Transformers Converge to Invariant Algorithmic Cores

Anonymous Authors[1]

Abstract

Training selects for behavior, not circuitry: many weight configurations can implement the same function. Studying any single trained neural network thus risks describing accidents of one training run rather than the computation itself. This work shifts focus from what transformers happen to do to what they must do by extracting *algorithmic cores*, compact subspaces that are necessary and sufficient for a task and that recur across independently trained models. Here, Algorithmic Core Extraction (ACE) is introduced to isolate these subspaces, causally validate them, and recover the algorithms they implement across settings ranging from synthetic tasks to large-scale pretrained models. Markov-chain transformers embed three-dimensional cores in nearly orthogonal subspaces yet recover identical transition spectra. Modular-addition transformers form compact cyclic cores at grokking that later inflate under continued regularization, redundantly distributing the same computation across many functionally equivalent modes. This functional redundancy is found to accelerate the transition from memorization to generalization, yielding an inverse scaling law for grokking time. In six language models spanning two orders of magnitude in scale (GPT-2 Small/Medium/Large, LLaMA-3.1, Gemma-2, and Qwen2.5), subject–verb agreement is governed by a single, steerable axis that aligns across architectures. Flipping this axis inverts grammatical number throughout open-ended generation. Together these results suggest that beneath the apparent complexity of trained transformers lies a simpler, shared computational structure, and that targeting invariants rather than parameterizations may offer a more tractable path to mechanistic understanding and control.

Code: anonymous.4open.science/r/cores-C008

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. Introduction

A key obstacle to mechanistic interpretability (Elhage et al., 2021; Sharkey et al., 2025) is underdetermination: while training constrains model behavior – how inputs are mapped to outputs – it generally does not constrain how behavior is realized internally. This poses a fundamental challenge for interpretability: if mechanisms don't generalize across realizations, which explanations are real?

Such *functional equivalence* is routinely observed among independently trained artificial neural networks, and has been investigated in loss landscape geometry and model merging (Draxler et al., 2018; Garipov et al., 2018; Ainsworth et al., 2022), the nonidentifiability of mechanistic circuits (Méloux et al., 2025), representational similarity (Kornblith et al., 2019), and in the Rashomon effect (Breiman, 2001). Yet, this phenomenon is not restricted to neural networks and has been explored across scientific disciplines. In biology, it appears as *degeneracy* (Edelman & Gally, 2001) (e.g., in the genetic code), and in evolution as *system drift* (Schiffman & Ralph, 2022), where the wiring of a gene network changes but its function does not. In control theory, different *realizations* (Kalman, 1962; 1963) induce identical observable dynamics, and in physics, *gauge symmetry* indicates that many mathematical descriptions represent the same state. A natural response is to shift focus from individual realizations to equivalence classes, studying the invariants shared across them. If mechanistic explanations of language models are to generalize across random seeds (Gurnee et al., 2024), checkpoints, and architectures, they should be tethered to stable, implementation-invariant quantities rather than idiosyncratic details that vary across training runs.

To explore this perspective, Algorithmic Core Extraction (ACE) is introduced to isolate *algorithmic cores*: low-dimensional subspaces that are necessary and sufficient for a task and shared across independent realizations. Applying ACE across three settings of escalating complexity demonstrates that functionally equivalent models can converge on compact, invariant mechanisms. In single-layer transformers (Vaswani et al., 2017), ACE recovers ground-truth Markov chain dynamics. In modular addition, it isolates the emergence of rotational dynamics at grokking. Finally, in six pretrained language models (spanning GPT-

2, LLaMA-3.1, Gemma-2, and Qwen2.5) (Radford et al., 2019; Grattafiori et al., 2024; Riviere et al., 2024; Yang et al., 2025), ACE identifies a shared, one-dimensional core that causally steers subject–verb agreement during open-ended generation.

This work contributes: (1) a conceptual framework for mechanistic interpretability that shifts focus from realization-specific circuitry to invariants; (2) ACE, a method for isolating compact subspaces that are causally necessary and sufficient for task performance; (3) evidence that these cores isolate interpretable mechanisms; (4) a theory linking functional equivalence, regularization, and grokking time; and (5) a steerable one-dimensional subject–verb agreement core shared by six distinct LLMs.

