

Checkpoint Surgery: Eval-Gated, Agent-Proposed Tensor Repair of Trained Models

Ibrahim Ahmed

Mohammed A. Kaleelurrahman

Abstract

A frontier model is no longer the output of one training run: it is a fleet of specialized policies collapsed into one artifact. Kimi K3 consolidates nine RL experts by multi-teacher distillation; Command A merges twelve at a measured 1.8 % average drop. By the time that damage is measured, the run that produced it is over — there is no training loop left to rewind.

We ask whether it can be recovered by operating on the checkpoint’s tensors instead: substituting or interpolating selected tensors between two checkpoints of one lineage, eval-gated and recorded as content-addressed lineage, with the selection proposed by an agent reading only weight-space telemetry — a loop we demonstrate end to end *at 0.5B only*. We pre-registered and scored twenty-three predictions on artifacts we did not train — **eleven held, six falsified, two not scoreable as registered, four unscored when a pre-registered gate of our own refused to fund their lane** — for \$96.6 of measured spend; two more were withdrawn with their arm.

A finished 70B RLVR checkpoint whose `lm_head` a saving bug destroyed is repaired by lineage-parent graft to 100.00 % / 100.73 % / 100.26 % of the released model on MMLU / GSM8K / HellaSwag, for \$16.6, against 2,880 published H100-hours to rerun that stage. At 120B, class-wise substitution between `gpt-oss-120b` and its safeguard post-train is *additive*: attention carries about 57 % of the MMLU delta at 2.93 % of the bytes, at a resolution of ± 14 pp, while the 93.5 % of the checkpoint that is MXFP4 experts carries nothing.

The falsifications are contributions and are reported at equal weight: drift magnitude does not localize behaviour, falsified twice on two model families; substitution composes only near the donor; and the one consolidation we performed ourselves traded *one-sidedly*, losing 37.2 pp and 52.1 pp of code against its code ingredient while giving up nothing on MMLU or IFEval. In the one public 65B loss-spike incident no measurable deficit existed at the restart point, on the authors’ own trail or on our probes, so the rollback was unnecessary at the moment

it was made — though the model it eventually produced is ahead of the branch it discarded. That decision, part of an incident its authors price at 345,600 A100-hours, can now be audited for \$33.

1 Introduction

1.1 What changed since the instability literature was written

A frontier model is no longer the output of one run. It is a *fleet of specialized policies that get collapsed into one artifact*, and the collapse is the last thing that happens before release.

The current reference point is Kimi K3 — a 2.8T-parameter MoE with 104B active parameters, open weights, released 2026-07-27 [35] — whose post-training section states the recipe plainly: “a three-stage paradigm: ... developing specialized domain experts at varying reasoning effort via Reinforcement Learning (RL), and **consolidating these domain-specific policies into a single model** using Multi-Teacher On-Policy Distillation (MOPD)” (§4.1). Three RL domains crossed with three reasoning-effort levels is nine experts, all collapsed into the checkpoint that ships.

This is a three-lab fact at the 2026 frontier, not one lab’s idiosyncrasy. DeepSeek-V4’s own card states the recipe almost verbatim — “a two-stage paradigm: independent cultivation of domain-specific experts ... followed by unified model consolidation via on-policy distillation” [6] — and NVIDIA’s Nemotron 3 Ultra (550B-A55B) trains “more than ten domain-specialized teacher models” and runs *two iterations* of the same MOPD consolidation [30]. Behind them is a two-year lineage: Command A’s **twelve experts across two merge stages** (§3.1 of [4]), Llama 3’s averaging at its preference stages — “at each RM, SFT, or DPO stage” [9] — SmolLM3’s soup stage [13]. The practice has moved from the edge of the pipeline to its centre.

The consolidation step is the damage event, and it is unavoidable by construction. Every lab has an answer —

merge, soup, average, distil — and every answer is uniform in the parameter space: **one scalar coefficient per ingredient model, applied across every tensor** (Section 2.2). K3’s and Nemotron’s answer is the expensive end of that range: MOPD is a *training run*, not weight arithmetic, which is what sets our efficiency baseline. And when a consolidated model gives something up, the remedy on offer is more of the same — consolidate again, or run *more* training on top of the merge (Command A §4.9.2).

And the cost of the collapse should be quoted, not asserted. Cohere measure their own: “linear merging is sufficient to preserve expert performance with only a **1.8% average drop**”, most metrics within 2.5 % of the best expert score, some improving, and “code performance the least well preserved during merging” (§4.9 of [4]). The trade is real and *concentrated rather than uniform* — not the “best at nothing in particular” story, which the primary source contradicts and which we therefore do not tell.

And the fleet keeps branching after release. One open 20B base, `openai/gpt-oss-20b`, already carries **1,241 published first-generation derivatives** — 546 full fine-tunes, 454 adapters, 234 quantizations and 7 merges, counted against the Hugging Face model API on 2026-09-01¹ — none of which inherits a fix applied to its parent.

That is the setting this paper is about, and it is not the setting the rollback literature describes. The closed frontier is not evidence for it: **we could find no published weights and no stage checkpoints** for Sonnet 5, Opus 5 or Muse Spark (checked against the Hugging Face model API [12], 2026-09-01 — an absence is a search result, not a fact, and this is the search), and a press release is not an instrument.

1.2 The damage that actually happens now

Three classes, none of which is a pretraining loss spike, and each carrying its own receipt from this campaign.

1. Consolidation trades capabilities, and it trades them asymmetrically. Nine K3 experts become one model; a stage buys one thing and sells another. Frontier reports do not publish the per-domain cost of their own collapse, so the receipt comes from a lab that publishes stage by stage: Tulu-3-70B goes MMLU 78.9 → 83.3 → **83.1** across SFT → DPO → final while IFEval and MATH rise [20]. **And the trade is now measured at class granularity on a second family:** between `gpt-oss-120b` and its safeguard post-train, **attention** carries about 57 % of the MMLU delta at 2.93 % of the bytes and the output head about 34 % at 1.78 %, while the 93.5 %

¹One `/api/models?filter=base_model:R:... query per relation class R`, paginated to exhaustion; the classes are disjoint and sum to the `base_model: total`. Our own census read 1,235 on 2026-08-27 (`measurements/g5-selection-survey.json`): the count drifts upward by about one repository per day.

of the checkpoint that is MXFP4 experts carries **nothing** (Section 5.2). **The heading says *asymmetrically* rather than *by construction* because of the one consolidation this paper performed under control:** it lost 37.2 pp and 52.1 pp of code against its code ingredient and gave up nothing on MMLU or IFEval, so consolidation damage need not be two-sided (Section 6).

2. RL post-training is where instability moved — at dense scale. Pretraining was stabilized: QK-norm, z-loss, careful initialization, auto-skip on overflow, and the shipped dense runs mostly say so — OLMo 2’s mitigations are deliberate and are reported as working [32], and Marin 8B’s own retrospective calls its training “fairly stable with very few spikes” [26]. Neither is a claim of zero rollbacks: OLMo 2’s own §3.3.3 records abandoning a fork and re-training from the point of divergence, and Marin’s retrospective states an absence rather than a count. **The qualifier is not decoration.** Nemotron 3 Ultra hit “two instances of training divergence” in 2026 at 550B-A55B [30]; the second, around 16T tokens, was mitigated by **rolling back to the 15T checkpoint** and cutting the run’s horizon to 20T — 1T tokens of a 550B pretraining run, discarded in 2026.

RL, meanwhile, was never stabilized: policy entropy collapses early and monotonically across RL runs absent intervention [5]. And what the one public artifact taken from *inside* a real 65B spike says is sharper than “it is bad”: at checkpoint 240 the branch the humans discarded is ahead of *the checkpoint they rolled back to* on all seven metrics LLM360 record per checkpoint on their public dashboard, and on both of our own probes. So at the moment the restart was ordered — one of the two spikes for which LLM360 price an extra 30 days on 480 A100s (345,600 A100-hours, 30 % of the base run [24], [PUBLISHED]) — **there was no measurable deficit to recover.** That is a claim about the decision and not about the model the restart eventually produced, which is ahead of the discarded branch and which no arm of ours could see (Section 5.1.2).

3. Faults land in the weights and survive there. Silent data corruption on unhealthy nodes makes models “converge to different optima with different weights,” sometimes with no distinguishing signal in the loss [25]. Its public instance is `allenai/Llama-3.1-Tulu-3-70B-broken`, a finished 70B RLVR checkpoint whose LM head was destroyed by a checkpoint-saving bug, and the class is no longer hypothetical: we localized the damage to **exactly** `lm_head` and repaired it by lineage-parent graft to indistinguishability from the released model — 100.00 % / 100.73 % / 100.26 % recovery on MMLU / GSM8K / HellaSwag against a 95 % bar, at 141 GB, in someone else’s training run (Section 5.1).

1.3 Why rollback is the wrong baseline — and we say so first

For an in-flight pretraining run, rollback got cheap *and stayed available*: checkpointing is dense and sharded resumption is routine, so a spike costs tokens, not the two months of 992-A100 restarts the OPT-175B report records (§2.4–§2.5 of [45]). Nemotron’s two 2026 divergences make that point for us rather than against us: the run was still in flight, so rollback was still the remedy and was taken. We concede it rather than let a reviewer find it.

But it does not rescue the three classes above, because **in all three the run is over**. A released checkpoint, a merged enterprise model, a fault discovered after the stage finished — there is no training loop left to rewind, and the options are to rerun the stage, at GPU-days and a fresh data-mixture gamble, or to accept the regression. That is the baseline surgery competes against, and unlike pretraining rollback it has not gotten cheaper.

1.4 The material question, and how it turned

Surgery needs at least two checkpoints of one lineage, and for most of 2026 the frontier shipped one: K3 publishes the consolidated model and not the nine experts; DeepSeek-V4, Qwen3.8 and Mistral publish base-and-final at best; Command A publishes the merge and not its ingredients; Meta’s Muse Glimmer, our guess at the nearest candidate, is a **clean negative** — four repositories, one branch each, no stage of any kind (both checked against the Hugging Face model API, 2026-09-01). That constraint chose this paper’s arms: AI2’s Tulu-3 and OLMo chains, `gpt-oss` base → safeguard, LLM360’s per-checkpoint branches, the only material on which the claim could be tested at all.

The constraint loosened while the campaign was running. NVIDIA has since published five of Nemotron 3 Ultra’s ten-plus pre-consolidation MOPD teachers alongside the base and the consolidated final [31]; Ling-3.0 publishes a merged base stage with the ingredient it was merged from [16]; AI2’s TMax publishes RL-step branches carrying a visible capability trade at 2B–27B [17]. **None of it is measured here**: it is the target material the granularity claim was built for, and [Section 7](#) carries it as costed future work. What keeps the current arms from being a bait-and-switch is that the mechanism does not depend on scale — a consolidation trade at 70B and one at 2.8T are the same weight-space event — which is an argument rather than a measurement, and marked as one there.

Where the frontier *does* bear on the design is architecture rather than lineage: K3 activates 16 of 896 routed experts per token and ships MXFP4 weights under quantization-aware training from SFT onward, which makes router- and expert-granular surgery the architecturally current question and MXFP4 arithmetic a **precondition** rather than a `gpt-oss`

curiosity — one that now has its receipt ([Section 5.2](#)).

1.5 The claim

The asymmetry this paper exploits is not “rollback moves along time, not tensors.” It is sharper: **the time axis is gone, and the tensor axis is all that is left**. The material is there — every stage’s full checkpoint is already on disk, its lineage already recorded — and almost nobody operates on it at tensor granularity, on identified damage, under a verification gate. *Almost*: [Section 2.1](#) names the one paper that got to the operation three days before this one, and says so before the claim rather than after it.

Tensor-granular repair — substituting or interpolating selected tensors between two checkpoints of the same lineage, proposed by an agent reading only weight-space telemetry, verified by re-running the evals that define “repaired”, and recorded as content-addressed lineage — recovers capability that a finished training stage damaged or traded away, at a small fraction of the cost of rerunning that stage, and does so **selectively, verifiably, and under a gate that can refuse the repair** — where the alternatives on offer are rerunning the stage or applying one uniform whole-model coefficient, for which no published pipeline we found documents a refusal path.

Three scope notes travel with that sentence and are not fine print. **“Proposed by an agent” is demonstrated end to end at 0.5B and nowhere else**: the 70B graft selection came from the localizer, the 120B ladder is a hand-designed sweep, and [Section 6](#) dispatched no agent ([Section 5.4](#)). **“Damaged or traded away” is carried by different arms and they are not interchangeable**: what is *repaired* above 0.5B is a fault — the 70B `lm_head` — while what is measured about a consolidation *trade* at 120B is attribution, not repair, and the one arm built to repair a trade stopped at its own gate before any repair ran ([Section 6](#)). **No consolidation trade is repaired at scale in this paper**, and [Section 6](#) names the experiment that would do it. And **the third clause is not a claim that selection beats blending**. Two head-to-heads in this campaign’s own evidence went the other way — Cohere’s uniform embedding merge beat their selective one ([Section 2.2](#)), and our own class-selective graft at $p = 0.337$ — p is the fraction of a checkpoint’s bytes an arm moves — lost catastrophically to a uniform blend at $p = 0.50$ ([Section 5.2.3](#)). What is claimed is the *discipline* (selective, verified, gated), not the superiority of selection as an operation.

The claim decomposes into four sub-claims, each with its own falsifier, and [Section 5](#) is organised one subsection per sub-claim.

Accuracy. Surgery or blend recovers the damaged or traded capability to at least the best prior checkpoint, and can

clear bars no ancestor cleared. *Falsifier*: a repair that trails the best recorded version on the damaged metric.

Granularity. Attribution at tensor-class granularity is well-posed: class-wise substitution between two checkpoints of one lineage is *additive*, so “which class carries the trade” has an answer at equal eval budget, and the answer is measurable. *Falsifier*: class-wise substitution is not additive — complementary class sets do not recover the whole delta — or the class our model-free localizer ranks most-moved is also the class whose substitution moves the most eval. **The stronger form this sub-claim was originally registered as — that tensor-selective repair beats the whole-model coefficient production merging uses — is not claimed here.** Its registered falsifier named a global- α baseline, and that baseline was never put to an instrument at 70B, 120B or 11.6B (Section 5.2.3); the claim is lowered to what \$5.2 measured, and the head-to-head is carried as owed work in Section 7.

Efficiency. Repair costs CPU-minutes plus one eval run, against GPU-days to rerun the damaged stage. *Falsifier*: a repair loop costing within an order of magnitude of the rerun. This sub-claim originally had a continuation half; the one arm carrying it was removed and its two predictions withdrawn rather than scored (Section 6). **The permanent wording is: surgery for deployment; this paper makes no continuation claim.**

Agency. An agent reading registry telemetry — drift, localization, gate history, never activations — proposes the correct selection and α , at latency and accuracy competitive with documented human practice. *Falsifier*: agent selections that a whole-model interpolation baseline matches or beats. **This sub-claim is the one exception to the ledger discipline, and we name the exception rather than let it be found: no Agency prediction was ever registered, so Table 2 carries no Agency row and Section 5.4 is two 0.5B lanes and their negative space.**

1.6 What we concede up front

Pretraining loss spikes are largely engineered away at dense scale, and we do not motivate the paper with them. They survive in the design for one reason that is about evidence rather than framing: LLM360’s K2-Spike-1 and K2-Spike-2 are the only public checkpoints taken from *inside* a real 65B-scale spike.

And the concession got worse in the direction that hurts. Our own detection prediction was **falsified on both halves**: a real 65B spike is not a weight-drift magnitude anomaly at two-checkpoint granularity, and the rule we had shipped false-positived on 25.9 % of healthy transitions. On the instrument we brought to it, the spike class is *invisible*. That published

null is the strongest available argument for the framing above: the class this paper was originally written around is both the one the industry stopped having and the one our detector cannot see, while the class it is now written around is the one we repaired to indistinguishability at 141 GB.

