VL-2026-04 JUN 2026 measured map

PROPOSED program (Part IV is not results)

toy / 1.5B-scale where stated

Verifier-Centric Capability Growth: A Measured Mechanism Map and a Research Program

How capability can be grown into a frozen pretrained model, where it stops, and a proposed portfolio for moving the boundary — with the proposals firewalled from the results.

Annalea Layton

Vext Labs, Inc. · alayton@tryvext.com

SCOPE · READ FIRST

Parts I–III report measured results; every number maps to a committed artifact. Part IV is a research program and must not be read as results. In an internal audit, 35 of 35 draft claims required correction before they survived; the surviving forms are what appear here.

PUBLIC SCOPE (vextlabs.ai): `measured map` + **PROPOSED** research program · toy / 1.5B-scale where stated · **Part IV is not results** · no unbacked Gate-as-product claim.

Reading contract (the firewall). This paper separates, explicitly and structurally, what we have **measured** from what we **propose**. **Parts I–III report measured results** — every number maps to a committed artifact (script + JSON) under `research/right_ai/` or the training logs, and every claim is stated in its honestly-corrected form (in an internal audit, **35/35** draft claims required correction before they survived; the surviving forms are what appear here). **Part IV is a research program** — a portfolio of mechanisms we have **not** yet validated; it is labeled PROPOSED throughout and must not be read as results. Where a result is toy-scale, 1.5B-scale, single-seed, or reconstructed, we say so inline. The negative results are not an embarrassment to be minimized; they are the contribution.

