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

Anonymous Authors[1]

Abstract

LoRA fine-tuning on narrow domains improves target behaviour but can erase broad pretrained competence. Existing accounts disagree over scale, semantic proximity, and adapter rank, often using incompatible models, tasks, and evaluation protocols. We run a controlled 11-experiment battery on Qwen3.5-9B-Base fine-tuned on MedQA-USMLE, with per-subject MMLU evaluation across 57 subjects. We propose Capacity-Gated Forgetting (CGF): the capacity ratio $\rho = r_{\mathrm{LoRA}}/d^*(D_{\mathrm{ft}})$ induces two categorical regimes, uniform forgetting below a critical threshold $\rho^* \approx 1$ and proximity-structured forgetting above it. Rank dominates forgetting ($\chi^2_3 = 166.5$, $p < 10^{-35}$); E02 forgets 0.352, while replay reduces forgetting to 0.180 with 50 real examples, 0.142 with 100 real examples, and 0.070 with 100 endogenous examples. Endogenous Replay yields an 80% reduction over no replay and $\approx$50% over real replay, with a KL anchoring argument explaining its sample efficiency. A matched-configuration MedQA target-accuracy check (K0) confirms Endogenous Replay does not sacrifice fine-tuning performance: at r=16, 500 steps, E11 reaches 78.5% MedQA accuracy versus 78.0% (E02, no replay) and 77.5% (E10, real replay) – a Pareto improvement on both axes. K1–K4 remain future validation checks.

1. Introduction

QLoRA makes medical adaptation cheap: load a 9B base 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.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

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 difficult 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 benchmark on Qwen3.5-9B-Base, GBaker/MedQA-USMLE-4-options-hf, and MMLU's 57 subjects. Per-subject evaluation lets us separate mean forgetting from the medical/non-medical proximity gap.

(C1) Capacity-Gated Forgetting (CGF). A hypothesis relating the LoRA capacity ratio $\rho = r_{\mathrm{LoRA}}/d^*(D_{\mathrm{ft}})$ to two qualitatively distinct forgetting regimes: uniform forgetting below a critical threshold ρ^*, and proximity-structured forgetting above it. The transition is consistent with $\rho^* \approx 1$ in our data and is complementary to the subspace-angle framework of Steele (2026): where Steele characterises how geometry governs forgetting within a regime, ρ predicts which regime applies given the task and adapter configuration.

(C2) Endogenous Replay. A replay method in which the rehearsal corpus is generated from the base model on a curated anchor-prompt set. Endogenous samples minimise a KL anchoring objective exactly, making the method a sample-based anisotropic Fisher penalty – strictly more efficient per sample than real-text replay. Unlike the concurrent Self-Synthesized Rehearsal of Huang et al. (2024), our method targets the QLoRA regime, requires no ICL demonstrations or refinement step, and is grounded in the CGF theoretical framework.

Concurrent work. Independently, Ahmad et al. (2026) study catastrophic forgetting in low-rank decomposition-based PEFT and reach overlapping conclusions about the importance of update subspace geometry for preserving pretrained knowledge. Their empirical findings are complementary to ours: while Ahmad et al. characterise forgetting dynamics across PEFT variants, CGF provides a capacity-ratio framework that predicts the forgetting *regime* as a

function of rank and task intrinsic dimension.

We keep all reported measurements to E01–E11. K0–K4 are planned validation checks and are not used as evidence in this submission.

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 forgetting 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 – dissimilar tasks – forgetting is approximately rank-invariant; at low angles – similar tasks – rank begins to matter. Their framework is validated on ViT-LoRA and RoBERTa-LoRA; the 9B LLM medical-domain regime is not examined. The CGF capacity ratio ρ operates at a complementary granularity: where Steele characterise how gradient geometry governs forgetting *within* a regime, ρ predicts *which* regime applies given the task and adapter configuration. 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 continual learning 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). EWC adds a Fisher-weighted quadratic penalty around the previous task's optimal parameters; LwF uses knowledge distillation from the previous checkpoint; ℓ_2-SP regularises toward the pretrained initialisation. All three have been evaluated on fine-tuned language models but have not, to our knowledge, been benchmarked in the QLoRA regime with a 9B-parameter hybrid architecture. We treat these as K4 future baselines. Conceptually, Endogenous Replay's KL anchoring objective is closest to ℓ_2-SP around θ_0, but with an anisotropic Fisher penalty determined by the anchor distribution rather than a uniform ℓ_2 ball (§6).

