Retro-DARC: Function-Preserving Residual-Memory Adapters for Pretrained Language and World Models

Queryable Residual Evidence for Planners — Litepaper (conference-format short version)

Ada Cyborg — July 2026

This is the short version of the full technical report. The report (30 pp, with the complete related-work survey, theory, typed world-model interface, verification plan, and revision record), all code, raw per-run JSONs, and adapter checkpoints are in the released repository. Preregistration-style program: predictions and baselines were committed to the public repository before the runs that tested them.

Abstract

Standard pretrained Transformers accumulate depth uniformly: every layer's update enters the residual stream with coefficient one, and the stream goes unqueried. We introduce **Retro-DARC**, a family of function-preserving adapters that turns a model's own layer-to-layer updates — its *innovations*, in the filtering-theory sense — into typed, queryable, causally auditable memory. The adoption core, **Retro-DARC-Lite**, inserts into an existing causal LM as an exact no-op: a zero-gated, null-reserving read over recent layer deltas that provably preserves the model's function at insertion and trains gate-first.

We ground the design in a measured, replicated record on public pretrained checkpoints: **bitwise logit identity at insertion** on three open models; delta-key addressing that stays near ceiling under noise and int4-style corruption while output-key addressing collapses with depth; a registered twin pretrain in which from-scratch reads at full initial gate hurt on every seed — pure gate-initialization shock, by factorial ablation — while the zero-gated retrofit added its full increment even on an architected-retrieval base; and, at matched parameters on a converged public checkpoint, **depth reads that beat LoRA** and support an audit weight-touching adapters cannot express by construction: content-only destruction removes 70–107% of the gain, and **zeroing the bank returns the frozen loss to machine precision**, replicated across seeds and accelerators. **Four registered predictions failed and are printed beside their diagnoses**; two returned design rules; one marks the boundary the next run resolves.

---

1. Introduction

Transformer language models use attention to select information across sequence positions. The same selectivity is usually absent across network depth. In a standard PreNorm residual block, with $h_l(t) \in \mathbb{R}^d$ the token representation at position t and depth l,

$$\Delta_l(t) = f_l(\mathrm{RMSNorm}(h_{l-1}(t))), \qquad h_l(t) = h_{l-1}(t) + \Delta_l(t), \tag{1}$$

and unrolling gives $h_L(t) = h_0(t) + \sum_{l=1}^{L} \Delta_l(t)$. Every previous layer contributes with fixed coefficient one. This residual stream is essential for optimization [He16; Vaswani17], but as a *memory* mechanism it is crude: depth is accumulated, not queried. In filtering-theory terms, each Δ_l is an *innovation* — the part of layer l's computation its predictive context did not already contain [Kalman60; Kailath68] — and the standard architecture sums its innovations instead of keeping them addressable.

Recent work challenges the fixed-aggregation design: Attention Residuals replace uniform accumulation with softmax aggregation over previous layer outputs [AttnRes26]; Delta Attention Residuals route over layer deltas rather than redundant cumulative states [DeltaAR26]; OASIS adds null-aware routing

[OASIS26]; MoDA, MUDDFormer, DeepCrossAttention, and Depth-Attention independently confirm the axis [MoDA26; MUDD25; DCA25; DepthAttn26]. The axis is now settled at frontier scale: **Kimi K3** [KimiK3-26], a 2.8-trillion-parameter open MoE, ships Attention Residuals and Block Attention Residuals as core architecture alongside Kimi Delta Attention [KimiLinear25], reporting ~2.5× scaling efficiency over its predecessor (vendor-self-reported). The sharper question is therefore not *whether* depth routing works, but:

> ***Can a pretrained Transformer — including a frontier model that does not yet have depth memory — be upgraded with it without changing its original function at insertion, and can that memory be made typed, queryable, and causally auditable?***

This matters because most valuable models already exist as checkpoints: K3 baked Attention Residuals in at pretraining, but the tens of thousands of already-trained checkpoints cannot. Meanwhile the 2026 memory-benchmark cluster measures exactly the missing capability — MemoBench finds none of ten leading world models exceeds an Object-Reappearance-Score of 0.6 and concludes memory must be explicit in model design [MemoBench26]; Genie 3's stated limitation is interaction duration in minutes [Genie3-25].