2. Methods

The structure–function relationship is often many-to-one, but how many different structures can implement the same function? In linear system theory this can be answered with the *Kalman decomposition* (Kalman, 1962; 1963; Anderson et al., 1966; Kalman et al., 1969), which guarantees the existence of a *minimal realization* – a dynamical system that can be empirically recovered via *balanced truncation* (Moore, 1981). *Algorithmic Core Extraction* (ACE) operationalizes this principle for transformers by first extracting activation subspaces that are both highly active and relevant, then causally validating them with ablations, and finally fitting operators to identify the computations they perform (Appendix A).

Extract. Fix a transformer layer with hidden dimension D. For N inputs, let $\mathbf{H} \in \mathbb{R}^{N\times D}$ denote the mean-centered activations, with rows $\mathbf{h}_i^\top$, and let $f : \mathbb{R}^D \to \mathbb{R}^K$ map activations to K task-relevant outputs. Stack the Jacobians as $\mathbf{J} \coloneqq [(\partial f/\partial \mathbf{h}_1)^\top \cdots (\partial f/\partial \mathbf{h}_N)^\top]^\top \in \mathbb{R}^{NK\times D}$. ACE finds directions that are jointly active and relevant by computing the SVD of their interaction:[1]

$$\mathbf{H}\mathbf{J}^\top \;=\; \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top.$$

The singular values quantify the joint activity and relevance of each direction and provide a principled criterion for rank selection. The *algorithmic core* is obtained by mapping the leading r columns of $\mathbf{U}$ back into activation space:

$$\mathcal{C} \;\coloneqq\; \mathrm{span}\big(\mathbf{H}^\top \mathbf{U}_r\big)\,,$$

and QR decomposition yields an orthonormal basis $\mathbf{Q} \in \mathbb{R}^{D\times r}$ and core projector $\mathbf{P} \coloneqq \mathbf{Q}\mathbf{Q}^\top$.

Validate. A core is *sufficient* if the projection $\mathbf{Ph}$ preserves task performance, and *necessary* if its complement $\mathbf{h} - \mathbf{Ph}$ reduces it to near chance.

[1] When $NK \gg D$, use SVD of $\mathbf{L}^\top\mathbf{\Gamma} \in \mathbb{R}^{D\times D}$ instead, where $\mathbf{L}\mathbf{L}^\top = \mathbf{H}^\top\mathbf{H} + \varepsilon\mathbf{I}$ and $\mathbf{\Gamma}\mathbf{\Gamma}^\top = \mathbf{J}^\top\mathbf{J}$.

Identify. A core's computational structure is recovered by examining its coordinates $\mathbf{z} = \mathbf{Q}^\top\mathbf{h}$ directly, or by fitting an operator $\mathbf{A}$ (e.g., $\mathbf{z}_{t+1} \approx \mathbf{A}\mathbf{z}_t$ by least squares) and inspecting its spectrum.

3. Algorithmic Core Necessity and Sufficiency

The central goal of this manuscript is to determine whether low-dimensional subspaces, or *algorithmic cores*, within higher-dimensional trained transformers exist that are functionally necessary and sufficient for task performance. If so, are such cores shared across independently trained models, and do they admit simple mechanistic characterizations?

Recovering algorithmic cores. The analysis begins in a fully controlled setting: three single-layer transformers ($d_{\text{model}} = 64$, $d_{\text{ff}} = 256$, $|V| = 4$) trained with independent random seeds on a four-state Markov chain (Appendix B). Although each reached near Bayes-optimal test accuracy, their learned weights exhibited near-zero cosine similarity, indicating highly divergent parameterizations (**Fig. 1A**). To search for a shared internal representation, ACE was applied to each model's 64-dimensional hidden state, successfully isolating a 3-dimensional algorithmic core. Ablations using all test data confirmed these cores were both necessary (removing the core drops accuracy to chance) and sufficient (retaining only the core preserves baseline accuracy) for the task (**Fig. 1B**; **Table A1**).

