Transformers Converge to Invariant Algorithmic Cores

Joshua S. Schiffman New York Genome Center · jschiffman@nygenome.org

Mechanistic Interpretability Workshop · ICML 2026 · Seoul, South Korea

github.com/joshseth/cores

1 The Problem: There's More Than One Way to Wire a Network

- Training constrains behavior, not circuitry: independently trained models solve identical tasks with near-orthogonal weights — so which mechanistic explanation is "real"?
- **Proposal: study equivalence classes — the invariants shared across realizations — rather than any single trained network.**
- **Method — Algorithmic Core Extraction (ACE):** an SVD of the activity–relevance interaction $\mathbf{HJ}^{\top}$ isolates a compact activation subspace; causal ablations certify it (removal → chance, alone → full performance); operators fit inside expose the algorithm through their spectra.

Shift the explanatory target: **from what transformers happen to do → to what they must do.**

Why invariants? *Sparsity and circuits are basis-dependent descriptions of one implementation; spectra and causal subspaces are preserved across implementations — like eigenvalues under diagonalization.*

2 Markov Chains: Different Weights, Same 3D Core

Three 1-layer transformers, independent seeds, on 4-state Markov next-token prediction: near-orthogonal weights, yet ACE recovers the same 3D core from each 64D hidden state.

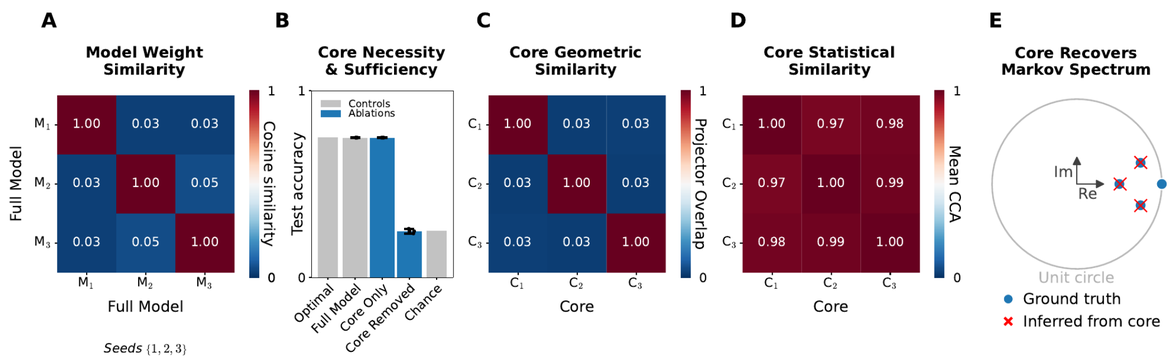

Fig. 1 — Divergent weights (A), causal 3D cores (B), orthogonal geometry (C), statistical alignment (D), recovered Markov spectrum (E).

- Keeping only the 3D core preserves full task accuracy; removing it collapses performance to chance — necessary and sufficient.
- Core subspaces are nearly orthogonal across seeds (overlap ≈ 0) yet statistically identical (CCA ≈ 0.99) — same computation, different coordinates.
- Operators fit inside each core recover the true Markov transition spectrum (Fig. 1E) — an internal model of the data-generating process.

3 Grokking: Cores Crystallize, Then Inflate

Two-layer transformers on $a + b \equiv c \ (mod\ 53)$ with weight decay; ACE traces the computation across training, with no mechanism hypothesized a priori.