The measured record in brief (details in §5). Bitwise insertion identity on gpt2 / pythia-410m / Qwen2.5-0.5B; output-key condition numbers growing with depth to 76 while interior delta keys stay < 6.7, with per-slot normalization holding full-stack conditioning at 3.0–3.7 on five models; top-1 retrieval under corruption **0.86–0.99 for delta keys** (one disclosed exception) vs. 0.04–0.75 for output keys; the registered twin pretrain answering the retrofit-vs-from-scratch question with a diagnosed failure; and the matched-parameter comparison — **depth reads beat LoRA**, trail the best weight-touching adapter by 0.006–0.024 CE with the gap shrinking over a 16× budget, with content-destroying audits and a two-machine replication agreeing to $\leq 6\times10^{-4}$ CE.

Contributions. (1) A function-preserving depth-memory retrofit with exact insertion identity, verified bitwise on public checkpoints, through MoE routing, early exit, and recurrent state (Props. 1, 3′), training gate-first (Prop. 2′). (2) Delta keys as measured-better depth addresses on real models, with the top-of-stack failure mode localized to norm growth and cured by per-slot key normalization. (3) The retrofit-vs-from-scratch question executed: the from-scratch penalty is pure gate-initialization shock, and retrofit adds its full increment even on an architected-retrieval base. (4) Causal auditability as a measured property: content-destroying, parameter-preserving interventions attribute the gain to memory content, with a machine-checkable zero-out integrity test. (5) A falsification-first record: four registered predictions failed and are reported verbatim beside their diagnoses.

2. Related work (compressed; full survey in the report)

Depth routing. The from-scratch line — AttnRes/Block AttnRes [AttnRes26], deltas as routing targets [DeltaAR26], Softmax1 null routing [OASIS26], depth-KV mixing [MoDA26], dense or input-dependent layer combinations [MUDD25; DCA25], in-cache value mixing as our systems bar [DepthAttn26], and K3's production deployment [KimiK3-26; KimiLinear25] — designs architectures for pretraining. Retro-DARC targets the orthogonal problem: *safe retrofit of a queryable, typed, intervenable memory with causal attribution* into checkpoints that already exist. The closest LLM-side prior art (LongMem, CAMELoT, Larimar, and the 2026 retrofit papers arXiv:2601.15324 and 2603.22329) claims reversible frozen-model memory but neither bitwise insertion identity nor causal audits nor robustness measurement.

Adapters and adaptive compute. LoRA [LoRA21] and bottleneck adapters [Houlsby19] are our capacity-matched controls; Mixture-of-Depths and LayerSkip motivate the optional route-entropy compute axis [MoD24; LayerSkip24].

Memory benchmarks. MemoBench [MemoBench26] and its siblings independently establish that persistent state must be an explicit design target — the external restatement of this paper's motivation.

3. Method: Retro-DARC-Lite

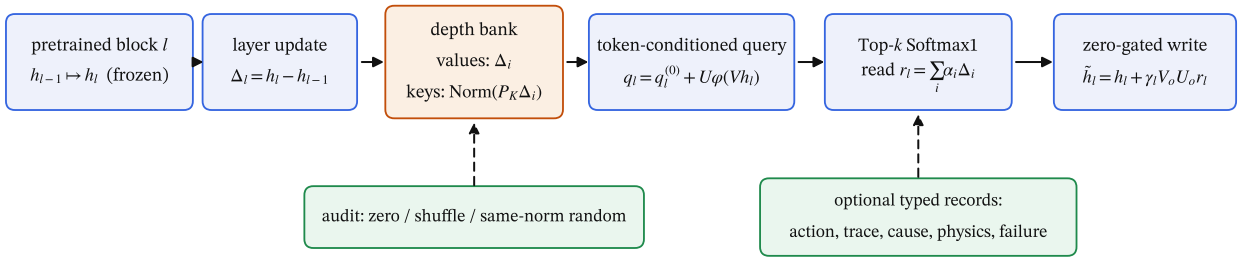

Figure 1: The Retro-DARC-Lite read path. The base block stays frozen; recent layer updates Δ are stored as full-width values and addressed through low-rank, per-slot normalized keys. The read returns through a low-rank output projection behind a scalar gate initialized at zero, so insertion is exact identity. Because memory content is an explicit runtime object, it supports the zero / shuffle / same-norm interventions of §5 without touching any learned weight.

