 0.76; binomial SE at N=200 is ≈3.0 pp, so the conditions are within noise. The point is the falsifier — E11 does *not* fall below no-replay. MMLU $\bar{f}$ reproduced from Table 1. (A separate K0 run on the `-Base` checkpoint gives 0.78/0.775/0.785; we report the chat-model run here because it matches the model used for E01–E12.)

Condition	Replay	MedQA acc. (K0)	MMLU $\bar{f}$
E02	none	0.76	0.352
E10	100 real	0.76	0.142
E11	**100 endogenous**	**0.76**	**0.070**

6.2. Target-task accuracy check (E13/K0)

We measure MedQA target accuracy for E02, E10, and E11 at the matched r=16, 500-step configuration, scoring the argmax answer letter over the first N_{test}=200 examples of the `GBaker/MedQA-USMLE-4-options-hf` test split (Appendix A).

All three conditions score 0.76 on the chat model (Table 4). With ≈3.0 pp SE at N=200 these differences are not resolvable, but the result serves as a falsifier check: E11 does *not* sacrifice MedQA accuracy relative to no replay (the required condition for a Pareto improvement); a drop exceeding 5 pp would have refuted it, and the observed change is 0 pp. The flat profile is consistent with the KL-anchoring argument of §6.1. Definitive ordering would require $N_{\text{test}} \geq 1000$, left to future work.

Table 5. ℓ_2-SP Pareto sweep (E12) on `Qwen/Qwen3.5-9B`, r=16, 500 steps, seed 42. Med acc. is mean fine-tuned accuracy over the nine medical MMLU subjects (§3 partition; retained-domain proxy). ℓ_2-SP traces a monotone retention/learning frontier. Values are from `l2sp_pareto_summary.csv`.

Method	$\bar{f}\downarrow$	$\mathrm{SE}_{\bar{f}}$	non-med $\bar{f}\downarrow$	med acc.↑
E02 (no regularisation)	0.352	0.013	0.352	0.464
ℓ_2-SP, λ=10^{-3}	0.042	0.005	0.047	0.747
ℓ_2-SP, λ=10^{-2}	0.023	0.004	0.026	0.758
ℓ_2-SP, λ=10^{-1}	0.009	0.003	0.011	0.763
E11 (Endogenous Replay)	*0.070*	*0.007*	*0.068*	—

7. ℓ_2-SP Pareto Baseline (E12)

We run an ℓ_2-SP (Xuhong et al., 2018) baseline on `Qwen/Qwen3.5-9B`, the same model and configuration as E02 (seed 42), adding a weight-space penalty $\frac{\lambda}{2}\|\Delta\theta\|_2^2$ on LoRA parameters only and sweeping $\lambda \in \{10^{-3}, 10^{-2}, 10^{-1}\}$. The no-regularisation reference is E02 ($\bar{f}$=0.352; fine-tuned mean accuracy on the nine medical subjects =0.464; the pre-fine-tuning medical-subject baseline is ≈0.803). Appendix C gives implementation details.

E02 is strictly Pareto-dominated by every ℓ_2-SP point: $\bar{f}$ falls from 0.352 to 0.009 as λ rises two orders of magnitude, while medical-subject accuracy rises from 0.464 to 0.763, approaching the pre-fine-tuning medical baseline of ≈0.803. The frontier is monotone in both directions. For reference, E11 ($\bar{f}$=0.070) lies *above* all three ℓ_2-SP points on the forgetting axis — between the no-regularisation baseline E02 (0.352) and the weakest ℓ_2-SP setting λ=10^{-3} (0.042). At this single seed, weight-space ℓ_2-SP therefore attains lower forgetting than Endogenous Replay. Because E11 and the ℓ_2-SP sweep were run in separate pipelines, were not re-seeded, and inherit the seed-sensitivity of §4.7, we report this as positioning on a shared conceptual frontier, not a controlled head-to-head.

8. Discussion

8.1. What is robust and what is not

The one finding that survives re-seeding is the negative one: under a matched protocol, mean forgetting at r=16 ranges from ≈ 0 (seeds 0,7) to 0.352 (seed 42), so neither the absolute forgetting nor the replay benefit is a stable property of the configuration (§4.7). The seed-42 patterns — the rank sweep (χ^2_3=166.5), the weak proximity gap ($\leq$ 0.04), the non-monotonic step dynamics, and the 80% replay reduction — are real measurements at that seed but have either not been re-seeded (rank, proximity, steps; E01–E08 ran only at seed 42) or have been re-seeded and collapsed (replay). We therefore advance them as hypotheses, not benchmarks, and treat the seed-sensitivity itself as the result. A practical corollary: any LoRA-forgetting number reported at a single seed — ours or others' — should be assumed to carry seed variance of the same order as its mean until shown otherwise.