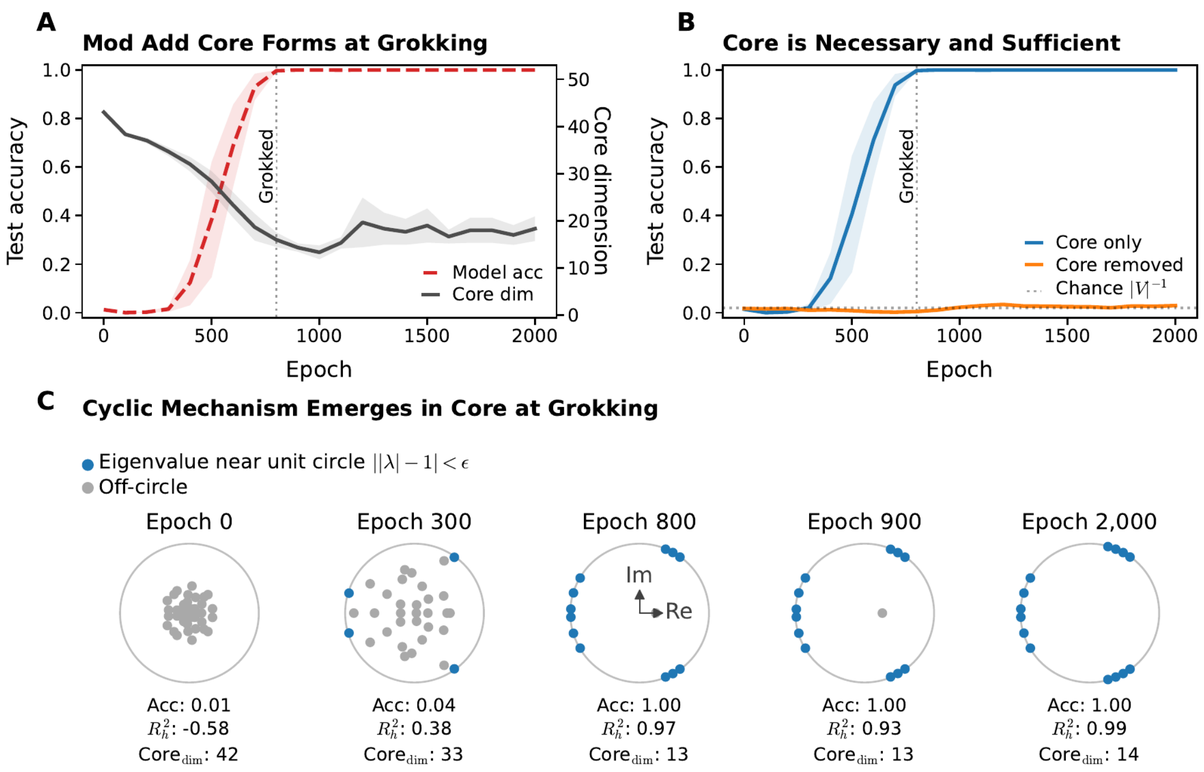

Fig. 2 — Cores condense exactly at grokking (A,B); operator eigenvalues snap onto the unit circle, blindly revealing the rotational "clock" mechanism (C).