3.1 Layer-delta memory. The memory unit is a residual delta $\Delta_l(t) = h_l(t) - h_{l-1}(t)$: what layer l *contributed*, versus a cumulative state containing all history. With local window w, the bank is $\mathcal{M}_l^{\text{lite}}(t) = \{\Delta_{l-w}(t), \ldots, \Delta_{l-1}(t), \mathbf{0}_\varnothing\}$.

3.2 Low-rank token-conditioned query.

$$q_l(t) = q_l^{(0)} + U_l\, \varphi\big(V_l\, \mathrm{RMSNorm}(h_l(t))\big), \quad V_l \in \mathbb{R}^{r\times d},\ U_l \in \mathbb{R}^{d_r \times r},\ r \ll d.$$

3.3 Null-aware top-k routing. Each candidate $M_m(t)$ is scored

$$s_{l,m}(t) = \frac{q_l(t)^\top \mathrm{RMSNorm}(P_K M_m(t))}{\sqrt{d_r}} + b_{\text{type}(m)} + b_{\text{age}(l,m)},$$

with $I = \mathrm{TopK}(s_l, k)$ and Softmax1 normalization, which reserves mass for a null route and lets the model decline retrieval. The read is $r_l(t) = \sum_{m\in I} \alpha_{l,m}(t)\, M_m(t)$.

3.4 Zero-gated injection.

$$h_l^{\text{new}}(t) = h_l(t) + \gamma_l\, W_O\, r_l(t), \qquad \gamma_l = 0 \text{ at insertion},\ W_O \neq 0.$$

The distinction matters: $\gamma_l = 0$ with $W_O = 0$ also preserves identity but kills first-step gate learning (Proposition 2).

3.5 What you add. Four small linears and one scalar per insertion point: U_q (d→64, bias), P_K (d→64), U_o (d→64), V_o (64→d, init $\mathcal{N}(0, 0.02)$), $\gamma = 0$. Trainable parameters = 4·64·d + 65: 196,673 (gpt2), 229,441 (Qwen2.5-0.5B), 262,209 (pythia-410m). Two measured design rules attach (from the failures in §5): $\ell 2$-normalize each delta before P_K, and initialize γ at 0 always — including from-scratch builds.

4. Guarantees (proofs in the report)

Proposition 1 (Exact function preservation). Insert adapters $h_l^{\text{new}} = h_l + \gamma_l W_O r_l$ at any set of layers of pretrained F_θ. If $\gamma_l = 0$ for all inserted adapters, all subsequent hidden states and final logits equal the base model's for every input, up to numerical precision.

Proposition 2 (First-step gate trainability). At $\gamma = 0$: $\partial \mathcal{L}/\partial \gamma = \langle \partial \mathcal{L}/\partial h', W_O r \rangle$ — a zero gate is trainable at step one iff $W_O r$ has nonzero projection onto the downstream gradient; $W_O = 0$ makes it identically zero.

Proposition 3′ (Composability). Zero-gated insertion is exact through sparse MoE routing, data-dependent early exit, and weight-tied recurrent iteration — the iterates, exit decisions, and outputs of a looped model are unchanged for any iteration count.

Proposition 6 (Softmax1 null-mass bounds). With k selected logits in [m, M], the null mass satisfies 1/(1+k e^{M}) ≤ p_∅ ≤ 1/(1+k e^{m}): never starved, never saturated (verified: 0 violations in 2×10^5 draws).

5. Measured results on public checkpoints

All rows below ran on real open weights (gpt2, pythia-410m/1.4b, Qwen2.5-0.5B/1.5B) on an Apple M5 Max (CPU/MPS) with a DGX Spark GB10 (CUDA) replication node; scripts, per-run conditions, and raw JSONs are in the repository. Identity checks run in float32 on CPU for bitwise comparison.