1.7 Contributions

1. **A system** (Section 3) in which lineage is an operand: a content-addressed delta registry carrying per-tensor drift telemetry, a CPU-only surgery executor, a gate that runs *the customer’s* evals and refuses the push on failure, and an agent that may propose but never act.
2. **A pre-registration protocol and its full ledger** (Section 4): twenty-three predictions, each registered at a named commit before its run and reported at its recorded verdict — eleven held, six falsified, two not scoreable as registered, and four unscored because a pre-registered spend gate stopped their own lane, for \$96.6 of measured spend. Falsifications are typeset at the same weight as confirmations.
3. **Accuracy evidence** (Section 5.1) on artifacts we did not train: a 70B released-and-broken RLVR checkpoint repaired to 100.00 % / 100.73 % / 100.26 % of the released model, and a 65B hybrid beating the human-chosen roll-back on both purchased probes.
4. **Granularity evidence, including its two falsifications** (Section 5.2): class-wise substitution at 120B is additive, which makes “which class carries the post-train” answerable — and the answer is not the one our localizer predicted, because **drift magnitude localizes where the optimizer wrote, not where the behaviour lives**. We report $R_{\text{res}} = \text{band}/|D|$ — the run-to-run noise band divided by the paired endpoint difference the attribution is a fraction of — as a required column: a gate that makes a ratio *defined* does not make it *precise*. **And a boundary condition arrived as a surprise** (Section 5.2.3): substitution composes only near the donor.
5. **An efficiency accounting with published baselines** (Section 5.3), not estimates: 1,216 H100-hours to rerun Tulu-3-70B’s *DPO* stage (2,880 for its RLVR stage) and LLM360’s own 345,600 A100-hours for the K2 incident, against \$25–50 for surgery plus one gated eval — a denominator taken from a published OLMES rental quote, and one our own Table 6 measurement of \$16.6 for that exact operation on that exact model makes *conservative*.

2 Related work

Every ingredient of this paper exists in isolation. What follows is organised so that the closest neighbour to each of our claims

is named before the claim is made, and the two places where the literature actively contradicts us are stated in our own words rather than left for a reviewer.

2.1 The closest prior art, and it is three days older than this paper

Locked at the Entrance, Open Inside: Where RLVR Narrows the Solution Space [47], submitted 2026-08-29, blends layers 20–28 of a collapsed PPO checkpoint with those of an early checkpoint of the same run at $\alpha = 0.5$ and reports **a 37 % relative gain in solution coverage** ($0.111 \rightarrow 0.152$) at no pass@1 cost. *We state that in the source’s own denominator rather than ours: the 37 % is a relative gain on the collapsed checkpoint, not a fraction of what the collapse destroyed. Against the lost coverage ($0.337 - 0.111 = 0.226$) the confirmed +0.041 recovery is 18 %. Understating the one paper we nominate as prior art would be the worst place in this paper to be imprecise.* From its §6: “layer-targeted weight interpolation, i.e., linearly blending layers 20–28 of the late checkpoint with their step-50 weights, boosts solution coverage to 0.152 at an uncompromised pass@1 of 0.305.” From Appendix E: “Layer interpolation replaces blocks 20–28 by the midpoint of their step-50 and step-275 parameters,” under a fixed 50-problem tuning split and a 100-problem confirmation split with hyperparameters frozen before confirmation and pass@1 equivalence assessed by paired two one-sided tests at a pre-specified ± 0.02 margin. Confirmation result: coverage +0.041 [0.020, 0.069].

That is post-hoc, weight-space, sub-model-granular repair of a collapsed RL policy, selected under a pre-registered eval protocol. **The operation this paper performs is therefore not novel, and we say so here rather than in a limitations paragraph.** What remains, stated as narrowly as the evidence supports:

*Tensor-granular substitution and interpolation between checkpoints of one lineage, proposed from **weight-space telemetry alone**, eval-gated, human-approved, and recorded as content-addressed lineage — demonstrated live at 0.5B on controlled failure injection, replayed against finished public training runs at 6.7B and 65B, and executed at 70B dense and 120B MoE. The operation is prior art as of 2026-08-29; the granularity, the telemetry, the gate and the scale are not. **The global- α ablation is not in that list, because we did not run it** (Section 5.2.3).*

The gap is specific, and we state how we checked it rather than asserting it. Two of the differences are positive facts about the paper. Selection in [47] is a fixed layer band derived from *behavioural* localization — “per-token likelihood shifts are 11x–16x larger prior to the first arithmetic operation” —

i.e. from activations and generations, where our localizer reads weights only; and α is not searched, it is the midpoint, fixed. There is no agent, no provenance and no registry, and the weight-space result is a single Qwen2.5-3B PPO run on one synthetic arithmetic task.

The third difference is an *absence*, and an absence is a search result rather than a fact. **We could not find a whole-model- α baseline or a layer-band ablation in v1** (submitted 2026-08-29; PDF fetched and text-extracted 2026-09-01). The search was a grep of the extracted text for “all layers”, “full-model” and “whole model”, which returns nothing, plus a read of §6 and Appendix E, in which the only layer specification anywhere is “20–28”, stated three times. That is evidence of absence in an extraction we performed once, on one version; it is not proof, and a figure caption or a supplementary release could carry the arm we did not find. On what we can see, the granularity sub-claim is *untested* there rather than pre-empted. **If a revision adds the baseline, this section’s claim narrows but does not vanish.** What survives either way is the residual in the quote above: tensor-granular selection from weight-space telemetry, agent-proposed, eval-gated, at 70B and above, on damage we found rather than damage we induced. **Re-run the check before submission** — it is the one paragraph in this paper whose correctness expires.

One space is still unoccupied and it is worth one sentence: **no paper we found repairs a released collapsed artifact** — someone else’s finished checkpoint, rather than a checkpoint of one’s own run with siblings still on disk. That setting is this paper’s, and Section 5.1’s 70B repair is in it.

2.2 Weight arithmetic is production practice, and it is gated

Task vectors [14], TIES [40], DARE [43], model soups [38] and mergekit [8] establish that weight-space recombination works. It is not a laboratory technique: Command A merges twelve experts across two soup stages [4], Llama 3 averages across variants “at each RM, SFT, or DPO stage” [9], Ling-3.0 replaces learning-rate decay with a checkpoint merge [16, 36], and Kimi K3, DeepSeek-V4 and Nemotron 3 Ultra all consolidate expert fleets by distillation [6, 30, 35].

What they share is **granularity**: one scalar coefficient per ingredient model, applied uniformly across every tensor, constrained to sum to one. Cohere are explicit that this is a choice made against alternatives, not a default: “we employ linear merging (also known as weight averaging), with weights chosen by manual search”; SLERP and task vectors were tried and “found no significant performance improvements, at the cost of increased complexity” (§3.4.3.1 of [4]). Ling-3.0 states the motive in this paper’s own terms — weighted checkpoint merging “enabl[es] offline exploration of different decay profiles **without rerunning costly experiments for each strategy**” [16].

The correction our verification pass forced on us. We had written that the baseline is “one coefficient and no gate”. That is wrong about the gate, at least for the one pipeline whose full report is public. Cohere searched their merge coefficients with heuristics plus brute-force perturbation *against a 500-metric evaluation suite*, under a subsection titled **“Merging is cheap, evaluation is expensive”**:

“while model merging is cheap and fast, evaluating each merge requires significant inference time and compute. Evaluation is therefore a significant bottleneck when applying model merging in a production context.” (§3.4.3.3 of [4])

and “Our final model was informed by 500 separate evaluation metrics” (§3.4.3). The accurate contrast is therefore **granularity** (per-ingredient-model versus per-tensor) and **proposer** (human heuristic search versus an agent reading weight-space telemetry). The cost structure is *the same*, and that is not a weakness of our position — it is the best external, production, primary-source corroboration this paper will get for its own cost model. Merging is free; the eval is the bill.

And the counter-evidence in the same source, which we own rather than omit. Cohere tried tensor-selective merging on exactly one class and reported that it lost:

“some domain experts make use of special tokens (e.g., tool calls) whose embeddings otherwise remain untrained. We find that **selectively merging these embeddings only from checkpoints where they are trained is beneficial, but suboptimal.**” (§3.4.3.2 of [4])

A uniform merge from a shared initialization beat the selective one. This is the strongest published counter-evidence to our granularity sub-claim. The reason Cohere give is that the *initialization* was the real fix rather than that selectivity is worthless, and our own result at $p = 0.337$ (Section 5.2.3) is a second, independent instance of selective substitution losing — so we confront the pair of them together, in Section 7.3, where our own instance is on the table beside Cohere’s, and not here.

2.3 Patching by interpolation, at one global coefficient

PAINT patches open-vocabulary models by interpolating weights at a single global coefficient [15]. Alignment-calibration merging [11] interpolates a model’s weights *before and after alignment* at one global α and reports **Pareto-optimal interpolations that improve accuracy beyond both parents** — post-hoc, weight-space, between two checkpoints of one lineage, and the same “better than both parents” shape

as our own 0.5B blend result. It was public before this campaign began and we missed it. It is the exact PAINT-class foil the granularity claim is built to beat, and it belongs in the baseline column rather than merely in a citation list. Agent-Patch [34] is the closest by name: coarse-to-fine weak-task repair *for merging different agentic multimodal LLMs*, which is backbone selection plus model-wide repair — not per-tensor selection, and with no verification loop.

2.4 In-run practice is temporal, and mostly preventive now

The instability literature’s remedies act along time: restart-and-skip-batches as recorded in the OPT-175B report — “at least 35 manual restarts . . . over the course of 2 months” on “992 80GB A100 GPUs”, \$2.4–\$2.5 of [45] — GLM-130B’s embedding-gradient shrink [44], OLMo 2’s data-hygiene, initialization and z-loss mitigations [32], LAWA’s checkpoint averaging [18]. We treat this line as *solved for its own failure class* rather than as our foil, and Section 1 concedes that rollback for an in-flight run got cheap and stayed available. The foil is what happens after the run ends: nobody repairs a finished checkpoint in place by replacing only what broke.

2.5 RL post-training instability, and where post-hoc repair has reached

Policy entropy collapses early and monotonically across RLVR runs absent intervention [5]. We swept the 2025–2026 line on 2026-09-01 and it is overwhelmingly what Section 2.1 implies: objective, optimizer, advantage-shaping, clipping, sampling or data interventions applied *inside* the loop — OPEFO [39], Entrocraft [22], HEAL [23], SAF-OPD [7] and roughly thirty others. Three exceptions mark the real frontier, in increasing proximity to our operation. TS-OPSD [41] starts from an entropy-collapsed RL checkpoint and repairs it after the fact, but does so by *training* it — self-distillation from a temperature-scaled self-teacher. That costs GPU-hours where ours costs an eval, which makes it a *supporting* citation for our efficiency sub-claim rather than a threat to it. Alignment-calibration merging [11] interpolates whole models at one global α . And [47] is the one treated above.

2.6 Fault-injected damage is characterized but not repaired

Silent data corruption on unhealthy nodes drives models to “converge to different optima with different weights,” sometimes with no distinguishing signal in the loss [25]. The literature’s response is detection and step recomputation — rollback again — and nothing we found addresses corruption already resident in a released artifact. The one public instance of such an artifact, allenai/Llama-3.1-Tulu-3-70B-broken [1], is released

with the explicit invitation to “try and reconstruct an LM head” and is the subject of [Section 5.1](#).

2.7 Tensor-level storage exists, passively

Git-Theta [19], ZipLLM [37], TStore [21], MGit [10], DeltaZip [42] and Hugging Face’s Xet backend all store or serve models at sub-checkpoint granularity, and MGit in particular is a content-addressed tensor store with a lineage graph and parent-relative deltas. None of them uses the registry as the substrate for *repair*: the lineage graph is a storage optimization there, and an operand here.

2.8 Agents already babysit training runs, above the weights

Interactive Training [46] and EvoTrainer [3] put a model-driven loop around a training run, intervening on hyperparameters, data and harness. Neither touches weights. That is the boundary our agent layer crosses — under an approval gate and a deterministic executor ([Section 3.4](#)), and never with the agent itself holding the write path.

2.9 Knowledge editing is a different operation

ROME [27] and MEMIT [28] *synthesize* new weight values to change specific facts. We move existing, provenance-tracked bytes between two checkpoints that both exist, and gate on evals rather than on edit-success probes. The two lines share a substrate and share nothing else.

3 ARZACH: lineage as an operand

Tensor-granular repair needs four things that do not usually sit together: two checkpoints of one lineage addressable at tensor granularity, a way to say *which* tensors moved without loading weights, an operator that recombines recorded bytes deterministically, and a judge that is the customer’s own definition of “repaired” rather than a benchmark we chose. ARZACH is those four things. This section is short on purpose — every design decision below exists because a measurement in [Section 5](#) depends on it.

3.1 The artifact: a delta pack whose manifest is telemetry

A model version is stored as an OCI artifact carrying a manifest and, for a non-root version, compressed per-tensor deltas against a *resolved ancestor* rather than against a fixed base. Three decisions matter downstream.

Content hashing is over the canonical tensor set, not file bytes. Shard layout and tensor order are invisible to identity, so a 15-shard checkpoint and its single-file save resolve to the same model. This is what makes “are these two checkpoints the same model” a decidable question, and [Section 5.2](#) leans on it directly: two arms of the 120B ladder that were built from opposite directions turned out to have the *same content hash*, which is how we learned they were the same model and skipped paying for its evaluation.

Precision is part of identity, deliberately. The hash covers (name, dtype, shape, bytes). A BF16 save and an F32 save of the same weights are two models. Casting is directional — widening from BF16 is exactly reversible, narrowing is not — so a dtype-blind rule would declare an F32 master and its degraded BF16 release the same artifact, which is false. The one place this bit us is recorded honestly in [Section 5.1](#): the broken 70B checkpoint is serialized F32, so it had to be cast before it could be compared, and the cast turned out to be information-lossless for a reason that is itself a finding.

Every manifest entry carries per-tensor drift telemetry. Each tensor records $\{\ell_2, \text{rel}, \text{max}_{\text{abs}}, \text{cos}\}$ against the resolved ancestor, computed in the same encoder pass that compresses the delta. The manifest is therefore training telemetry as well as payload: “which classes did this stage move, and how hard” is answered from the artifact alone, with no weights loaded and no GPU. **This is the only input our agent is given for localization, and [Section 5.2](#) is the measurement of how far that input can be trusted.** The answer is: less far than we assumed.

3.2 Surgery: five operations over recorded bytes

Surgery takes a *candidate* version, a *reference* version of the same lineage, and a *selection* — a list of tensor-name globs — and produces a new version. It is CPU-only recombination of bytes the registry already holds; no GPU is involved and no training occurs. Five modes ship:

substitute selected tensors are replaced by the reference’s, byte for byte.

blend selected tensors become $(1 - \alpha) \cdot \text{candidate} + \alpha \cdot \text{reference}$.

layered blend one α per selection, the union as the selection — differential pull, in one recorded operation.

ties each parent’s delta from the shared base is trimmed to its *top-density* fraction by magnitude, signs are elected where they disagree, and the result is rescaled [40].

extrapolate selected tensors become candidate + $\beta(\text{reference} - \text{candidate})$ with β outside $[0, 1]$ — negation [14] at $\beta < 0$, continuation past a partially-recovered reference at $\beta > 1$.

Substitute and blend carry every result in this paper. **The other three shipped and were never exercised live**, and Section 7 says so rather than letting the feature list imply evidence.

Two properties of the executor are load-bearing for the results. The output version’s name is *derived* from the candidate’s plus the mode’s parameters, not chosen — a blend at $\alpha = 0.25$ and one at $\alpha = 0.5$ cannot collide at the immutability guard, so an iterating agent’s second attempt is a new artifact rather than an overwrite. And every derivation record carries both parents’ content hashes, per-class tensor counts, per-tensor SHA-256 provenance samples, a non-finite count and the assertion results: the 65B arm ladder in Section 5.2 passed 19 of 19 assertions with zero non-finite values, which is how we know its catastrophic result is the operation and not a bug.

3.3 The gate: the customer’s evals, and a refusal

A registered *eval policy* names a runner (any executable), a timeout, and a per-metric threshold table. The daemon invokes the runner as `<runner> --reference <path> --candidate <path>`, and the runner prints one JSON object of paired metric values. Thresholds are absolute floors (min, max) or paired deltas (min_delta, max_delta); the two ask different questions and the policy must say which it means.

Three things about the gate are deliberate and all three matter to Section 5.

The comparison is paired. Absolute scores move with sampling, prompt sets and engine versions, so “is this model good” is a question we cannot answer honestly and do not try to. “Did this operation change the answer” is one we can.

An eval that did not run is not an eval that passed. A runner that exits non-zero, prints a malformed report, times out, omits a gated metric, or reports NaN fails the gate, and the push is refused. A gate that goes green because the harness ran out of GPU memory is worse than no gate, because it produces a badge nobody earned.

