Not All Components Forget Alike: Component-Level Intervention Profiles in Continual Pretraining

Sangyeol Baek
School of Computer Science
Semyung University
218026@semyung.ac.kr

Suan Lee
School of Computer Science
Semyung University
suanlee@semyung.ac.kr

Abstract

Continual pretraining adapts a language model to new domains but can degrade performance on earlier ones. We study this through component-level interventions in Transformer blocks and find that forgetting has sharply different intervention profiles across Q/K, V/O, and MLP. In SmolLM2-360M over six domains, V/O rollback improves prior-domain Wiki perplexity while leaving new domains nearly unchanged (-5.72 Wiki; new-domain range -1.06 to $+0.76$), MLP rollback recovers Wiki but severely damages all new domains ($+13.95$ to $+20.75$), and Q/K rollback shows no recovery. Training-time V/O freeze reproduces the selective pattern: -2.98 ± 0.13 Wiki PPL (3 seeds, stress order) with no measurable current-domain learning cost and no compensatory MLP/Q/K weight change. Same-parameter-count Q/K freeze has the opposite sign ($+3.17 \pm 2.53$); random attention-subset freeze tracks the V/O tensors it includes. Inference-time activation transplantation modulates forgetting bidirectionally without weight updates (forward -6.27, reverse $+3.26$), though the activation-level effect is carried by o_proj output rather than the full V+O bundle. The V/O freeze effect replicates across scale and model family (SmolLM2-1.7B, OLMo-2-1B, and Gemma3-1B, 3 seeds each); across these and the 360M model, prior-domain relative mitigation stays within roughly 10–27% despite order-of-magnitude differences in absolute forgetting, and the V/O versus Q/K asymmetry also survives in a LoRA adaptation regime. We report these as a component-level intervention profile *in the settings tested* — decoder-only LMs, six English domains, 15M tokens per domain — rather than as proof that V/O is the unique causal forgetting locus; validation at larger scales, in multilingual settings, and under longer adaptation schedules remains open.

1 Introduction

Continual pretraining is a practical way to adapt a pretrained language model to a sequence of domains, but later domains can degrade performance on earlier ones. This phenomenon is usually discussed as catastrophic forgetting (McCloskey and Cohen, 1989; French, 1999). Many mitigation approaches aim to identify, protect, isolate, or adapt a restricted subset of parameters, including EWC-style regularization (Kirkpatrick et al., 2017), PackNet-style capacity allocation (Mallya and Lazebnik, 2018), and parameter-efficient adaptation such as LoRA (Hu et al., 2022).

We start from a different question: if a model forgets after sequential pretraining, which Transformer component can be intervened on to selectively recover the earlier-domain state? This is not the same as asking where the model changed the most. In our setup, the largest representational drift appears at the top layer, but rolling back that layer causes new-domain disruption rather than selective forgetting recovery. This motivates a component-level intervention analysis.

Our central finding is a three-way contrast. V/O interventions selectively recover previous-domain perplexity with little new-domain cost; MLP interventions also recover previous domains, but only by damaging new domains; Q/K interventions do not recover forgetting. This contrast appears in post-hoc weight rollback, training-time freeze, and inference-time activation transplantation, although the activation-level effect is more precisely localized to o_proj output and should not be conflated with the weight-level V/O projection bundle.

Contributions.

- We show that largest representational drift is not the same as the most effective recov-

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 33719–33739
October 24-29, 2026 ©2026 Association for Computational Linguistics

ery intervention point: rolling back the top layer where drift is largest barely recovers Wiki perplexity, and instead damages new domains.

- We identify a robust component-level intervention profile: V/O selective recovery, MLP entangled recovery, and Q/K no recovery (Q/K rollback is null, and same-parameter-count Q/K freeze trends in the opposite, harmful direction), which holds across rollback, training-time freeze, and inference-time activation transplantation.
- We verify the V/O profile with training-time freeze and controls for parameter count (same-count Q/K freeze gives the opposite sign), random subset selection (effect tracks V/O inclusion), and compensatory weight changes (no measurable MLP/Q/K compensation).
- We show that inter-checkpoint activation transplantation modulates forgetting bidirectionally without weight updates, while clarifying via a mechanism note that the activation-level effect is primarily o_proj-output transferability rather than a true V+O bundle effect.
- We compare against four EWC variants and report a V/O freeze + Full EWC stack that exceeds either intervention alone, and provide scale and family evidence on SmolLM2-1.7B (weight-level effect replicated across 3 seeds) and 3-seed pilots on two further model families (OLMo-2-1B, Gemma3-1B).[1]

2 Related Work

Continual learning and forgetting. Classic continual-learning methods such as EWC, SI, and PackNet mitigate forgetting by estimating parameter importance or allocating capacity (Kirkpatrick et al., 2017; Zenke et al., 2017; Mallya and Lazebnik, 2018). Recent LLM adaptation and expansion methods include LLaMA-Pro (Wu et al., 2024), while Li and Lee examine forgetting in continual pretraining of aligned LLMs (Li and Lee, 2024). Broader continual-pretraining studies show that self-supervised pretraining can behave differently from supervised continual learning (Cossu et al., 2022; Ke et al., 2023). Spurious-forgetting and representation-drift analyses further caution that measured forgetting can reflect multiple mechanisms rather than a single overwritten store (Zheng et al., 2025; Ramasesh et al., 2021). Closest in spirit to our intervention, Yamaguchi et al. (2025) mitigate forgetting during target-language adaptation by scoring parameter importance on source data and column-wise freezing the parameters deemed critical to source ability. Their selection is importance-driven and evaluated in a language-adaptation setting; we instead ask which architectural components admit selective recovery at all, and probe this with post-hoc and inference-time interventions in addition to freezing. Our work is not a new mitigation method; it is an intervention analysis of which Transformer components exhibit selective recovery profiles.

Transformer component roles. Prior work has associated MLPs with factual storage or editable knowledge (Geva et al., 2021; Dai et al., 2022; Meng et al., 2022), and attention with information extraction or routing (Geva et al., 2023). Transformer circuit analyses distinguish QK circuits, which determine attention routing, from OV circuits, which write selected content into the residual stream (Elhage et al., 2021). This distinction motivates our component grouping and fixes our notation: throughout the paper, **Q/K** denotes the query and key projections that instantiate the QK (routing) circuit, and **V/O** denotes the value and output projections that instantiate the OV (content-writing) circuit. Attention-head pruning studies show that heads can be highly redundant or unequally useful (Voita et al., 2019; Michel et al., 2019), while progressive layer freezing provides a related precedent for selective freezing (Brock et al., 2017). Recent skill-addition work also highlights component-sensitive adaptation behavior (Zhu et al., 2025). Weight-space intervention work such as task arithmetic provides related evidence that parameter changes can compose or interfere in structured ways (Ilharco et al., 2023).

Activation patching and localization. Activation patching estimates the effect of replacing internal states at inference time and is closely related to causal circuit analysis (Conmy et al., 2023; Geiger et al., 2021). However, successful intervention sites are not automatically causal storage sites: model-editing work has shown that localization and edit-success locations can diverge (Hase

[1] Code and per-run result summaries: https://github.com/baeksang1124/not-all-components-forget-alike

et al., 2023). We therefore frame our findings as component-level intervention profiles rather than direct proof of a unique causal forgetting locus.

3 Experimental Setup

Models. We primarily use SmolLM2-360M (Allal et al., 2025), a 32-layer GQA Transformer (15 query heads, 5 KV heads, head dim 64, hidden 960, FFN 2560). Scale and model-family observations use SmolLM2-1.7B (Allal et al., 2025) (24-layer MHA), OLMo-2-1B (Team OLMo et al., 2024) (16-head MHA with Q/K RMSNorm), and Gemma3-1B (`google/gemma-3-1b-pt`, hybrid sliding-window and full attention). All experiments start from publicly released pretrained checkpoints; we report intervention diffs relative to no-intervention continual-pretraining baselines, so the underlying pretraining corpus is not a confound for our intervention contrasts.

Domains and order. The model is continually pretrained on six domains in the stress order:

`Wiki` $\rightarrow$ `Law` $\rightarrow$ `Math` $\rightarrow$ `Patent` $\rightarrow$ `ArXiv` $\rightarrow$ `PubMed`

Wiki acts as the prior-knowledge anchor, and the five remaining domains are selected to be linguistically and topically distinct from Wiki (legal text, formal math, patent prose, scientific articles, biomedical literature). The five-domain tail induces measurable Wiki PPL degradation in the no-intervention baseline ($+10$ to $+20$ PPL across seeds), giving a non-trivial recovery margin to compare against. A swap order (Wiki $\rightarrow$ PubMed $\rightarrow$ ArXiv $\rightarrow$ Patent $\rightarrow$ Math $\rightarrow$ Law) is used as a robustness control.

Training. Each domain is trained for 15M tokens with AdamW, learning rate 5×10^{-4}, cosine schedule, batch 16, sequence length 1024. Main 360M results use three seeds (42, 1, 7). V/O freeze is additionally tested under the swap-order control, yielding 3 seeds $\times$ 2 orders. End-to-end runs fit on a single RTX 3090 in roughly 6 GPU-hours per 6-domain run (SmolLM2-360M) and approximately 28 GPU-hours per run (SmolLM2-1.7B). Full reproduction of all 360M experiments reported in the main paper takes on the order of 250 GPU-hours; the OLMo-2-1B pilot adds another roughly 65 GPU-hours.

Evaluation. Perplexity is computed on per-domain held-out splits (1k–3k sequences per domain, length 1024, stride 512). Q/K/V/O rollback and Fisher diagnostics additionally use a Wiki-only held-out split (457 sequences) to isolate prior-domain effects; appendix tables mark the split used. Downstream evaluation (Appendix F) uses zero-shot LM-eval-harness on ARC-Easy, HellaSwag, PIQA, SciQ, WinoGrande, plus domain-matched MMLU subsets for Law/Math/Medicine.

Interventions. We use three families. *(i) Weight rollback.* At the final checkpoint, selected component weights are replaced with their values from the Wiki-end checkpoint. We evaluate Q, K, V, O, Q+K, V+O, all MLP, all attention, layer bands, embeddings, LM head, and layer norms. *(ii) Training-time freeze.* Selected projections are frozen throughout sequential pretraining. The main condition freezes all V/O projection tensors. Controls freeze same-parameter-count Q/K tensors or randomly selected attention subsets. *(iii) Activation transplantation.* At inference time, a source checkpoint forwards the same batch and its selected module outputs are captured; a target checkpoint then forwards the batch while those module outputs are overwritten by the source activations via forward hooks. We report forward (Wiki-end source $\rightarrow$ PubMed-end target, "cure") and reverse (PubMed-end source $\rightarrow$ Wiki-end target, "elicit") transfers.

4 Results

4.1 Largest Change Is Not the Recovery Locus

Comparing runs with different domain orders shows that internal representations diverge as training proceeds. The largest drift appears at the top layer L31. Measuring hidden-state branching (L31 residual stream) and output branching (post-LM-head) at the end of training across 4 paired runs (front/back $\times$ stress/swap, seed=42), the L31-hidden-to-output drift ratio is $5.17\times$ on average (range 3.77–5.70). However, rolling back L31 gives only a small Wiki improvement (-0.59) while damaging new domains ($+1.0$ to $+6.2$). A middle-layer control has negligible effect.

Thus, the component or layer that changes the most is not necessarily the best recovery intervention point (Figure 1). Change magnitude alone is insufficient for identifying forgetting-relevant interventions.