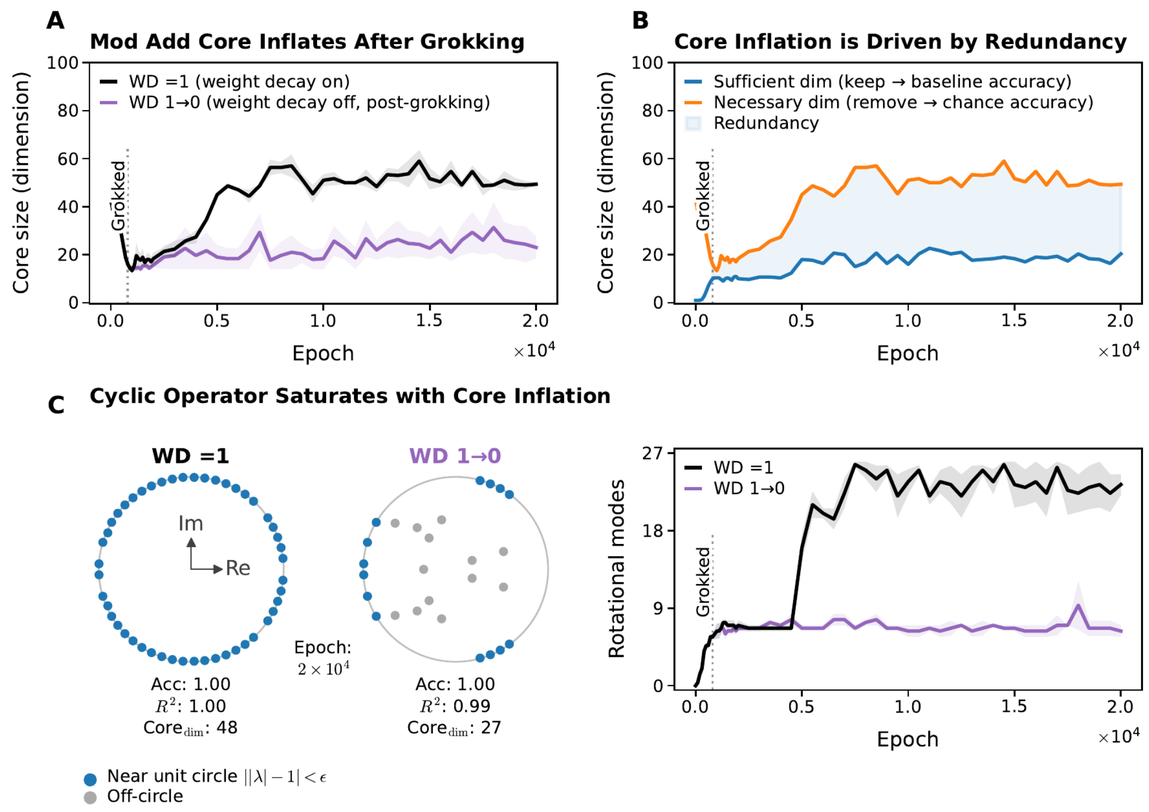

Fig. 3 — Under continued weight decay, cores inflate from ~15 to ~60 dims (A) via redundant encoding (B); the core operator saturates, with rotational modes climbing toward the maximum ⌊p/2⌋ = 26 (C).

4 Functional Equivalence Sets the Speed of Grokking

Post-grokking, weight decay solves a min-norm program over the ⌊p/2⌋ functionally equivalent Fourier modes α:

$$\min \|\alpha\|^2 \ \text{s.t.} \ \langle \alpha, \psi \rangle \geq \delta \implies \alpha^* \parallel \psi$$

— optimal only when every equivalent mode is active; each redundant mode adds drift toward generalization, so grokking delay shrinks with redundancy p and weight decay ω:

$$\tau_{\text{grok}}(p) = -\Omega \log\left(1 - \frac{p_{\text{crit}}}{p}\right) \propto \frac{1}{\omega p}$$

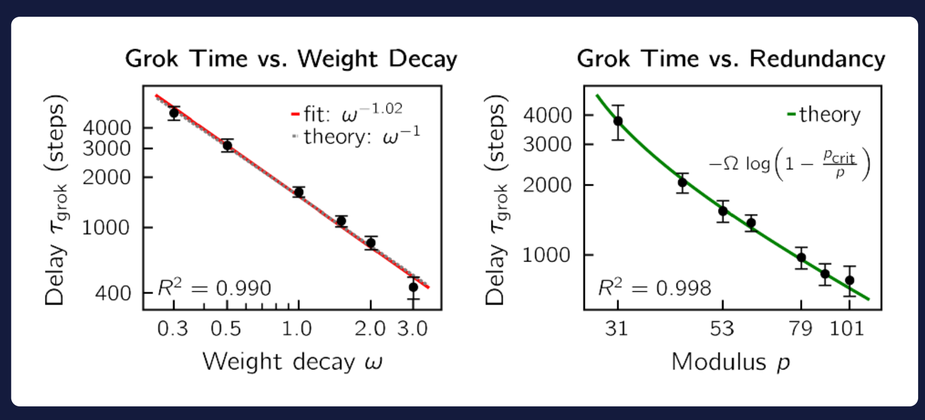

Fig. 4 — Delay scales as $\omega^{-1.02}$ (theory ω^{-1}); the closed-form solution fits the p-sweep with $R^2 > 0.99$ (12 seeds per condition). Practically: cores are most compact right after grokking — an interpretability window; anneal weight decay to zero to preserve the parsimonious solution.