Abstract

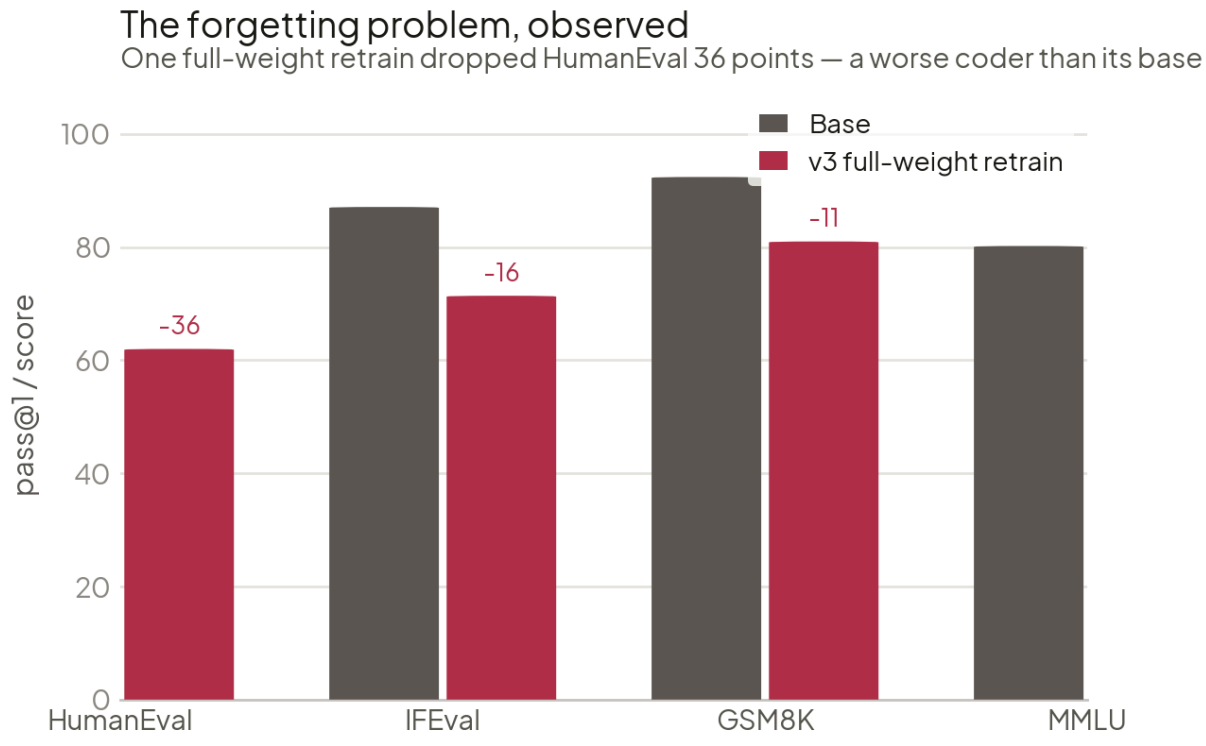
We present a measured map of how capability can be **grown** into a pretrained language model without retraining its weights, where that growth hits a computational ceiling, and a research program for the mechanisms that might move the ceiling. The paper has four parts. **(I) Training methodology (measured).** We describe the **Capability Injection Protocol** (CIP) — frozen-base depth-upscaling with low-rank adapters confined to newly inserted layers, gated by a non-regression benchmark — framed as the additive dual of **abliteration** ("reverse abliteration"). Instantiated on Qwen3-VL-32B-Instruct to produce Theron-Base v9 (41B, 80 layers) with 17 domain LoRAs, per-LoRA base regression is within ±1.2 pts (mean ≈ 0) across IFEval / HumanEval+ / GSM8K / MMLU, against a prior full-weight retrain that lost **36 HumanEval points**. **(II) Mechanism + ceiling (measured).** A small recurrent carry on a **position-free** substrate length-generalizes serial (NC^1) reasoning to 6× training length where fixed-depth (TC^0) readouts collapse to chance; on a real Qwen-2.5-1.5B the mechanism works on position-free **input embeddings** (0.996 @ L48) but **dies on deep hidden states (0.09)**, with **RoPE position-bleed** isolated as the precise blocker. **(III) Negative results + verifier frictions (measured).** A weight-tied "looped transformer" does **not** inherit the recurrent generalization; off-support "fuel" enlarges a flat-rule task but is **falsified** for serial reasoning and **null** at the 1.5B gold-standard enlargement test (0/30); a frozen embedding bridge transmits the LLM's **lookup table** but not its **semantic knowledge** (held-out number-words: Qwen 0.061 vs random 0.039); and a self-grading loop degrades 0.88→0.56 with same-model verification FPR ≈ 0.57. **(IV) Research program (proposed).** We lay out a portfolio of ~20 orthogonal, oracle-verified mechanism candidates (queries, register-programs, constraint-elimination, backward proof-search, latent planning, parallel voting, hierarchical DAGs) ranked for cheap kill-testing, plus a graveyard of falsified directions. The organizing thesis across all four parts is **verifier-centric**: a sound, different-execution-class oracle is the load-bearing component — for training acceptance (I), per-step certification (II), and mechanism verification (IV) — precisely because same-model self-verification is measured to be unreliable (III). The honest bottom line: **the mechanisms are real and the boundaries are measured at toy and 1.5B scale; a capability claim at frontier scale is not yet earned — and the paper is stronger for saying so.**

Part I — Training Methodology (Measured)

1. The forgetting problem, observed

The dominant way to make a domain expert is to continue-train a strong base on domain data. This works at small N and fails as N grows: **catastrophic forgetting** (McCloskey & Cohen, 1989) means each new specialty is partly financed by destroying earlier capability. We hit this in a directly observable form. An earlier model in our line — **v3**, a 72B cybersecurity specialist produced by full-weight continued SFT — gained domain skill and **dropped HumanEval pass@1 from 98% to 62%**. It became, literally, a worse coder than its base. That single observation reshaped our training program: we wanted a protocol in which prior capability is, **by construction**, untouched.

BENCHMARK	BASE	V3 (FULL-WEIGHT RETRAIN)	Δ
HumanEval	98.0	62.0	**−36.0**
IFEval	87.1	71.4	−15.7
GSM8K	92.4	81.0	−11.4
MMLU	80.2	76.8	−3.4



SOURCE: verifier-centric-capability-growth.md §1 (v3, 72B cyber specialist, full-weight continued SFT).

FIGURE 1 Full-weight continued training destroys prior capability while gaining domain skill: base vs v3 full-weight retrain on HumanEval (98→62), IFEval (87.1→71.4), GSM8K (92.4→81.0), and MMLU (80.2→76.8).

We do not claim **every** full-weight run regresses by 36 points; we claim the failure mode is on the table whenever the optimizer is permitted to move base parameters.