The gate has no opinion. There is no suite here, no leaderboard, no compression-quality score. The entire value is that the bar is the customer’s. The suites used in this paper (Section 4.3) are therefore *our* eval policies for *our* experiments, not a benchmark ARZACH endorses.

3.4 The agent layer: propose, never act

The daemon publishes a machine-readable event feed. Three model agents watch it and one deterministic executor acts.

triage wakes on a push whose gate did not pass. It reads the gate runs, walks the pack’s history to the last version that passed, localizes the candidate against it, and files one finding. When the diagnosis points at a repair *and* a dry run confirms the substitution would actually change tensors, the finding carries a machine-readable *action*.

patch wakes only on addressed request. It surveys a whole chain, builds a capability matrix over gate history, and proposes one of: a surgery, a blend, a layered blend, a TIES merge, an extrapolation, a rollback, or *null* — “the chain is healthy, and saying so is the deliverable”.

sentinel wakes on every push. It gathers statistics with no model involved, decides deterministically whether a class jumped, and only then calls a model to write prose. The model contributes the changelog’s title and body and the anomaly’s body, and nothing else; every structured field is written from the collected statistics afterwards and the model’s own arguments for those fields are read and discarded.

The anomaly rule the sentinel used at the start of this campaign flags a class when its hardest-moved module’s relative drift exceeds a fixed multiple of that class’s previous mean. **That rule was falsified by our own pre-registered prediction** at a 25.9% false-positive rate, replaced by a shape statistic, and shipped behind an environment switch with the old rule preserved row for row so the falsification stays reproducible against the product (Section 5.4).

The split between “writes down” and “acts” is the whole design. The three agents’ toolsets contain reads, one dry-run preview that writes nothing, and a terminal `submit_finding` that never reaches the daemon. The three calls that change what the registry serves — surgery, gate-run, finding-status — are in *no* toolset. The executor that performs an approved action **contains no language model at all**; a test parses its import table to keep that true. The status machine has four legal edges and **no path from open to executed**: a human clicks Approve, or nothing happens. An agent cannot file a finding that is already approved and cannot move one.

This is a bound on what a model can *do*. It is not a claim that what it *writes* is any good, and Section 5.4 is careful about the difference: the loop is proven end-to-end by a hermetic browser test in which the one fake is the model, and the quality of the model’s judgement rests entirely on two live 0.5B lanes.

Table 1: The five pre-registration receipts. Each is a commit in which the predictions were written down, with their falsifiers, *before* the corresponding run was funded. The wording binding on each prediction is the wording at its receipt.

Receipt	Predictions registered
363b8f3	P-C1, P-MX1, P-B1', P-B3, P-A1', P-A4'
5ecb6bc	the P-MX1 refinement, recorded pre-run
e756aeb	P-C2a-0... 3, P-C2b-Q, P-C2b-N
15889f2	P-B2'-0... 5
2ec8ef9	P-A'-0... 7, post-amendment

4 Experimental design

4.1 Pre-registration, and why the receipt is a commit

Every ratio this project published before 2026-08-10 came from a hand-picked pair, and the most interesting one arrived by accident. “We predicted X and measured Y ”, told after the fact, is not evidence: a skeptical reader cannot tell which numbers were predictions and which were rationalizations. Writing the prediction down first is the whole mechanism, and the mechanism has four parts.

Selection, stated before the run. Why this pair, what was rejected and why, and whether we produced the artifact ourselves. A search that surfaced one usable artifact out of eighty-six is itself a finding about the population; record the denominator.

Pinning. Both sides pinned by *commit sha*, never by tag or branch; the tool pinned by git commit and built from it; hardware and wall time recorded.

Registration. The prediction as a number or a range, the reasoning that produces it, **a specific measured value that would falsify it**, and what counts as a null result — with the commitment to publish it. Where a prediction can be derived from headers alone (tensor counts, dtypes, a LoRA’s target modules), it is derived, and the derivation is stated.

Reporting. The prediction and the outcome reported *together*, including when they disagree, and any prior claim of ours that turns out wrong corrected in place rather than in a quieter appendix.

The commit hash of the registration is the receipt, and it is binding as worded *there* rather than as re-described anywhere later. The five receipts behind this paper are given in [Table 1](#), and every verdict in [Section 5](#) names the one it answers to. Where a lane report and a later summary disagree on a number, the lane report wins and the disagreement is named.

We are explicitly *not* doing sampled-corpus statistics with confidence bounds over a population of published artifacts. The population of, for example, public checkpoints taken from inside a real 65B loss spike has one member. You cannot sample that. What you can do is pre-register, pin, and report nulls — and the honest headline of this campaign is that **six of twenty-three predictions were falsified, two were not scoreable on the instrument they were written for, and four were never put to an instrument at all, because a pre-registered spend gate stopped their own lane**. All twelve are reported at that verdict.

4.2 The arms, as run

Five arms were designed; four ran. Every artifact in every arm is one we did not train, except the injected-failure lane at 0.5B.

Arm C1a — released-and-broken, 70B dense.

allenai/Llama-3.1-Tulu-3-70B-broken [1], a finished RLVR checkpoint whose `lm_head` was destroyed by a checkpoint-saving bug, repaired by graft from its own lineage parent Llama-3.1-Tulu-3-70B-DPO, against the released Llama-3.1-Tulu-3-70B as denominator. Registered at 363b8f3.

Arm C2 — consolidation trade, 120B MoE.

openai/gpt-oss-120b against openai/gpt-oss-safeguard-120b [33]: a twelve-arm class-substitution ladder measuring which tensor class carries a safety post-train’s capability cost, plus an MXFP4 arithmetic precondition study. Registered at e756aeb; the MXFP4 round-trip half at 363b8f3 and 5ecb6bc.

Arm B — finished public training runs, 6.7B and 65B.

Two model-free sub-lanes and one counterfactual repair lane. **P-B3** mines LLM360’s Amber-7B eval trail for capability regressions the loss curve does not show. **P-B1'** tests whether our shipped drift-anomaly rule detects the first in-spike transition of LLM360’s K2-Spike-1 without false-positiving on healthy transitions. **P-B2'** builds an eight-arm p-frontier between the spiked and rolled-back branches of the same 65B incident, with the human rollback as a first-class arm. Registered at 363b8f3 and 15889f2.

Arm A' — consolidation under control, 11.6B. Two sibling specialists of one Olmo 2 7B lineage [2, 32] merged by us at one uniform coefficient over all 483 tensors, against a ladder of tensor-selective repairs and the unmerged ingredients as oracle ceilings. It is the only arm whose damage event we perform ourselves and the only one that is contamination-clean by construction. Registered at 2ec8ef9; it ran and **stopped at its own gate** ([Section 6](#)).

Arm A — injected failures with a training loop. The only arm where we train, and the only source of evidence for the *continuation* half of the efficiency sub-claim. **It did not run**; Arm A' replaced it (Section 6).

Two live end-to-end lanes at 0.5B, on controlled failure injection, carry the agent evidence and are described where they are used (Section 5.4).

4.3 Instruments

Weight-space instrument. `modelctl diff` over both checkpoints, producing per-tensor and per-class {mean_rel, max_rel, cos, identity rate, non-finite count} plus the compressed delta size. It reads weights and never activations. Raw `--json` output is committed verbatim for every diff in the paper.

Capability instrument. `lm_eval` 0.4.11 driving vLLM 0.11.0 on a pinned stack (torch 2.8.0+cu128, transformers 4.57.6, CPython 3.11.10), greedy decoding, seed 1234, with `--log-samples` on so every verdict is rerunnable over per-item rows rather than over a summary — 2,731 rows per run on the 120B ladder, and 7,683 paired items per arm across the 65B lane's *registered* suite, of which 1,983 were purchased (123 bpb_general + 1860 bpb_math; the seven-benchmark half was bought on `spike_240` alone, Section 5.1.2).

Noise bands, and the fact that they changed our readings.

No single-run comparison in this paper is treated as a measurement. Two band constructions are used, and which one applies is stated at every use. *Run-to-run bands* come from $N=3$ repeats of one arm on one suite; *paired bootstrap bands* come from 10,000 resamples over per-item records at seed 20260831. The bands did real work in both directions: the 70B band came in at exactly 0.000 on both loglikelihood tasks — *not* the spread the runbook expected — while on the 120B ladder GSM8K's $N=3$ band is 7.0 pp, *wider than the 6.5 pp effect it was meant to measure*, and the channel was refused rather than reported.

Resolution, not just separability. Passing a separability gate makes an attribution ratio *defined*; it does not make it *precise*. We therefore report

$$R_{\text{res}}(m) = \frac{\text{run-to-run band}(m)}{|D(m)|}$$

as a **required column** beside every attribution ratio, where D is the paired endpoint difference. On the 120B ladder this is ± 0.137 on MMLU but ± 0.500 and ± 0.508 on HellaSwag and IFEval — so every claim in that lane is made on MMLU and checked against the others, never the reverse. Adding the column changed the lane's own reading of its results.

Pre-registered aborts. Each runbook carries numbered abort conditions that fire before or during a run and are reported whether or not they fire. Two fired and both are in Section 5: a cross-instrument check disqualified HellaSwag on the 65B lane at -10.70 pp against LLM360's own published log while five other metrics landed inside 0.8 pp; and a byte-identity abort caught that two arms of the 120B ladder were the same model and refused to evaluate the second, saving one suite run.

4.4 Cost accounting

Spend is reconciled against the provider's own `clientBalance` deltas rather than estimated from posted rates, and both are reported. Every pod and network volume in every lane was deleted and **404-verified** after the lane, and the resources belonging to the machine's owner were verified untouched byte-for-byte on their status records. Waste is on the record too, itemised in Section 5.3: it is small, and reporting it is cheaper than being asked.

External costs — what it would cost to rerun a damaged stage instead — carry a confidence label at every use: [PUBLISHED] where the source states the number literally, [DERIVED] where it is arithmetic on published inputs with the chain shown, [ESTIMATED] where no primary source exists and the bound and its basis are stated. Nothing in this paper quotes an external cost without one.

5 Results

Table 2 is the whole ledger: every prediction registered in this campaign, at the verdict its lane recorded, before any of them is discussed.²

5.1 Accuracy

5.1.1 A released, broken 70B checkpoint, repaired to indistinguishability

`allenai/Llama-3.1-Tulu-3-70B-broken` [1] is a finished 70B RLVR checkpoint that AI2 released with the LM head missing — “This is a model missing the LM head, caused by an unfortunate bug in checkpoint saving” — and an explicit invitation to reconstruct one. It has 722 tensors where its lineage parent has 723. We pinned all three checkpoints by revision sha, re-verified against the Hugging Face API immediately before renting, and fetched 560.2 GB in 47 minutes.

²The headline count — nine held, six falsified, two not scoreable as registered, four unscored because a pre-registered gate stopped their lane — covers the twenty-one predictions of the five sub-claim tables. The ledger below carries four further resolved items that count separately: P-B2'-0 and P-A'-0 are pre-spend gates rather than claims, P-B2'-3 was registered in the same design but is not a sub-claim row, and C2b-0a extends P-MX1's finding to the second parent. P-B2'-3 is counted as held because its structural clause held; its row and Section 5.1.2 both name the registered *value* it missed.

Table 2: The full pre-registration ledger. Verdicts are the exact ones recorded by the lanes; nothing is smoothed, and a falsification is a row like any other. Receipts are the commits of Table 1.

Prediction	Receipt	Verdict	Number
<i>Accuracy (Section 5.1)</i>			
P-C1	363b8f3	HELD, stronger	100.00 % / 100.73 % / 100.26 % recovery; $N=3$ band 0.000 / 0.010 / 0.000, full-width in absolute score
P-B2'-0	15889f2	passed, \$0, pre-spend	gsm8k separates -6.7 against σ 2.18; mathqa -2.1 against σ 0.82
P-B2'-1	15889f2	HELD	graft-embed -0.017658 / -0.001549 ; blend-a0.25 -0.012178 / -0.001556 ; both CIs exclude zero
P-B2'-2	15889f2	NOT SCOREABLE AS REGISTERED	gsm8k unbought for that arm; on bpb_math it closes 8.8 % of the gap, consistent with the 25 % bar
P-B2'-3	15889f2	HELD on its structure, missed on its value	no tested hybrid is indistinguishable from rollback, which is the clause that held; the bracket $p^* \in (0.75, 1.0]$ sits <i>above</i> the registered $p^* \approx 0.60$
P-B2'-4	15889f2	HELD	-0.033022 [-0.037942 , -0.028600] and -0.019372 [-0.020268 , -0.018485] bits/byte
P-B2'-5	15889f2	NOT SCOREABLE AS REGISTERED	LLM360's trail supports it (5 of 7 inside noise); our two probes contradict it (0 of 2). Both reported; the lane does not pick
<i>Granularity (Section 5.2)</i>			
P-C2a-0	e756aeb	HELD	3 of 5 separate: mmlu $ D = 0.0383$ vs band 0.0053; hellaswag 0.0120 vs 0.0060; ifeval 0.0400 vs 0.0203
P-C2a-1	e756aeb	FALSIFIED	$R(A1) = -0.031$ (mmlu), $+0.167$, $+0.400$ — max below 0.5 on every gated metric
P-C2a-2	e756aeb	HELD	$+1.046$ vs -0.031 on mmlu; the router's complement recovers the whole delta
P-C2a-3	e756aeb	FALSIFIED	argmax is attention ($+0.565$) > head ($+0.336$) > experts ($+0.061$) > router (-0.031)
P-C2b-Q	e756aeb	FALSIFIED on its falsifier	5 distinct byte outputs on the interior grid (7 of 7 overall), not 2; mechanism clause held at 91.54 %
P-C2b-N	e756aeb	HELD numerically	72 BF16 router tensors, 26,551,296 B = 0.0407 %, 5 of 5 distinct, 0 non-finite. Refusal path untested
P-MX1	363b8f3, 5ecb6bc	HELD	bit-exact under preserve <i>and</i> recompute-up; max_abs_err 0.0
C2b-Oa	e756aeb	HELD	72/72 byte-identical under both policies over 114,661,785,600 elements
<i>Detection and gates (Section 5.4, Section 5.5)</i>			
P-B1'	363b8f3	FALSIFIED on both halves	spike invisible to drift magnitude; 25.9 % false-positive rate on 54 healthy transitions
P-B3	363b8f3	CONFIRMED	predicted ≥ 3 , measured 5 , for \$0
<i>Consolidation, performed under control (Section 6)</i>			
P-A'-0	2ec8ef9	passed, \$0, pre-spend	the publisher's matrix separates them in opposite directions: Math 55.8 vs 26.9, Code 22.1 vs 50.4
P-A'-1	2ec8ef9	FALSIFIED	below <i>one</i> ingredient, not both: -37.20 / -52.12 on Code, $+18.73$ / $+5.40$ on Math, all four separated
P-A'-2	2ec8ef9	UNSCORED AT THE GATE	G3 stopped the lane before any repair arm ran. This is the gate working, not a gap
P-A'-3, P-A'-4	2ec8ef9	UNSCORED AT THE GATE	no repair arm was evaluated; the instrument that would score them exists and is tested
P-A'-5	2ec8ef9	UNSCORED AT THE GATE	the patch agent reads gate history and was never dispatched. No proposal exists
P-A'-6	2ec8ef9	FALSIFIED	on its condition, not its falsifier: above the floor on both Math channels ($+15.01$, $+5.00$), below on both Code (-4.88 , -8.73). <i>nonfinite</i> = 0 holds
P-A'-7	2ec8ef9	HELD	construction + verification \$0.840 against \$18.04 of eval = 4.4 %, bar 10 %
<i>The whole-model baseline, run after review (Section 5.2.3)</i>			
P-C2-GA	8eaf457	HELD	the global $\alpha=0.5$ blend sits below A9 on the deciding channel: R $+0.489$ vs $+1.046$, 4.1 resolutions apart
P-C2-GA-L	8eaf457	HELD	R landed 0.011 from the linear-mixture prediction 0.5 — global α is a near-linear dial on this pair
<i>Efficiency (continuation half) (Section 6)</i>			
P-A1', P-A4'	363b8f3	withdrawn, not scored	their arm was removed before they ran; the wording stands at its receipt and is binding there. The paper makes no continuation claim

Localization first: the damage does not extend beyond 1m_head. The drift between broken and its DPO parent is $\text{mean_rel } 1.4264 \times 10^{-5}$ with cosine 0.99999999986, and the worst class max_rel *anywhere* in the localization — across both the DPO and the released comparison — is 5.7×10^{-5} , that is 0.0057 %, about $88\times$ below the runbook's 0.5 % damage bar. The $\text{dpo} \leftrightarrow \text{final}$ control reproduced an earlier independent measurement exactly: delta

$3,133,681,469 \text{ B}, 45.0\times, \text{mean_rel } 1.426 \times 10^{-5}$.