5 LLMs: A Universal 1D Grammar Core

Six pretrained LMs across four families — GPT-2 S/M/L, LLaMA-3.1 (8B), Gemma-2 (9B), Qwen2.5 (32B) — on subject–verb number agreement.

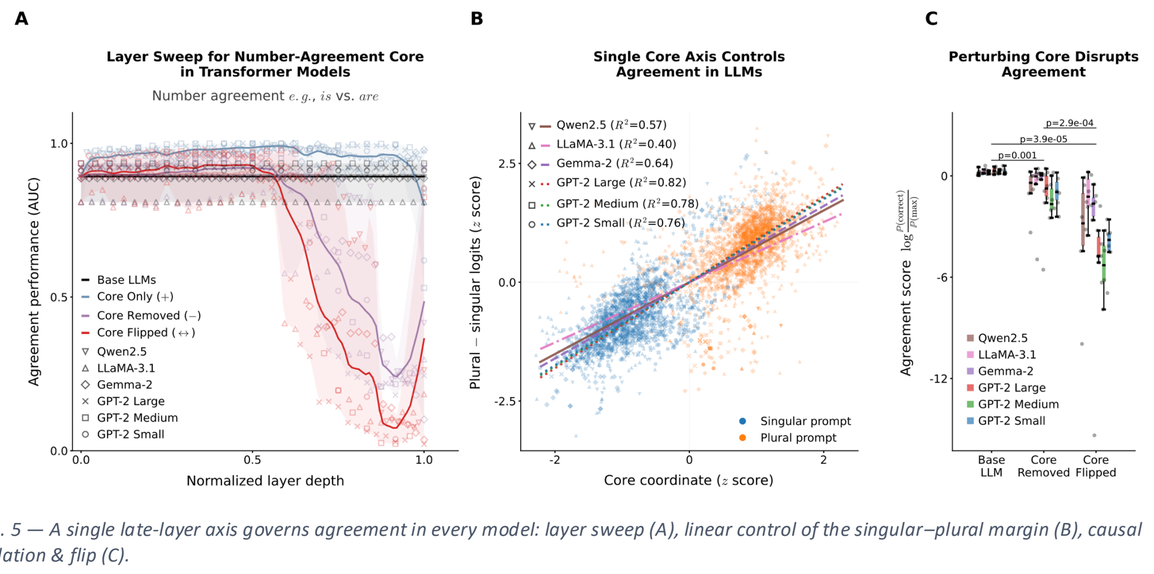

Fig. 5 — A single late-layer axis governs agreement in every model: layer sweep (A), linear control of the singular–plural margin (B), causal ablation & flip (C).

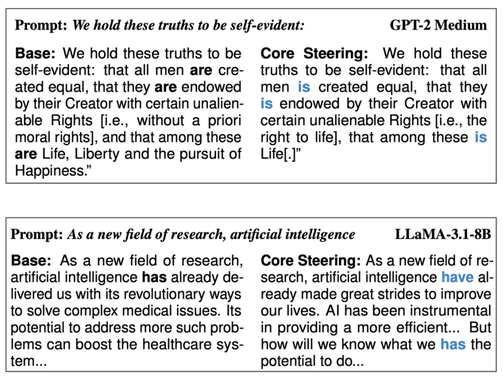

Fig. 6 — Adaptive core steering flips grammatical number mid-generation.

- 1D core: sufficient (AUC ≥ 0.92), necessary (≤ 0.24); flipping it inverts preference (AUC ≤ 0.09).
- Coordinates align across families (Spearman ρ up to 0.93) — independent of architecture, tokenizer, scale (117M → 32B).
- "The key next to the cabinets": inversion drives P(is) 0.51 → 0.01, P(are) 0.06 → 0.71.

Takeaways

- ACE is automated and causal — one pipeline runs unchanged from toy models to production LLMs.
- A universal, steerable 1D grammar core aligns across six LLMs spanning 117M → 32B parameters.
- Functional redundancy explains core inflation and yields an inverse scaling law for grokking, $\tau_{grok} \rightarrow 1/(\omega p)$.
- **Beneath divergent parameterizations lies shared, low-dimensional structure — target invariants, not parameterizations.**