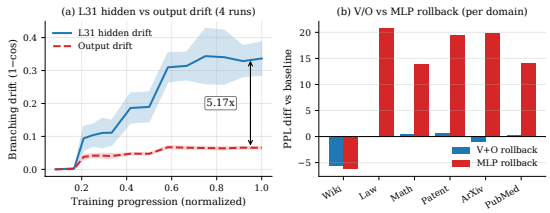


Figure 1: Change magnitude is not the recovery locus. **(a)** L31 drifts most across 4 paired runs, yet rolling it back barely helps Wiki. **(b)** Per-domain rollback: MLP recovers Wiki but damages every new domain; V/O recovers selectively.

4.2 Rollback Reveals a Three-Way Component Contrast

Component-wise rollback shows sharply different profiles for MLP, V/O, and Q/K.

Domain	V+O rollback diff	MLP rollback diff
Wiki	-5.72 ± 0.08	-6.29 ± 0.12
Law	-0.12 ± 0.02	$+20.75 \pm 0.45$
Math	$+0.44 \pm 0.02$	$+13.95 \pm 0.25$
Patent	$+0.76 \pm 0.03$	$+19.56 \pm 0.45$
ArXiv	-1.06 ± 0.03	$+19.89 \pm 0.41$
PubMed	$+0.29 \pm 0.00$	$+14.14 \pm 0.14$

Table 1: Per-domain PPL diff for V+O vs MLP rollback (3 seed mean $\pm$ std).

MLP rollback recovers Wiki but destroys all five later domains. V/O rollback recovers Wiki with small new-domain effects. This indicates that MLP and V/O have different intervention profiles: MLP recovery is entangled with new-domain disruption, while V/O recovery is comparatively selective. Two reference-point controls support this reading: rolling V/O back to the original pretrained weights (rather than the Wiki-end checkpoint) recovers essentially nothing, and per-domain anchor rollback recovers each prior domain selectively while the same-count Q/K rollback does not (Appendix J); an MLP sub-component decomposition further shows that no gate/up/down matrix alone provides damage-free separation, and that their effects are strongly non-additive (Appendix K).

Q/K does not show the same recovery behavior. On the Wiki-only split:

Rollback condition	Wiki PPL diff	Interpretation
Q only	$+0.12 \pm 0.22$	no recovery
K only	$+0.31 \pm 0.29$	no recovery
Q+K	$+0.44 \pm 0.29$	no recovery
V only	-2.68 ± 0.05	recovery
O only	-3.08 ± 0.08	recovery

Table 2: Q/K/V/O decomposition on the Wiki-only held-out split.

The V+O vs Q+K difference has bootstrap 95% CI $[-5.21, -4.72]$.[2] The effect is concentrated in layers L16-L30. Importantly, V/O weight-change magnitude is smaller than Q/K weight-change magnitude (V 0.0182, O 0.0185 vs Q 0.0201, K 0.0194), so the recovery profile is not explained by larger raw weight drift (Figure 2).

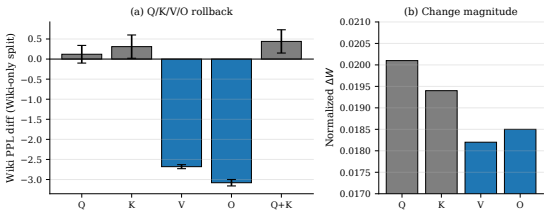


Figure 2: Q/K/V/O rollback decomposition (Wiki-only split, 3 seeds). **(a)** V and O individually recover Wiki; Q and K do not. **(b)** V/O weight change is *smaller* than Q/K, so recovery does not track raw drift.

4.3 V/O Freeze Preserves Prior Domains Without Current-Domain Cost

Rollback is a post-hoc intervention and may introduce co-adaptation mismatch. We therefore freeze V/O from the start of continual pretraining. Across all 360M runs, V/O freeze improves Wiki PPL relative to baseline (Table 3):

Condition	Wiki PPL diff
V/O freeze, stress order, 3 seeds	-2.98 ± 0.13
V/O freeze, 3 seeds $\times$ 2 orders	-2.84 ± 0.29

Table 3: V/O freeze Wiki PPL diff at 360M.

Current-domain learning efficiency is unchanged. In a representative stress-order seed:

Training domain	Baseline	V/O freeze	Diff
Law	4.9	4.9	0.0
Math	8.8	8.7	-0.1
Patent	1.6	1.6	0.0
ArXiv	10.1	10.1	0.0
PubMed	12.0	11.9	-0.1

Table 4: Current-domain PPL: baseline vs V/O freeze.

The effect also accumulates across transitions: Wiki PPL is already improved after Law training and remains better through PubMed. Freezing V/O does not cause visible compensation in other components: MLP, Q, and K normalized weight-change ratios remain within 1% of baseline (MLP

[2]The V-only and O-only values in Table 2 sum to -5.76, close to the -5.72 reported for V+O in Table 1. This apparent additivity should not be over-interpreted: the two are measured on different evaluation splits (the Wiki-only held-out split here, the all-domain split in Table 1), so their numerical agreement is not evidence that the V and O contributions compose linearly.

1.007, Q 0.996, K 0.997). Thus, under this setup, V/O changes are not required to obtain the observed current-domain PPL improvement, while their absence reduces previous-domain degradation.

4.4 Controls Rule Out Simple Parameter-Count Explanations

Freezing a same-parameter-count Q/K subset does not reproduce the V/O effect. Q/K freeze gives Wiki PPL diff $+3.17 \pm 2.53$ across three seeds, with all seeds positive ($+0.26$, $+4.73$, $+4.52$). V/O freeze and Q/K freeze freeze the same number of tensors but have opposite directions.

Random attention-subset freeze further supports V/O specificity. A random 64-tensor attention subset containing 36 V/O and 28 Q/K tensors gives Wiki PPL diff -2.54 ± 0.10, about 85% of the full V/O-freeze effect despite including only 56% of V/O tensors (Table 5, Figure 3).

Freeze condition	Wiki Δ PPL	V/O %
V/O freeze	-2.98 ± 0.13	100
Random attention subset	-2.54 ± 0.10	56
Q/K freeze	$+3.17 \pm 2.53$	0

Table 5: V/O specificity controls (3 seeds each). V/O % denotes the fraction of V/O tensors among the 64 frozen tensors.

Finally, recovery fine-tuning after V/O rollback preserves the Wiki improvement (-4.34 after 100 steps, -4.33 after 300 steps), suggesting that the rollback result is not merely a transient mismatch artifact. As a supplementary continual-learning baseline, Full EWC improves Wiki PPL by -4.76 ± 0.42, V/O-targeted EWC gives an anti-cure of $+2.65 \pm 0.03$, and V/O freeze + Full EWC reaches -7.13 ± 0.15; details are in Appendix E. The V/O vs. Q/K freeze asymmetry also survives in a LoRA adaptation regime with matched trainable-parameter counts (Appendix I).

4.5 Activation Transplantation Modulates Forgetting Bidirectionally

Without changing weights, inter-checkpoint activation transplantation modulates forgetting. Injecting Wiki-end activations into the PubMed-end model improves Wiki PPL; injecting PubMed-end activations into the Wiki-end model worsens Wiki PPL (Table 6).

This activation-level result is related to, but not identical with, the weight-level V/O result. In the activation intervention, V/O-combined patching numerically matches O-only patching. This is expected from the hook mechanism: once o_proj output is overwritten, any upstream v_proj replacement becomes silent for that module output. Therefore, the activation-level result is best described as o_proj-output transferability, while the weight-level result is a V/O projection-bundle intervention profile.

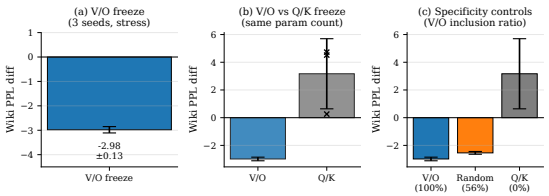


Figure 3: V/O freeze and controls (3 seeds each). **(a)** V/O freeze reduces Wiki forgetting. **(b)** Same-parameter-count Q/K freeze has the opposite sign. **(c)** Random attention subsets recover according to the V/O tensors they contain (56% of them already gives 85% of the effect).

Direction	Wiki Δ PPL	std
Wiki → PubMed (cure)	-6.27	0.13
PubMed → Wiki (elicit)	$+3.26$	0.09

Table 6: Bidirectional V/O activation transplantation (3 seeds). Negative Δ indicates reduced Wiki forgetting; positive Δ indicates elicited forgetting.

4.6 Scale Observations

On SmolLM2-1.7B, the original deterministic run points in the same direction as 360M: V/O freeze improves Wiki PPL by -6.1 ($35.5 \rightarrow 29.4$, -17.3%), and is non-positive for every other evaluation domain (Law -2.2, Math -3.3, Patent 0.0, ArXiv -2.2, PubMed -0.3; Appendix C). After identifying and fixing the data-ordering issue behind the seed-invariance of this configuration (Section 7), we re-ran baseline + V/O freeze across 3 seeds with dataset shuffling enabled: endpoint Wiki PPL improves in all seeds, $\Delta = -10.86 \pm 1.52$ (relative mitigation $20.2 \pm 3.5\%$; Appendix C), consistent in direction and relative magnitude across both data-ordering protocols. Activation transplantation gives forward -5.90 and reverse $+4.70$, and reverse Patent / Law transfers reproduce 360M with high fidelity (Patent -23.54 vs -25.70; Law -20.20 vs -23.48). The one disagreement is V-alone forward patching, which goes from -2.25 at 360M to $+9.86$ at 1.7B; the V/O-combined and O-only patches remain numerically identical at both scales, so the disagree-

ment is isolated to V proj acting alone (Figure 4b, Appendix D). The weight-level V/O freeze effect at 1.7B is thus supported by 3 seeds, while the activation-level 1.7B checks remain single-run scale observations.

In an OLMo-2-1B 3-seed pilot, V/O freeze is lower than baseline in all 108 cross-domain PPL measurements (6 checkpoints $\times$ 6 eval domains $\times$ 3 seeds), with no current-domain harm on the diagonal (6/6 freeze $\leq$ baseline). The 3-seed averaged Wiki PPL reduction is -76.83 ± 8.11 ($-27.4 \pm 2.6\%$ vs. baseline, Appendix C); the other prior domains improve as well (Law -36.2%, Math -26.3%, Patent -54.9%, ArXiv -21.6%). Absolute effects scale with the much larger baseline forgetting in OLMo-2-1B (Wiki endpoint baseline PPL 280.78 ± 4.24, versus 35.5 for SmolLM2-1.7B). The OLMo architecture differs from SmolLM2 in attention configuration (16-head MHA with Q/K RMSNorm) and in training corpus, so consistency across both families argues that the V/O intervention surface is not an artifact of a single architecture or tokenizer family. The pilot uses a single stress-order schedule and a smoke verification cycle, so we report it as supporting evidence rather than as multi-order replication; a multi-order extension is left to future work (Section 7).

On a third model family, Gemma3-1B (hybrid sliding+full-window attention, a different tokenizer and pretraining corpus), a 3-seed replication shows freeze below baseline in 35 of the 36 cross-domain cells, with no current-domain harm; at the endpoint Wiki improves by -61.25 ± 11.18 (-19.7%) and the four other prior domains by -15.5% to -26.0%. The single exception cell and its seed-level variance are reported in Appendix C. Across SmolLM2-360M/1.7B, OLMo-2-1B, and Gemma3-1B, absolute forgetting spans an order of magnitude while endpoint relative mitigation of the anchor domain stays within roughly 10–27%.