Two facts about the saving bug that AI2 does not document. Both came out of the localization rather than being looked for.

First, -broken is F32-serialized but BF16-valued. The runbook predicted the diffs would carry a BF16 representation floor near 1.1×10^{-3} relative and would be “a destruction

Table 3: C1a, 70B dense. Six runs on one GPU pod, TP=2, seed 1234, greedy; item counts identical across arms (mmlu 1,140, gsm8k 200, hellaswag 500).

Arm	Role	MMLU	GSM8K	HellaSwag
broken-randhead	floor	0.1632	0.0000	0.2460
hybrid-dpo rep1	result	0.8035	0.910	0.770
hybrid-dpo rep2	repeat	0.8035	0.920	0.770
hybrid-dpo rep3	repeat	0.8035	0.920	0.770
hybrid-final	oracle ceiling	0.8035	0.915	0.770
final	denominator	0.8035	0.910	0.768

detector, not a drift detector”. Measurement says otherwise: observed drift is 1.4×10^{-5} , about $77\times$ below that floor, and 14.52 % of tensors came out *byte-identical* to a natively-BF16 donor. A lossy cast cannot produce byte-identical output. The checkpoint-saving bug serialized at F32 without adding precision, so the cast down is information-lossless and these diffs resolve ordinary training movement after all — which makes the localization *stronger* than pre-registered.

Second, it was saved from the inner module. `-broken’s` tensors carry no model. prefix: `embed_tokens.weight`, `layers.0.self_attn.q_proj.weight`, `norm.weight`. It was saved from the inner `LlamaModel` rather than from `LlamaForCausalLM` — which is almost certainly the same bug that dropped the head, since `lm_head` lives on the outer module. After adding the prefix the two name sets differ by exactly one element: `['lm_head.weight']`.

The repair, and its floor. The graft substitutes the 722 body tensors of `broken` into a head donor, producing `hybrid-dpo` (the result: DPO’s head, broken’s body) and `hybrid-final` (an *oracle ceiling*, not a result: the released model’s head, which a repairer would not have). 19 of 19 derivation assertions passed and per-tensor SHA-256 provenance confirms each hybrid’s head is byte-identical to its intended donor’s. A fourth arm, `broken-randhead`, redraws the head from $\mathcal{N}(0, 0.02)$ under `numpy.default_rng(1234)`: **it is chance performance by construction and is labelled as such, with the seed, wherever it appears.** Table 3 is the whole lane.

P-C1 HELD, on all three gated tasks, by a margin the noise cannot reach. Recovery is 100.00 % / 100.73 % / 100.26 % on MMLU / GSM8K / HellaSwag against a 95 % bar — margins of +5.00, +5.73 and +5.26 pp. The measured $N=3$ band is **exactly zero** on both loglikelihood tasks and 0.55 pp on the one generative task — *that figure is a half-width, expressed as a percentage of final’s score; Table 2 prints the same band in absolute score units at full width, where it reads 0.010* — so on the only task where run-to-run noise exists at all the margin is $10.4\times$ it. The right statement is not that the graft recovers 95 % of the released model but that **it is indistinguishable from it**, and nominally above it on two of three tasks by one item each, inside or at the band.

Three qualifications travel with that result. The oracle ceiling `hybrid-final` does *not* beat the achievable repair past the band on any task — on GSM8K it is nominally *below* it — which means RLVR barely moved the LM head on this chain, and which is a fact about this lineage rather than about grafting. The denominator `final` was run once: the honest reading is that its 0.910 sits inside `hybrid-dpo’s` $[0.910, 0.920]$, i.e. the two are indistinguishable, and the +0.73% is *not* a claim that the repair beats the released model. And on this suite, at this scale, the body carries the score and the head choice is worth at most one item — provided the head is a real head at all, which is exactly what the 0.1632/0.0000/0.2460 floor arm is there to price.

5.1.2 A 65B loss-spike incident, replayed as a ρ -frontier

LLM360’s K2 pretraining hit two malignant loss spikes and the team rolled back. They published both branches — IFM/K2-Spike-1 spiked, IFM/K2 mainline — and, unusually, the recorded eval trail of both. Three zero-cost identity controls fixed the material before any GPU was rented: the fork is exact, the healthy arm *is* the model LLM360 shipped (27 of 27 shards byte-identical), and the spiked branch at 240 is a distinct model (0 of 27 identical).

Their numbers first, and they are unambiguous: at the moment the rollback was made there was no deficit to recover. We read LLM360’s own trail for both branches off their public dashboard, anonymously, for \$0. At checkpoint 240 the rollback *target* is behind the checkpoint it replaced on all seven metrics they record per checkpoint on that dashboard, **including both damaged ones:**

Metric (shots)	rollback	damaged	counterfact.
	ckpt_160	spike_240	ckpt_240
GSM8K (5)	29.0	33.9	40.6
MathQA (5)	34.0	35.2	37.3
MMLU (0)	54.4	59.0	59.2
HellaSwag (10)	81.6	83.5	83.6
WinoGrande (5)	76.2	78.9	80.0
TruthfulQA-mc2 (0)	38.6	40.1	41.7
ARC-Challenge (25)	56.3	59.6	61.3

`ckpt_160` is the branch point, byte-identical on both branches, so it is one row; the spiked branch kept learning maths across those 80 checkpoints, just more slowly than the restarted one. The table is also the pre-spend gate P-B2’-0, passed on the record rather than on hope: gsm8k separates by -6.7 against a trail σ of 2.18, mathqa by -2.1 against 0.82.

P-B2’ asked whether a *partial* revert can beat a *full* rollback; on their own record the **empty** revert already does.

And the third column is the one a reviewer should press, so we press it first. `ckpt_240` is not an arm and no repair here may see it, but it *is* the artifact the rollback decision

actually produced, 30 days later, and it beats `spike_240` on all seven trail metrics above and on both of our own probes — by 0.0147 and 0.0195 bits/byte in [Table 4](#), margins outside every confidence interval in that table. So the claim this lane supports is narrow, and it is the only form in which we make it:

The rollback was unnecessary at the moment it was made. At the restart point there was no measurable deficit to recover — not on LLM360’s own instrument, and not on ours. This is a statement about the *decision*, at the information available when it was taken. It is *not* a statement that the rollback’s eventual output was no better than the branch it replaced. It was better, on all nine metrics either instrument could read.

We make no claim that LLM360 decided wrongly — they had a loss curve and a qualitative read, we have their checkpoints, their evals and two years. The claim is about the *instrument*, and [Section 5.3](#) prices it.

Our own probes corroborate it, and this is exactly what they are. We built an eight-arm ladder in ρ , the byte fraction reverted from `spike_240` toward `ckpt_160`, with the rollback as a first-class arm and `ckpt_240` as an oracle no arm may see. The registered gated suite is nine metrics — K2’s seven benchmarks plus two bits-per-byte probes — and we bought the probes on all eight arms and the benchmark half on `spike_240` alone, cut on measured cost at \$9.05 per arm against \$21.9 of headroom. **Every hybrid statement below therefore rests on 2 of the 9 registered gated metrics, the capability-confirmation leg is unpurchased, and no run-to-run band was measured in this lane at all** — the 70B lane’s 0.000/0.010 band is cited as *prior*, never adopted here. That scope is stated once and governs [Table 4](#) entire — whose first row is the trail’s finding on an independent instrument: `spike_240` is -0.018324 bits/byte *better* than the rollback on general text with the CI separated, and $+0.000168$ on the math probe with the CI straddling zero.

Tied on the damaged capability and better on everything else, then, on both instruments. Three registered predictions resolve on the ladder itself. **P-B2’-1 HELD:** two of the three qualifying arms — qualifying means *a hybrid at $\rho \leq 0.34$* , as registered, which is `graft-embed`, `blend-a0.25` and `graft-attn-vocab-norm` — beat the rollback on both available gated metrics with nothing worse: `graft-embed` at $\rho = 0.004019$ and `blend-a0.25`, both CIs excluding zero. So reverting the embedding matrix alone, 0.4 % of the bytes, moves maths in the right direction. (`blend-a0.75` also beats the rollback on both, and is not one of the three: it is above the registered ceiling.) **P-B2’-4 HELD:** `ckpt_240` beats `ckpt_160` past band on both metrics, by 0.033 and 0.019 bits/byte, so the 80 discarded checkpoints are not free. **P-B2’-3 HELD on**

its structure and missed on its value. The clause that held is the one it was written to test: no tested hybrid is indistinguishable from the rollback, because every rung separates from it on at least one gated metric, including $\rho = 0.75$, and since $\rho = 1$ is the rollback, $\rho^* \in (0.75, 1.0]$. **That bracket lies entirely above the registered $\rho^* \approx 0.60$** , so the number the prediction named was missed, and we score it that way rather than let HELD stand beside a missed value. A bracket is what a discrete ladder honestly supports.

And two the purchased instrument cannot score. Both are reported NOT SCOREABLE AS REGISTERED rather than re-specified onto it. P-B2’-2 asked about `gsm8k`, never run on `graft-embed`; on `bpb_math` that arm closes 8.8 % of the damage gap, consistent with its 25 % bar but not scored on it. P-B2’-5 asked about the seven benchmarks, and the evidence points both ways: five of the seven above sit inside their own noise floors at 240 with the deficit confined to `gsm8k` and `mathqa`, while our two probes put 0 of 2 within band. Both are reported; the lane does not pick. (HellaSwag was disqualified by pre-registered abort A3, -10.70 pp against LLM360’s own log while five other metrics landed inside 0.8 pp.)

5.1.3 End to end at 0.5B, with the agent in the loop

Two live lanes at 0.5B, on controlled failure injection, ran the whole loop — localize, propose, human-approve, execute, re-gate — with a real model driving the agent. The graft repair produced **v5-fix** at perplexity 6.533 and QA 0.833, better than any recorded version of that chain, because the epoch-5 MLP and embedding progress was kept, for \$0.82. The blend repair produced **v6-fix-a40** at $\alpha = 0.4$ — perplexity 6.819, QA 0.867, math 0.85 — **the first version in the chain’s history to clear all three capability floors**, for \$0.22. These are the paper’s only end-to-end results and their scale is stated every time they are used.

5.2 Granularity

5.2.1 Which class carries a consolidation trade? Not the one that moved

`gpt-oss-120b` and `gpt-oss-safeguard-120b` are the same architecture, same tensor count (687), same `total_size` (65,248,815,744 B), differing only by a safety post-train. Both are MXFP4 in the experts and BF16 in attention, router, embeddings and head. We ran the base three times and the safeguard once on a five-channel suite, and gated the attribution on separability.

P-C2a-0 HELD: three of five channels separate. MMLU $|D| = 0.0383$ against a band of 0.0053; HellaSwag 0.0120 against 0.0060; IFEval 0.0400 against 0.0203. Two channels were **refused rather than reported**: GSM8K’s own $N=3$

Table 4: P-B2', 65B. Paired bootstrap over per-item records (123 for bpb_general, 1860 for bpb_math), 10,000 resamples, seed 20260831. $\Delta = \text{arm} - \text{ckpt_160}$; negative is better than the human rollback. Every arm byte-verified, nonfinite 0 throughout.

Arm	Role	p	Δ bpb_general	95% CI	Δ bpb_math	95% CI
spike_240	damaged	0.000	-0.018324	[-0.020484, -0.016345]	+0.000168	[-0.000754, +0.001061]
graft-embed	hybrid	0.004	-0.017658	[-0.019755, -0.015766]	-0.001549	[-0.002443, -0.000651]
blend-a0.25	hybrid	0.250	-0.012178	[-0.014972, -0.009474]	-0.001556	[-0.002366, -0.000719]
graft-attn-vocab-norm	hybrid	0.337	+ 0.761713	[+0.708740, +0.818196]	+ 1.267118	[+1.252220, +1.282355]
blend-a0.50	hybrid	0.500	+0.000569	[-0.002333, +0.003547]	+0.003349	[+0.002632, +0.004071]
blend-a0.75	hybrid	0.750	-0.001505	[-0.002998, -0.000054]	-0.003296	[-0.003741, -0.002860]
ckpt_160	rollback	1.000	—	—	—	—
ckpt_240	oracle	—	-0.033022	[-0.037942, -0.028600]	-0.019372	[-0.020268, -0.018485]

Table 5: C2a, 120B MoE. Twelve arms built, eleven distinct, fifteen suite runs. R_{res} in the header row is the resolution of each column, not an error bar on a single cell.

Arm	% bytes	MMLU ± 0.137	HellaSwag ± 0.500	IFEval ± 0.508
A9-everything-but-router	99.9593	+1.046	+0.833	+0.400
A10-attn-graft	2.9297	+ 0.565	+0.167	+0.262
A4-router-and-head	1.8159	+0.336	+0.500	+0.354
A3-head-graft	1.7752	+0.336	-0.167	+0.077
A11-experts-6layers	15.5796	+0.313	-0.000	+0.215
A7-router-mid	0.0136	+0.176	+0.333	-0.200
A5-router-early	0.0136	+0.153	-0.833	+0.215
A6-router-late	0.0136	+0.107	-0.500	+0.262
A8-experts-graft	93.4785	+0.061	+0.167	-0.292
A1-router-graft	0.0407	- 0.031	+0.167	+0.400
A12-experts-18layers	46.7388	-0.031	-0.500	+0.215

band is 0.0700, *wider* than the 0.0650 endpoint separation it was meant to measure (the base scored 0.285/0.280/0.215 under `--batch-size auto`), and minerva_math floors at 0.000 for both parents outside harmony format and is reported as a dead channel.

For each substitution arm we report

$$R(\text{arm}) = \frac{\text{score}(\text{arm}) - \overline{\text{score}(\text{base})}}{\text{score}(\text{safeguard}) - \overline{\text{score}(\text{base})}}$$

— so $R(\text{arm})$ is the fraction of the paired base $\rightarrow$ safeguard delta that substituting that arm’s tensor class recovers, and $R(A1)$ is that fraction for the router-only arm — with $R_{\text{res}} = \text{band}/|D|$ beside it: ± 0.137 on MMLU, ± 0.500 on HellaSwag, ± 0.508 on IFEval. Every claim below is made on MMLU and checked against the others. Table 5 is the full ladder.

P-C2a-1 FALSIFIED. Router-only substitution recovers $R(A1) = -0.031$ on MMLU, $+0.167$ on HellaSwag, $+0.400$ on IFEval — **below 0.5 on every gated metric**, so the first branch of the registered falsifier fires. The second branch does not: $R(A3)$ ’s maximum is $+0.336 < +0.400$. But on the one metric where both are resolved, the ordering *reverses*: head $+0.336$ against router -0.031 .

P-C2a-3 FALSIFIED, and this is the transferable one. The registered claim was that the class our model-free localizer ranks most-moved is also the class whose substitution moves the most eval. The router is the localizer’s top-ranked class (mean_rel 7.375×10^{-3}) and moved 83.79 % of its elements against the MXFP4 experts’ 0.13 %. On eval it ranks **last of four**: attention $+0.565$, head $+0.336$, experts $+0.061$, router -0.031 . **The class the localizer ranked first of six is last of four on eval.**

P-C2a-2 HELD, and it is what makes the question well-posed. A1 and A9 are exact complements over all 687 tensors, and

$$R(A1) + R(A9) = 1.015 / 1.000 / 0.800$$

on MMLU / HellaSwag / IFEval. Perfect additivity gives 1 by construction; the departure measures interaction, and on two of three channels it is *at most* 1.5 % — MMLU’s departure is exactly that, and HellaSwag’s is zero. **Class-wise substitution on this pair is additive**, which is what turns “which class carries the post-train” from a metaphor into a question with an answer.