5.1 Insertion is exactly free (E10.1). Zero-gated insert at layer 2L/3 on gpt2 / pythia-410m / Qwen2.5-0.5B, 8 prompts each: max abs logit deviation **0.0 on all 3 models × 8 prompts (bitwise equal)**; $\gamma=10^{-3}$ departs ($1.1–3.7\times10^{-2}$); hook removal restores the base bitwise. Every zero-gated run that followed re-verified the gate-first gradient order exactly, on CPU, MPS, and CUDA.

5.2 Delta keys are better-conditioned addresses — and the one failure taught the design rule (E10.2, E13, E13b). Output-key condition numbers grow ≈linearly with bank size (7.8→75.9) while interior delta keys stay ≤ 6.6 (three models, wikitext-2). The registered O(1)-everywhere sweep *failed at the top of the stack*: the final-delta norm ratio is 4.9–15.9× (5/5 models), the registered spike-not-drift bet held on **0 of 5** models, and single-layer exclusion held on only **2 of 5** — the ill-conditioning is gradual norm growth over the top ~3 layers. Per-slot $\ell 2$ normalization removes 92–99% of the κ excess, holding full-bank κ at 3.0–3.7 on all five models. The addressing consequence, measured directly: top-1 retrieval under gaussian key noise (σ = 0.25–1.0×RMS) and simulated int4 corruption is **0.86–0.99 for delta keys** (one disclosed exception: gpt2 k=11 with the anomalous final layers in-bank, 0.57–0.63 — still ≥6× the alternative) vs. **0.04–0.75 for output keys**, generally collapsing with depth (pythia-1.4b k=23: 0.91–0.98 vs. 0.08–0.14).

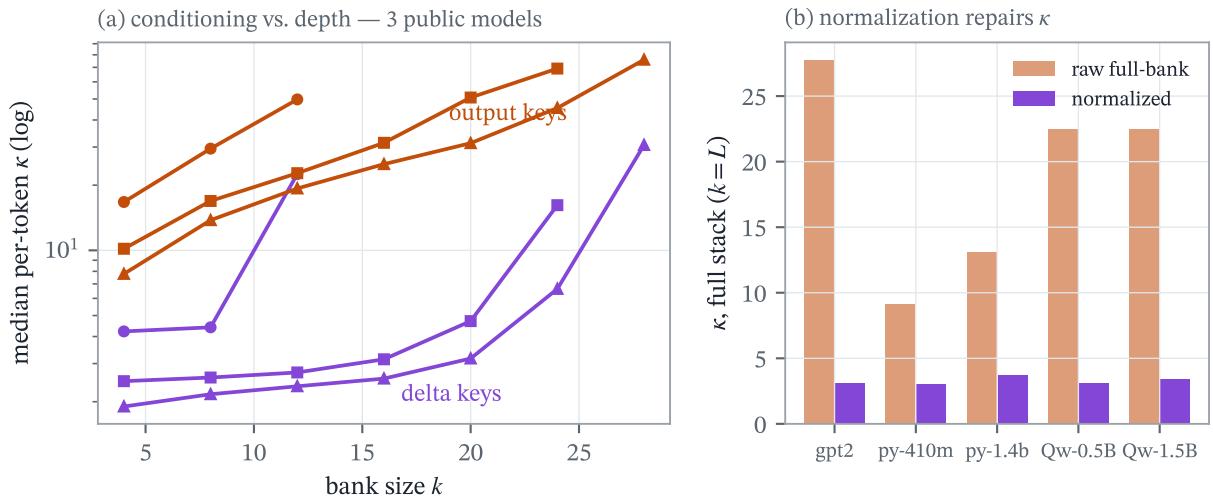

Figure 2: (a) Output-key condition numbers grow with bank size while interior delta keys stay flat (three public models; the top point of each curve is the anomalous full stack). (b) Per-slot $\ell 2$ normalization repairs full-stack conditioning on all five models (E13).