2. The Capability Injection Protocol (CIP)

CIP is a four-step composition of known ingredients — depth upscaling, parameter freezing, low-rank adaptation, and benchmark-gated acceptance — with explicit invariants.

1. **Merge best prior adapter** (bf16) into the base, anchoring the strongest pre-existing capability before the protocol begins. (Skipped for the v9 build; the input base was used unmodified.)
2. **Organic upscale: +N zero-init residual layers.** Insert N transformer blocks across the tower, each wrapped as `f̃(h) = h + γ·(f(h) − h)` with gate `γ⁽⁰⁾ = 0`, so the upscaled model is **bitwise identical** to the base at t=0. For v9, N=16 (64→80 layers, 32B→41B).
3. **Freeze old layers; LoRA-only on new layers.** Every pre-existing parameter is `requires_grad=False` with no optimizer state; trainable set = LoRA factors on the new layers + the scalar gates.
4. **Benchmark gate (non-negotiable).** Accept a checkpoint only if the merged form is non-regressive: IFEval ≥ 85%, HumanEval+ ≥ 90%, GSM8K within −1 pt.

Three invariants. (I) **Frozen base** — base parameters cannot move, so catastrophic forgetting of base capability cannot occur in the strict sense. (II) **Identity initialization** — the model cannot be worse than its base at t=0. (III) **Acceptance gate** — catches degenerate-contribution failures the structural invariants do not rule out.

3. Reverse abliteration: the additive dual

Abliteration (Arditi et al., 2024) **removes** a behavior by projecting a learned "refusal direction" out of weight matrices in closed form: $W' = W - \hat{r}\,\hat{r}^{\top} W$. It is geometric, subtractive, and base-preserving. CIP is the **additive dual** — it **adds** capability by inserting identity-initialized structure and confining learning there. Both rest on one discipline: **do not retrain what already works.** To modify an open-weight base, either **subtract** a direction or **add** structure; do not push gradients through the bulk tensor. The v9 build uses **both** in sequence: CIP grow-and-train (additive), then Arditi-style abliteration (subtractive) to remove inherited refusal of authorized-use security queries.

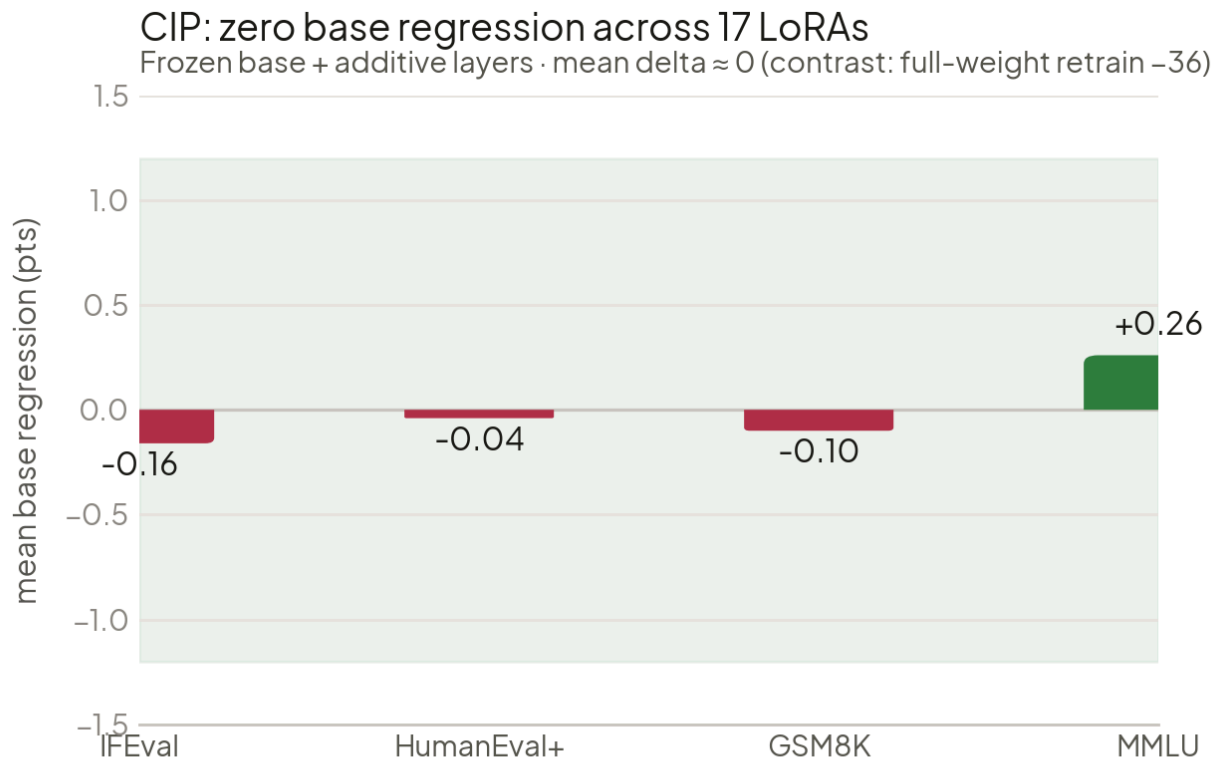
ASPECT	ABLITERATION	CIP
Polarity	Subtractive	Additive
Mechanism	Projection along a learned direction	Gradient descent on an isolated subset
Form of $\Theta - \theta_0$	Rank-one per affected layer	Zero on base layers; low-rank + gates on new
Trainable params	None (closed form)	LoRA (B,A) on new layers + gates γ