4.7 Internal Representation Probes and Downstream Sanity Checks

Two auxiliary diagnostics support the interpretation that V/O interventions act on a recoverable output-side bottleneck rather than wiping previous-domain knowledge.

Probe accuracy decays much less than output PPL. On the forgotten Wiki-end checkpoint, a single-linear-layer next-token probe trained on hidden states drops by only 0.8–2.0 accuracy points relative to the pre-forgetting Wiki checkpoint (L13 24.3 $\rightarrow$ 23.5, L20 29.1 $\rightarrow$ 27.5, L29 33.8 $\rightarrow$ 31.8), while the model's output PPL degrades from 11.3 to 28.9. The hidden representation retains most of the information needed for the next-token decision; the gap between probe accuracy and output PPL is consistent with forgetting acting primarily on the readout/output pathway, which is exactly the locus that V/O interventions touch (Appendix A).

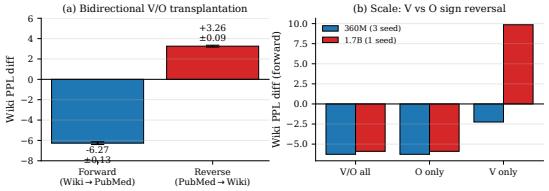


Figure 4: Activation-level results and scale. **(a)** Bidirectional V/O transplantation cures or elicits forgetting without weight changes. **(b)** At 1.7B, V-alone patching reverses sign, while V/O-combined and O-only stay consistent.

Token-rank dynamics agree. Per-token analysis shows the correct token is rarely lost: median correct-token rank moves from 0 to 2, top-5 share from 74% to 62%, and the dominant wrong predictions are generic function words rather than new-domain tokens (Appendix A). This is the rank-pushing failure mode one would expect from a perturbed readout, again consistent with the V/O intervention profile.

Downstream sanity. V/O freeze does not damage general capability. Across five zero-shot benchmarks (ARC-Easy, HellaSwag, PIQA, SciQ, WinoGrande, 3 seeds), V/O freeze mean accuracy is $+0.7$ over the no-freeze baseline (61.7 vs 61.0), with $3/5$ tasks significant before multiple-comparison correction and ARC-Easy and HellaSwag surviving it (m=5; Appendix F). On the domain-matched MMLU subsets, MMLU-Medicine improves by $+5.48$ (uncorrected $p = 0.022$, 22.84 $\rightarrow$ 28.32; trend level after correction), while MMLU-Law and MMLU-Math are within noise. So the PPL recovery co-occurs with downstream non-degradation rather than being a domain-specific PPL artifact.

5 Discussion

5.1 What the Component Contrast Means

The main result is not simply that V/O matters. The result is that three component groups have distinct intervention profiles:

- V/O: previous-domain recovery with little measured new-domain PPL cost.
- MLP: previous-domain recovery coupled with severe new-domain damage.
- Q/K: no previous-domain recovery.

This contrast argues against a purely uniform overwriting story in which reverting any important changed component should affect old and new domains together. It also argues against simple change-magnitude explanations, because Q/K changes more than V/O but does not recover forgetting. A natural reading of these three profiles, consistent with the Transformer circuit decomposition of Elhage et al. (2021), is that the QK and OV circuits play distinct functional roles in continual pretraining: QK routing is largely orthogonal to the recovery axis, OV content writing carries the bulk of the recoverable signal, and MLP stores new-domain knowledge in an entangled fashion that cannot be selectively undone without losing it. Our experiments do not directly verify these mechanistic interpretations, but the asymmetric intervention profile they predict is exactly what we measure.

The V/O freeze effect on Wiki is also strongest when V/O is frozen earliest (A=-3.03) and weakens by the latest freeze point (E=-1.58); the full sequence (A=-3.03, B=-2.46, C=-2.86, D=-2.46, E=-1.58, Appendix B, seed=42 single run) is not strictly monotonic due to a rebound at C, but the overall timing dependence is consistent with V/O changes accumulating throughout training rather than being a simple per-domain byproduct. The earliest-freeze advantage suggests that the V/O drift relevant to forgetting accumulates progressively across the domain sequence, rather than concentrating at any single domain boundary.

5.2 Intervention Surface vs Causal Locus

These experiments identify where interventions are effective, not necessarily where forgetting is stored. A component can be a high-sensitivity intervention surface without being the unique causal locus of forgetting. This distinction is especially important because the weight-level and activation-level results point to related but different objects: V/O projection weights at the weight level, and o_proj output at the activation level.

We therefore avoid claims such as "V/O is the forgetting locus" or "V/O stores previous-domain knowledge." The supported claim is narrower: V/O is a component whose intervention selectively shifts previous-domain PPL under our continual-pretraining setup. This narrower framing is what model-editing work such as Hase et al. (2023) repeatedly motivates — causal-tracing localization need not coincide with edit-success localization, so reporting an intervention profile is methodologically safer than asserting a storage site. Two converging pieces of evidence further constrain the claim. First, V/O combined activation patching is numerically indistinguishable from O-only patching due to the module-level hook (Appendix D), so the activation-level effect is narrower than the weight-level V/O bundle. Second, V-alone patching changes sign with scale (360M -2.25 vs. 1.7B $+9.86$, Appendix C), suggesting that V proj's contribution is architecture-conditional and that the o_proj output is the more architecturally stable surface across SmolLM2 GQA and MHA configurations.

5.3 Why the Recovery Is Partial

At 360M, V/O freeze recovers roughly 10–18% of the forgetting gap depending on metric. This should not be read as the full explanatory fraction of V/O: MLP rollback produces larger previous-domain recovery, but at unacceptable new-domain cost. The V/O-freeze number is better interpreted as the recovery achievable by a coarse component-level intervention that preserves new-domain performance. The freeze-timing analysis (Appendix B, seed=42) supports the same reading: across all five timepoints Wiki improves by only -3.03 to -1.58, with new-domain costs in a tight band (-1.35 to $+0.04$). The decline from earliest to latest freeze is gradual — freezing after the first new domain still captures 81% of the earliest-freeze benefit, and freezing after ArXiv still captures 52% — so the recoverable drift accumulates across the sequence rather than concentrating in one transition. Since even the best timing recovers about -3 PPL, this bounds what a freeze-only intervention can buy back, and suggests that finer-grained interventions — direction-level de-

composition of V/O subspaces, head-level masking informed by per-head Fisher, or MLP-targeted editing that separates new-domain content from interference-relevant features — could close more of the gap without paying MLP's plasticity tax. Because the effect transfers across the SmolLM2 family and to OLMo-2-1B and Gemma3-1B (Appendix C), this floor is unlikely to be model-family-specific. We leave a quantitative ceiling estimate to future work; the V/O freeze + Full EWC stack of Section 5.5 (-7.13 ± 0.15, recovering most of the -7.74 linear sum) indicates that V/O freeze composes with other recovery mechanisms rather than substituting for them.

5.4 Relation to Prior Work

Our findings connect to several existing lines of mechanistic and continual-learning research without subsuming any of them. Compared to MLP-as-knowledge-store evidence (Geva et al., 2021; Dai et al., 2022; Meng et al., 2022), our MLP rollback profile is consistent with that view — MLP carries new-domain content that cannot be removed without losing it. Compared to OV-circuit framing (Elhage et al., 2021), the V/O intervention surface aligns with the content-writing direction that the OV circuit is hypothesized to implement, although our experiments measure intervention sensitivity rather than circuit-level causal flow. Convergent evidence comes from Dong et al. (2025): transformers whose key and query weights are frozen at initialization still form induction heads and remain competitive at language modeling, with knowledge storage attributed largely to MLPs. Their question differs from ours — what stays learnable when QK is never trained, versus what is forgotten and recoverable during continual pretraining — but both point to the QK pathway not being where domain content is written, which is what our Q/K-null profile measures from the other direction. Relative to existing continual-learning methods such as EWC (Kirkpatrick et al., 2017) and PackNet (Mallya and Lazebnik, 2018), our V/O freeze acts as a structural, component-targeted analogue to importance-weighted regularization: it forgoes the per-parameter Fisher signal but achieves a similar recovery direction on Wiki with a simpler intervention. The contrast with parameter-importance methods is not adversarial; it is informative, because it shows that a coarse component selection can recover a non-trivial fraction of the forgetting gap without paying the new-domain plasticity tax incurred by either MLP rollback or aggressively tuned Fisher regularization (Appendix E).

5.5 Comparison with EWC Variants

To check whether the V/O profile is reproduced by parameter-importance regularization, we ran four EWC variants on the same 360M setup (Appendix E, λ swept on the seed=42 grid, best λ extended to 3 seeds where applicable). Summary in Table 7.

Condition	cure	n-max	n-mean
V/O freeze (3s)	$\mathbf{-2.98}$	~ 0	~ 0
Full EWC λ1k (3s)	$\mathbf{-4.76}$	$+0.98$	$+0.42$
V/O-EWC λ1k (1s)	$+4.49$	$+1.91$	$+0.98$
V/O-EWC λ50k (3s)	$+2.65$	$+6.4$	$+1.6$
Q/K-EWC λ1k (1s)	$+4.84$	$+2.25$	$+1.20$
MLP-EWC λ1k (1s)	$+3.60$	$+2.45$	$+1.12$
V/O freeze + Full EWC (3s)	$\mathbf{-7.13}$	~ 0	~ 0

Table 7: V/O freeze vs EWC variants on Wiki cure / new-domain max damage / new-domain mean damage (PPL diff vs baseline; lower cure = better recovery; lower n-max/n-mean = less plasticity tax). Full data + seed=1 outlier in Appendix E.

Full EWC at $\lambda = 1000$ is the only sweep value that simultaneously satisfies Wiki cure ≤ -1.0 and new-domain max damage $\leq +2.0$, improving Wiki by -4.76 ± 0.42 at $+0.42$ mean new-domain damage; beyond $\lambda \approx 20$k it trades plasticity for recovery roughly one-for-one (full sweep in Appendix E). V/O-targeted EWC has the opposite sign at every λ in our grid (Wiki $+2.6$ to $+4.5$): anchoring V/O to its Wiki-end values while letting MLP/Q/K drift freely is structurally different from preventing V/O drift by freezing, and induces compensatory MLP changes that hurt Wiki. V/O freeze plus Full EWC, in contrast, stacks to -7.13 ± 0.15, exceeding either intervention alone and recovering most of the -7.74 linear sum. We therefore use EWC here as a composition probe — a way to test whether the V/O surface is redundant with parameter-importance regularization — rather than as a state-of-the-art baseline to be beaten. Two takeaways: (1) the V/O intervention surface is not redundant with Fisher regularization and can be combined with it; (2) the success of V/O *freeze* against the failure of V/O-targeted *EWC* shows that what matters is preventing drift in V/O, not anchoring V/O to a reference value — the latter creates a stiffness asymmetry that other components compensate for, which our random-subset and same-parameter-count controls ruled out for

freeze (Section 4.4).