8.2. Approximate capacity bound from the rank transition

If ρ^*=1, the observed transition between r=16 (uniform) and r=64 (structured) implies $d^*(D_{\mathrm{ft}}) \in (16, 64)$ for MedQA-USMLE on this model and LoRA coverage. This is an approximate bound, not a measurement, and motivates the sub-1-GPU-hour d^* estimation of Open Problem 1.

8.3. Deployment implications

Taken at face value (and only at seed 42 on the chat model), practitioners in the sub-critical regime ($r \leq 16$) can expect substantial but uniformly distributed forgetting, while those in the super-critical regime ($r \geq 64$) should monitor task-adjacent MMLU subjects most closely. Because forgetting is measured through the chat template, the safe practical recommendation is to (a) evaluate the *deployed* model variant with the *deployed* prompting format, and (b) treat replay budgets as hypotheses to validate per model and seed.

9. Open Problems and Falsifiers

Open Problem 1. *Measure $d^*(D_{\mathrm{ft}})$ directly for MedQA-USMLE using random-subspace optimisation (Aghajanyan et al., 2021). If d^* is far outside the rank window implied by E01–E04, CGF's capacity-ratio interpretation is wrong.*

Open Problem 2. *Identify a representation-space distance that predicts the within-medical forgetting gradient, not just the medical/non-medical partition. MiniLM similarity is insufficient at the current resolution.*

Open Problem 3. *Separate instruction-format degradation from knowledge forgetting. Because forgetting is scored through the chat template, an unknown fraction of the measured drop may be answer-format breakage rather than knowledge loss. Re-scoring the same checkpoints with a format-robust scorer (constrained A–D decoding without the chat template) alongside the chat-template scorer would quantify the split.*

K0 is separate from Table 6: it checks MedQA target accuracy for E02, E10, and E11 rather than testing a CGF prediction.

10. Limitations

10.0.1. Only the replay sweep is re-seeded; seed and protocol are confounded

The multi-seed evidence (§4.7) covers the replay sweep (E02/E09/E10/E11) at three seeds. The rank, proximity, and

Table 6. Planned validation checks. These experiments have not been run and are not used as evidence here.

Test	Prediction	Falsified by
K1 (cross-family)	Same fine-tuning data on an attention-only 9B model (e.g. `google/gemma-2-9b`) should produce a comparable critical-rank window ($r \in [16, 64]$ for MedQA-USMLE).	critical rank outside $[16, 64]$
K2 (task complexity)	Simpler tasks should shift the critical rank downward; harder medical QA upward.	no rank-window shift across tasks
K3 (anchor scaling)	Anchor benefit should improve sublinearly with budget and prompt-set diversity.	no monotone budget or diversity effect
K4 (EWC/LwF, multi-seed)	Endogenous Replay should match or beat EWC and LwF at matched budget across ≥ 3 seeds (cf. ℓ_2-SP in E12).	any baseline below E11; or no cross-seed effect

step results (E01–E08) were run only at seed 42 and have not been re-seeded; given the variance we document, they should be read as single-seed observations. Two seeds (0,7) is also a small sample for estimating variance. Finally, as noted in §4.7, the seed-42 E02 number came from the rank-sweep configuration while the seed-0/7 runs used the replay configuration, so the cross-seed contrast is confounded with protocol; the two factors should be separated by running each configuration at all three seeds (Appendix D).

10.0.2. Pipeline inconsistencies

The rank/step runs (effective batch 16, warmup ratio 0.05) and the replay runs (effective batch 8, 50 warmup steps) were executed in separate notebooks; the K0 and ℓ_2-SP runs use additional throughput optimisations. These do not change rank, step count, or training data, but a single unified pipeline is preferable and is future work.

10.0.3. Single base family

`Qwen/Qwen3.5-9B` only. The hybrid DeltaNet/attention architecture is atypical; K1 addresses cross-family replication on an attention-only model.

10.0.4. Step dynamics is suggestive, not causal

The 500-step dip coincides with the cosine LR trough; we do not separate schedule-driven from signal-driven causation. A flat-LR comparator is the cleanest test and is left to future work.