The label **reverse abliteration** is a research-direction frame, not a claim of closed-form equivalence (Part III, §11.3 of the limits discussion in the source methodology paper).

4. Results: zero base regression across 17 LoRAs

On the merged + abliterated v9 base we trained 17 domain LoRAs (cyber, code, math, reasoning, medical, legal, finance, business, science, engineering, creative, education, humanities, language, reconciler, inquisitor, long_context) via DPO on primary-source preference pairs (~2.94M pairs total; no teacher distillation). Each LoRA was evaluated by **merging** into a v9 checkpoint copy and measuring the delta from the unmerged-v9 baseline.

- **Per-LoRA regression (mean across the fleet):** IFEval **−0.16**, HumanEval+ **−0.04**, GSM8K **−0.10**, MMLU **+0.26**; **all deltas within ±1.2 pts** across all four benchmarks and all 17 LoRAs. We read this as zero regression in the protocol's stated sense.
- **DPO convergence (mean):** reward-accuracy 0.83 at step 1000; cyber strongest (0.94 @ 746 steps, densest primary-source corpus); creative/humanities weakest (0.75–0.77, honestly bounded by weaker preference signal).
- **Cost:** the full v9 pipeline (CIP grow-and-train 3.5 h + merge + abliterate 0.5 h + 17 sequential LoRAs 58 h) ≈ **65 GPU-hours on one 4×H200 SXM pod ≈ $700**; per-cluster marginal cost ≈ 4 GPU-hours.



SOURCE: verifier-centric-capability-growth.md §4 (v9 base, 17 domain LoRAs, evaluated by merge).

FIGURE 2 Mean per-LoRA base regression across 17 LoRAs, all within the ±1.2 pt band: IFEval −0.16, HumanEval+ −0.04, GSM8K −0.10, MMLU +0.26.

Scope honesty. This is **per-LoRA, evaluated by merge.** We do **not** claim composed multi-LoRA inference is non-regressive; a benign {cyber+code} composition is shown (within 0.5 pt of the better single-LoRA on SecQA and HumanEval+), but the full 17×17 interference matrix is open (Part IV). CIP adds **modules**; it does not, on its own, **bind** them into one integrated reasoner.

Part II — Mechanism and Ceiling (Measured)

5. The computational-class question

A single transformer forward pass is in **TC^0** (Merrill & Sabharwal, 2024). Serial state-tracking — the word problem of the symmetric group S_5; cumulative modular sums — is **NC^1**-complete and provably outside TC^0 for unbounded length. Chain-of-thought escapes this by externalizing steps into tokens. We ask the complementary question: can a **recurrent module** attached to a model internalize serial reasoning the model lacks in one pass — and does it do so by harnessing the model's **knowledge** or merely its **embeddings as a lookup table**? This is the exact "computational-class ceiling →

representational shift required" event that recent position work (DeepMind, **From AGI to ASI**, arXiv:2606.12683) names as a canonical post-AGI pathway; our results are a concrete, measured instance, **not** evidence that any ceiling has been cleared.