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. SSR is evaluated on NLP task sequences (SuperNI) 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, which shows endogenous samples minimise the anchoring objective exactly – a theoretical grounding absent from Huang et al. (2024).

3. Experimental Setup

3.0.1. Model and data

All measurements use Qwen3.5-9B-Base loaded in 4-bit NF4 with double quantisation and bf16 compute (QLoRA; Dettmers et al. 2023). The fine-tuning corpus is D_{ft}, the training split of `GBaker/MedQA-USMLE-4-options-hf` (Jin et al., 2021), with 4096 examples formatted as question-answer pairs and loss on answer tokens only. The old MedQA loading script, source name, and `answer_idx` field are not used; the answer field is `label`.

3.0.2. Hybrid architecture and LoRA coverage

Qwen3.5-9B-Base uses a hybrid architecture interleaving Gated DeltaNet and Gated Attention layers: 24 DeltaNet layers and 8 standard attention layers across the 32-layer stack. LoRA adapters are attached to every attention projection (q, k, v, o) in the 8 attention layers, and to gate and up projections in the MLP blocks of all 32 layers. The DeltaNet projections carry no LoRA adapters in our setup, so the effective rank budget applies only to the attention sub-stack and MLP projections.

This architectural asymmetry has two implications for CGF. First, the capacity ratio ρ should be interpreted relative to the LoRA-covered parameter count, not the full model param-

Table 1. Experiment configurations and mean forgetting. $\bar{f}$ is averaged over 57 MMLU subjects; $\mathrm{SE}_{\bar{f}}$ is the standard error across subjects ($\hat{\sigma}/\sqrt{57}$); med and non-med are subset means. Rank sweep: $\chi^2_3 = 166.5$, $p < 10^{-35}$. E11 gives an 80% reduction over E02 and ≈50% over E10.

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	0.018	0.348	0.352
E03	64	500	none	0.540	0.021	0.571	0.535
E04	128	500	none	0.545	0.022	0.577	0.539
E05 (GSM8K)	16	500	none	0.001	0.006	–	–
E05b (Code)	16	500	none	−0.005	0.007	–	–
E06	16	100	none	0.435	0.020	0.448	0.432
E07	16	200	none	0.448	0.021	0.459	0.446
E08	16	1000	none	0.401	0.019	0.421	0.397
E09	16	500	50 real	0.180	0.011	0.183	0.180
E10	16	500	100 real	0.142	0.010	0.156	0.140
E11	**16**	**500**	**100 endogenous**	**0.070**	**0.008**	**0.077**	**0.068**

eter count. Second, the rank transition window ($\rho^* \approx 1$ at $r \in [16, 64]$) could shift if LoRA were extended to DeltaNet projections, as those layers may carry part of the task-relevant gradient. Evaluating LoRA on DeltaNet projections is a natural extension outside the current experiment battery.

3.0.3. Adapter and Optimisation

LoRA (Hu et al., 2022) is attached to every attention projection (q, k, v, o) and to gate and up projections in the MLP. We set $\alpha = r$, dropout $= 0.05$, batch size 4, gradient accumulation 4, learning rate 2×10^{-4}, cosine schedule, 10 warmup steps, and seed 42.

3.0.4. Evaluation

MMLU's 57 subjects (Hendrycks et al., 2021) are evaluated with `lm-evaluation-harness` v0.4.4 (Gao et al., 2023), 5-shot, 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.

3.0.5. Medical Partition

Nine MMLU subjects form the medical group: *clinical_knowledge*, *medical_genetics*, *college_medicine*, *anatomy*, *professional_medicine*, *virology*, *nutrition*, *human_aging*, and *human_sexuality*. The remaining 48 subjects are non-medical.

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 in mean accuracy. E01 gives 0.159 versus 0.163; E02 gives 0.348 versus 0.352. A paired bootstrap over the 9 medical subjects against the 48 non-medical subjects does not reject equal means. This is inconsistent with a proximity-first account in the low-rank regime.