10.0.5. d^* is invoked, not measured

CGF is stated in terms of $d^*(D_{\text{ft}})$ but we do not estimate it. The empirical content is the rank window $r \in (16, 64)$; the ρ formulation is an interpretive overlay.

10.0.6. Proximity metric is crude

Subject-to-corpus proximity uses `all-MiniLM-L6-v2` (Reimers & Gurevych, 2019); the near-zero Pearson correlations (§4.6) are consistent with no low-rank proximity effect but also with an insufficient metric.

10.0.7. K0 power, and CL baselines

K0 at N=200 ($\pm$2.9 pp) confirms no accuracy sacrifice but cannot rank conditions. EWC and LwF are K4 future baselines.

11. Conclusion

A controlled per-subject battery on `Qwen/Qwen3.5-9B`, MedQA-USMLE, and MMLU makes LoRA forgetting at a single seed (42) look orderly: rank dominates, a weak proximity gap appears above r=16, step dynamics are non-monotonic, and Endogenous Replay removes 80% of forgetting at no target-accuracy cost. The headline result of this paper is that this order is fragile. A matched two-seed replication of the replay sweep finds essentially no forgetting and no replay benefit at seeds 0 and 7 — the same configuration that forgets 0.352 at seed 42. Catastrophic forgetting in this setting is therefore better described as a high-variance, seed-sensitive event than as a property of the rank/step/replay configuration, and single-seed measurements of it (ours and the wider literature's) should be treated with corresponding caution. We retain CGF and Endogenous Replay as falsifiable seed-42 hypotheses, document exactly where they do not replicate, and release the matched protocol (Appendix D) so that re-seeding the rank sweep, disentangling seed from training configuration, and separating format degradation from knowledge loss (Open Problem 3) can settle what, if anything, is robust.

Impact Statement

Forgetting is a safety concern in deployed medical assistants: a model that loses general competence may also lose broad safety behaviours acquired during pretraining and alignment. Endogenous Replay is a low-cost candidate mitigation. A safety-relevant caveat from our setup is that forgetting measured through a chat template is sensitive to model variant and evaluation format, so deployment decisions should use the deployed configuration and confirm results across seeds. The dual-use concern is that rank and proximity measurements could guide targeted capability removal; we disclose this because falsifiable forgetting benchmarks make both retention and removal more auditable.

Reproducibility & Availability

All code and results are available at `https://github.com/AkankshaNarula/Major-Project-Group-8-`.

References

Aghajanyan, A., Gupta, S., and Zettlemoyer, L. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pp. 7319–7328, 2021. URL https://aclanthology.org/2021.acl-long.568/.

Ahmad, B. et al. Catastrophic forgetting in low-rank decomposition-based parameter-efficient fine-tuning. *arXiv preprint arXiv:2603.09684*, 2026. URL https://arxiv.org/abs/2603.09684.

Biderman, D., Portes, J., Ortiz, J. J. G., Paul, M., Greengard, P., Jennings, C., King, D., Havens, S., Chiley, V., Frankle, J., Blakeney, C., and Cunningham, J. P. LoRA learns less and forgets less. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2405.09673.

Biderman, S., Prashanth, U. S., Sutawika, L., Schoelkopf, H., Anthony, Q., Purohit, S., and Raff, E. Emergent and predictable memorization in large language models. *arXiv preprint arXiv:2304.11158*, 2023. URL https://arxiv.org/abs/2304.11158.

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2305.14314.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2009.03300.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2106.09685.

Huang, J., Cui, L., Wang, A., Yang, C., Liao, X., Song, L., Yao, J., and Su, J. Mitigating catastrophic forgetting in large language models with self-synthesized rehearsal. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1416–1428, Bangkok, Thailand, 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.77/.

Jin, D., Pan, E., Oufattole, N., Weng, W.-H., Fang, H., and Szolovits, P. What disease does this patient have? A large-scale open-domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021. URL https://www.mdpi.com/2076-3417/11/14/6421.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., and Hadsell, R. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. URL https://www.pnas.org/doi/10.1073/pnas.1611835114.

Li, Z. and Hoiem, D. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2018. URL https://arxiv.org/abs/1606.09282.

Luo, Y., Yang, Z., Meng, F., Li, Y., Zhou, J., and Zhang, Y. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*, 2024. URL https://arxiv.org/abs/2308.08747.

