Escaping the saturated game: plasticity, harder benchmarks, and retrieval in continual learning

W. Schulz

Jul 26, 2026

Abstract

Thirty pre-registered experiments. Plasticity loss behaved as an Adam/weight-norm pathology that plain weight decay fixed better than specialized methods at this scale; representation-hard benchmarks left class-IL collapse dominated by cross-task selection; retrieval over external memory worked weight-free. The selection decomposition is Kim et al. 2022's theorem.

⚠ ε CORRECTION (2026-07-26)

Claims in this document about the ε law being a novel cross-domain law or a universality class are superseded. $\varepsilon(\sigma)=\Phi(-\text{gap}/(\sigma\sqrt{2}))$ is Thurstone's Law of Comparative Judgment, Case V (1927) — an identity, not a discovery — and it is not a universality class (it breaks 723× under heavy tails). The class-IL row is a proved theorem (Kim et al., NeurIPS 2022). The measured constants, the shipped tools, and the formula $\varepsilon^*=p/(1-p)$ all stand. Canonical correction: **`CORRECTION-epsilon-thurstone.md`**.

2026-06-27. 30 pre-registered run-once experiments across the three frontiers proposed for "actually breaking ground" on continual learning. Plans + full results: `EXPERIMENTS-frontier1-plasticity.md`, `EXPERIMENTS-frontier2-hard-benchmark.md`, `EXPERIMENTS-frontier3-pretrained-cl.md`. 29 run, 1 deferred.

The one finding that ties everything together

The selector-exploitability law **`ACC = A_sel · p_within`** (selection × within-task, with selection pinned by exploitability ε) is invariant across substrate, benchmark, and mechanism. It was found for the class-IL expert selector (prior campaign, SE3/SE6). This campaign shows it also governs:

- A representation-hard benchmark (Frontier 2, BF9): on a parity stream where frozen features provably fail, sequential finetune still learns each task (task-IL 0.829) — the class-IL collapse (0.163) is entirely cross-task selection, not representation drift. Harder representations did *not* change the bottleneck.
- Retrieval over external memory (Frontier 3, PM10): retrieval accuracy = A_retrieve (0.847) × p_within (0.989) = 0.838, exactly, with retriever ε = 0.153 that grows with memory diversity (#tasks: 0.028 → 0.079 → 0.153) — the identical ε-scaling law as the weight-based selector (SE6).

So catastrophic forgetting, expert-selection, and retrieval-augmented memory are one problem: near-perfect within-task competence gated by an exploitable *selector/retriever* whose ε rises with how many

things it must disambiguate. Weights, experts, or external memory — same law.

The groundbreaking direction this points to

The genuinely novel, valuable, and *owned* thread is memory-based / retrieval-augmented continual learning where the research object is the retriever's exploitability ε. Frontier 3 showed: - A frozen model + external memory + retrieval is a viable, weight-free CL method (PM9: 0.838, crushing finetune's 0.173, within 0.03 of replay) — and it is exactly how real systems keep models current (RAG, memory-augmented LLMs). - Its failure mode is selector-ε, so the entire verifier-budget toolkit (ε, consensus, N*(ε)) imports directly: a retrieval system's safe memory size is N*(ε); to scale memory you must lower retriever ε, not just add entries.

This is the one place all three durable assets of the program converge — the selector-ε theory, the verifier-budget tool, and a real, valuable system — into a single research question: how do you keep retriever exploitability low as external memory grows?

Secondary findings (honest)

Frontier 1 — Loss of plasticity. Plasticity loss is not intrinsic to backprop — it is an Adam + long-stream + small-net weight-norm pathology (LP1 SGD-robust; LP7 Adam loses 8× faster; LP9 width delays it). Mechanism = weight-norm growth → dead units in the *representation* (LP2 r=−0.75; LP8 features not head). Plain weight decay beats the specialized fixes (CBP, shrink-perturb, regenerative L2-init) at this scale (LP6 +211% of CBP's gain; LP10 near-dominant on both axes). Honest bound: Dohare et al. show L2 is insufficient at deeper-net/RL/longer-stream scale this harness can't reach — so this falsifies the fixes' *necessity in-regime*, not universally. Publishable angle: *plasticity loss as an optimizer pathology; weight decay as the strong baseline the fixes must beat.*

Frontier 2 — Escape the benchmark. A supervised task frozen features provably can't solve exists (BF3: parity k≥4 fails at *any* random-feature width; MLP succeeds). But making the benchmark representation-harder does not create a new CL problem: method ranking is preserved (BF8 ρ=+0.80), GDumb/replay still wins (BF7/BF10), selection still dominates (BF9). "Change the benchmark to escape the saturated game" largely fails — the prize was the unification above, not a new game.

Frontier 3 — Pretrained-model CL. Task-vector merging just works at small scale and naive ≈ TIES (PM3 kill — fancy merging unnecessary here; PM5 sign-conflict mechanism real); naive weight editing is catastrophic (PM7: one edit → chance — locality mandatory); retrieval is the standout (PM9/PM10).

Scorecard

frontier	confirmed	killed/failed	net
F1 plasticity (LP1-10)	LP2,3,4,6,8,9	LP1,5,7◐,10◐	weight decay beats the fixes
F2 hard benchmark (BF1-10)	BF1,3,6,10	BF2,4,7◐,8,9→	benchmark doesn't escape selection
F3 pretrained CL (PM1-10)	PM1,4,5,7,9,10	PM2◐,3,6,8	retrieval = selector-ε (the synthesis)

Bottom line for "how do we break ground": not by a better forgetting algorithm (saturated), not by a harder benchmark (still selection), and not by plasticity fixes (beaten by weight decay at this scale). The

opening is retrieval-augmented CL with retriever-ε as the lever — novel, valuable, and the convergence point of everything this program already owns.

Published by NeuroGenesys (AW3 Technology, Inc.). Canonical version: https://neurogenesys.si/research/three-frontiers-continual-learning