5.6 Practical Implications for Continual Pretraining

Three practical recommendations follow from the intervention profile, scoped to the continual-pretraining regime we study (decoder-only LMs, six text domains, 15M tokens per domain). *(a) Default to V/O freeze when prior-domain protection matters.* V/O freeze costs nothing at inference, requires no Fisher estimation, and is implementation-light: it flips a single `requires_grad` flag on V and O projections. In our setup it preserves $> 99\%$ of current-domain learning efficiency (Table 4) while reducing Wiki PPL gap by -2.84 to -2.98 and producing a $+5.48$ MMLU-Medicine signal (trend level after correction). The risk-adjusted profile is more favorable than aggressive Fisher regularization. *(b) Treat MLP as the new-knowledge layer.* Our MLP rollback experiments make it explicit that MLP changes carry new-domain content that cannot be reversed without losing it. Any continual-learning method that wholesale-protects MLP is implicitly choosing to under-learn the new domain. Methods like LoRA (Hu et al., 2022) or PackNet (Mallya and Lazebnik, 2018) that route new capacity through MLP-adjacent additions, rather than overwriting trunk MLP, are aligned with this asymmetry. *(c) Q/K is not a useful protection target.* Q/K rollback does not recover forgetting, and same-parameter-count Q/K freeze moves Wiki in the harmful direction ($+3.17 \pm 2.53$). Methods should therefore not expect protective gains from constraining Q/K, even though it remains available for unrelated purposes such as attention pruning.

5.7 Future Work

Several directions follow directly from the intervention profile. First, the activation-level result is best characterized as o_proj-output transferability rather than a full V/O bundle effect (Appendix D); the same forward-hook mechanism prevents us from cleanly separating the V and O contributions at the activation level. Resolving this would require either gradient-level patching (which avoids the module-output override) or a controlled comparison between pre-attention and post-attention residual-stream patches, neither of which we attempt here. Second, our freeze-timing analysis (Appendix B) uses a single seed; multi-seed verification of the A–E timing curve would establish whether the saturation pattern we observe is robust or seed-specific. Third, the V/O freeze + Full EWC stack (-7.13, Section 5.5) is the largest cure we measured without paying a new-domain damage penalty, but we have not explored more aggressive stacks (V/O freeze + replay; V/O freeze + LoRA-routed adaptation), which could plausibly close more of the gap. Fourth, the OLMo-2-1B pilot is single-order and smoke-verified; a multi-order replication across the OLMo family would directly test whether the V/O profile is invariant to attention configuration (GQA vs MHA) and tokenizer choice.

Finally, our results establish *where* interventions act but not *why* the OV pathway carries the recoverable component. A candidate account is gradient-geometric: the OV path writes directly into the residual stream, whereas QK gradients act on the routing distribution through the softmax. Testing whether that asymmetry drives the effect needs gradient-level analysis — for instance per-component gradient alignment across domain boundaries — which we leave as the central mechanistic question our profile raises.

6 Conclusion

Continual pretraining forgetting does not respond uniformly to Transformer component interventions. In our setup, V/O interventions selectively recover previous-domain PPL with little new-domain cost, MLP interventions recover previous domains only while damaging new domains, and Q/K interventions do not recover forgetting. Controls show that the V/O effect is not explained by parameter count, random subset selection, or compensatory MLP/Q/K weight changes; internal-representation probes and downstream evaluations are consistent with this picture; and the V/O freeze effect reproduces on SmolLM2-1.7B, OLMo-2-1B, and Gemma3-1B, combining complementarily with Full EWC. Activation transplantation gives a related inference-time result, though its effect is primarily o_proj-output transferability rather than a copy of the weight-level V/O result. These findings establish a component-level intervention profile and a practical intervention (V/O freeze) that costs essentially nothing in current-domain learning *in the regimes tested*. Validating it as a general default at larger scales, and separating causal loci from high-sensitivity intervention surfaces, remain future work.

7 Limitations

Appendix M expands most of the points below; Appendix L summarizes the seed coverage of the principal results.

Scale and generalization. Our evidence comes from SmolLM2-360M and SmolLM2-1.7B, with 3-seed single-order pilots on OLMo-2-1B and Gemma3-1B. The 1.7B weight-level result is replicated across 3 seeds after we identified and fixed a data-ordering issue that had made that configuration seed-invariant (Appendix C); our original explanation of that invariance, in terms of size-dependent numerical determinism, was incorrect. Verification at 7B and above, on multilingual or more heterogeneous domains, and under schedules longer than 15M tokens per domain remains future work, as does a sensitivity analysis for the relatively high learning rate (5×10^{-4}).

Statistical support. The weight-level 1.7B freeze is 3-seed, but the activation-level 1.7B transplantation is a single run, and the layer scan (Appendix G) and the component-targeted EWC variants (Appendix E) remain seed 42 only. Extending Q/K freeze to 3 seeds revealed that its original single-seed value was the lowest of the three ($+0.26$ against $+3.17 \pm 2.53$), so our remaining single-seed numbers should be read as indicative rather than established.

Claim scope. We do not distinguish whether V/O is the causal locus of forgetting or a high-sensitivity intervention surface. Because forward hooks override module outputs, our activation-level "V/O effect" is more precisely o_proj output transferability (Appendix D), and V-alone patching reverses sign between 360M and 1.7B (-2.25 versus $+9.86$) for reasons we do not establish. Our setting is continual *pretraining*, so comparison with continual fine-tuning results (Li and Lee, 2024; Zheng et al., 2025) is indirect, and the resemblance between inter-checkpoint transplantation and function- or steering-vector methods is conceptual rather than a direct mapping.

Evaluation. Results are PPL-centric, supplemented by next-token accuracy, intermediate-layer probes, and five zero-shot benchmarks. In the main 360M setup, V/O freeze recovers only 10–18% of the forgetting gap, and we do not separate MLP's forgetting contribution from the new-domain knowledge it stores. The MMLU-Medicine signal is trend-level once corrected for multiple comparisons, and the downstream accuracy of activation-patched models is unverified beyond that subset.

Baselines. We compare against EWC variants only (Appendix E) and verify that the V/O versus Q/K asymmetry survives in a LoRA adaptation regime (Appendix I). A head-to-head comparison against replay-based and LoRA-based continual-learning *methods* remains future work.

Societal impact. This is an analysis paper, and the intervention it motivates (V/O freeze) is capability-preserving rather than capability-adding. Two dual-use considerations apply. First, the same localization could be used to selectively *preserve* undesirable content: freezing components that carry biased or outdated associations would shield them from revision by later training, so component-level protection should be paired with bias and safety evaluation of what is being preserved. Second, inter-checkpoint component swapping is weight-level editing that leaves architecture and parameter count unchanged, and so is not detectable by casual inspection; provenance tracking of released checkpoints, such as per-component weight hashes, is a straightforward safeguard. Both require full access to weights and training checkpoints, which bounds the marginal risk beyond what open-weight releases already entail. We release our intervention code so that such modifications remain transparent and auditable.

References

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, and 3 others. 2025. SmolLM2: When smol goes big — data-centric training of a small language model. *arXiv preprint arXiv:2502.02737*.

Andrew Brock, Theodore Lim, James M. Ritchie, and Nick Weston. 2017. FreezeOut: Accelerate training by progressively freezing layers. *arXiv preprint arXiv:1706.04983*.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Andrea Cossu, Tinne Tuytelaars, Antonio Carta, Lucia Passaro, Vincenzo Lomonaco, and Davide Bacciu. 2022. Continual pre-training mitigates forgetting in language and vision. *arXiv preprint arXiv:2205.09357*.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Yihe Dong, Lorenzo Noci, Mikhail Khodak, and Mufan Li. 2025. Is random attention sufficient for sequence modeling? disentangling trainable components in the transformer. *arXiv preprint arXiv:2506.01115*.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, and 6 others. 2021. A mathematical framework for transformer circuits. https://transformer-circuits.pub/2021/framework/index.html. Transformer Circuits Thread.

Robert M. French. 1999. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. 2023. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. In *Advances in Neural Information Processing Systems (NeurIPS), Spotlight*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. In *International Conference on Learning Representations (ICLR)*.

Zixuan Ke, Yijia Shao, Haowei Lin, Tatsuya Konishi, Gyuhak Kim, and Bing Liu. 2023. Continual pre-training of language models. In *International Conference on Learning Representations (ICLR)*.

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. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences (PNAS)*, 114(13):3521–3526.

Chen-An Li and Hung-Yi Lee. 2024. Examining forgetting in continual pre-training of aligned large language models. *arXiv preprint arXiv:2401.03129*.

Arun Mallya and Svetlana Lazebnik. 2018. PackNet: Adding multiple tasks to a single network by iterative pruning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Michael McCloskey and Neal J. Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation*, 24:109–165.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems (NeurIPS)*.

Vinay Venkatesh Ramasesh, Ethan Dyer, and Maithra Raghu. 2021. Anatomy of catastrophic forgetting: Hidden representations and task semantics. In *International Conference on Learning Representations (ICLR)*.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, and 24 others. 2024. 2 OLMo 2 furious. *arXiv preprint arXiv:2501.00656*. Published at COLM 2025.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Chengyue Wu, Yukang Gan, Yixiao Ge, Zeyu Lu, Jiahao Wang, Ye Feng, Ying Shan, and Ping Luo. 2024.

LLaMA pro: Progressive LLaMA with block expansion. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Atsuki Yamaguchi, Terufumi Morishita, Aline Villavicencio, and Nikolaos Aletras. 2025. Mitigating catastrophic forgetting in target language adaptation of LLMs via source-shielded updates. *arXiv preprint arXiv:2512.04844*. Accepted to ACL 2026.

Friedemann Zenke, Ben Poole, and Surya Ganguli. 2017. Continual learning through synaptic intelligence. In *International Conference on Machine Learning (ICML)*.

Junhao Zheng, Xidi Cai, Shengjie Qiu, and Qianli Ma. 2025. Spurious forgetting in continual learning of language models. In *International Conference on Learning Representations (ICLR)*.

Zhen Zhu, Yiming Gong, Yao Xiao, Yaoyao Liu, and Derek Hoiem. 2025. How to teach large multimodal models new skills. *arXiv preprint arXiv:2510.08564*.

A Internal Representation Analysis (probe + token-level failure)

Probing. When a linear probe (a single linear layer, hidden_dim $\rightarrow$ vocab_size) is trained on intermediate-layer hidden states to predict the next Wiki token, probe accuracy after forgetting decreases by only $-0.8 \sim -2.0$ percentage points, whereas model output (PPL) degrades substantially from 11.3 to 28.9. Probes were trained at L13, L20, and L29 respectively (Wiki held-out 1,000 samples, AdamW, lr=1e-3, 10 epochs, weight decay=0.01), and the pre-forgetting absolute accuracies were L13: 24.3%, L20: 29.1%, L29: 33.8%. After forgetting these become 23.5%, 27.5%, and 31.8% respectively, so in all three layers we observe the same pattern: probe accuracy drops only marginally compared with the sharp degradation of the output PPL. Information remains internally but the output is heavily degraded, suggesting that forgetting may be closer to a weakening of the output pathway than to loss of knowledge itself.

Token-level failure analysis. After forgetting, the correct token does not disappear; rather, its rank is pushed down (median rank $0 \rightarrow 2$, top-5 share $74\% \rightarrow 62\%$). The wrong predictions are not recent-domain-specific tokens but regressions toward generic function words (the, comma, etc.).