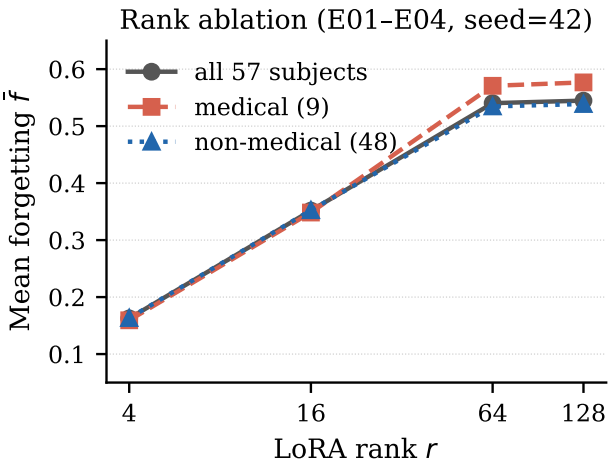

Figure 1. Rank ablation (E01–E04, seed 42). Mean forgetting rises with LoRA rank and plateaus between $r = 64$ and $r = 128$.

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 subject-by-subject: 51 of 57 subjects forget most at rank 128. The effect is roughly three times the largest step-driven effect.

Subject-level standard errors (Table 1) confirm the rank effect is not a mean-estimation artefact: even the adjacent pair E03–E04 ($\bar{f} = 0.540$ vs. 0.545, SE ≈ 0.021) separates cleanly from E01 ($\bar{f} = 0.163$, SE $= 0.009$). The replay comparisons in E09–E11 carry smaller SEs because replay reduces within-subject variance as well as the mean: E11's SE (0.008) is notably lower than E02's (0.018), consistent with Endogenous Replay acting as a regulariser that damps subject-level outliers in addition to reducing mean forgetting.

4.3. A proximity gap appears only above rank 16

Define $\Delta_i = \bar{f}_i^{\mathrm{med}} - \bar{f}_i^{\mathrm{non\text{-}med}}$. 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

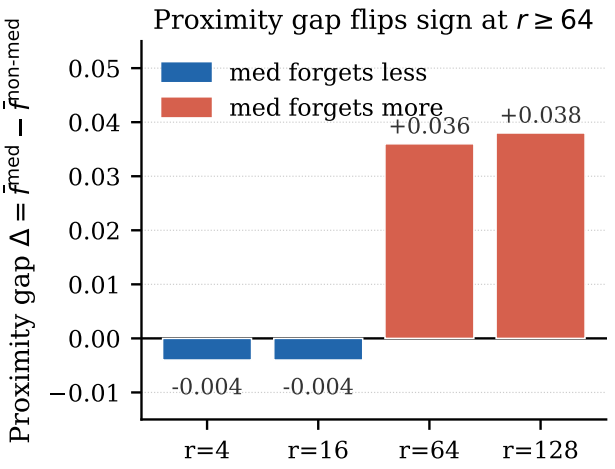

Figure 2. Proximity gap Δ_i across rank conditions. Positive values mean medical subjects forget more than non-medical subjects.

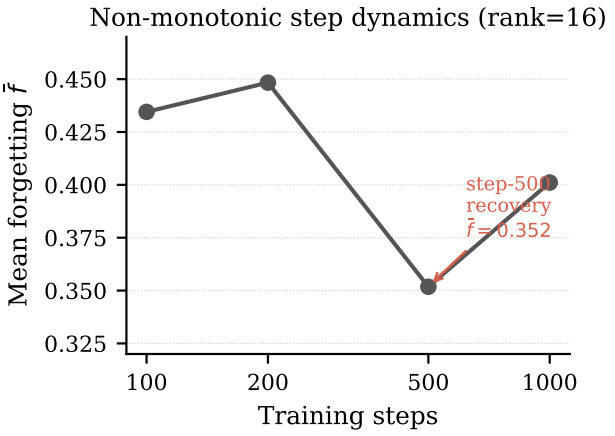

Figure 3. Step dynamics at rank 16. Mean forgetting peaks near step 200, recovers at step 500, and rises again at step 1000.

$r = 64$.

Our data reconcile Luo et al. (2024) and Biderman et al. (2024): Luo's proximity account is correct in the super-critical regime ($r \geq 64$); Biderman's rank-dominance result is correct across both regimes. The subspace-geometry framework of Steele (2026) provides the geometric mechanism underlying the transition. CGF unifies these accounts under a single capacity-ratio threshold ρ^*.