6. Recurrent carry vs fixed depth (toy + cross-model)

A small recurrent carry trained with per-step supervision length-generalizes three serial tasks to **6× training length** where a fixed-depth (TC^0) readout collapses to chance:

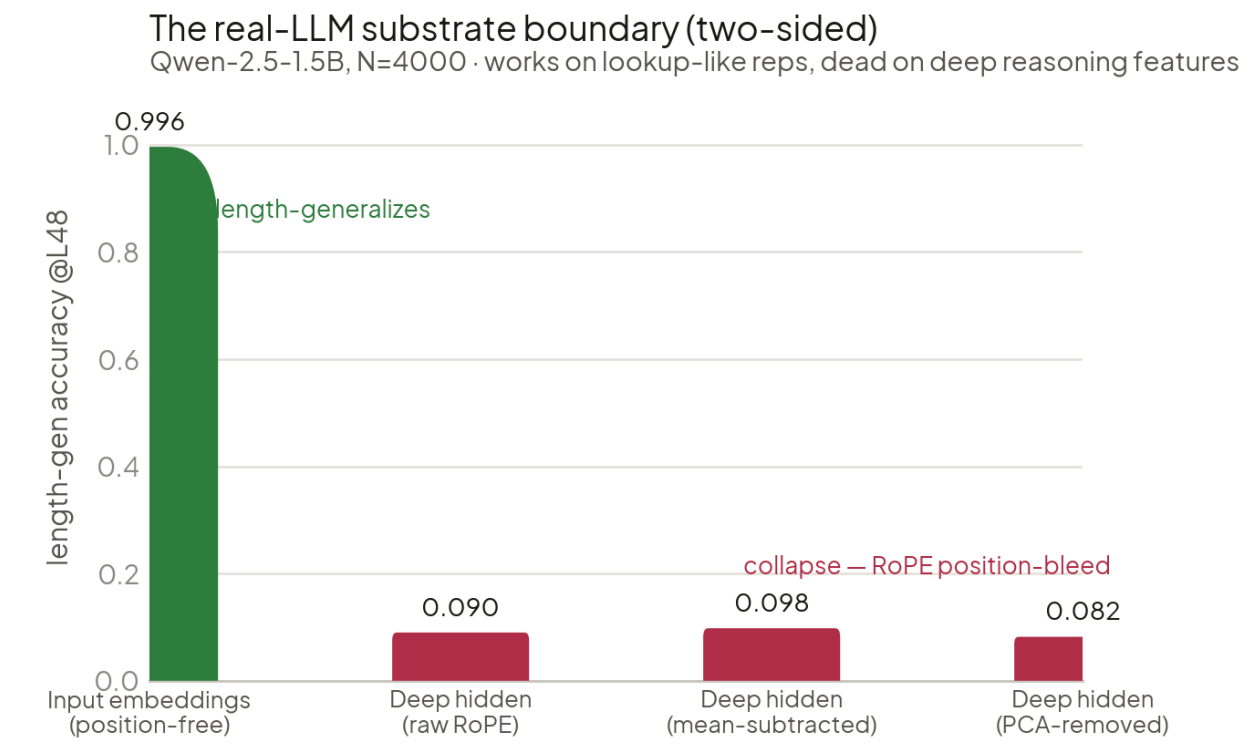
TASK	RECURRENT CARRY @L48	FIXED-DEPTH @L48
Cumulative cipher (GRU, from-scratch)	**0.98–0.99**	0.08–0.10
Cumulative cipher (GPT-2, cross-model)	**0.85**	0.10
S_5 word problem	**B = 1.000**	(collapses)

This reproduces the predicted TC^0 wall (fixed-depth arms 0.08–0.18 @L48 across tasks, consistent with Merrill–Sabharwal) and shows a recurrent carry escapes it on a position-free substrate. **Bound: from-scratch / GPT-2, synthetic serial tasks.**

7. The real-LLM substrate boundary (the novel core)

We then ask whether the mechanism transfers to a **real pretrained LLM** (Qwen-2.5–1.5B). With the gap battery closed on real hardware (RunPod, auto-cleanup; raw artifacts in `research/right_ai/method_runs/`), the result is decisive and **two-sided**:

SUBSTRATE	LENGTH-GEN @L48	VERDICT
Position-free **input embeddings** (`emb`)	**0.996**	length-generalizes
Deep hidden states, raw (RoPE)	**0.09**	collapses
Deep hidden states, mean-subtracted (debias)	0.098	insufficient
Deep hidden states, PCA-removed	0.082	insufficient



SOURCE: verifier-centric-capability-growth.md §7; ocrd_nope.json (position_bleed_confirmed_and_fixed=true).