Reimers, N. and Gurevych, I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pp. 3982–3992, 2019. URL https://aclanthology.org/D19-1410/.

Steele, B. Subspace geometry governs catastrophic forgetting in low-rank adaptation. *arXiv preprint arXiv:2603.02224*, 2026. URL https://arxiv.org/abs/2603.02224.

Sun, F.-K., Ho, C.-H., and Lee, H.-y. LAMOL: Language modeling for lifelong language learning. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1909.03329.

Xuhong, L., Grandvalet, Y., and Davoine, F. Explicit inductive bias for transfer learning with convolutional networks. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 2825–2834, 2018. URL https://arxiv.org/abs/1802.01483.

A. K0: MedQA Evaluation Protocol

We evaluate MedQA accuracy by scoring the answer letter (A–D) the model assigns highest probability at the final prompt position, over the first N_{test}=200 examples of the `GBaker/MedQA-USMLE-4-options-hf` test split, on the chat model `Qwen/Qwen3.5-9B` at the matched r=16, 500-step configuration of E02/E10/E11 (a separate run on `-Base` gives 0.78/0.775/0.785 and

is not used). Optimiser, batch sizes, learning rate, and seed match the main runs; throughput optimisations (TF32, FlashAttention 2, fused AdamW, the `fla` kernel for `chunk_gated_delta_rule`) do not change rank, step count, or training data. The ± 2.9 pp binomial SE at $N{=}200$ makes the check a falsifier (no drop > 5 pp) rather than a ranking.

B. Informal Arguments for the CGF Two-Regime Picture

Geometric argument. Below $d^*(D_{\mathrm{ft}})$, the LoRA update is a rank-constrained approximation to the task gradient and cannot align with a small set of nearby pretrained directions. Above d^*, spare rank can reduce the principal angle to task-adjacent representations, so forgetting becomes structured by proximity.

Bayesian argument. Endogenous Replay approximates a local prior around θ_0 under $\mathcal{P}_{\mathrm{anchor}}$. In the sub-critical regime the likelihood term is too rank-limited to express selective drift; in the super-critical regime it has enough degrees of freedom to move selectively, and the anchor prior determines which directions are penalised.

C. ℓ_2-SP Implementation Details

The penalty is $\mathcal{L}_{\mathrm{SP}} = \frac{\lambda}{2} \sum_{\ell \in \mathrm{LoRA}} \|W_\ell - W_\ell^{(0)}\|_2^2$, with $W_\ell^{(0)}$ the LoRA initialisation (A: Kaiming uniform, B: zero). It is added to the SFT cross-entropy before the backward pass; only LoRA parameters carry the penalty. We sweep $\lambda \in \{10^{-3}, 10^{-2}, 10^{-1}\}$ at seed 42 and report per-subject MMLU forgetting and mean accuracy on the nine medical subjects (Table 5).

D. Multi-Seed Replication Protocol

The matched replication of §4.7 (seeds 0, 7 completed; seed 42 of the *replay* configuration is the immediate next run) follows this protocol, and the runnable script is released with the code. (i) **Model:** load `Qwen/Qwen3.5-9B` — the post-trained *chat* checkpoint, identical to E01–E12 — in 4-bit NF4 (float16 compute), *not* the `-Base` checkpoint. (ii) **Conditions:** E02 (no replay), E09 (50 real), E10 (100 real), E11 (100 endogenous), at $r{=}16$, $\alpha{=}32$, 500 steps, effective batch 8, linear schedule, 50 warmup steps, sequence length 256 (the replay-sweep configuration). (iii) **Seeds:** $\{0, 7, 42\}$; the same seed governs `set_seed`, the data-sampling RNG, the LoRA initialisation, and the trainer's `data_seed`. (iv) **Evaluation:** the identical MMLU protocol of §3.0.4 — the `Qwen/Qwen3.5-9B` chat template, greedy decoding, `max_new_tokens=5`, all 57 subjects — with forgetting computed against the *chat-model* baseline (mean 0.776). The pre-registered decision rule: if the seed-averaged E10→E11 gap is ≥ 0.04 with the ordering $\mathrm{E11} < \mathrm{E10} < \mathrm{E09} < \mathrm{E02}$ preserved at all three seeds, the seed-42 Endogenous-Replay result is confirmed; a seed-averaged gap < 0.04 downgrades C2 to a directional trend. This is the single experiment that converts the headline from a single-seed observation into a benchmark.