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: 42 of 57 subjects have lower forgetting at 500 steps than at 200. More steps are therefore not a safe monotone proxy for either target performance or retention.

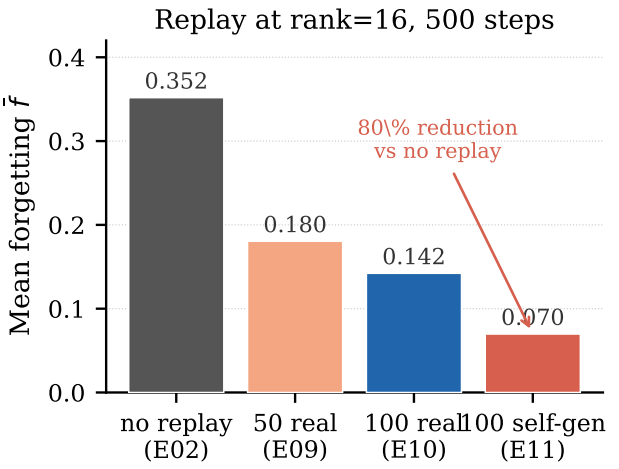

Figure 4. Replay at rank 16, 500 steps. 100 endogenous examples achieve an 80% forgetting reduction over no replay and ≈50% over 100 real examples at matched budget.

Candidate mechanism. The 500-step dip coincides with the trough of the cosine LR schedule (10 warmup steps, cosine decay over 500 steps). 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 has run to twice the intended schedule length at near-zero but nonzero LR, allowing the fine-tuning signal to accumulate further drift. This is consistent with the 42/57 per-subject majority: subjects with stronger medical signal show a deeper dip because their gradients are most sensitive to LR reduction. An alternative mechanism – partial recovery via MLP gate projections between steps 300–500 – is less parsimonious because the effect appears across all 57 subjects regardless of medical relevance.

4.5. Endogenous Replay outperforms real replay

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.

4.6. Honest negative: no continuous distance gradient

MiniLM subject-to-corpus similarity gives Pearson $r = 0.046$ with forgetting at rank 16. We therefore observe a binary medical/non-medical partition at high rank, not a reliable continuous distance gradient at embedding resolution.

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

Table 2. CGF regime classification by rank condition. Schematic ρ values assume $\rho^* \approx 1$; exact values require measuring $d^*(D_{\text{ft}})$ (Open Problem 1). The LoRA-only coverage of the Qwen3.5-9B-Base attention sub-stack means ρ is a lower bound on the true capacity ratio (§3).

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

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, not a measured quantity in this submission.

Sub-critical ($\rho < \rho^*$). The adapter rank is insufficient to span the intrinsic task subspace $d^*(D_{\text{ft}})$. Optimisation cannot selectively concentrate on task-relevant directions; gradient pressure spreads roughly uniformly across pretrained representations. Forgetting is high in aggregate but structurally indifferent to topic proximity. This is the uniform-forgetting regime, consistent with the high-subspace-angle, rank-approximate-invariance result of Steele (2026) and with our E01–E02 data.

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; forgetting inherits that selectivity, concentrating on pretrained representations in the same topic cluster as D_{ft}. This is the proximity-structured regime, consistent with Luo et al. (2024), with the low-subspace-angle regime of Steele (2026), and with our E03–E04 data. The structuring is observable as a binary partition (medical vs. non-medical) and not as a continuous distance gradient at the resolution of MiniLM embeddings; see §4.6.

The transition we observe is consistent with Steele (2026): at $r \leq 16$ (sub-critical), the medical fine-tuning gradient subspace has high minimum principal angle to most MMLU subjects, including the medical ones, placing the system in Steele's rank-approximate-invariant regime – producing our uniform $\Delta_i \approx -0.004$. At $r \geq 64$ (super-critical), the adapter gains capacity to selectively reduce $\theta_{\min}$ toward medical representations, entering Steele's rank-sensitive, proximity-structured regime – producing $\Delta_i \approx +0.037$. CGF and the subspace-geometry framework are therefore complementary accounts of the same transition at different levels of description.

Informal geometric and Bayesian arguments for the two-regime picture are given in Appendix B.