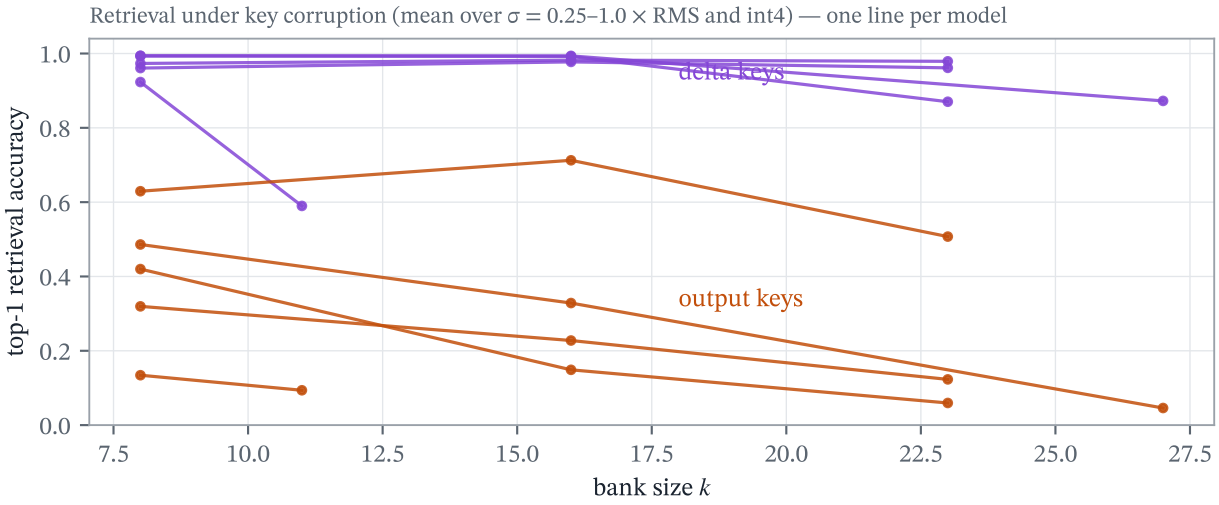

Figure 3: Top-1 retrieval under key corruption (mean over noise levels and simulated int4), one line per model: delta keys hold near ceiling while output keys collapse with depth; the one delta dip is gpt2's anomalous full-stack bank (E13b).

5.3 Retrofit vs. from-scratch, executed (E11, E14; 3 seeds). The registered twin pretrain falsified its own prior: building the read in at full initial gate (γ_0=1) *hurt* on all 3 seeds (B−A = +0.0142 CE, range +0.0100…+0.0206). The E14 keys × routing × init factorial isolates the cause to **gate-initialization shock**: zero-init beats γ_0=1 on 3/3 seeds in both key/routing configurations, zero-init reads recover to ≈parity with plain pretraining (2/3 seeds within the registered +0.005 band; mean gap +0.0024), and keys and routing each tie at the registered contrasts. Meanwhile the zero-gated retrofit adds its full increment (+0.0313 ≈ the plain-base +0.0321) even on top of an architected-retrieval (AttnRes-pretrained) base — with ~100% shuffle sensitivity and exact zero-out restoration. Architected retrieval and typed retrofit memory are complements, not substitutes.