The answer. Of the base $\rightarrow$ safeguard capability delta, measured on the only channel with the resolution to carry the claim: **attention carries about 57 % at 2.93 % of the bytes**, the output head about 34 % at 1.78 %, the router about 0 % at 0.041 %, and the MXFP4 experts — 93.5 % of the checkpoint — carry $+0.061 \pm 0.137$, which is nothing. Two things survive for the router and both are real: its complement recovers 105 % of the delta, and per byte it remains the most efficient class in the checkpoint on IFEval at 9.83 R/B against attention’s 0.089 — roughly 110 $\times$. But that channel’s own resolution is ± 0.508 , so **it is reported as a direction, not a measurement**.

5.2.2 MXFP4 is a quantizer fixed point, and a blend of it measures rounding

Both parents are **fixed points of their own quantizer**: de-quantize then requantize is byte-identical under the preserve

policy *and* under the real pipeline’s recompute-up exponent derivation, over 3,583,180,800 blocks and 114,661,785,600 elements each, with `max_abs_err` exactly 0 on 72 of 72 tensors. The negative control (recompute-down) damaged all 72, moving 45 % of codes with a worst error of 2^{11} , so the nulls are informative rather than vacuous. P-MX1 and C2b-0a both HELD.

The consequence is stated exactly: the GPU “eval-invisibility” half of P-MX1 is **discharged by construction** — a byte-identical model needs no benchmark, at a line-item cost of \$0 — and **this licenses substitution and exponent re-derivation, not interpolation.**

P-C2b-Q FALSIFIED on its falsifier, vindicated on its mechanism. The registered falsifier was “three or more distinct outputs” from the α grid. We measured five distinct byte outputs on the interior grid and seven of seven overall. But the mechanism clause held: 91.54 % of moved elements lie one E2M1 code apart within exponent-matched blocks, and **nothing anywhere moved more than three codes.** And the byte-distinctness gate passes on its wording while failing on its substance — of the five interior grid points, **two are the parents in disguise** ($\alpha = 0.1$ reproduces the candidate exactly, $\alpha = 0.9$ the reference), leaving three genuinely interior models.

Worse than confounded: on this pair the requantization error *equals* the interpolation signal across the whole useful half of the grid, $\epsilon/\text{signal} = 1.00, 0.93, 0.88$ at $\alpha = 0.1, 0.3, 0.5$. **An MXFP4 expert blend is not a measuring instrument; it is rounding with an α -shaped label.** The amendment — require $\epsilon \ll \text{signal}$ on at least one interior α before any MXFP4 blend arm is built — and the fallback to graded *substitution* were both recorded in the runbook before the GPU stage, not after the result.

Beside that, the shape of a safety post-train seen through MXFP4: 95.869 % of expert blocks are byte-identical between the parents, no block moved its shared E8M0 exponent by more than one power of two, and 0.1322 % of elements changed value at all. **P-C2b-N HELD numerically** — the 72 routers are BF16, 26,551,296 B, 0.0407 % of the checkpoint, five of five α points distinct, zero non-finite — with the refusal branch of the blend path recorded as *untested* rather than inferred.

5.2.3 The boundary: lineage distance, not byte fraction

This is the single most transferable measurement in the campaign and it arrived as a surprise. From Table 4:

Arm	ρ	Δ gen.	Δ math
graft-attn-vocab-norm	0.337	+0.761713	+1.267118
blend-a0.50	0.500	+0.000569	+0.003349

A class-granular graft moving a **third** of the bytes all the way back across an **80-checkpoint gap** destroys the model —

2.2 $\times$ and 2.8 $\times$ worse than the rollback it was competing with — while a uniform blend moving *more* distance on *every* byte does not. The graft is correctly built: every selected tensor byte-identical to `ckpt_160`, every `mlp` tensor byte-identical to `spike_240`, nonfinite 0, 19 of 19 assertions passed, and the 483 substituted tensors decompose exactly as $320 + 1 + 1 + 161$. **So the failure is not a bug, it is the operation.**

A byte-weighted fraction does not predict damage. And this is what reconciles the result with Section 5.1: **the 70B graft succeeded because the damage was one tensor and the donor was one RLVR step away.** Same operation, opposite outcomes, and the difference is donor distance. That is the boundary condition for the whole substitution claim, and this is its first measurement. The inter-layer-consistency reading of *why* is an interpretation and was not tested; the house boundary is to localize, not explain.

Along the blend segment itself the curve is nearly flat with a shallow midpoint barrier: $\alpha = 0.25$ and $\alpha = 0.75$ are both *below* both endpoints on the math probe while $\alpha = 0.50$ is above, excursions of 0.003–0.005 bits/byte against bands of ~ 0.001 . No blend fabricated a single non-finite value across 723×5 tensors.

The negative space, and the baseline this subsection does not have. Three of the five surgery modes — layered blend, TIES and extrapolation (including negation, which is *extrapolate* at $\beta < 0$ rather than a sixth mode) — are shipped and **never exercised live.** The head-to-head of layered- α against single- α against whole-model- α on the 70B chain was designed and **decided against**, on three measured grounds: the 120B ladder already carries the granularity sub-claim with cleaner and falsification-hardened evidence; the 65B lane already supplies the blend-beats-graft contrast; and the separability risk is fatal, because three different 70B checkpoints score **0.8035 on MMLU to all printed digits** on the exact suite it would have gated on.

The consequence is that the whole-model- α baseline the Granularity sub-claim was originally registered against was never put to an instrument at any scale this paper headlines, and we state that plainly rather than let it be inferred. All eleven distinct arms of Table 5 are substitutions: there is no interpolation arm and no global- α arm on the 120B pair. Section 6’s repair ladder *did* contain one — a uniform $\alpha = 0.25$ re-merge over all 483 tensors, and P-A’-2 was registered against exactly that comparison — and it was built, byte-verified and UNSCORED AT THE GATE, because that lane’s own gate stopped it. The 70B head-to-head was decided against on the grounds just given. **What this subsection therefore delivers is an attribution result — that class-wise substitution is additive, and which class carries the delta — and not the claim that tensor selection beats a global coefficient.** That claim is lowered in Section 1 to match, and the missing arm is carried as owed work in Section 7. The pre-registered global- α baseline (P-C2-GA, receipt 8eaf457)

Table 6: Measured cost of the loop. Every row is a receipt from a lane, not an estimate, and each was reconciled against provider balance deltas.

Operation	Measured
localize + graft + verify a 70B checkpoint (CPU)	≈\$5.7
full gated suite at 70B, 6 runs incl. $N=3$ band	≈\$10.9
one suite run at 120B MoE, TP=1, one 96 GB card	≈\$0.38, ~11 min
seven-benchmark suite at 65B, 4 GPUs	\$9.05/arm
two bits-per-byte probes at 65B	\$1.80/arm
the whole K2 counterfactual audit (8 arms)	\$33.27, ~11 h

has since run. The whole-model $\alpha=0.5$ blend of the same pair recovers $R = +0.489$ on MMLU while moving 100 % of the checkpoint’s bytes; the attention-only graft recovers +0.565 from 2.93 % — a **39× per-byte gap** — and the full-complement graft sits 4.1 resolution-widths above the blend. The registered falsifier does not fire. Scope travels with the number: the companion prediction landed within 0.011 of the linear-mixture value ($R = 0.4885$ against a predicted 0.5), so the global coefficient is a near-linear dial on this pair, and the comparison is therefore pinned to *the coefficient production merging actually uses*, not to every α — a tuned-up coefficient would climb toward the 99.96 % graft at the price of moving essentially the whole checkpoint.

Two structures inside the ladder that resolve to nothing. Depth structure inside the router is unresolved (early +0.153, mid +0.176, late +0.107, all against ± 0.137) and **no depth claim is made**. The graded expert ladder is flat and non-monotone (6 layers +0.313, 18 layers −0.031, 36 layers +0.061): substituting six times more expert bytes does not move the eval six times more, or at all.

5.3 Efficiency

5.3.1 The substrate, and what surgery therefore costs

Stage deltas within a lineage are small and the ratios are measured, not modelled: Tulu-3-70B DPO → final is $45.03\times$ **loss-less whole-model** (141.1 GB → 3.13 GB); gpt-oss-120b’s own safeguard post-train is $64.55\times$ (65.2 GB → 1.01 GB); materialization runs at $\sim 1.6\text{GBs}^{-1}$ and is scale-invariant (65.2 GB reconstructed and verified in 41.0 s). *Every byte figure in this paragraph is decimal GB, taken from total_size in the checkpoint index; the same 120B checkpoint is 60.8 GiB if a reader prefers binary units, and mixing the two is how a $64.55\times$ becomes a $60\times$.* Surgery itself is CPU-only recombination of recorded bytes. **The loop’s cost is therefore the eval, full stop** — which is the same cost structure Cohere report for production merging (Section 2.2), and Table 6 is our side of it.

5.3.2 What it costs to rerun the stage instead

AI2 state the Tulu-3 training settings outright [20]: “the 70B DPO model is trained for 19 hours on 64 interconnected H100s” (\$5) and “the final 70B RL run takes 60 hours on 48 GPUs” (\$6). LLM360 state their incident cost outright [24]: “We used 480 A100s, and the entire pre-training process lasted approximately 100 days. ... We used an additional 30 days to handle loss spikes”, and both stated energies (430.8 MWh and 129.3 MWh) reconcile exactly against that cluster.

Table 7 is the comparison, and it is deliberately unfavourable to us in its denominator.

The headline is the first row, not the last. The DPO number is [PUBLISHED], modest, on this paper’s own experimental chain, and immune to “you cherry-picked a catastrophe”; and the ratio is *conservative*, because it charges the rerun nothing for the hyperparameter search that produced its settings. The K2 figure sits beside it as the upper end of what the class costs when it goes wrong, and it is **LLM360’s own arithmetic, not ours** — 30 % of their base run.

The honest headline is still not “we beat their restart”. Section 5.1.2 shows that at the restart point there was no measurable deficit to recover — and it also shows that the model the restart eventually produced is ahead of the branch it replaced on every metric either instrument could read. Both readings survive the same number. What does not depend on either is the one that matters: **a decision priced at \$411k–688k can now be checked for \$33 and eleven hours.**

Two frontier consolidation stages, and what we refuse to claim about them. Kimi K3’s MOPD cost is **not derivable** from its report [35]: no GPU count, no steps, no tokens, no wall time, and Figure 8’s FLOP axis is unlabelled. The only quotable form is a floor built on “a few hundred GPUs”: ≥ 600 GPU-days, i.e. $\geq 3.5\times$ the entire Tulu-3-70B DPO+RLVR bill, [ESTIMATED]. **And the label is not enough disclosure on its own, so we spell the gap out:** that quote is K3’s §5.3.1 statement about its *1M-context RL experiments* (“we adopt co-located RL training to keep each 1M-context Kimi K3 RL experiment within a few hundred GPUs”), not about MOPD. Carrying it across to the consolidation stage is *our* inference, it is the weakest link in the floor, and a reader who rejects the transfer should treat the K3 row as absent rather than as loose. Nemotron 3 Ultra’s teacher card carries an unredacted internal compute block (1.15e23 post-training FLOPs, 299,600 kWh) from which two derivations disagree by roughly $6\times$ — 35,000–222,000 GB200-hours, [DERIVED] — and **the disagreement is the honest answer**, because RL and MOPD are rollout-dominated and low-MFU. We report the range with both routes shown, and flag that the block may be removed before submission.

Table 7: Rerunning the damaged stage, against surgery plus one gated eval (\$25–50; the denominator is a published OLMES rental quote sourced 2026-08-26, and Table 6’s measured \$16.6 for that same operation on that same model makes every ratio here conservative). **The hardware × wall column is [PUBLISHED]; the GPU-hour and GPU-day columns are [DERIVED]**, being our own multiplication of those published inputs — neither product appears in either source paper. Dollar columns are 2026-08-26 rental quotes — what it would cost a reviewer to redo it — not what the original labs paid; both ran on owned or granted hardware. Ratios are computed from the unrounded quotes (\$2,420–\$3,636 and \$5,731–\$8,611), not from the rounded dollars printed here.

Stage to rerun	Hardware × wall [PUBLISHED]	GPU-hours [DERIVED]	GPU-days	\$	vs. surgery
Tulu-3-70B DPO	64× H100, 19 h	1,216	50.7	\$2.4k–3.6k	48–145×
Tulu-3-70B RLVR	48× H100, 60 h	2,880	120.0	\$5.7k–8.6k	115–344×
K2, both loss spikes, total	480× A100, 30 extra days	345,600	14,400	\$411k–688k	~10 ⁴ ×

5.3.3 Spend, waste included

The campaign cost ≈\$96.6 against a re-based \$200 ceiling: ≈\$30.0 for Sprint 1 (against a \$70–90 estimate), \$10.39 for the 120B lane (against ~\$40), \$33.27 for the 65B lane (against a \$42.94 worksheet and a \$70 cap), and \$22.93 for the consolidation lane (Section 6, against a \$46.32 worksheet and a \$60 cap). Also on the record, because reporting it is cheaper than being asked: \$0.20 of stray-pod waste from a capacity-hunting loop that parsed `create`’s stdout instead of listing pods afterwards, \$0.03 of orphaned volume, \$5.6 spent building an eval virtualenv on a \$8.36/h GPU pod that should have been built on a \$0.49/h CPU box, and — the largest single item in the campaign — **\$7.56 of idle GPU in Section 6** after a queue died unsupervised against a watchdog deadline sized for the work rather than for the work having stopped.

One methodological result about estimates. Every estimate written *before* the campaign came in 2.3×–29× over. The pattern breaks exactly where the estimate was built from receipts instead of from intuition: the 65B lane’s design priced itself from Sprint 1’s own measured unit costs, forecast “\$28–38 actual against \$43 planned”, and landed at \$33.27 — inside its own forecast, at 77 % of plan. **The padding is not a house style; it is what happens when nobody has yet measured the thing.**

The negative space, and it is permanent. The continuation half of this sub-claim — “where a loop still exists, the repaired hybrid continues training as well as a rollback does” — has **no evidence, and will not acquire any.** The only arm of this campaign with a training loop was removed (Section 6) and its two predictions withdrawn rather than scored. This is not a gap awaiting a future run: **surgery for deployment; this paper makes no continuation claim**, and every use of the efficiency result is scoped to deployment in exactly those words.

5.4 Agency

This subsection rests on two 0.5B lanes, and it says so in those words. No agency prediction has been registered and resolved at scale: the selection-versus-global- α matrix and the sentinel false-positive-rate prediction were both deferred, and neither is claimed here.

5.4.1 What the agent did, at 0.5B

In the surgery lane, the agent correlated the eval collapse with the damaged tensor class, took the localizer’s 24 attention globs verbatim rather than authoring its own, and **predicted the outcome before the run** — “perplexity falls back to about v4’s 6.578” — which landed at 6.533. In the blend lane it chose interpolation over substitution *with the reason stated*, picked $\alpha = 0.4$ from the metrics alone, measured the α -frontier under a three-preview budget, and **declined its third preview with a stated reason**. Both lanes ran the full loop including human approval and the deterministic executor, and both produced a registry version that cleared the gate.

5.4.2 The instrument the agent reads was falsified, and replaced

The agent’s localization input is drift telemetry (Section 3.1), and the sentinel’s anomaly rule is a threshold on it. We registered a prediction that this rule detects the first in-spike transition of a real 65B incident with zero lag while false-positiving on at most 2 % of ≥ 50 healthy transitions.

P-B1’ FALSIFIED, on both halves. The first in-spike transition is drift-*indistinguishable* from the matched healthy transition at the same steps: class ratios 0.99–1.15, pack ratio 1.6711 against 1.6749. The shipped rule “detects” it at 9.76 against a healthy range of 7.23–9.51 — **the spike sits one step above a contiguous healthy range, and the rule flags all seven transitions**. It is not discriminating; it is on. On a labelled set of 54 healthy transitions (50 OLMo, judged leave-one-out within site so the LR schedule is held fixed, plus 4 K2), it false-positived **14 of 54, 25.9 %**, and every false

Table 8: Sentinel redesign. **separation is the column that decides, not FP**: with one positive label, a false-positive count at a threshold chosen after seeing the positive is worthless. Four magnitude statistics, including a trend-corrected one, all leave the spike inside the healthy cloud. †: the rule in production when P-B1' was registered; its 25.9 % false-positive rate is what falsified it.