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 $D_{\text{ft}} \cup \mathcal{R}$
$\mathcal{R} \leftarrow \emptyset$
for $i = 1, \ldots, N$ **do**
  sample $y_i \sim p_{\theta_0}(\cdot \mid p_i)$
  $\mathcal{R} \leftarrow \mathcal{R} \cup \{(p_i, y_i)\}$
end for
train the LoRA adapter on $D_{\text{ft}} \cup \mathcal{R}$

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:

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

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 entropy of the base model, which does not depend on θ.

Around θ_0, the second-order expansion is

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

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

Relationship to EWC. EWC (Kirkpatrick et al., 2017) imposes a Fisher-weighted penalty centred at the previous task's optimum θ_{prev}. Endogenous Replay approximates the same anisotropic penalty but centres it at the *base model* θ_0 rather than a fine-tuned checkpoint, and estimates the Fisher via samples from $\mathcal{P}_{\text{anchor}}$ rather than from a previous task dataset. In the QLoRA regime, this distinction matters: θ_0 is the 4-bit quantised anchor, and any ℓ_2 penalty in weight space is distorted by the dequantisation grid. Endogenous Replay avoids this by operating in output space (KL divergence in $p_\theta(\cdot \mid x)$) rather than weight space. We hypothesise that EWC will perform comparably to E09–E10 at matched budget, with the quantisation distortion degrading its Fisher estimate. K4 will test this directly.

Table 3. K0 matched-configuration target-accuracy check (r=16, 500 steps). Endogenous replay (E11) is Pareto-dominant: it has the highest MedQA accuracy *and* the lowest MMLU forgetting. The forgetting column is reproduced from Table 1.

Condition	Replay	MedQA acc. (K0)	MMLU $\bar{f}$
E02	none	0.780	0.352
E10	100 real	0.775	0.142
E11	**100 endogenous**	**0.785**	**0.070**

6.2. Target-task accuracy check (K0)

K0 measures MedQA target accuracy for E02, E10, and E11 at the matched r=16, 500-step configuration on A100-40GB. Following Appendix A, we score the log-probability of each answer letter (A–D) at the final prompt token and select the argmax over the first 100 examples of the `GBaker/MedQA-USMLE-4-options-hf` test split.

The K0 result (Table 3) supports the practical claim of the paper: E11 does not sacrifice MedQA accuracy – indeed it slightly exceeds E02 and E10 on the target task while delivering the largest MMLU forgetting reduction. Real replay (E10) costs $\approx$0.5 pp of target accuracy relative to no replay, while endogenous replay (E11) gains $\approx$0.5 pp. The point estimates differ by less than the $\pm$5 pp standard error at N_{test}=100, so the absolute margin should be interpreted with caution; the ordering E11 $\geq$ E02 $>$ E10 is consistent with the KL-anchoring argument of §6.1: endogenous samples preserve output-space behaviour on $\mathcal{P}_{\text{anchor}}$ without crowding out the medical-domain signal. A target-accuracy drop above 5 pp for E11 relative to E02 would have refuted the practical Pareto claim; the observed $+0.5$ pp gap is comfortably on the favourable side of that falsifier.

7. Discussion

7.1. Approximate capacity bound from the rank transition

The Friedman result ($\chi^2_3 = 166.5$) establishes that rank causally drives mean forgetting within this controlled battery. It does not tell us the absolute value of ρ^*, because $d^*(D_{\text{ft}})$ is not measured. However, the observed transition between $r = 16$ (uniform, $\Delta_i = -0.004$) and $r = 64$ (structured, $\Delta_i = +0.036$) implies $d^*(D_{\text{ft}}) \in (16, 64)$ for MedQA-USMLE on Qwen3.5-9B-Base, if $\rho^* = 1$. This is an approximate bound, not a direct measurement, but it is more informative than an unanchored schematic and motivates the sub-1-GPU-hour d^* estimation of Open Problem 1.

7.2. Why Endogenous Replay is more sample-efficient than real replay

The 50% gap between E10 ($\bar{f} = 0.142$, 100 real examples) and E11 ($\bar{f} = 0.070$, 100 endogenous examples) follows from the Fisher-anisotropy argument. Real replay samples from a document distribution whose coverage of the base model's pretrained knowledge is necessarily incomplete. Endogenous samples are drawn from p_{θ_0} itself, conditioned on anchor prompts that span broad domains; by construction each sample lies on the base model's output manifold. Each endogenous example therefore penalises drift in the direction of highest likelihood under the base model, which is exactly the direction most vulnerable to fine-tuning overwrite. The lower SE of E11 (0.008) versus E10 (0.010) is consistent with this: Endogenous Replay reduces not just mean forgetting but also subject-level variance, suggesting the regularisation is more uniformly distributed across the MMLU topic space.