FIGURE 3 Length-generalization @L48 by substrate on Qwen-2.5-1.5B: input embeddings 0.996 versus deep hidden raw 0.09, mean-subtracted 0.098, and PCA-removed 0.082 — RoPE position-bleed isolated as the blocker (N=4000, `ocrd_nope.json`).

`ocrd_nope.json` (N=4000): `position_bleed_confirmed_and_fixed = true`. **RoPE position-bleed** in the deep features is the precise blocker; neither mean-subtraction nor PCA removes enough of it. So the mechanism length-generalizes a serial NC^1 algorithm on a **real LLM's position-free representations**, reproduced at N=4000 — a main-track-grade positive — **but only on representations that behave like a vocabulary substrate.**

Part III — Negative Results and Verifier Frictions (Measured)

These are stated prominently because they are the credibility of the work.

8. What is falsified or null

1. **"Looped transformer" inherits recurrent generalization — FALSIFIED.** A NoPE weight-tied looped transformer scores 0.12 @L48 (collapses like fixed-depth). Weight-tying a forward block is not the same as a trained recurrent carry.

2. **Off-support "fuel" enlarges serial reasoning — FALSIFIED.** SGDF-R: A(long fuel) = 0.199 vs B(short-first elicitation) = 1.000; short-first elicitation already saturates, so the "fuel" adds nothing for serial tasks. (Flat-rule enlargement **does** hold in a toy: cipher oracle-depth delta ≈ 0.467 — see artifact caveat below.)
3. **SGDF enlargement at the 1.5B gold standard — NULL.** `sgdf_keystone.json`: base off-support confirmed (held_p1 = 0.000, easy = 0.633); **all arms enlarged 0/30**; `sgdf_enlarges = FALSE`, `non_destructive = true`. Even the toy flat-rule enlargement did **not** replicate on real 1.5B-Instruct. Frozen-base LoRA **consolidates**; it does not **enlarge** (this vindicates the CIP principle of Part I and bounds it: CIP is for non-regressive capability **injection**, not for pushing past the base's support).
4. **Frozen-embedding bridge transmits knowledge — NULL (a new, clean negative).** `llm_knowledge_harness_2026-06-20.json`: train the cell to sum number-**words** ("one"…"five"), test on **held-out** words ("six"…"nine"). Both arms learn train perfectly (1.000); on held-out words **Qwen embeddings 0.061 barely beat random 0.039** (gap 0.021, far below the 0.20 "harnesses-knowledge" bar). **The cell exploits embeddings as a lookup table, not as harness-able semantic knowledge** — converting our most important caveat from assertion to measurement.
5. **Factored growth vs naive warm-start.** Factored, frozen-block growth preserves capability where naive warm-start exhibits **negative transfer** (OGM: WARM 0.59→0.28 vs FACTORED preserved); a shared carry across lanes induces ~99pp forgetting. **(Toy.)**

9. Verifier frictions — why the oracle must be external

The single most consequential measured finding for architecture: **same-model self-verification is unreliable.**

- A self-grading recursive-improvement loop **degrades 0.88 → 0.56**.
- Same-model self-verification **false-positive rate ≈ 0.571** (worse than a coin on the thing that matters).
- A **sound, different-execution-class** oracle has ε_fp = 0.0 on GSM8K — but is **void on prose** (no sound checker exists for open-ended natural-language claims).

This is why every mechanism in this paper — training acceptance (Part I §2 step 4), per-step certification (Part II §6), and the Part IV portfolio — routes verification through an external oracle, **never** through the model judging itself. It also bounds the program: the verifier-centric approach is strong exactly on the **formalizable** slice and silent off it.