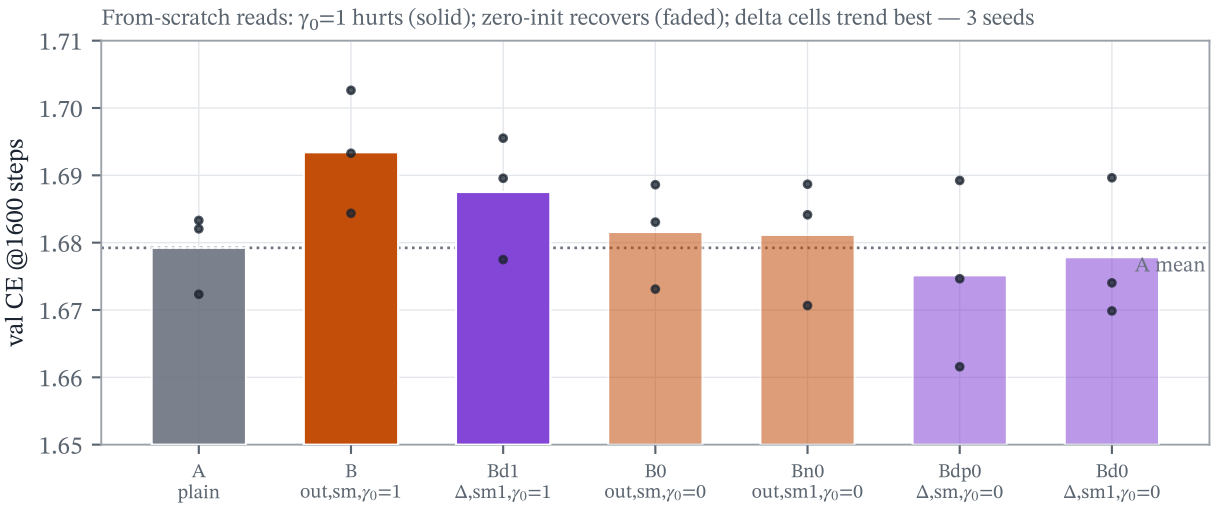

Figure 4: The from-scratch factorial (E11/E14, 3 seeds; dots are seeds). Full-gate cells (solid) sit above the plain base; zero-init cells (faded) recover to ≈parity, delta-key cells trending best. The E11 violation was initialization shock, not architecture.

5.4 At matched budgets the depth reads are competitive and uniquely auditable (E12, E15). On frozen pythia-410m (INS=16, wikitext-103, matched ~262k trainable params), at 1,500 steps ≈3.1M tokens (3 seeds, MPS): LoRA r18 / MLP / AttnRes-style / RD-Lite val CE = **2.9907 / 2.9528 / 2.9655 / 2.9764** (frozen = 3.3400), same ordering on all 3 seeds (range ≤ 0.004) — **both depth reads beat LoRA**. At 24,414 steps (50M tokens, 16× budget; seed 0 on M5 Max/MPS, seed-1 replication on DGX Spark GB10/CUDA, reported separately, never pooled; batch stream digest-identical across machines): MLP / AttnRes-style / RD-Lite = **2.9334 / 2.9395 / 2.9481**, and the weight-toucher's lead *narrows* with budget (0.0110→0.0062 vs. AttnRes-style; 0.0252→0.0147 vs. RD-Lite). Every verdict reproduces on the second machine, with final CEs agreeing to $\leq 6\times10^{-4}$.

The audit is the property no weight-touching adapter can express: with every trained parameter left in place, shuffling memory content removes 97–107% (output keys) vs. ~70–80% (delta keys) of the gain; same-norm random content lands *worse* than frozen; and zeroing the bank returns the frozen loss **bit-exactly in all 16 CPU/MPS tests** across E9/E11/E12/E15 and to $\leq 1\times10^{-7}$ in both CUDA tests (kernel reduction-order noise).