Rule	FP/54	Sep.	70B ctrl
shipped-max-vs-mean [†]	14	1.03×	flagged , 101.93×
max-vs-max	0	1.01×	clear
detrended-magnitude-z	0	0.84×	unscoreable
profile-deviation	0	2.88×	clear
median-ratio-deviation	0	4.03×	clear

positive is the same tensor class. The secondary mean-based rule reaches 0 of 54 but never fires on the spike either. **No reading of the shipped rule both catches the spike and clears the 2 % bar.**

The replacement was built, scored against the falsification, and shipped — at \$0 and no pods. Eight candidate rules were scored on the 54-transition labelled set the falsification produced, at a threshold computed from healthy transitions alone ($\tau = 2 \times$ worst healthy excursion), fixed before any spike number was looked at. Table 8 is the scoreboard.

The winner, median-ratio-deviation, measures how much harder each tensor class moved than the *median class* did, in log space — a size-factor normalization borrowed from RNA-seq. It is scale-free against the learning-rate schedule (OLMo's raw attention drift falls $2.4 \times$ over the run), it clears the 70B known-good control that the production rule flags at $101.93 \times$, and it **keeps the direction**: on the zero-lag transition the signed deviations are embed $+0.1303$, lm_head $+0.0642$, attn 0.0000 , mlp -0.0079 , norm -0.0217 — the two tied-vocabulary tensors moving harder than the model around them. It shipped behind an environment switch, with the old rule preserved row for row so the falsification's published numbers stay reproducible against the product.

The caveat is load-bearing and travels with every use of this result: there is one positive label. A separation of $4.03 \times$ at $n=1$ is an encouraging effect size and **not a detection rate**. No sensitivity, no ROC and no confidence interval is computable from it, and none is quoted. Worse for us: the two later in-spike transitions score 0.0091 and 0.0110, inside the healthy cloud. **The rule sees the onset, not the state.** The honest next experiment is more positives, not more statistics.

5.4.3 What has never been tested

The contamination protocol — anonymized pack and version names, dates and tokenizer hints stripped, reported beside the named condition — is designed, documented and **unexecuted**.

No agent lane was funded on the finished public runs, and that arm's work is model-free by construction, which is exactly why it was affordable. **Agency is therefore a 0.5B existence proof with a contamination mitigation that has never been tested**, and Section 7 repeats that in its own words.

The arm that was designed to change this was Section 6, and it did not. Its contamination condition is clean by *construction* — the artifact under repair is a consolidation *we* performed, so it post-dates every training cutoff — but its pre-registered spend gate stopped the lane before the agent was dispatched, so **no proposal, no action and no stated reasoning exists from it. This subsection remains the whole of the evidence for Agency**, unchanged in scope by the largest arm the campaign ran after it.

5.5 Cross-cutting: what the campaign measured that nobody asked for

5.5.1 Drift magnitude localizes where the optimizer wrote, not where the behaviour lives

Two independent lanes, on two model families, killed the same assumption. At 65B (Section 5.4) the first in-spike transition is drift-indistinguishable from a matched healthy one, class ratios 0.99–1.15. At 120B (Section 5.2) the router — the most-moved class by mean_rel, and the one that moved 83.79 % of its elements — ranks last of four on eval at $R = -0.031$.

One sentence carries both: **drift magnitude localizes where the optimizer wrote, not where the behaviour lives.** That is a result about a technique the merging literature uses implicitly, obtained from two falsified predictions of our own. It also bounds this paper: **no localization claim here is a claim about where capability lives.** The localizer says where the optimizer wrote, and Section 5.2's attribution — which *is* behavioural — had to be bought with fifteen suite runs to establish it.

5.5.2 Eval gates catch what loss curves miss

LLM360 released 360 checkpoints of Amber-7B *together with the evals to prove they were fine*. **P-B3 CONFIRMED: the trail contains five capability regressions past 3σ of a trailing-MAD floor that the loss curve does not show**, against a registered prediction of at least three, for \$0. They are named individually in Table 9.

Checkpoint 162 → 164 is the headline: two independent capabilities fall together — TruthfulQA at 8.96σ and Hel-laSwag at 3.32σ — while held-out Pile perplexity *improves* from 14.514 to 13.963. **The filter does real work**: 48 % of loss transitions show perplexity rising, so “loss improved or flat” is a genuinely $\sim 50\%$ -selective condition, and it rejected exactly half the candidate drops. Our reduction of the committed eval files matches LLM360's own published dashboard on **356 of 356 checkpoints for all four metrics**.

Table 9: The five silent regressions in Amber-7B’s own published eval trail. Values are percentage points on LLM360’s numbers; the noise floor is $\text{MAD} \times 1.4826$ over the ten preceding deltas. Δ loss negative means held-out Pile perplexity improved across the transition.

Transition	Metric	Drop	σ	Δ loss
162 $\rightarrow$ 164	TruthfulQA	1.932	8.96	−0.552
56 $\rightarrow$ 58	TruthfulQA	1.787	6.46	−0.853
172 $\rightarrow$ 174	MMLU	1.088	4.41	−0.683
162 $\rightarrow$ 164	HellaSwag	0.637	3.32	−0.552
30 $\rightarrow$ 32	TruthfulQA	2.160	3.15	−0.070

Sensitivity is on the record in both directions rather than in our favour. P-B3 holds for $k_\sigma \leq 4.0$ and **fails for trailing windows of 20 or more** — heteroscedasticity, stated rather than hidden. The alternate checkpoint-id alignment yields 8 rather than 5. And every finding survives detrending: the two weakest get *stronger*, not weaker.

What these five are not. They are *recorded* evals, not runs. Amber’s harness ran once per checkpoint with no repeats, so the noise floor is estimated from the checkpoint-to-checkpoint series itself, which conflates evaluation noise with genuine step-to-step movement. That makes the floor generous and the five findings conservative — and it means **none of the five is confirmed damage**. Confirming one means re-evaluating ckpt_162 and ckpt_164, a GPU lane nobody funded.

5.5.3 Three numbers we had wrong, corrected rather than quietly dropped

One. We had claimed the router is the most-moved class “on four independent MoE artifacts”. That phrasing does not survive audit and is retired: router dominance holds on gpt-oss-20b/120b (0.68 % and 0.74 % against 0.05 %) — **one lab, MXFP4-confounded** — directionally on DeepSeek under a different metric, and is **contradicted** on OLMoE (parity) and reversed on a third artifact. Section 5.2 then removed the behavioural reading of it entirely.

Two. Section 1 quotes AI2’s published MMLU of 83.1 for the Tulu-3-70B final model; our own re-measurement under the pinned stack reads 0.8035 on a 20-per-subject, 1140-item subsample. These are different instruments on the same model, not a discrepancy — but **83.1 and 80.35 never appear in one table without saying which harness produced which**.

Three. An earlier derivation priced the K2 incident at “ ~ 34 checkpoints lost $\Rightarrow \approx 100k$ A100-hours”. **Both are withdrawn:** no primary source gives a spike index range, and IFM/K2-Spike-1 carries 36 branches spanning indices 160 $\rightarrow$ 240 — an 80-chunk span, not 34. The published 345,600 A100-hours makes the derivation unnecessary, and

their ratio still says something worth keeping: against *both* spikes’ recompute ($2 \times 108,800$ A100-hours) the published total is $1.59\times$, so 37 % of a real incident’s cost is not token recompute at all; against *one* spike’s recompute it is $3.18\times$, i.e. 69 %. We quote the two-spike form, because that is the incident LLM360 priced.

6 Arm A’: consolidation performed under control

6.1 Why this arm

Section 1.2 names consolidation as damage class one, and every receipt this paper had for it was second-hand: AI2’s published stage table, Command A’s 1.8 % average drop [4], and the gpt-oss safeguard pair standing proxy for a consolidation event that is really a safety post-train. Arm A’ measures the trade on ingredients we hold, because we perform the collapse ourselves. Ground truth is exact — every byte the merge moved, we moved — and the merged artifact post-dates every training cutoff, so the arm is contamination-clean by construction rather than by the mitigation Section 5.4 records as never tested.

It replaces the campaign’s only arm with a training loop, removed because its damage class is the one Section 1 concedes, our instrument was falsified on that class at 65B, and a continuation result at 124M would not show that a repaired 70B hybrid trains on. That arm’s two predictions are **withdrawn, not scored** — their wording stands at 363b8f3 and binds there — and the cost is permanent: **surgery for deployment; the paper makes no continuation claim**.

6.2 Method

AI2’s BAR set [2, 29] publishes five domain specialists on one Olmo 2 7B lineage [32], the base they fork from, and the lab’s own router-trained consolidation of all five. We merged the sibling pair BAR-2x7B-Math $\oplus$ BAR-2x7B-Code — 11,626,795,008 parameters, 483 tensors — as a linear weight-space blend at $\alpha = 0.5$, **uniform over every tensor, one coefficient, no gate**: the production shape (Section 2.2), deliberately the crude one. All seven revisions were pinned by sha and re-resolved against the Hugging Face API [12] immediately before renting.

The merge is byte-verified, and the assertions verifying it were wrong twice before any money was spent. Twenty-three whole-model provenance assertions passed: 483/483 tensors moved, content hash distinct from both parents, *nonfinite* 0, the artifact marked synthetic by the registry rather than by a promise. Both corrections would have failed a *correct* merge. The design derived its anchor tolerance from real-valued arithmetic — an $\alpha = 0.5$ average sits half-way between its parents — in a regime where 90.9 % of

the relevant elements are within one bf16 ULP, so rounding returns the average *onto* a parent; we predicted $0.5\text{--}0.707\times$ the pair separation pre-spend and measured $0.625\times$, so the bound holds at a real margin of $1.90\times$, not the claimed $2.37\times$. The fixture written to test that assertion then used a flat 9×10^{-4} relative offset, which bf16 cannot represent at all.

The “frozen” anchor was not frozen, and a \$0 gate caught it before any rental. The design rested on the ingredients sharing a byte-identical anchor expert. A ranged-read probe over 272 MB of remote checkpoint — no pod — found they do not: BAR unfreezes all shared parameters during RLVR, so the merge surface went from the design’s 62.77 % to 100 %. The design was amended and re-registered before any prediction bound (2ec8ef9); the paid whole-model pass later reproduced the \$0 finding exactly.

The instrument and the gate. Four arms were scored at full split with per-item logging. Every claim below is a **paired** bootstrap over identical items with McNemar beside it; the $N=3$ run-to-run band, entire configuration including seed held fixed, came in at ≤ 0.0132 everywhere and twice at exactly 0.0000 — which Section 5.2’s rule treats as a trap, not a licence. The pre-registered gate **G3** required both that the ingredients separate on *our* instrument and that the merge fall below *both* past the band, before any repair arm could be built.

6.3 Results

The trade is real, large, and asymmetric. Against each ingredient on its own channels, paired: HumanEval+ -37.20 [$-45.12, -29.88$] and MBPP+ -52.12 [$-57.41, -46.56$] against Code; GSM8K $+18.73$ [$+15.69, +21.68$] and MATH-500 $+5.40$ [$+2.40, +8.60$] against Math. All four separate past the band, in *opposite directions*. One coefficient over every tensor routed code away while giving up nothing on MMLU or IFEval, and it moved both maths channels upward — *but abort A3 fired on exactly those two channels and bounds what the movement means; see below, and read no capability claim into it here or anywhere else*. **Consolidation damage need not be two-sided.** The registered reasoning — Cohere measure 1.8 %, these ingredients have no cross-domain regulariser, so the trade should be far larger — got the magnitude right and the sign structure wrong. **P-A’-1 FALSIFIED**, on its own words (“not one”), by 18.73 pp in the direction it excluded.

G3 therefore failed, and the machinery this paper argues for refused to spend \$23.80 on us. The gate’s second clause is the shape the repair ladder was built to repair, and that shape is absent, so the lane stopped where it was written to stop: **no repair arm was evaluated and no agent was dispatched**. Seven constructed, byte-verified, serve-ready arms

— the whole repair ladder plus the re-merge baseline — carry no benchmark number, among them a 32-tensor router graft at $p = 0.0000225$, **the smallest selection in this campaign’s history**, verified as moving the tensors it named and nothing else, and never run. **P-A’-2 — the central claim, registered against our own preferred outcome — is UNSCORED AT THE GATE, with P-A’-3 and P-A’-4. P-A’-5 is UNSCORED AT THE GATE because the patch agent reads Stage-2a gate history and was never dispatched: there is no proposal, no action and no stated reasoning, and this arm supplies the Agency sub-claim with nothing.** We state the symmetry rather than let a reader find it: **the gate that stopped us is the gate we propose**, run against our own budget.

Abort A3 fires, and it bounds the interesting half. BAR’s published ordering does not reproduce on our instrument for the math channels. Their card puts the Math expert 28.9 points ahead of the Code expert on Math; we measure it behind on GSM8K by -4.25 [$-7.20, -1.21$], separated, and behind BAR-2x7B-Base, the model it was fine-tuned from. Its HumanEval+ score is exactly 0.00 over 164 items with a measured band of 0.0000, again below that base; we record this as an unexplained anomaly and build nothing on it. The Code ordering reproduces. Three explanations are available — our chat-template and filter departure, an aggregate we do not run, or the artifact — and **none is adjudicated here**. The pre-registered consequence applies: **this lane’s Math-channel numbers license no claim about the Math ingredient’s capability**, the striking `merge > math` result included. G3’s stop needs only that the merge not fall below both, so it stands; the interpretation does not.

P-A’-6 FALSIFIED on its condition; its falsifier does not fire. The merge is above the floor control on both Math channels ($+15.01, +5.00$) and below it on both Code channels ($-4.88, -8.73$), all four separated, with `nonfinite = 0` across all 483 tensors as the condition’s second clause requires; the prediction demanded *both*, and so did the falsifier the other way. What that settles matters more than the verdict: we did **not** reproduce AI2’s catastrophic dense merge of these same domains — “Dense model merging after mid-training fails catastrophically (6.5 overall), producing a nearly non-functional model” [29], a sentence that is in the BAR paper and on none of the BAR cards. `merge-a50` is a functional model: it is above *both* ingredients on MMLU and IFEval, above the base floor everywhere except the two code channels, and it has lost exactly one specialization. It is nominally above both ingredients on the two maths channels too — *abort A3 bounds that and we do not count it*, which is why this sentence says two tasks and not four.

P-A’-7 HELD: merging is cheap, evaluating it is not. Cohere state that as a production fact; a paper quoting it should

Table 10: Arm A', 11.6B MoE. Four arms at full split, greedy, seed 1234, apply_chat_template on for every task and arm; item counts identical across arms (gsm8k 1,319, MATH-500 500, HumanEval+ 164, MBPP+ 378, MMLU 2,280, IFEval 541). The ingredients are *per-channel oracle ceilings*, not results; BAR-5x7B, the consolidated reference, was fetched and never evaluated and is therefore absent. merge-a50 is the uniform $\alpha = 0.5$ merge of the two. **Abort A3 fired on the two maths channels**, and §6.3 bounds every reading of them. Levels are not comparable to BAR's own published table; the deltas are paired.

Arm	Role	GSM8K	MATH-500	HumanEval+	MBPP+	MMLU	IFEval
BAR-2x7B-Math	oracle	30.63	4.60	0.00	30.16	60.22	69.50
BAR-2x7B-Code	oracle	34.87	6.80	39.02	70.63	62.24	68.76
BAR-2x7B-Base	floor	34.34	5.00	6.71	27.25	63.33	64.88
merge-a50	damage event	49.36	10.00	1.83	18.52	63.46	71.16

measure it. Construction, materialization and whole-model byte verification of every arm cost \$0.840 of CPU against \$18.04 of eval — a construction share of **4.4 %** against a registered 10 % bar, and still 7.4 % if the idle GPU below is stripped out. One line inside it travels with the claim: the most expensive CPU operation in the arm was not a merge but the whole-model byte verification, at 1,579 s, more than both merges combined. **“Merging is cheap” is true here; “verifying the merge is cheap” is not.**

Spend, waste included. The lane cost **\$22.93** against a \$46.32 worksheet and a \$60 cap, differenced against provider balance across a \$30 top-up and net of the account's own standing burn: \$1.28 for the whole CPU stage and \$21.86 for the GPU stage, each differenced over its own window and so not summing to the whole-lane figure exactly. **\$7.56 bought nothing.** A queue died on a GPU-memory leak we had documented and not fixed, the script correctly refused to continue, and nothing noticed for 3.6 hours because the watchdog deadline had been sized for the work rather than for the work having stopped. Both fixes are in the artifact rather than in prose, and the leak recurred and was caught on the first run after the patch.