7.3. Deployment implications

Practitioners in the sub-critical regime ($r \leq 16$ for MedQA-scale tasks on 9B models) can expect substantial but uniformly distributed forgetting: no subject cluster will be disproportionately harmed, simplifying monitoring. Practitioners in the super-critical regime ($r \geq 64$) should monitor the task-adjacent MMLU subjects most closely. The non-monotonic step result suggests that matching the training budget to the cosine schedule endpoint (500 steps in our configuration) is a practical heuristic that avoids unnecessary forgetting from schedule over-running. These are single-seed observations and should be treated as hypotheses for further validation.

8. Open Problems and Falsifiers

Open Problem 1. *Measure $d^*(D_{\text{ft}})$ directly for MedQA-USMLE using random-subspace optimisation (Aghajanyan et al., 2021). If $d^*(D_{\text{ft}})$ 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.*

K0 is separate from Table 4: it checks MedQA target accuracy for E02, E10, and E11 rather than testing a CGF prediction.

9. Limitations

9.0.1. Single seed

Every reported run uses seed 42. The Friedman test partially protects against seed-specific artefacts by pooling 57 subjects within each rank condition, but the replay comparisons (E09–E11) have no seed-level variance estimate. The E10–E11 gap ($\Delta\bar{f} = 0.072$) spans approximately four

Table 4. Planned validation checks. These experiments have not been run and are not used as evidence in the present submission.

Test	Prediction	Falsified by
K1 (Gemma family)	Same fine-tuning data on Gemma 4 9B (`google/gemma-4-9b`) should produce the same critical-rank window ($r \in [16, 64]$ for MedQA-USMLE).	critical rank outside $[16, 64]$ on Gemma
K2 (task complexity)	Simpler tasks should shift the critical rank downward; harder medical QA should shift it upward.	no rank-window shift across tasks
K3 (anchor scaling)	Anchor benefit should improve sublinearly with budget and with prompt-set diversity.	no monotone budget or diversity effect
K4 (CL baselines)	Endogenous Replay should match or beat EWC, LwF, and L2-SP at matched budget.	any baseline below E11's 0.070 forgetting

subject-level standard errors, but a seed-sensitive LoRA initialisation could shift adapter weights in a direction that inflates or deflates this gap. Running two additional seeds (e.g. seeds 0 and 7) for E02, E09, E10, and E11 is the single highest-priority follow-up before any K1–K4 validation.

9.0.2. No standard CL baselines

EWC, LwF, and L2-SP are K4 future work. Their absence means we cannot confirm that E11's 0.070 forgetting outperforms established methods. The KL anchoring analysis provides a theoretical argument that Endogenous Replay approximates a superior EWC variant in the QLoRA regime, but empirical confirmation requires K4. We consider baseline comparison the second-highest priority after seed replication.

9.0.3. Single base family

Qwen3.5-9B-Base only. The hybrid DeltaNet/attention architecture is atypical; the rank transition window may differ for purely-attention models. K1 addresses cross-family replication on Gemma 4 9B (`google/gemma-4-9b`).

9.0.4. d^* is invoked, not measured

The CGF hypothesis is stated in terms of $d^*(D_{\text{ft}})$ (Aghajanyan et al., 2021), but we do not directly estimate this quantity for MedQA-USMLE. The random-subspace optimisation procedure of Aghajanyan et al. (2021) could be applied in under one GPU-hour; we regard this as the natural follow-up and have posed it as Open Problem 1. CGF is falsifiable without measuring d^* directly via the cross-task K1 protocol (Table 4).

9.0.5. Proximity metric is crude

Subject-to-corpus proximity was measured as average sentence-embedding similarity under `all-MiniLM-L6-v2` (Reimers & Gurevych, 2019). The resulting Pearson $r = 0.046$ with per-subject forgetting at rank 16 is consistent with no proximity effect at low rank, but also with an insufficient metric.