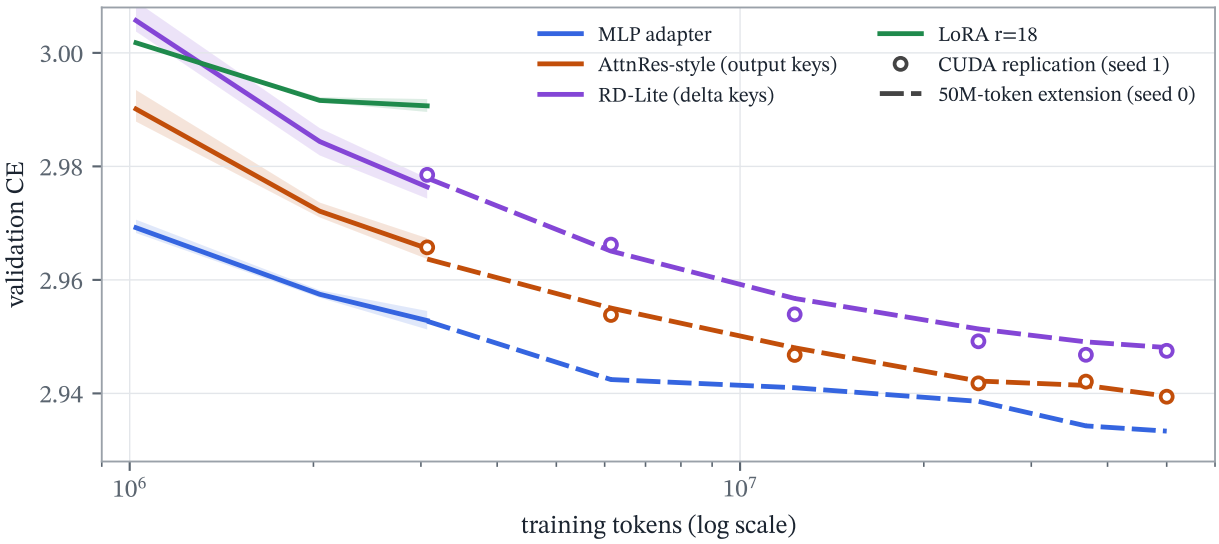

Figure 5: Matched-parameter adapters on frozen pythia-410m. Solid: 3-seed mean ± range at the E12 budget; dashed: the 50M-token extension (seed 0); open circles: the CUDA replication (seed 1), agreeing to $\leq 6\times10^{-4}$ CE. Frozen base CE 3.3400 (off-scale above). Both depth reads beat LoRA; the MLP lead narrows with budget.

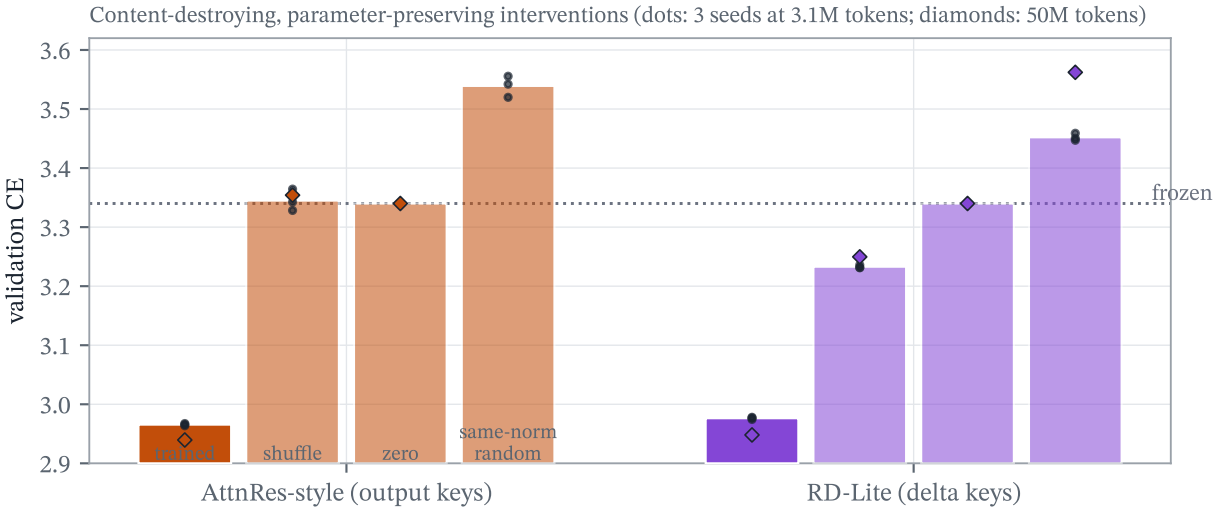

Figure 6: The causal audit. Zeroing returns the frozen loss (bit-exactly on CPU/MPS); shuffling removes 97–107% (output keys) vs ≈70–80% (delta keys) of the gain; same-norm random lands above frozen. Dots: 3 seeds at 3.1M tokens; diamonds: 50M tokens.