6.4 What this arm does not establish

The scale is 11.6B, not the frontier. **The consolidation is ours**, so its coefficient was chosen by us to be the crude production shape rather than tuned by anyone with a product at stake — which makes the damage legible and also stops it standing in for a lab's own decision. Abort A3 bounds every Math-channel sentence above. **No repair arm, no coefficient search and no agent proposal exists in this lane at all:** P-A'-2's baseline was never located, the Granularity sub-claim gains nothing here, and Section 5.4's negative-space statement — Agency is a 0.5B existence proof with a contamination mitigation that has never been tested — **stands exactly as written.** BAR-5x7B, the lab's own router-trained consolidation of all five experts and the *consolidated reference* this arm would be measured against, was fetched and never evaluated, so Table 10 has a floor, two per-channel oracle ceilings, and no consolidated reference.

The repair question this arm was built to answer therefore remains open, for a stated reason rather than a budgetary one: it needs a pair whose trade is two-sided. One candidate sits in BAR's own set — the Tool-Use and Safety siblings are SFT-only, so no reinforcement stage has moved their shared parameters apart — and we name it as the next experiment without promising its outcome. What this arm did establish is smaller than what it was designed for and is not nothing: **consolidation damage is real, it can be one-sided, and it is detectable for about \$20 by the gate rather than by the repair.**

7 Discussion and limitations

7.1 The house boundary: localize, never explain

No activations are read anywhere in this paper. No routing is instrumented, and no behavioural-semantics claim is made about why any tensor class matters. The localizer reads weights; the gate reads evals; nothing reads the space between them. Synthetic artifacts — anything surgery produced — are marked as such in the registry and blocked from ordinary promotion. The agent proposes, a human approves, a deterministic executor acts, and the customer's own evals are the only judge of “repaired”.

7.2 What we do not claim

The operation is not claimed as novel. Post-hoc weight-space repair of a damaged checkpoint was published on 2026-08-29 [47]; whole-model post-alignment interpolation, in 2025 [11]. The claim is tensor granularity from weight-space telemetry, agent-proposed at 0.5B, gated, provenanced, and executed at 70B and above — and the paper says so in Section 2.1, before the results, not after them. **It is not “defended against the global- α ablation”** as a universal: the ablation has now been run once, at 120B at the production coefficient, and did not fire (Section 5.2.3); the α -sweep and any 70B instance remain owed, and an earlier draft of this paragraph claimed the defended form before any instance existed. The correction is here rather than in a footnote.

Substitution composes only near the donor. *Verbatim from the campaign’s own boundary list:* a class-granular graft across an 80-checkpoint gap destroys the model while a uniform blend at larger ρ does not (Section 5.2.3). We record *that*; the inter-layer-consistency reading is an interpretation and was not tested. No claim that surgery repairs loss spikes in general: one incident, one model, one scale, one failure class.

MXFP4 licenses substitution and exponent re-derivation, not interpolation. *Verbatim:* any α -sweep on packed MXFP4 must carry its own quantization-error accounting and report ϵ/signal ; on this pair no interior α satisfies the readability condition. The consequence for a reader who wants to apply this at the current frontier is direct: Kimi K3 applies quantization-aware training from SFT onward and ships MXFP4 weights, so a naive expert blend on a K3-class checkpoint would be measuring rounding.

Drift magnitude is not behavioural load. *Verbatim:* and therefore no localization claim in this paper is a claim about where capability lives. The localizer says where the optimizer wrote (Section 5.5).

Detection at two-checkpoint granularity remains open. *Verbatim, and it travels with every citation of the sentinel redesign:* there is one positive label — a single transition from a single documented incident in a single 65B run. A separation of $4.03 \times$ at $n=1$ is an encouraging effect size and **not a detection rate**. No sensitivity, no ROC and no confidence interval is computable from it, and none is quoted. The rule sees the onset, not the state: the two later in-spike transitions score inside the healthy cloud.

And the detection result is calibrated on the wrong domain for this paper’s damage classes. Every one of the 54 healthy transitions in the labelled set, and the one positive, is a *pretraining* transition — OLMo-2-7B pretraining checkpoints at ten LR-schedule sites and K2’s 65B pretraining chain. The three damage classes this paper is written around (Section 1.2) are all *post-training* events: a consolidation trade, an RL collapse, a checkpoint-saving fault. Nothing establishes that a rule calibrated on pretraining drift transfers to a merge or an RL stage, whose drift profiles differ by more than an order of magnitude in every class we measured. **The detection instrument and the repair claim are calibrated on different domains, and closing that is the honest next experiment** — more positives, from post-training incidents, not more statistics on the ones we have.

No continuation claim, permanently. Nothing in this paper establishes that a repaired hybrid *trains on* as well as a rollback does. The only arm with a training loop was removed,

its two predictions were withdrawn rather than scored, and the kill-test it would have answered — “hybrids may not train on” — is **retired untested** rather than left pending. A reviewer is entitled to treat this as the paper’s largest open question, and it is the honest reason the efficiency result is stated as *deployment* efficiency everywhere it appears. Establishing it would need a training loop at a scale this paper claims at, which was never in budget.

Agency at scale is unevidenced. The agent claims rest on two 0.5B lanes costing \$0.82 and \$0.22. The anonymized-versus-named diagnosis condition has never been run (Section 5.4).

The 65B hybrid claims rest on 2 of 9 registered gated metrics, and no run-to-run band was measured in that lane. The 70B lane’s 0.000/0.010 band is cited there as *prior*, never adopted as that lane’s number.

Amber’s five silent regressions are recorded evals, not re-runs. None is *confirmed* damage.

No claim that LLM360 decided wrongly. They had a loss curve; we have their checkpoints and two years.

No number is claimed for K3’s or Nemotron’s consolidation stage. K3’s MOPD cost is not derivable from its report and we say so rather than manufacture one; Nemotron’s rests on an unredacted internal block and two derivations that disagree by $6 \times$, so it ships as a range with both routes shown, or not at all.

Command A’s merge is not “ungated”. Its coefficients were searched against 500 eval metrics. The contrast this paper draws is granularity and proposer, not the presence of a gate, and the 1.8% average drop is *their* measurement of consolidation damage, quoted rather than rhetorically inflated.

The gpt-oss suite numbers are paired deltas, not benchmarks. Both endpoints are out of distribution in the same way (no harmony format), which is acceptable for a delta between arms cut from the same two parents and is **not** a gpt-oss benchmark number. None should be quoted as one.

7.3 Threats to validity a reviewer should press hardest

Scale. The strongest results are at 70B dense and 120B MoE; the frontier is at 550B and 2.8T. Our defence is mechanical rather than empirical: a consolidation trade at 70B and one at 2.8T are the same weight-space event, and Section 5.2’s additivity result is a property of the substitution

operator rather than of the model. That is an argument, not a measurement, and we mark it as one.

$n = 1$ on the interesting incidents. There is one public 65B loss spike, one public released-and-broken 70B checkpoint, and one gpt-oss safeguard pair. Every headline in [Section 5](#) is a single incident measured carefully rather than a population measured roughly, and the pre-registration protocol ([Section 4.1](#)) is what we offer in place of statistical power. It is not a substitute for it.

The granularity claim’s strongest counter-evidence is published, and it is ours too. Cohere found uniform embedding merging beat selective embedding merging ([Section 2.2](#)); we found a class-selective graft losing catastrophically to a uniform blend at larger ρ ([Section 5.2.3](#)). Two independent instances of selection losing. Our reading is that both are distance effects — Cohere’s selective merge discards a shared initialization, ours discards inter-layer consistency across 80 checkpoints — and that selection earns its keep *near the donor*, which is exactly where the 70B and 120B results sit. **That is a boundary, not a collapse**, and it is a better section than a claim that selection always wins.

7.4 Future work, costed rather than promised

The five Nemotron 3 Ultra MOPD teachers plus base plus final [31] are the first frontier-scale consolidation event with published ingredients, and they are the natural successor to this paper’s gpt-oss proxy. At ~ 1.1 TB per teacher in BF16 that is a multi-terabyte ingest, affordable only under the registry’s delta storage — which is itself the substrate claim, so the experiment and its enabling measurement are the same measurement. allenai/tmax-* (2B–27B, RL-step branches, Apache-2.0, with a published capability trade) [17] is the cheap version of the same experiment and should come first.

One baseline is owed and it is the one the Granularity sub-claim’s original falsifier named. [11]’s global- α before/after-alignment interpolation must be run as a *named baseline* rather than merely cited, and the same shape — one whole-model coefficient, scored on the same suite as [Table 5](#)’s eleven substitution arms — is what would license the strong form of the sub-claim that [Section 5.2.3](#) explicitly declines to make. One pre-registered arm of exactly that shape has now run on the gpt-oss-120b pair — the $\alpha=0.5$ blend, which the selective arms beat $39\times$ per byte ([Section 5.2.3](#), P-C2-GA) — so what remains owed is the sweep across α and an instance on a second pair, not the first data point. Command A §3.4.3.2’s selective-versus-uniform embedding finding is confronted in [Section 7.3](#) above, beside our own instance of the same failure, rather than omitted.

And the honest next detection experiment is not a better statistic. It is more positive labels, from post-training incidents.

8 Conclusion

The instability literature was written for runs that were still running. The damage that costs labs capability now happens after the loop closes: a consolidation stage trades one capability for another, an RL stage collapses a policy, a saving bug destroys a tensor in an artifact that then ships. In all three the time axis is gone and the tensor axis is all that is left — and the tensors are already on disk, with their lineage already recorded.

We asked whether operating on them works, pre-registered twenty-three predictions about it, and reported all twenty-three at the verdict their lanes recorded: nine held, six falsified, two not scoreable as registered, four unscored because a gate of our own stopped a lane of our own, for \$96.6. A released, broken 70B checkpoint was repaired to indistinguishability from the model its authors shipped. At 120B, class-wise substitution is additive, and about 57 % of a safety post-train’s capability cost lives in 2.93 % of the bytes — at a resolution of ± 14 pp — while the 93.5 % that is MXFP4 experts carries nothing.

The falsifications are the parts we would keep if we could keep only one thing. Drift magnitude localizes where the optimizer wrote, not where the behaviour lives — falsified twice, on two model families, against a technique the merging literature uses implicitly. Substitution composes only near the donor: a graft moving a third of the bytes across an 80-checkpoint gap destroys a 65B model while a blend moving more distance on every byte does not. Consolidation damage need not be two-sided: the one we performed ourselves routed code away — 37.2 pp and 52.1 pp against its code ingredient — and gave up nothing on MMLU or IFEval, and the gate we propose read that and refused to fund the repair we had designed. *What that merge did to the maths channels is bounded by a pre-registered abort and is not claimed here* ([Section 6](#)). And in the one public 65B loss-spike incident there was no measurable deficit at the restart point — on the authors’ own eval trail and on our two probes — so the human rollback was **unnecessary at the moment it was made**. It is not that the rollback bought nothing: the model it eventually produced, 30 days later, is ahead of the branch it discarded on all nine metrics we could read. It is that at the point of decision there was nothing to recover. That decision — part of an incident its authors price at 345,600 A100-hours — is now checkable for \$33.

That last number is the paper’s real argument. Not that surgery always beats a rerun, but that the question stopped being unaffordable to ask.

Acknowledgments

Every experiment in this paper ran on stage checkpoints published by AI2, LLM360/IFM, OpenAI, NVIDIA and Ant

Group. None of the artifacts we operated on is ours, and the paper does not exist without their release policies.

This work operates entirely on publicly released checkpoints. We are indebted to the teams who publish stage-by-stage — and in particular to those who publish their *failures*: allenai/Llama-3.1-Tulu-3-70B-broken and LLM360’s spike branches are two of the very few artifacts in the field that make post-hoc repair a measurable question rather than a hypothetical one. Releasing a checkpoint that is known to be broken is a scientific contribution, and it is one the field currently does not reward.

References

- [1] Allen Institute for AI. Llama-3.1-Tulu-3-70B-broken. Hugging Face model repository allenai/Llama-3.1-Tulu-3-70B-broken, 2024. “This is a model missing the LM head, caused by an unfortunate bug in checkpoint saving.” Pinned at revision c6fd2da664301ce6fb7135757aaee82e75463217.
- [2] Allen Institute for AI. BAR: a released expert set on one Olmo 2 7B lineage. Hugging Face model repositories under allenai/. Nine BAR-* repositories are published, seven of them BAR-2x7B-* (five domain experts plus two -SFT intermediates). Seven revisions were pinned for this paper: BAR-2x7B-Math@772eb58b9dbb, -Code@dc5c4003957a, -Base@80987e28715d, -Safety@020eb765c730, -Tool-Use@94fea3d085ba, BAR-7B@bee49db02284 and BAR-5x7B, 2026. Apache-2.0. Cited *only* for the published capability matrix on the BAR-5x7B card (Math 55.8/26.9, Code 22.1/50.4), which we verified on the card itself. The “6.5 overall / nearly non-functional” dense-merge result is in the BAR *paper* and is cited there (barpaper), not here. Pins re-resolved four times, the last immediately before renting.
- [3] Guhong Chen et al. EvoTrainer: Co-Evolving LLM Policies and Training Harnesses for Autonomous Agentic Reinforcement Learning. *arXiv preprint arXiv:2606.03108*, 2026. arXiv:2606.03108, published 2026-06-02; PDF fetched and text-extracted 2026-09-01.
- [4] Team Cohere et al. Command A: An Enterprise-Ready Large Language Model. *arXiv preprint arXiv:2504.00698*, 2025. arXiv:2504.00698, published 2025-04-01; PDF fetched and text-extracted 2026-09-01.
- [5] Ganqu Cui et al. The Entropy Mechanism of Reinforcement Learning for Reasoning Language Models. *arXiv preprint arXiv:2505.22617*, 2025. arXiv:2505.22617, published 2025-05-28; PDF fetched and text-extracted 2026-09-01.
- [6] DeepSeek-AI. DeepSeek-V4-Pro model card. Hugging Face model repository deepseek-ai/DeepSeek-V4-Pro, 2026. Card fetched raw 2026-09-01; repository created 2026-04-22, sha b5968e9190ef, one branch. RE-VERIFY: shas drift.
- [7] Yifan Ding et al. SAF-OPD: Stable Advantage Fusion for On-Policy Distillation. *arXiv preprint arXiv:2607.29209*, 2026. arXiv:2607.29209, published 2026-07-31; PDF fetched and text-extracted 2026-09-01.
- [8] Charles Goddard, Shamane Siriwardhana, Malikeh Ehghaghi, Luke Meyers, Vlad Karpukhin, Brian Benedict, Mark McQuade, and Jacob Solawetz. Arcee’s MergeKit: A Toolkit for Merging Large Language Models. *arXiv preprint arXiv:2403.13257*, 2024. arXiv:2403.13257, published 2024-03-20; PDF fetched and text-extracted 2026-09-01.
- [9] Aaron Grattafiori et al. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, 2024. arXiv:2407.21783, published 2024-07-31; PDF fetched and text-extracted 2026-09-01.
- [10] Wei Hao, Daniel Mendoza, Rafael da Silva, Deepak Narayanan, and Amar Phanishaye. MGIT: A Model Versioning and Management System. *arXiv preprint arXiv:2307.07507*, 2023. arXiv:2307.07507, published 2023-07-14; PDF fetched and text-extracted 2026-09-01.
- [11] Tiancheng Hu, Benjamin Minixhofer, and Nigel Collier. Navigating the Alignment-Calibration Trade-off: A Pareto-Superior Frontier via Model Merging. *arXiv preprint arXiv:2510.17426*, 2025. arXiv:2510.17426, published 2025-10-20; PDF fetched and text-extracted 2026-09-01.
- [12] Hugging Face. Hugging Face Hub HTTP API. /api/models, /api/models/{id}/refs; OpenAPI description at /.well-known/openapi.json, 2026. Every model-availability claim in this paper was measured against this API on 2026-09-01, paginated to exhaustion, never against a search summary. Note for a reproducer: /api/models/{id}/refs is documented, but the /api/models?filter= route the derivative census uses is *not* in the current OpenAPI description, though it serves. We record the route we exercised rather than the one that is documented.
- [13] Hugging Face TB Research. SmolLM3. Hugging Face model repository and accompanying blog post, 2025. Cited only for the existence of its model-soup stage.
- [14] Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing Models with Task Arithmetic. *arXiv preprint arXiv:2212.04089*, 2022. arXiv:2212.04089, published 2022-12-08; PDF fetched and text-extracted 2026-09-01.
- [15] Gabriel Ilharco, Mitchell Wortsman, Samir Yitzhak Gadre, Shuran Song, Hannaneh Hajishirzi, Simon Kornblith, Ali Farhadi, and Ludwig Schmidt. Patching open-vocabulary models by interpolating weights. *arXiv preprint arXiv:2208.05592*, 2022. arXiv:2208.05592, published 2022-08-10; PDF fetched and text-extracted 2026-09-01.
- [16] inclusionAI. Ling-3.0-flash-base-midtrain model card. Hugging Face model repository inclusionAI/Ling-3.0-flash-base-midtrain, 2026. Card fetched raw 2026-09-01; released 2026-08-11, MIT. Publishes pretrain → midtrain → WSM-merge base stages at two sizes.
- [17] Hamish Ivison, Junjie Oscar Yin, Rulin Shao, Teng Xiao, Nathan Lambert, and Hannaneh Hajishirzi. Tmax: A simple recipe for terminal agents. *arXiv preprint arXiv:2606.23321*, 2026. arXiv:2606.23321, published 2026-06-22; PDF fetched and text-extracted 2026-09-01.
- [18] Jean Kaddour. Stop Wasting My Time! Saving Days of ImageNet and BERT Training with Latest Weight Averaging. *arXiv preprint arXiv:2209.14981*, 2022. arXiv:2209.14981, published 2022-09-29; PDF fetched and text-extracted 2026-09-01.
- [19] Nikhil Kandpal et al. Git-Theta: A Git Extension for Collaborative Development of Machine Learning Models. *arXiv preprint arXiv:2306.04529*, 2023. arXiv:2306.04529, published 2023-06-07; PDF fetched and text-extracted 2026-09-01.
- [20] Nathan Lambert et al. Tulu 3: Pushing Frontiers in Open Language Model Post-Training. *arXiv preprint arXiv:2411.15124*, 2024. arXiv:2411.15124, published 2024-11-22; PDF fetched and text-extracted 2026-09-01.
- [21] Tingfeng Lan, Zirui Wang, Yunjia Zheng, Zhaoyuan Su, Juncheng Yang, and Yue Cheng. TStore: Rethinking AI Model Hub with Tensor-Centric Compression. *arXiv preprint arXiv:2604.17104*, 2026. arXiv:2604.17104, published 2026-04-18; PDF fetched and text-extracted 2026-09-01.
- [22] Bolian Li, Yifan Wang, Yi Ding, Anamika Lochab, Ananth Grama, and Ruqi Zhang. Addressing Performance Saturation for LLM RL via Precise Entropy Curve Control. *arXiv preprint arXiv:2604.26326*, 2026. arXiv:2604.26326, published 2026-04-29; PDF fetched and text-extracted 2026-09-01.
- [23] Zhanyu Liu, Qingguo Hu, Ante Wang, Chenqing Liu, Zhishang Xiang, Hui Li, Delai Qiu, and Jinsong Su. HEALing Entropy Collapse: Enhancing Exploration in Few-Shot RLVR via Hybrid-Domain Entropy Dynamics Alignment. *arXiv preprint arXiv:2604.17928*, 2026. arXiv:2604.17928, published 2026-04-20; PDF fetched and text-extracted 2026-09-01.