B Freeze Timing Comparison

We freeze V/O at 5 distinct training timepoints (A: right after Wiki, B: after Law, C: after Math, D: after Patent, E: after ArXiv) and measure 6-domain PPL at the end of training (after PubMed) for each freeze run (360M, seed=42, stress order). The full 5×6 matrix of PPL diffs vs no-freeze baseline (Figure 5):

Freeze	Wiki	Law	Math	Pat.	ArX.	PubM.
A (Wiki)	$\mathbf{-3.03}$	-0.49	-0.79	-0.10	-1.05	-0.04
B (Law)	$\mathbf{-2.46}$	-1.35	-0.69	-0.13	-1.09	$+0.00$
C (Math)	$\mathbf{-2.86}$	-1.04	-1.07	-0.10	-1.10	-0.01
D (Pat.)	$\mathbf{-2.46}$	-0.56	-0.58	-0.38	-1.17	$+0.03$
E (ArX.)	$\mathbf{-1.58}$	-0.26	-0.18	-0.14	-0.96	$+0.04$

Table 8: V/O freeze timing $\times$ evaluation domain PPL difference vs baseline (seed=42, single run). Freeze A–E indicates the training checkpoint at which V/O is frozen.

V/O freeze improves Wiki perplexity at all 5 timepoints (Wiki PPL diff range -3.03 to -1.58). The Wiki improvement is strongest at A (right after Wiki) and weakest at E (after ArXiv); the full sequence is A $-3.03 \rightarrow$ B $-2.46 \rightarrow$ C $-2.86 \rightarrow$ D $-2.46 \rightarrow$ E -1.58, which is not strictly monotonic because of a rebound at C. This overall timing dependence is consistent with V/O changes accumulating throughout training: the earlier V/O is frozen, the more accumulated drift is prevented. New-domain costs remain small at all 5 timepoints (range -1.35 to $+0.04$). We report this as a single-seed measurement (seed=42); multi-seed verification of the freeze-timing pattern is a future task.

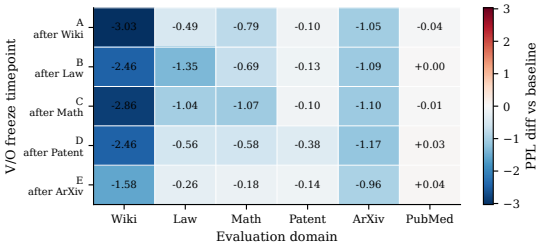


Figure 5: V/O freeze timing $\times$ evaluation domain (5×6 PPL difference vs baseline, seed=42 single run). Wiki improvement is strongest at all timepoints, with overall weakening A$\rightarrow$E (not strictly monotonic).

C Scale and Family Details: SmolLM2-1.7B, OLMo-2-1B, and Gemma3-1B

C.1 SmolLM2-1.7B Detailed Results (Section 4.6 consolidation)

Domain	Baseline	V/O Freeze	diff
Wiki	35.5	29.4	$\mathbf{-6.1}$
Law	11.8	9.6	-2.2
Math	20.0	16.7	-3.3
Patent	2.1	2.1	0.0
ArXiv	17.4	15.2	-2.2
PubMed	10.4	10.1	-0.3

Table 9: SmolLM2-1.7B V/O Freeze per-domain final-checkpoint PPL (1 run, seed=42).

(a) Weight-level — V/O Freeze (1 run). For every domain, V/O freeze yields a PPL equal to or lower than baseline.

Condition	360M	1.7B	Match
Forward V/O all (cure)	-6.27 ± 0.13	-5.90	✓
Reverse V/O all (elicit)	$+3.26 \pm 0.09$	$+4.70$	✓
Forward O only	-6.27 ± 0.13	-5.90	✓
Reverse Patent transfer	-23.54 ± 0.23	-25.70	✓
Reverse Law transfer	-20.20 ± 0.15	-23.48	✓
Forward V only	-2.25	$+9.86$	×

Table 10: SmolLM2-1.7B activation transplantation vs 360M (3 seed mean for 360M; seed=42 single run for 1.7B). ✓=consistent direction, ×=sign reversal.

(b) Activation-level — Inter-checkpoint Activation Transplantation (1 run, Section 4.5 setup).

C.2 SmolLM2-1.7B Multi-Seed Replication (data-shuffle fix)

After identifying the cause of the seed-invariance described in Section 7 (deterministic ordering of the HF streaming datasets) and fixing it with an environment-guarded dataset shuffle (disabled by default, so results produced before the fix are reproduced byte-exact), we re-ran the 1.7B baseline + V/O freeze across 3 seeds with shuffling enabled, using the same domains, token budget, and schedule as the original run.

Seed	Baseline	V/O Freeze	Δ	Rel.
42	56.34	47.57	-8.77	-15.6%
1	54.25	42.76	-11.49	-21.2%
7	51.53	39.20	-12.33	-23.9%
mean	54.04	43.18	$\mathbf{-10.86 \pm 1.52}$	$\mathbf{-20.2 \pm 3.5\%}$

Table 11: SmolLM2-1.7B endpoint Wiki PPL with dataset shuffling enabled (3 seeds). Δ = freeze − baseline; Rel. = relative mitigation vs. baseline.

Endpoint Wiki PPL improves in all 3 seeds. Together with the original no-shuffle run (Table 9: -6.1, -17.3%), the weight-level V/O freeze effect at 1.7B is consistent across both data-ordering protocols. Note that with shuffling enabled the baseline forgetting itself is larger (endpoint Wiki PPL ∼54 vs. ∼35.5 without shuffling), so we compare directions and relative mitigation rather than absolute values.

C.3 Gemma3-1B: Third Model Family (3 seeds)

To test the V/O freeze effect on a third model family, we ran baseline + V/O freeze × 3 seeds (42, 1, 7) on Gemma3-1B (hybrid sliding+full-window attention, a different tokenizer and pretraining corpus from both SmolLM2 and OLMo-2), with the same six domains, token budget, and stress order. A smoke test confirmed that Gemma3-1B is multi-seed effective (seeds diverge from the first measured step), unlike the pre-fix 1.7B configuration of C.2.

Domain	Baseline	V/O Freeze	Δ	Rel.
Wiki	309.5 ± 9.2	248.2 ± 7.3	$\mathbf{-61.25 \pm 11.18}$	-19.7%
Law	128.7 ± 2.1	108.7 ± 1.3	-20.01 ± 3.23	-15.5%
Math	143.1 ± 0.7	120.0 ± 2.6	-23.11 ± 2.18	-16.2%
Patent	41.9 ± 0.5	31.0 ± 0.8	-10.90 ± 0.72	-26.0%
ArXiv	68.8 ± 3.0	56.8 ± 2.6	-12.00 ± 2.70	-17.4%
PubMed	21.4 ± 0.1	19.7 ± 0.1	-1.68 ± 0.11	(current)

Table 12: Gemma3-1B endpoint (post-PubMed) per-domain PPL, 3 seed mean ± std. Δ = freeze − baseline; Rel. = relative mitigation. Mean mitigation over the five prior domains is -19.0%; the $-19.7 \pm 3.3\%$ figure is the Wiki-only 3-seed statistic.

Across the full cross-domain PPL matrix (6 checkpoints × 6 eval domains), **V/O freeze lowers the 3-seed mean PPL in 35 of the 36 cells**, with no current-domain harm on the diagonal (6/6, per-cell std < 0.11). The single exception is the ArXiv-checkpoint → Wiki cell, where the freeze−baseline diff is $+28.70 \pm 63.33$: the positive mean is produced by one seed (seed 7, $\approx +118$) while the other two seeds remain below baseline in that cell. We report this anomaly as-is — it does not affect the endpoint results above, and whether it relates to Gemma's sliding-window attention and long-range information integration is left to future analysis. Absolute forgetting in Gemma3-1B (endpoint Wiki baseline PPL ∼310) is an order of magnitude larger than in SmolLM2, while the relative mitigation remains in the same range as the other models.

C.4 OLMo-2-1B Pilot (3 seeds)

The OLMo-2-1B pilot (`allenai/OLMo-2-0425-1B` (Team OLMo et al., 2024); 16-head MHA with additional Q/K RMSNorm, a different pretraining corpus and tokenizer from SmolLM2) uses the same six

domains, token budget, and stress order, with seeds 42, 1, 7 after two smoke verification cycles.

Prior domain	Δ Wiki-equivalent	Rel.
Wiki	-76.83 ± 8.11	$-27.4 \pm 2.6\%$
Law	—	-36.2%
Math	—	-26.3%
Patent	—	-54.9%
ArXiv	—	-21.6%

Table 13: OLMo-2-1B endpoint (post-PubMed) forgetting mitigation, 3 seed mean ± std where available. Wiki endpoint baseline PPL is 280.78 ± 4.24.

Across the full cross-domain PPL matrix (6 checkpoints × 6 eval domains × 3 seeds), **V/O freeze is below baseline at all 108 measurements**, with no current-domain harm on the diagonal (6/6 freeze ≤ baseline). Absolute forgetting here is roughly an order of magnitude larger than in SmolLM2, while the relative mitigation stays in the same band as the other families. The pilot uses a single domain order, so we report it as supporting evidence rather than a multi-order replication.

D V vs O Decomposition and Mechanism Note

This section reports the V vs. O decomposition results of *activation patching* in Section 4.5, separately from the V/O separation in *weight rollback* of Section 4.2 (V -2.68, O -3.08, Q+K $+0.44$). Since the two experiments use different intervention modes (weight vs. inference-time activation), absolute values are not directly comparable, but their directions are.

Component	Wiki diff	Note
V/O all	$\mathbf{-6.27 \pm .13}$	ref
O only	$\mathbf{-6.27 \pm .13}$	100% of V/O
V only	-2.25	$n = 1$

Table 14: V vs O decomposition, forward (360M, 3 seed except V only n=1).

The same asymmetry holds in the reverse direction:

Component	Wiki diff	Note
V/O all	$+3.26 \pm .09$	ref
O only	$+3.26 \pm .09$	100% of V/O
V only	$+1.51$	$n = 1$, weak

Table 15: V vs O decomposition, reverse (360M, 3 seed except V only n=1).

Mechanism note — V/O combined patching is numerically identical to O alone. In both directions, on both 360M and 1.7B, the V/O combined patching result and the O alone patching result agree to the decimal place (e.g., 1.7B forward Wiki -5.9025 is identical). This is because the attention computation order is $q, k, v \rightarrow$ softmax-weighted sum $\rightarrow$ o_proj $\rightarrow$ residual, and the forward hook force-overrides the module output with the source value. Once o_proj output is overwritten with the source value, the v_proj swap becomes silent inside the attention computation. Therefore, what this paper calls the "V/O effect" is, more precisely, *o_proj output transferability*.

To see the independent contribution of V proj, V alone must be patched, and in this case it is -2.25 (weak cure) on 360M and $+9.86$ (destruction) on 1.7B — *the direction depends on scale*.

Control (360M, s.42)	Wiki diff vs. all	
V/O random 32 (v17, o15)	$\mathbf{-4.41 \pm 1.16}$	V/O all $= -6.27$
Q/K random 32 (q17, k15)	-0.43 ± 0.75	Q/K all $= +2.76$

Table 16: Random subset control (3 random seeds, n=32 picks from pool=64) showing V/O specificity at the activation level.

Random subset control (V/O specificity). For the Q/K random subset (-0.43 ± 0.75), the std exceeds the mean, so the effect is indistinguishable from zero; the V/O random subset (-4.41 ± 1.16) shows a significant effect. Within the random sample, the larger the O proportion, the larger the effect size (rng=0 v=19, o=13: -3.18; rng=1 v=17, o=15: -5.48), so the O ≫ V asymmetry is reproduced even within random selection.