5.5 The standing negative, stated with equal prominence (E15 gate). With *unnormalized* keys at mid-stack insertion, output keys win raw CE by 0.0086/0.0081 (two machines) at 50M tokens — the shallow-redundancy regime at real budget; the registered deltas-beat-outputs prediction **failed** and is reported verbatim. The delta-utility claim is therefore conditional: **delta keys buy conditioning, auditability, and depth/noise robustness; output keys buy shallow-depth redundancy; production designs should choose per regime — or route over both with the null option arbitrating.** The E15 probe was specced before E13's normalization fix; the normalized-key rerun is registered next.

6. Limitations

No S3-scale (≥0.6B base, ≥0.5B-token) utility results yet — every claim at that scale remains registered and gated; the measured record is mechanism- and miniature-scale on ≤1.5B-parameter frozen public bases with ≤50M training tokens. E15's S2-gate negative binds the unnormalized configuration only (replicated: two seeds on two independent machines agreeing to $\leq 6\times10^{-4}$ CE); E13b's int4 is simulated rounding, not kernel arithmetic. K3's efficiency figures are vendor-self-reported, its weights announced but unreleased at the time of writing. The from-scratch comparisons use an internally trained proxy. The full report's §14 carries the complete ten-item statement, including the citation-verification sweep.

7. Conclusion

Retro-DARC turns the residual stream into queryable memory at exactly zero behavioral cost at insertion: identity is bitwise on public checkpoints, delta addressing is measured-better under the corruption quantized serving imposes (raw keys, one disclosed exception), the retrofit path dominates naive from-scratch integration at miniature scale with the failure diagnosed and cured by a design rule, and at matched budgets the depth reads beat LoRA while supporting content-destroying, parameter-preserving audits with a machine-checkable zero-out integrity test — replicated across independent hardware. Four registered predictions failed and are printed beside their diagnoses. The scaled target: same checkpoint, same trainable parameters, same tokens — beat parameter-matched adapters, matched-FLOP recurrence, and a faithful AttnRes-style retrofit control, with memory interventions removing the advantage, at 0.6B scale and beyond.

References

[He16] He et al., Deep residual learning, CVPR 2016. · [Vaswani17] Vaswani et al., Attention is all you need, NeurIPS 2017. · [Kalman60] Kalman, A new approach to linear filtering and prediction problems, J. Basic Eng. 1960. · [Kailath68] Kailath, An innovations approach to least-squares estimation, IEEE TAC 1968. · [AttnRes26] Kimi Team, Attention residuals, arXiv:2603.15031. · [DeltaAR26] Luo, Cai, Hu, Delta attention residuals, arXiv:2605.18855. · [OASIS26] Luo et al., Attention sinks and outliers in attention residuals, arXiv:2605.17887. · [MoDA26] Zhu et al., Mixture-of-depths attention, arXiv:2603.15619. · [MUDD25] Xiao et al., MUDDFormer, arXiv:2502.12170. · [DCA25] Heddes et al., DeepCrossAttention, ICML 2025. · [DepthAttn26] Zeng et al., Depth-Attention, arXiv:2606.05014. · [KimiK3-26] Moonshot AI, Kimi K3: a 2.8-trillion-parameter open MoE model, released July 16, 2026 (kimi.com/blog/kimi-k3; weights announced for July 27, 2026; architectural figures from launch materials pending the technical report). · [KimiLinear25] Kimi Team, Kimi Linear (Kimi Delta Attention), arXiv:2510.26692. · [LoRA21] Hu et al., LoRA, arXiv:2106.09685. · [Houlsby19] Houlsby et al., Parameter-efficient transfer learning, ICML 2019. · [MoD24] Raposo et al., Mixture-of-Depths, arXiv:2404.02258. · [LayerSkip24] Elhoushi et al., LayerSkip, arXiv:2404.16710. · [MemoBench26] Chen et al., MemoBench: benchmarking world modeling in dynamically changing environments, arXiv:2606.27537, ECCV 2026. · [Genie3-25] Google DeepMind, Genie 3, Aug 2025 (deepmind.google). Memory-adapter prior art: LongMem, arXiv:2306.07174 · CAMELoT, arXiv:2402.13449 · Larimar, arXiv:2403.11901 · Prometheus Mind, arXiv:2601.15324 · Trained Persistent Memory, arXiv:2603.22329.