- [24] Zhengzhong Liu et al. LLM360 K2: Building a 65B 360-Open-Source Large Language Model from Scratch. *arXiv preprint arXiv:2501.07124*, 2025. arXiv:2501.07124, published 2025-01-13; PDF fetched and text-extracted 2026-09-01.
- [25] Jeffrey Jian Ma, Hengzhi Pei, Leonard Lausen, and George Karypis. Understanding Silent Data Corruption in LLM Training. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025. ACL Anthology 2025.acl-long.996.
- [26] Marin Community. Marin 8B Retrospective. Marin project documentation, “Marin 8B Retro”, 2025. Fetched 2026-09-02. Cited only for its own words on pretraining stability — training was “fairly stable with very few spikes”. The retrospective reports no rollback; it does not state a rollback count, and we do not read one into it.
- [27] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and Editing Factual Associations in GPT. *arXiv preprint arXiv:2202.05262*, 2022. arXiv:2202.05262, published 2022-02-10; PDF fetched and text-extracted 2026-09-01.
- [28] Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-Editing Memory in a Transformer. *arXiv preprint arXiv:2210.07229*, 2022. arXiv:2210.07229, published 2022-10-13; PDF fetched and text-extracted 2026-09-01.
- [29] Jacob Morrison, Sanjay Adhikesaven, Akshita Bhagia, Matei Zaharia, Noah A. Smith, and Sewon Min. Train Separately, Merge Together: Modular Post-Training with Mixture-of-Experts. *arXiv preprint arXiv:2604.18473*, 2026. arXiv:2604.18473, published 2026-04-20; the paper behind the allenai/BAR-* model repositories. Cited for the dense merge result: “Dense model merging after mid-training fails catastrophically (6.5 overall), producing a nearly non-functional model” — a sentence that appears in the paper and *not* on any BAR-* card.
- [30] NVIDIA. NVIDIA Nemotron 3 Ultra Technical Report. Technical report, NVIDIA, June 2026. Report dated 2026-06-09. PDF fetched and text-extracted 2026-09-01. RE-VERIFY BEFORE SUBMISSION: the compute figures this paper quotes come from an unredacted internal “Computational Load” block on the companion nvidia/NVIDIA-Nemotron-Labs-Teacher-STEM model card, which may be removed.
- [31] NVIDIA. NVIDIA Nemotron Labs Teacher models (STEM, Competition-Coding, Chat, Instruction-Following, General-Reasoning). Five Hugging Face model repositories under nvidia/NVIDIA-Nemotron-Labs-Teacher-*, 2026. All created 2026-08-14, licence OpenMDW-1.1; verified against the Hugging Face model API 2026-09-01. Pre-consolidation MOPD teachers of Nemotron 3 Ultra (550B-A55B). RE-VERIFY: shas drift.
- [32] Team OLMo et al. 2 OLMo 2 Furious. *arXiv preprint arXiv:2501.00656*, 2024. arXiv:2501.00656, published 2024-12-31; PDF fetched and text-extracted 2026-09-01.
- [33] OpenAI et al. gpt-oss-120b & gpt-oss-20b Model Card. *arXiv preprint arXiv:2508.10925*, 2025. arXiv:2508.10925, published 2025-08-08; PDF fetched and text-extracted 2026-09-01.
- [34] Zibo Shao et al. AgentPatch: Coarse-to-Fine Weak-Task Repair for Merging Agentic Multimodal Large Language Models. *arXiv preprint arXiv:2608.06699*, 2026. arXiv:2608.06699, published 2026-08-07; PDF fetched and text-extracted 2026-09-01.
- [35] Kimi Team et al. Kimi K3: Open Frontier Intelligence. *arXiv preprint arXiv:2607.24653*, 2026. arXiv:2607.24653, published 2026-07-27; PDF fetched and text-extracted 2026-09-01.
- [36] Changxin Tian et al. WSM: Decay-Free Learning Rate Schedule via Checkpoint Merging for LLM Pre-training. *arXiv preprint arXiv:2507.17634*, 2025. arXiv:2507.17634, published 2025-07-23; PDF fetched and text-extracted 2026-09-01.
- [37] Zirui Wang, Tingfeng Lan, Zhaoyuan Su, Juncheng Yang, and Yue Cheng. ZipLLM: Efficient LLM Storage via Model-Aware Synergistic Data Deduplication and Compression. *arXiv preprint arXiv:2505.06252*, 2025. arXiv:2505.06252, published 2025-04-30; PDF fetched and text-extracted 2026-09-01.
- [38] Mitchell Wortsman et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. *arXiv preprint arXiv:2203.05482*, 2022. arXiv:2203.05482, published 2022-03-10; PDF fetched and text-extracted 2026-09-01.
- [39] Huimin Xu, Shuai Zhao, Xiaobao Wu, and Anh Tuan Luu. Understanding and Preventing Entropy Collapse in RLVR with On-Policy Entropy Flow Optimization. *arXiv preprint arXiv:2605.11491*, 2026. arXiv:2605.11491, published 2026-05-12; PDF fetched and text-extracted 2026-09-01.
- [40] Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-Merging: Resolving Interference When Merging Models. *arXiv preprint arXiv:2306.01708*, 2023. arXiv:2306.01708, published 2023-06-02; PDF fetched and text-extracted 2026-09-01.
- [41] Xuewei Yang, Jiachen Yu, Jie Wu, Shaoning Sun, Junjie Wang, and Yujie Yang. Internalize the Temperature: On-Policy Self-Distillation as Policy Reheater for Reinforcement Learning. *arXiv preprint arXiv:2606.00755*, 2026. arXiv:2606.00755, published 2026-05-30; PDF fetched and text-extracted 2026-09-01.
- [42] Xiaozhe Yao, Qinghao Hu, and Ana Klimovic. DeltaZip: Efficient Serving of Multiple Full-Model-Tuned LLMs. *arXiv preprint arXiv:2312.05215*, 2023. arXiv:2312.05215, published 2023-12-08; PDF fetched and text-extracted 2026-09-01.
- [43] Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language Models are Super Mario: Absorbing Abilities from Homologous Models as a Free Lunch. *arXiv preprint arXiv:2311.03099*, 2023. arXiv:2311.03099, published 2023-11-06; PDF fetched and text-extracted 2026-09-01.
- [44] Aohan Zeng et al. GLM-130B: An Open Bilingual Pre-trained Model. *arXiv preprint arXiv:2210.02414*, 2022. arXiv:2210.02414, published 2022-10-05; PDF fetched and text-extracted 2026-09-01.
- [45] Susan Zhang et al. OPT: Open Pre-trained Transformer Language Models. *arXiv preprint arXiv:2205.01068*, 2022. arXiv:2205.01068, published 2022-05-02; PDF fetched and text-extracted 2026-09-01.
- [46] Wentao Zhang, Yang Young Lu, and Yuntian Deng. Interactive Training: Feedback-Driven Neural Network Optimization. *arXiv preprint arXiv:2510.02297*, 2025. arXiv:2510.02297, published 2025-10-02; PDF fetched and text-extracted 2026-09-01.
- [47] Qiancheng Zhou and Ruizhe Li. Locked at the Entrance, Open Inside: Where RLVR Narrows the Solution Space. *arXiv preprint arXiv:2608.29188*, 2026. arXiv:2608.29188, published 2026-08-29; PDF fetched and text-extracted 2026-09-01.

A Artifact and reproducibility

A.1 What is released

Every number in [Section 5](#) traces to a committed artifact. The release is organised so that a reader can re-derive a verdict from raw output without rerunning a GPU, and rerun a lane from its runbook if they want to.

Pins. Every checkpoint by revision sha, every tool by git commit of the repository it was built from, and hardware and wall time for every run.

Raw weight-space output. modelctl diff --json verbatim, never hand-edited, with its provenance block, for every diff in the paper.

Per-arm derivation records. Both parents’ content hashes, per-class tensor counts, per-tensor SHA-256 provenance samples, non-finite counts and assertion results for every surgery arm.

Eval suite JSONs with per-item rows. 7,683 paired items per arm across the 65B lane’s registered suite, of which 1,983 were purchased; 2,731 rows per run on the 120B ladder. The verdicts are scripts over these, and the scripts are released with them, so a reader can disagree with a scorer without re-renting a GPU.

Cost records. Balance and time TSVs, reconciled against provider `clientBalance` deltas.

A lane report per lane, recording every deviation from that lane’s own pre-registered design — including the ones that cost money.

Agent transcripts. The two live 0.5B lanes’ full transcripts: thoughts, tool calls, findings, as the daemon stored them.

Withdrawn from this list, and named so it is not looked for.

The failure-injection suite and the anonymization scripts for the contamination conditions were staged for the arm that [Section 6](#) replaced. That arm was removed, so neither ships. The anonymization scripts are not merely unshipped but *unnecessary* for the arm that replaced it, whose contamination condition is clean by construction.

[Section 6](#)’s consolidation script — the exact uniform merge, its derivation record, and the 23 byte-provenance assertions, so a reader can rebuild the damaged artifact from two published ingredients and check that our damage is the damage. The seven repair arms that were built, byte-verified and never evaluated ship as manifests and selection receipts, clearly marked as carrying no benchmark number.

Packaging. Everything above ships as one archive: the lane directories under `measurements/` verbatim (raw `--json` diffs, derivation records, per-item eval JSONs, cost TSVs, lane reports), the scorers of [Table 11](#) under `scripts/`, and the run-books. Nothing stays repository-only. Each row of [Table 11](#) reproduces by running that row’s scorer against that row’s lane directory — CPU-only, no checkpoint download, no GPU — and the scorers print the table’s numbers rather than a summary, so a reader can disagree with a scorer’s choices without re-renting anything. Weights are not redistributed: what ships is manifests, hashes, diffs, eval outputs and scripts.

Availability. The archive is published at <https://arzach.ai/artifact>. A DOI is minted at camera-ready and will supersede this URL as the citable reference; until then the archive carries a content hash so a reader can verify they have the same bytes this paper was written against.

A.2 Reproducing each result

Table 11: Result to artifact map: one row per table in [Section 5](#). Lane reports are relative to `measurements/`; scorers live under `scripts/`, except `gpu-cla-score.py` and `gpu-c2a-score.py`, which ship inside their own lane directory beside the JSONs they read. Each scorer re-derives that table’s numbers from the committed per-item output, CPU-only and without a checkpoint download.

Result	Lane report	Scorer
Table 3 , Table 2 P-C1	<code>sprint1-cla/LANE.md</code>	<code>gpu-cla-score.py</code>
Table 4	<code>sprint2-pb2/LANE.md</code>	<code>pb2_score.py</code>
Table 5	<code>sprint2-c2/LANE.md</code>	<code>gpu-c2a-score.py</code>
Table 8	<code>sprint1-detection/redesign-eval.md</code>	<code>sentinel_redesign.py</code>
Table 9	<code>sprint1-amber-mining.md</code>	<code>amber_mine.py</code>
Table 10	<code>sprint2-arma/LANE.md</code>	<code>arma_score.py</code>

A.3 Environment

The capability instrument is pinned: vLLM 0.11.0, torch 2.8.0+cu128, `lm_eval` 0.4.11, transformers 4.57.6, CPython 3.11.10. [Section 6](#) is the one lane that moves a pin — vLLM 0.11.1, scoped to that arm alone, because its architecture is not in 0.11.0’s model registry — and the house pin does not move with it. The weight-space instrument is `modelctl` 0.1.0 built from a named commit. Two environment facts cost us time and are recorded so they do not cost a reproducer any: on Blackwell-generation consumer-class cards the MXFP4 path falls through to a Marlin backend rather than the FlashInfer one, and multithreaded BLAS inside a single dot product collapses CPU-stage parallelism by a factor that only shows up under concurrency — pinning the thread-count environment variables to 1 tripled throughput.

A.4 Re-verification required before submission

Three of this paper’s external sources are volatile, and one was published by accident. None of these is optional.

1. **Every Hugging Face sha and branch count in [Section 1](#).** Branch counts drift and repositories are re-uploaded. The 1,241-derivative figure in particular is a moving target: our own census seven days earlier read 1,235.
2. **Whether the unredacted Computational Load block is still present** on `nvidia/NVIDIA-Nemotron-Labs-Teacher-STEM`. The 35,000–222,000 GB200-hour range in [Section 5.3](#)

rests on it. If it is gone, fall back to the Kimi K3 floor and say so.

3. **Whether [47] has acquired a v2 with a whole-model- α baseline.** This is the load-bearing one: [Section 2.1](#) claims the granularity gap is untested there, and a v2 that adds that baseline *closes the gap this paper is claiming*. Check it last, and check it again on the day of submission.

The remaining external facts in [Section 2](#) were verified 2026-08-26 and re-swept 2026-09-01; a third sweep before camera-ready is cheap.

A.5 Ethics and compute

Every experiment ran on rented cloud GPUs and CPUs for a total of $\approx \$96.6$. All checkpoints operated on are publicly released under permissive or research licences and none was redistributed; the artifacts we publish are manifests, diffs, eval outputs and scripts, not weights. No human subjects, no user data, no scraped corpora. The artifacts we produced that could be mistaken for released models — the repaired 70B hybrid, and the merged 11.6B consolidation of [Section 6](#) together with the repair arms built beside it — are marked synthetic in the registry by construction and none is published as a model.