10. Honest bounds (do not overclaim)

- Almost everything in Parts II–III is **toy/CPU or 1.5B-scale, synthetic tasks** — not frontier scale, not real reasoning tasks.
- The real-LLM win (`emb` 0.996) uses **position-free input embeddings ≈ vocabulary lookup**; it proves the mechanism on a real model's **embeddings**, **NOT** that the cell harnesses the LLM's **deep reasoning**. Deep RoPE features remain dead (0.09). `genuinely_new_ai at LLM scale = NOT YET.`

- **Artifact caveat:** the toy flat-rule enlargement delta (≈0.467) and some early `ocrd_killtest` values were once reconstructed; the 2026-06-20 battery regenerated raw JSON for the core claims (`ocrd_killtest` A@L48=0.99 vs B=0.08, `ocrd_nope`, `sgdf_keystone`, `sgdf_r_qwen_gpu`). Any number without a surviving JSON is flagged at use and should be re-run before formal submission.
- Several oracle / self-verification results are **n = 10–23**; some G1 GPU runs were single-seed and were subsequently multi-seeded (the per-step-oracle robustness result was **corrected** by multi-seeding — A>C across 3 seeds at M=10, reversing an earlier outlier).

Part IV — A Research Program (PROPOSED — not results)

> Everything in Part IV is a **proposal**: a portfolio of mechanisms we have designed and ranked for cheap kill-testing but have **not** validated. None is a measured capability claim. The value here is a disciplined, falsifiable map of what to try and why — and an equally explicit map of what we have already ruled out.

11. The verifier-centric thesis, generalized

The measured spine (Parts I–III) generalizes to a design stance: **the oracle is the hub.** Across training, inference, and search, a sound external verifier is the load-bearing component, and the model's job is to **propose** into a space the verifier can **certify**. The program asks: which proposal mechanisms, beyond next-token generation, can a frozen base drive into an oracle-certifiable space?

12. The candidate portfolio (ranked, $0-kill-test-first)

We enumerate ~20 mechanically **orthogonal** candidates — no two share an execution primitive, so they compose rather than compete. The four "pillars" (Tier 1, ~$0, <1 h, designed kill-tests exist) are:

PILLAR	EXECUTION MODEL	ORACLE ROLE	CAPABILITY TARGET
IAMB	Query–response loops (queries, not tokens)	gates retrieval	multi-hop lookup
TRAM	Token→register-program→symbolic execution	verifies execution	arithmetic
ASEC	Constraint **elimination** + resampling	proves infeasibility	CSP
SPAS	Backward proof search	verifies decompositions	proof reasoning

Amplification (Tier 2, $5–10 on a frozen 1.5B): **VDYM** (uncertainty-gated latent planning), **PIRM** (oracle-trained parallel voting). Reasoning pair: **NEPS** (hierarchical DAG / topological execution), **SCAT** (sparse concept activation). Refinement pair: **DYER** (bidirectional convergence), **REIF** (embedding interpolation + oracle ranking). Each has a one-line explicit scope (arithmetic, CSP, proof, lookup, planning, …); none claims to solve everything. The proposed orchestrator **tries several in parallel and keeps the oracle-verified winner**, emitting a decision trace — it is not a routing gate.

13. The graveyard (already falsified — do not retry)

A research program is only as honest as its discard pile:

- **Oracle as router/selector (VRLA)** — measured **worse than random**. The oracle verifies; it must not select.
- **Oracle injection into the recurrent carry (VRLA, RCWM)** — destabilizes length-generalization.
- **Learned graders without a sound oracle (the "ratchet")** — self-deceive (the 0.88→0.56 collapse of §9).
- **Generic bolt-on composition** (MoE + world-model + multimodal stacked without mechanical orthogonality) — no measured lift.
- **Frozen-base LoRA for enlargement** — consolidates only (Part I/III); use off-support methods for enlargement, and note even those were null at 1.5B (§8.3).

14. What would move the ceiling (the make-or-break experiments)

In cheapest-first order, the experiments that would convert this from a **boundary-map** into a **capability** paper:

1. **Re-run any remaining reconstructed artifacts** (~$0–1) — remove the only reproducibility risk.
2. **Recurrent head on a real LLM's unfrozen / non-position-bleeding deep features at a real downstream task** — the §7 boundary says frozen RoPE features are dead; the open question is whether **unfrozen** features (or a relative-position / NoPE-from-scratch base) transmit deep knowledge. A clean measured outcome here — positive **or** null — is the difference between a workshop preprint and a main-track result.
3. **Kill-test the four pillars** (IAMB/TRAM/ASEC/SPAS, ~$0) — convert Part IV's first row from proposed to measured.