Terminology note. Because in the weight-level results (Section 4.2–4.4: rollback Δ V -2.68, O -3.08; freeze -2.98 ± 0.13) both V and O projections show an effect, in weight-level descriptions we use the bundled term *V/O projection*. However, the activation-level results (Appendix G, Section 4.5, and this Appendix: V/O combined patch ≡ O alone patch; V alone is scale-dependent) are primarily explained by o_proj output. We therefore interpret the activation-level result not as a 'joint V/O mechanism' but as 'o_proj output transferability.' Accordingly, we present the weight-level and activation-level results not as *convergent evidence for a single mechanism* but as *related observations at two levels* that contribute to forgetting recovery.

E EWC Details (layer Fisher / λ sweep / seed robustness)

E.1 Layer-wise Fisher distribution (Component $\times$ Layer band)

Component	front (L0-L15)	back (L16-L30)	top (L31)
V	**8.77e-8**	2.68e-8	2.88e-8
O	**4.83e-9**	2.13e-9	1.46e-9
MLP	1.29e-7	2.09e-8	**1.40e-6**
K	5.00e-9	6.60e-9	6.75e-9
Q	2.98e-9	2.49e-9	1.62e-9

Table 17: Layer-wise Fisher distribution (mean Fisher per component, by layer band).

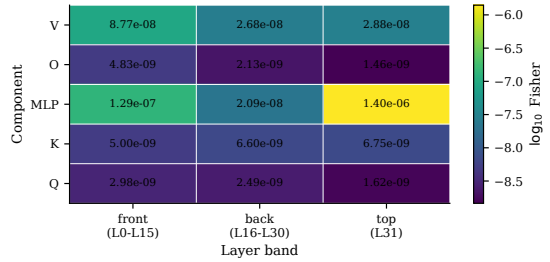


Figure 6: Fisher Information heatmap.

E.2 λ Sweep per-domain results (seed=42, Full EWC)

baseline: Wiki=28.85, Law=11.08, Math=14.24, Patent=2.58, ArXiv=15.62, PubMed=11.99

λ	Wiki	Law	Math	Pat.	ArX.	PM.
1k ✓	**−4.58**	−0.16	+0.97	+0.53	+0.76	+0.14
5k	−8.80	+0.70	+0.75	+3.77	+1.89	−0.11
20k	−11.59	+3.37	+0.69	+9.55	+6.11	−0.17
100k	−13.95	+6.38	+0.71	+14.07	+11.92	+0.09

Table 18: Full EWC λ sweep, seed=42 (Wiki cure + 5 new-domain diff). ✓=$\lambda = 1000$ uniquely satisfies cure ≤ -1.0 AND new-max $\leq +2.0$; cure/damage ratio drops 4.74, 2.34, 1.21, 0.99 across λ.

As λ increases, Wiki cure increases monotonically, but simultaneously the per-domain plasticity damage rises disproportionately. Per-domain monotonicity: Patent ($+0.53 \rightarrow +14.07$), ArXiv ($+0.76 \rightarrow +11.92$), and Law ($-0.16 \rightarrow +6.38$) rise the most steeply. Math and PubMed are barely affected. Derived summaries: new-max progression $+0.97, +3.77, +9.55, +14.07$; new-mean $+0.45, +1.40, +3.91, +6.63$; cure/damage ratio $4.74, 2.34, 1.21, 0.99$ across $\lambda \in \{1\text{k}, 5\text{k}, 20\text{k}, 100\text{k}\}$. $\lambda = 1000$ **is the only value that simultaneously satisfies cure** ≤ -1.0 **AND new-max** $\leq +2.0$.

E.3 Fisher Seed-robustness

Fisher is computed at each seed's own Wiki-end checkpoint (Wiki held-out 457 samples, batch=4, max_length=256).

seed	V Fisher mean	O Fisher mean	V/O ratio
42	5.73e-8	3.46e-9	**16.58**×
1	5.65e-8	3.41e-9	**16.58**×
7	5.70e-8	3.44e-9	**16.59**×

Table 19: Fisher seed-robustness: the V/O ratio agrees to within $0.01\times$ across 3 seeds.

All 3 seeds give V/O ratio = $16.58 \sim 16.59\times$, identical. The Fisher distribution is robust.

E.4 V/O-targeted EWC λ-sweep + 3-seed extension

(1) seed=42, 5-λ grid PubMed-end PPL diff vs baseline:

λ	Wiki	Law	Math	Pat.	ArX.	PM.
1k	+4.49	+1.51	+1.91	+0.25	+0.85	+0.36
5k	+3.62	+1.50	+1.66	+0.23	+0.80	+0.29
20k	+2.65	+1.22	+1.38	+0.15	+0.46	+0.24
50k ⋆	+2.64	+1.08	+1.43	−0.01	+0.33	+0.26
100k	+3.00	+0.84	+1.41	−0.06	+0.63	+0.29

Table 20: V/O-targeted EWC 5-λ grid (seed=42). ⋆=best behavior (lowest Wiki diff + new-mean). New-max ranges $+1.38$ to $+1.91$; new-mean $+0.62$ to $+0.97$.

Per-λ derived summaries (omitted from Table 20 for compactness): new-max $+1.91, +1.66, +1.38, +1.43, +1.41$; new-mean $+0.97, +0.89, +0.69, +0.62, +0.62$ across $\lambda \in \{1\text{k}, 5\text{k}, 20\text{k}, 50\text{k}, 100\text{k}\}$. $\lambda = 50000$ is selected as best behavior because it minimizes Wiki diff jointly with new-mean.

(2) $\lambda = 50000$ 3-seed (PubMed-end PPL diff vs baseline):

seed	Wiki	Law	Math	Pat.	ArX.	PubM.
42	+2.64	+1.08	+1.43	−0.01	+0.33	+0.26
1	+2.63	+16.28	+1.53	+0.03	+0.54	+0.26
7	+2.68	−0.02	+1.43	−0.02	+0.48	+0.27
mean	**+2.65**	+5.78	+1.46	0.00	+0.45	+0.26
std	**0.03**	9.45	0.06	0.03	0.11	0.01

Table 21: V/O-targeted EWC $\lambda = 50000$ 3-seed extension. For the seed=1 Law outlier, see Appendix M, caveat (4).

E.5 4-variant comparison table + seed=1 outlier details

Detailed 4-variant comparison table, footnote, and seed=1 outlier analysis for the EWC comparison.

Setup: Section 3 (same SmolLM2-360M, 6 domains, 15M tokens/domain). For λ, we adopt $\lambda = 1000$, the only value in the Full EWC seed=42 grid sweep ($\{1000, 5000, 20000, 100000\}$) that uniquely satisfies the cure ≤ -1.0 + new-max $\leq +2.0$ criteria. Only Full EWC is extended to 3 seeds (42, 1, 7); the Q/K-targeted and MLP-targeted variants are seed 42 single-seed diagnostics with the same $\lambda = 1000$ for specificity control. The V/O-targeted variant is additionally swept on a 5-value grid (Appendix E.4 table 20) + 3-seed extension at best behavior $\lambda = 50000$ (Appendix E.4 table 21).

Condition	n	cure	n-mx	n-mn
Baseline	—	0	0	0
V/O freeze	3	**−2.98**	∼0	∼0
Full EWC	**3**	**−4.76**	+.98	+.42
V/O-EWC λ1k	1	+4.49	+1.91	+.98
V/O-EWC sweep	5	+2.6–+4.5	+1.4–+1.9	+.6–+1.0
V/O-EWC λ50k	**3**	**+2.65**	**+6.4**†	**+1.6**†
Q/K-EWC λ1k	1	+4.84	+2.25	+1.20
MLP-EWC λ1k	1	+3.60	+2.45	+1.12

Table 22: 4-variant EWC comparison vs V/O freeze (Wiki cure / new-domain max / new-domain mean). 3-seed entries report mean; std bands omitted (Full EWC ±.42/.03/.07; V/O-EWC λ50k ±.03/8.57/1.99). †Driven by seed=1 outlier (Law +16.28 at Patent step).

seed=1 outlier details. At V/O-targeted EWC $\lambda = 50000$ seed=1, the Law trajectory showed unusually large forgetting at the Patent training step (PubMed-end Law diff +16.28, while the other seeds were around +1.0). This is independent of the Wiki cure result ($+2.65 \pm 0.03$), but suggests that when only V/O proj is anchored and the other components (MLP/Q/K) undergo free training, unconstrained drift may induce additional forgetting in some seed/domain combinations (Appendix M, caveat (4)).

F Downstream Details

F.1 General zero-shot benchmarks (5 tasks)

We evaluated three conditions (Wiki boundary / Final baseline / Final V/O freeze) on ARC-Easy, HellaSwag, PIQA, SciQ, WinoGrande across 3 seeds. Per-cell standard deviations are small (Wiki boundary ±0.0 to ±0.4; Final baseline ±0.1 to ±0.7; Final V/O freeze ±0.0 to ±0.8) and omitted from Table 23 for compactness.

V/O freeze − baseline mean diff **+0.7**. Paired t-test: ARC-Easy $p = 0.008$, HellaSwag $p = 0.008$, SciQ $p = 0.021$ (3 significant); PIQA $p = 0.556$, WinoGrande $p = 0.713$ (not significant). The improvement direction observed on PPL is also confirmed on downstream tasks; the mean diff of +0.7 is smaller in magnitude than the PPL improvement (−2.84), suggesting that general English benchmarks only partially reflect the recovery of the 6-domain specific capability.

Condition	ARC-E	HSwag	PIQA	SciQ	WinoG	Mean
Wiki bndy	66.7	39.0	69.9	89.5	58.9	64.8
Final base	59.9	38.2	65.7	85.1	56.1	61.0
Final V/O	60.8	38.8	66.0	86.9	55.8	61.7

Table 23: Zero-shot benchmark accuracy (3 seed mean; std ≤ 0.8 all cells, omitted). HSwag=HellaSwag, WinoG=WinoGrande, Wiki bndy=pre-forgetting Wiki boundary checkpoint.

F.2 Domain-matched MMLU evaluation (12 subsets)

condition	MMLU-Law	MMLU-Math	MMLU-Med
Final base	25.21	23.04	22.84
Final V/O	24.57	22.98	**28.32**
diff	−0.64	−0.05	**+5.48**
p-value	0.636	0.924	**0.022** ✓

Table 24: MMLU evaluation by training-domain group (Law/Math/Med = 3/4/5 subsets; 3 seed mean; std ≤ 1.97 all cells).

On MMLU-Medicine, which corresponds to PubMed training, we observe a +5.48 improvement over baseline ($p = 0.022$, single-test basis; baseline 22.84 ± 0.81, V/O freeze 28.32 ± 1.97). MMLU-Law baseline is 25.21 ± 1.36 / V/O freeze 24.57 ± 0.74; MMLU-Math baseline 23.04 ± 1.06 / V/O freeze 22.98 ± 1.14 (omitted from Table 24 for compactness). In particular, `mmlu_professional_medicine` shows a strong signal of +16.79 (baseline 19.36 → V/O freeze 36.15, $p = 0.021$), and `mmlu_anatomy` shows +3.46 ($p = 0.034$). That is, the PPL recovery for PubMed is also observed at the *trend level* of medical-domain task accuracy (see F.3 for multiple-comparison correction).