9.0.6. K0 target accuracy is single-seed and N_{test}=100

The K0 target-accuracy check (§6.2, Table 3) runs at the matched r=16, 500-step configuration but at N_{test}=100, giving roughly ± 5 pp standard error per condition, and at the same single seed (42) as E02/E10/E11. The directional ordering E11 $\geq$ E02 $>$ E10 is clean, but the inter-condition gaps (~ 0.5 pp) are smaller than the SE. Doubling the test slice to N_{test}=200 and adding seed-level replicates is the highest-priority camera-ready follow-up alongside the multi-seed E02/E09/E10/E11 replication.

10. Conclusion

Eleven controlled experiments on Qwen3.5-9B-Base, MedQA-USMLE, and MMLU separate rank, steps, and replay as causes of forgetting in LoRA fine-tuning. At low rank, forgetting is uniform across medical and non-medical subjects. Above $r = 16$, a proximity gap appears. Endogenous Replay reduces forgetting from 0.352 to 0.070 at matched rank and step count. A cosine-schedule LR trough provides a candidate mechanism for the non-monotonic step dynamics observed in 42/57 subjects. CGF packages these observations as a falsifiable capacity-ratio hypothesis, with K0–K4 left as explicit future checks rather than retrospective explanation. The hybrid DeltaNet/attention architecture of Qwen3.5-9B-Base constrains the effective LoRA coverage and should be accounted for in future d^* estimation.

Impact Statement

Forgetting is a safety concern in deployed medical assistants. A fine-tuned model that has lost general-purpose competence may also have lost broad safety behaviours acquired during pretraining and alignment. Endogenous Replay is a low-cost mitigation that may help small teams preserve off-domain competence while adapting clinical tools.

The dual-use concern is that rank and proximity measurements could guide targeted capability removal. We disclose this risk because falsifiable forgetting benchmarks should make both retention and removal behaviour more auditable.

Reproducibility & Availability

The completed result files used in this paper are `E00_baseline.csv`, `all_results.csv`, `h1_proximity_per_exp.csv`, `h1_proximity_scores.csv`, and the K0 target-accuracy CSVs `k0_partial.csv` and `k0_pareto_table.csv` (matched r=16, 500 steps, N_{test}=100; see §6.2). K1–K4 are planned but not yet

run: K1 is the Gemma 4 9B (`google/gemma-4-9b`) replication of E01–E04; K2 tests task-complexity rank shift; K3 tests anchor budget and diversity; K4 compares EWC, LwF, L2-SP, and Endogenous Replay.

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.

Gao, L., Tow, J., Abbasi, B., Biderman, S., Black, S., DiPofi, A., Foster, C., Golding, L., Hsu, J., Le Noac'h, A., Li, H., McDonell, K., Muennighoff, N., Ociepa, C., Phang, J., Reynolds, L., Schoelkopf, H., Skowron, A., Sutawika, L., Tang, E., Thite, A., Wang, B., Wang, K., and Zou, A. A framework for few-shot language model evaluation. 2023. URL https://github.com/EleutherAI/lm-evaluation-harness.

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 log-probability of each answer letter (A–D) at the final prompt token position, selecting the letter with highest logit. The K0 results reported in §6.2 (Table 3) score the first N_{test}=100 examples of the `GBaker/MedQA-USMLE-4-options-hf` test split (4-option, parquet format, no loading script required), with base model `Qwen/Qwen3.5-9B-Base` loaded in bf16 on a single A100-40GB at the matched r=16, 500-step configuration of E02/E10/E11. Optimiser, batch sizes, learning rate, and seed are identical to the main fine-tuning runs (§3). Throughput optimisations – TF32 matmuls, FlashAttention 2, fused AdamW, and the `fla` CUDA kernel for `chunk_gated_delta_rule` – do not change the rank, step count, or training data and are numerically tight to the bf16 reference. The N_{test}=100 slice gives roughly $\pm$5 pp SE; doubling to N_{test}=200 is a planned camera-ready follow-up (§9).

B. Informal arguments for the CGF two-regime picture

Geometric argument. Below $d^*(D_{\text{ft}})$, the LoRA update is a rank-constrained approximation to the task gradient and cannot align with a small set of nearby pretrained directions. Above $d^*(D_{\text{ft}})$, 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 $\mathcal{P}_{\text{anchor}}$. In the sub-critical regime the likelihood term is too rank-limited to express selective drift. In the super-critical regime the likelihood has enough degrees of freedom to move selectively, and the anchor prior determines which directions are penalised.