15. Conclusion

We have mapped, with measured evidence, a coherent picture of capability growth on pretrained language models: it can be **injected** non-destructively (CIP / reverse abliteration, Part I), it can **escape the TC^0 wall** via a recurrent carry on a position-free substrate (Part II), and it **stops** at a precise, reproduced boundary — the LLM's deep RoPE features do not transmit harness-able knowledge, and the model cannot reliably verify itself (Part III). The unifying lesson is **verifier-centric**: a sound external oracle is the load-bearing component, and the honest frontier is the **formalizable** one. Part IV turns that lesson into a falsifiable research program rather than a promise.

The honest statement of where we are: **the mechanisms are real and the boundaries are measured at toy and 1.5B scale; the claim that any of this harnesses a frontier LLM's deep reasoning is not yet earned.** The paper is stronger for saying so — its negative results and measured boundaries are its contribution, and its research program is disciplined by the same oracle that bounds its claims.

References

Arditi, A., Obeso, O., Syed, A., Paleka, D., Rimsky, N., Gurnee, W., & Nanda, N. (2024). Refusal in language models is mediated by a single direction. `arXiv:2406.11717`.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., et al. (2022). Constitutional AI: Harmlessness from AI feedback. `arXiv:2212.08073`.

Cobbe, K., Kosaraju, V., Bavarian, M., et al. (2021). Training verifiers to solve math word problems. `arXiv:2110.14168`.

DeepMind. (2026). From AGI to ASI: pathways and open problems. `arXiv:2606.12683`.

Failspy. (2024). Refusal direction ablation: reference implementation of Arditi et al. **GitHub repository**.

Hendrycks, D., Burns, C., Basart, S., et al. (2020). Measuring massive multitask language understanding. `arXiv:2009.03300`.

Hu, E. J., Shen, Y., Wallis, P., et al. (2021). LoRA: low-rank adaptation of large language models. `arXiv:2106.09685`.

Kim, D., Park, C., Kim, S., et al. (2023). SOLAR 10.7B: scaling LLMs with simple yet effective depth up-scaling. `arXiv:2312.15166`.

Liu, J., Xia, C. S., Wang, Y., & Zhang, L. (2023). Is your code generated by ChatGPT really correct? Rigorous evaluation of LLMs for code generation. **NeurIPS 2023**.

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks. **Psychology of Learning and Motivation**, 24, 109–165.

Merrill, W., & Sabharwal, A. (2024). The expressive power of transformers with chain of thought. **ICLR 2024** (and related circuit-complexity results, TC^0 characterization).

Sheng, Y., Cao, S., Li, D., et al. (2023). S-LoRA: serving thousands of concurrent LoRA adapters. `arXiv:2311.03285`.

Wang, X., et al. (2024). Block expansion for LLM capacity growth. `arXiv:2401.02415`.

Yue, Y., et al. (2025). Does RLVR enlarge reasoning capability beyond the base model? (elicits-not-enlarges). **Nature 2025 / arXiv**.

Zhou, J., Lu, T., Mishra, S., et al. (2023). Instruction-following evaluation for large language models. `arXiv:2311.07911`.

Correspondence: Annalea Layton, Vext Labs, Inc., `alayton@tryvext.com`.

Provenance & artifacts: Part I from the CIP methodology paper (`docs/papers/reverse_abliteration_cip.md`); Parts II–III from measured runs under `research/right_ai/` and `research/right_ai/method_runs/` (SHAs in each JSON), distilled in `docs/research/THERON_RESEARCH_PAPER_ASSESSMENT_2026-06-20.md`; Part IV from the Rounds 1–5 ideation corpus (`docs/research/ROUND*_2026-06-22.md`, `SYNTHESIS_BEST_LEVERS_RANKED_2026-06-22.md`). Measured numbers are stated in their audited/corrected form; reconstructed numbers are flagged at use.

Author contributions: A.L. designed the protocols and the research program, supervised the training and the measurement battery, and wrote the paper.