MMLU-Law and MMLU-Math have baseline accuracy near chance level (25%) (23∼25%), with null p-values (0.636, 0.924) — owing to the limitations of SmolLM2-360M's size and the format mismatch between the training data and the MMLU task format, the domain task-accuracy effect cannot be measured in these two domains. We therefore report honestly that *quantitative claims are possible only in the Medicine domain*. ArXiv / Patent / Wiki are excluded from this evaluation because there is no clear corresponding standard MMLU subset (Section 7).

F.3 Multiple-comparison correction (Bonferroni)

Appendix F tests multiple hypotheses across 5 zero-shot benchmarks + 12 MMLU subsets, so we leave the raw p-values in the F.1/F.2 tables and additionally report Bonferroni-corrected results. **MMLU subset family** ($m = 12$): `professional_medicine` $p = 0.021 \times 12 \approx \mathbf{0.252}$, `anatomy` $p = 0.034 \times 12 \approx \mathbf{0.408}$, MMLU-Medicine group $p = 0.022 \times 12 \approx \mathbf{0.264}$ — none retains significance at the single-test threshold ($\alpha = 0.05$). We therefore **report the MMLU-Medicine signal at a *trend* level and defer quantitative conclusions**. **5-benchmark family** ($m = 5$): ARC-Easy $p = 0.008 \times 5 \approx \mathbf{0.040}$ (retains significance), HellaSwag $p = 0.008 \times 5 \approx \mathbf{0.040}$ (retains significance), SciQ $p = 0.021 \times 5 \approx \mathbf{0.105}$ (does not retain). Under the most conservative family-union basis ($m = 17$), even ARC-Easy / HellaSwag fail at 0.136, so the 5-benchmark significance depends on the family choice ($m = 5$). We conservatively report the V/O freeze general English-benchmark effect as *the significant improvement of ARC-Easy / HellaSwag under family-specific correction ($m = 5$)*, and leave the quantitative emphasis on MMLU-Medicine to future multi-seed × multi-subset extension validation (Section 7).

G Activation Patching Layer Scan Details

We inject healthy V/O activations into the forgotten model and measure the change in PPL via a layer scan, reporting 360M results as 3 seed mean ± std.

Patch range (360M)	Wiki PPL	diff
no patch (s.42)	28.85	—
All (L0-L31)	$22.73 \pm .10$	$\mathbf{-6.27 \pm .13}$[1]
L16-L30 (back)	23.97	−4.89 (s.42)
L0-L15 (front)	28.50	−0.33 (s.42)
L31 (top)	26.77	−2.08 (s.42)

Table 25: Activation patching layer scan (360M, 3 seed for All layers; seed=42 single for bands). [1] $n = 3$ = seeds 42, 1, and 7, using the same evaluation configuration as Section 4.5. The band rows subtract from the seed-42 no-patch value in the same column; the All-layer diff is a 3-seed mean and so does not equal $28.85 - 22.73$.

In the forward direction, L16-L30 accounts for 79% of the total effect, matching the layer-concentration pattern of weight rollback. L31 patching alone (-2.08) shows a limited effect, consistent with the Section 4.1 observation that L31 is a store for new-domain learning.

Is the layer effect specific to V/O or a general Transformer property? We applied the same layer scan to MLP to dissociate this.

MLP patch (s.42)	Wiki diff	New-dom
front (L0-L15)	$+\mathbf{3.59}$ (worse)	+1.4 to +7.9
back (L16-L30)	$-\mathbf{9.49}$ (largest)	+2.1 to +12.1

Table 26: MLP layer scan for comparison with V/O (seed=42, single run).

MLP also shows a back > front pattern, with an effect size larger than V/O (MLP back -9.49 vs. V/O back -4.89). However, **MLP back damages new domains by** $+2 \sim +12$, while V/O back damages them only by $+0.2 \sim +4.4$ — making it clear that *layer position is a general Transformer property and V/O's distinguishing feature is component selectivity (recovering only previous domains, with minimal impact on new ones).*

H Inter-checkpoint Activation Transplantation Details

Extending the forward-direction layer scan of Appendix G to both directions and reproducing on 3 seeds, we also measured per-domain PPL gap recovery.

Direction	Wiki	std	Effect
Forward Wiki→PubMed	$-\mathbf{6.27}$	.13	cure
Reverse PubMed→Wiki	$+\mathbf{3.26}$	.09	elicit

Table 27: Bidirectional V/O activation transplantation, 3 seed mean (Forward injects healthy source → recovers forgetting; Reverse injects forgotten → elicits forgetting in healthy).

Bidirectional main result (V/O all, 3 seed). The forward direction shows the same direction at the activation level as the V+O weight rollback (-5.72) in Section 4.2, while the reverse direction is a new result of this work. In the reverse direction, the healthy Wiki-end model has its Wiki PPL worsen by $+3.26$ from injection of the forgotten model's V/O activations alone, while parts of the new-domain PPL gap are simultaneously recovered.

Forward Wiki cure number consolidation. Activation-level forward cure numbers vary slightly depending on measurement condition; we consolidate them as follows.

Location	Condition	n	Wiki diff
App. G (layer scan)	V/O all	3	$\mathbf{-6.27 \pm .13}$
Sec. 4.5 (transplant)	V/O all	3	$\mathbf{-6.27 \pm .13}$
App. D (V/O ≡ O)	V/O all	3	$\mathbf{-6.27 \pm .13}$
App. D (V/O ≡ O)	O only	3	$\mathbf{-6.27 \pm .13}$
App. G (all-layer patch)	All, s.42	1	$\mathbf{-6.12}$

Table 28: Forward Wiki cure consolidation across measurement conditions.

Domain	W-end	patched	diff	PM-end	ratio*
Wiki	11.3	14.6	+3.3	28.99	(elicit)
Law	37.7	17.5	**−20.20**	11.0	**76%**
Math	19.4	15.0	−4.47	14.3	86%
Patent	30.0	6.4	**−23.54**	2.6	**86%**
ArXiv	27.4	21.6	−5.84	15.5	50%
PubMed	26.3	17.3	**−9.06**	12.0	64%

Table 29: PPL difference recovery by domain (reverse direction). W-end=Wiki-end baseline, PM-end=PubMed-end value. *ratio = (W-end − patched) / (W-end − PM-end).

PPL difference recovery by domain (reverse direction, 3 seed mean). Two hypotheses for the inter-domain recovery-ratio difference (Math 86% vs. ArXiv 50%): (i) saturation due to the source-target PPL gap, and (ii) ArXiv's higher MLP dependence (the larger reverse MLP transfer for ArXiv (−3.78) than Math (−0.85) is partially consistent). This measurement is PPL-based; downstream task accuracy extensions are discussed in Section 7.

Target ckpt	gap	Wiki diff
Math-end	6.5	−0.40 (n=1)
ArXiv-end	13.0	$\mathbf{-4.91 \pm .10}$
PubMed-end	17.5	$\mathbf{-6.27 \pm .13}$

Table 30: Forgetting-gap dependence (source=Wiki-end). Gap=target Wiki PPL − healthy Wiki PPL; cure effect grows monotonically with gap (3 seed except Math-end).

Forgetting-gap dependence (forward direction, source=Wiki-end). The cure effect of V/O activation transplantation increases in a *monotonic trend* with the forgetting gap between source and target. With $n = 3$ data points strict monotonicity is hard to claim, so we report this only as a *trend*.

Recency bias (reverse direction, target=Wiki-end). Read by column, each target domain is best restored by the earliest source checkpoint that has already trained it — the checkpoint in which that domain is freshest. Law and Math are best restored from the Math-end source (−24.96, −7.67), Patent and ArXiv from the ArXiv-end source (−24.30, −9.28), and PubMed from the PubMed-end source (−9.06). As training proceeds, later domains encroach on the recovery effect that earlier ones leave in V/O activations, and we name this *recency bias in V/O activations*. The fact that PubMed transfer from the ArXiv-end source is nearly zero ($+0.04 \pm 0.13$) further supports the recency bias hypothesis — the ArXiv-end checkpoint is in a state in which PubMed *has not yet been trained*, so the PPL recovery effect of PubMed is not accumulated in V/O.

Source	Wiki	Law	Math	Pat.	ArX.	PubM.
PubMed-end	+3.26	−20.20	−4.47	**−23.54**	−5.84	−9.06
ArXiv-end	+3.44	−17.37	−2.54	**−24.30**	−9.28	+0.04
Math-end	+1.35	**−24.96**	−7.67	−8.55	−3.84	−6.78

Table 31: Recency bias: 3 source ckpts × 6 target domains. Wiki col reports elicit effect; new-domain cols report transfer. 3 seed mean; per-cell standard deviations are at most ±0.32 and are listed in the paragraph below.

Per-cell std (Table 31). Std values are small relative to means (≤ 0.32 across all 18 cells); omitted from Table 31 for compactness. PubMed-end: Wiki ±0.09, Law ±0.15, Math ±0.02, Patent ±0.23, ArXiv ±0.23, PubMed ±0.04. ArXiv-end: Wiki ±0.11, Law ±0.32, Math ±0.05, Patent ±0.06, ArXiv ±0.08, PubMed ±0.13. Math-end: Wiki ±0.02, Law ±0.06, Math ±0.02, Patent ±0.26, ArXiv ±0.27, PubMed ±0.06.

I LoRA-Regime Replication of the V/O vs. Q/K Freeze Asymmetry

To test whether the V/O > Q/K freeze asymmetry (Sections 4.3–4.4) is specific to full fine-tuning, we repeat the freeze comparison in a LoRA adaptation regime: identical data, domain order, learning rate (5×10^{-4}), and schedule as the main 360M experiments, with LoRA adapters ($r = 16$, $\alpha = 32$, dropout 0) attached uniformly to the q/k/v/o projections. "Freezing" a component is implemented as not attaching its adapter, mirroring the zero-deviation semantics of full-regime freezing; owing to the GQA weight shapes, V/O-freeze (adapters on q/k only) and Q/K-freeze (adapters on v/o only) train exactly the same number of parameters (1,638,400), preserving the parameter-count control of Section 4.4. The two freeze conditions are thus parameter-matched *to each other*; the LoRA baseline (adapters on all four projections, twice the trainable parameters) serves as a reference point rather than a matched comparison. The MLP is never adapted in this regime.

Seed	Base	V/O-fr.	Q/K-fr.	ΔV/O	ΔQ/K
42	18.29	16.52	19.05	−1.77	+0.76
1	17.82	16.45	19.13	−1.38	+1.30
7	18.57	16.66	18.82	−1.92	+0.25
mean				$\mathbf{-1.69 \pm 0.23}$	$\mathbf{+0.77 \pm 0.43}$

Table 32: LoRA regime, endpoint Wiki PPL (3 seeds). Δ = freeze condition − LoRA baseline.

ΔV/O $< 0 <$ ΔQ/K in every seed. Measured as Wiki forgetting (endpoint minus post-Wiki minimum), the ordering vo_freeze ($+4.53/+4.45/+4.67$) $<$ baseline ($+7.03/+6.56/+7.31$) $<$ qk_freeze ($+7.72/+7.79/+7.49$) also holds in every seed: V/O protection reduces forgetting by $\approx$35% (per-seed 32.2–36.1%), while Q/K freeze worsens it. The asymmetry therefore survives in a regime where the MLP — the primary new-domain knowledge store (Section 4.2) — never changes, suggesting it is intrinsic to the attention write-path rather than mediated by MLP adaptation. Two caveats: the smaller absolute magnitude relative to full fine-tuning (-1.69 vs. -2.98) is proportional to the smaller total forgetting in this regime, but the regime effect cannot be fully separated from the MLP-not-adapting difference with this experiment alone; and the Q/K freeze seed variance is smaller here (std 0.43) than in full fine-tuning (std 2.53).

J Rollback Reference-Point Controls

J.1 Per-domain anchor rollback

Section 4.2 uses the Wiki-end checkpoint as the primary rollback anchor. Here, at the PubMed endpoint, we roll V/O — or Q/K as the parameter-matched selectivity control — back to *each* prior domain's own boundary checkpoint and measure that domain's PPL change (3 seeds, same protocol as Table 1). PubMed, the final domain in the sequence, is not a forgotten prior at the endpoint and therefore has no anchor row.

Anchor	V/O rollback	Q/K rollback
Wiki	$\mathbf{-5.72 \pm 0.08}$	$+0.22 \pm 0.08$
Law	$\mathbf{-2.60 \pm 0.00}$	-0.51 ± 0.01
Math	$\mathbf{-1.66 \pm 0.01}$	-0.43 ± 0.01
Patent	$\mathbf{-0.61 \pm 0.04}$	-0.15 ± 0.01
ArXiv	$\mathbf{-2.54 \pm 0.06}$	-0.02 ± 0.05

Table 33: Own-domain PPL diff vs. the endpoint baseline when rolling back to each domain's own boundary checkpoint (negative = recovery; 3 seed mean ± std).

V/O rollback recovers every prior domain, while the same-parameter-count Q/K rollback recovers at most a small fraction of the V/O effect ($|\Delta| \leq 0.51$ PPL): the V/O-selective recovery profile is not specific to Wiki. Recovery magnitude tracks how much each domain was forgotten (Patent retains a very low endpoint PPL of $\sim$2.6, leaving little to recover). No non-anchor domain is degraded beyond the small shifts reported for Wiki-end rollback; the largest cross-domain side effects are additional *recoveries* of earlier domains (down to -3.47), expected because later checkpoints still carry the earlier domains' V/O adaptations, and current-domain (PubMed) shifts stay within $+0.1$ to $+0.2$.

J.2 Rollback to the original pretrained weights

To distinguish restoration of learned content from removal of accumulated drift, we replace the endpoint V/O with the *original pretrained* weights instead of the Wiki-end checkpoint (3 seeds, same protocol).

Domain	Wiki-end rollback	Original-weight rollback
Wiki	$\mathbf{-5.72 \pm 0.08}$	-0.26 ± 0.16
Law	-0.12 ± 0.02	$+0.49 \pm 0.05$
Math	$+0.44 \pm 0.02$	$+0.71 \pm 0.05$
Patent	$+0.76 \pm 0.03$	$+0.81 \pm 0.02$
ArXiv	-1.06 ± 0.03	$+0.11 \pm 0.08$
PubMed	$+0.29 \pm 0.00$	$+0.57 \pm 0.01$

Table 34: Per-domain PPL diff vs. the endpoint baseline for the two V/O rollback reference points (3 seed mean ± std).

The original-weight rollback recovers essentially nothing on Wiki (-0.26 ± 0.16 vs. -5.72 ± 0.08), even though it is the slightly *larger* weight perturbation (Frobenius norm over the 64 V/O tensors, seed 42: $\|\theta_{\text{final}} - \theta_{\text{orig}}\| = 125.1$ vs. $\|\theta_{\text{final}} - \theta_{\text{wiki}}\| = 115.3$, a $1.08\times$ ratio). The recovery effect of Section 4.2 therefore reflects restoration of Wiki-adapted V/O content acquired during continual pretraining, not removal of generic drift or a perturbation artifact. For reference, the pretrained model itself scores Wiki PPL 16.13 on the same eval set.

K MLP Sub-Component Rollback (gate/up/down)

Unlike Q/K/V/O, which admit an established functional decomposition (QK routing vs. OV content), the MLP's gate/up/down matrices are coupled through the nonlinearity and lack an analogous per-matrix functional account; the main text therefore treats MLP as a single block. Here we test the decomposition empirically, rolling each

MLP matrix back to the Wiki-end checkpoint separately (3 seeds, same protocol as Table 1).

Condition	Wiki diff	Worst new-domain diff
Full MLP	-6.29 ± 0.12	$+20.82 \pm 0.36$
gate_proj only	-2.44 ± 0.11	$+1.55 \pm 0.08$
up_proj only	-0.76 ± 0.22	$+3.61 \pm 0.06$
down_proj only	$\mathbf{-8.34 \pm 0.09}$	$+2.88 \pm 0.07$
V/O (reference)	-5.72 ± 0.08	$+0.76$

Table 35: MLP sub-component rollback to the Wiki-end checkpoint: Wiki recovery vs. worst-case new-domain damage (3 seed mean $\pm$ std).

Two findings. (i) No sub-component provides the damage-free separation seen for V/O: down_proj recovers Wiki most strongly (-8.34, exceeding even full-MLP and V/O rollback) — consistent with its residual-stream-writing role, analogous to o_proj — but degrades *every* new domain by $+1.4$ to $+2.9$. The damage criterion here is directional consistency: V/O rollback's new-domain shifts are small and sign-mixed (-1.06 to $+0.76$, Table 1), whereas down_proj degrades all five new domains in the same direction; the recovery-vs-damage trade-off remains far more favorable for V/O ($5.72/0.76 \approx 7.5$) than for down_proj ($8.34/2.88 \approx 2.9$). (ii) Sub-component effects are strongly non-additive: the individual worst-case damages sum to roughly $+8$, whereas rolling all three matrices back jointly produces $+20.8$, indicating that gate/up/down act as a coupled unit. This non-additivity is direct empirical support for treating the MLP as a single block; a finer mechanistic account of the down_proj write-path is left to future work.

L Multi-Seed Coverage Summary

For transparency about which claims carry which statistical support, Table 36 lists the seed/configuration coverage of the principal results. Results not listed are described with their configuration in the corresponding sections.

Result (location)	Coverage
Rollback profile: V/O, MLP, Q/K (§4.2)	3 seeds
Per-projection rollback (§4.2)	3 seeds
Rollback reference controls (App. J)	3 seeds
MLP gate/up/down rollback (App. K)	3 seeds
V/O freeze (§4.3)	3 seeds, 2 orders
Q/K freeze control (§4.4)	3 seeds
Random-subset freeze (§4.4)	3 seeds
Activation transplant (§4.5)	3 seeds
— V-only patch (App. D)	1 run
Recency-bias transfer (App. H)	3 seeds
Downstream zero-shot (App. F)	3 seeds
EWC variants, freeze stack (App. E)	3 seeds
— Q/K-, MLP-targeted; λ sweep	seed 42
Freeze timing (App. B)	seed 42
Layer scan (App. G)	seed 42
Representation branching (§4.1)	4 pairs, seed 42
Probes / token analysis (App. A)	1 configuration
LoRA regime (App. I)	3 seeds
SmolLM2-1.7B V/O freeze (App. C)	1 run + 3 seeds
— activation-level checks	1 run
OLMo-2-1B pilot (App. C)	3 seeds, 1 order
Gemma3-1B (App. C)	3 seeds, 1 order

Table 36: Seed/configuration coverage of the principal results.

M Extended Limitations

This appendix expands the points summarized in Section 7.

The 1.7B seed-invariance and its correction. Five independent launches produced bit-exact identical trajectories, which we first attributed to size-dependent numerical determinism. The actual cause was the deterministic iteration order of the streaming datasets, and an environment-guarded shuffle restores seed diversity while leaving pre-fix results byte-exact reproducible. Appendix C (C.2) reports the corrected diagnosis and the resulting 3-seed replication.

Q/K freeze variance, and why MLP-subset freeze is not a usable control. The Q/K freeze standard deviation (2.53) is roughly $20\times$ that of V/O freeze and warrants further analysis in its own right, although all three seeds share the same sign. A natural additional control would be freezing a comparable subset of MLP tensors, but because MLP is where new-domain knowledge is stored (Section 4.2), freezing it damages new-domain learning directly, which makes it unsuitable as a control for prior-domain recovery.

Continual pretraining versus continual fine-tuning. Prior work on continual fine-tuning (Li and Lee, 2024; Zheng et al., 2025) differs from our setting in training paradigm, model scale, and eval-

uation, so quantitative comparison is not meaningful. It remains possible that the components implicated in forgetting differ between paradigms, which would make our V/O intervention effect paradigm-conditional; testing this would require running both paradigms on the same model.

A working hypothesis for the V-alone sign reversal. V-alone patching improves Wiki at 360M (-2.25) but harms it at 1.7B ($+9.86$). One hypothesis is that this follows from the attention configuration: the 360M V projection is compressed, with 5 KV heads shared across 15 query heads, whereas the 1.7B V projection carries independent per-head representations. Replacing V alone may coherently substitute a compact KV representation in the former while simultaneously disturbing many specialized heads in the latter. Consistent with this, V/O-combined patching agrees in direction at both scales (-6.27 / -5.90), which the module-level hook explains: the effective intervention point is the o_proj output (Appendix D), a structure common to both architectures. We do not verify this hypothesis; doing so would require paired GQA and MHA models at matched size.

Relation to function- and steering-vector work. Inter-checkpoint transplantation is conceptually analogous to function-vector, persona-vector, and steering-vector methods (Section 5.2), but it is not a direct instance of them. Our contribution is confined to transplantation between checkpoints along a single continual-pretraining trajectory.

Evaluation caveats. (i) The MMLU-Medicine improvement ($+5.48$, uncorrected $p = 0.022$) does not survive Bonferroni correction over the 12-subset family ($p \approx 0.264$), so we report it at trend level and leave multi-seed, multi-subset verification to future work. (ii) The 50–86% PPL-gap recovery of activation transplantation is a perplexity-based measurement; we did not verify the patched model's downstream accuracy outside the MMLU-Medicine result, because Law and Math sit near chance and ArXiv, Patent, and Wiki have no corresponding MMLU subset. (iii) Forgetting-gap dependence is a monotonic pattern observed over three domain pairs (Math, ArXiv, PubMed against Wiki) and is unverified for other combinations; recency bias is consistent in direction across three source checkpoints but has not been checked on baselines with different training orders.

Caveats of the EWC comparison (Appendix E). (1) Only Wiki-anchored single-task EWC is compared; multi-task variants that accumulate Fisher at every domain boundary were not run. (2) The λ grid sweep is a single trial at seed 42, and the best λ may differ for other seeds. (3) Full EWC $\lambda = 1000$ and the V/O freeze + Full EWC combination are reported as 3-seed mean $\pm$ std, and V/O-targeted EWC was extended with a 5-point λ sweep plus a 3-seed extension at its best-behaved setting $\lambda = 50000$ (Wiki $+2.65 \pm 0.03$ anti-cure); the Q/K- and MLP-targeted variants remain seed 42 only. Their consistency across the four variants (Wiki $+3.60$ to $+4.84$) supports the qualitative finding, but quantitative comparison between them is future work. (4) At V/O-targeted EWC $\lambda = 50000$, seed 1, the Law trajectory showed unusually large forgetting during Patent training ($+16.28$, against roughly $+1.0$ for the other seeds): when only V/O is anchored and MLP and Q/K train freely, unconstrained drift appears able to induce additional forgetting in some seeds. This is independent of the Wiki cure result. (5) The EWC comparison is 360M only. (6) Fisher estimation uses $n = 457$ samples, the size of the Wiki held-out set.