Geometric dissimilarity, statistical equivalence. To assess universality, each core recovered from the independently trained transformers was compared geometrically and statistically. Despite meeting equivalent causal criteria, cores were embedded in nearly orthogonal subspaces: projector overlap was 0.02–0.04, and principal angles ranged from 75°–90° (**Fig. 1C**; **Table A2**). Yet canonical correlation analysis (CCA) (Morcos et al., 2018) revealed nearly exact statistical alignment, with mean CCA correlations near 0.99 (**Fig. 1D**; **Table A2**). This suggests the cores encode the same information in different geometric coordinates – a signature of functionally equivalent yet structurally divergent realizations.

Algorithmic cores encode Markov dynamics. To interpret what algorithm the cores implement, a linear operator was fit to next-token dynamics inside each core, and relative to "oracle" prediction, these operators achieved strong fits: $R^2_{\text{core}}/R^2_{\text{oracle}} > 0.98$ (Appendix B). Eigenvalues (the spectrum) of a linear operator determine its dynamics – such as oscillations and growth rates – so matching eigenvalues can indicate matching dynamics. Remarkably, the eigenvalues of each fit operator matched the non-trivial eigenvalues of the true Markov transition matrix to within 1% (**Fig. 1E**; **Table A3**). This suggests that the recovered cores learned

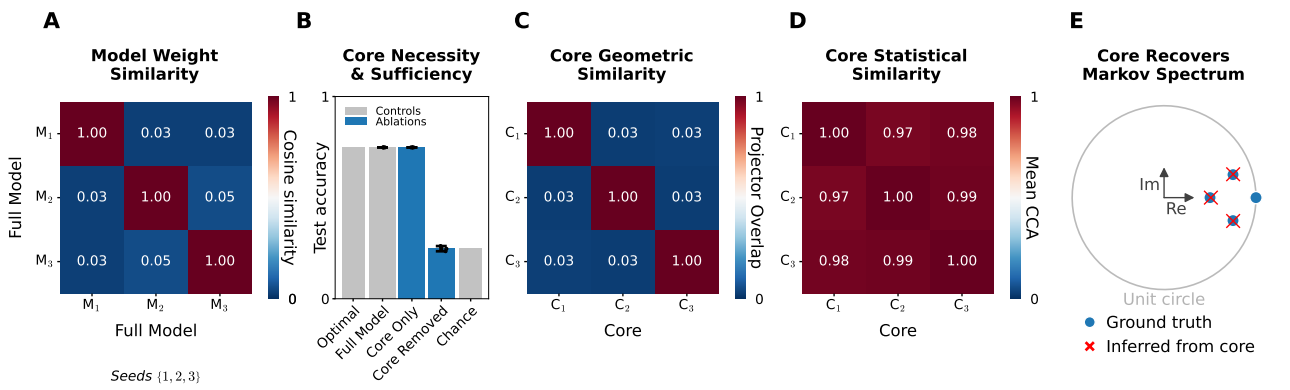

Figure 1. **Transformers trained on the same Markov task converge to a shared 3D causal core.** Three one-layer transformers were trained with different random seeds on next-token prediction for a four-state Markov chain (Appendix B). **(A)** Learned weights differ substantially across runs measured by cosine similarity. **(B)** 3D core extracted from each 64D hidden state are necessary and sufficient under ablation (**Table A1**), and compared to optimal $\sum_i \pi_i \max_j T_{ij}$ and chance $\max(\boldsymbol{\pi})$ theoretical controls, with transition matrix $\mathbf{T}$ and stationary distribution $\boldsymbol{\pi}$; points show individual accuracies and bars denote mean±sem. **(C)** Cores geometrically diverge with projector overlaps near zero and principal angles nearly orthogonal (**Table A2**). **(D)** Cores statistically align with mean cross-core CCAs reaching near unity (also see **Table A2**). **(E)** Dynamics fit in core recover the Markov chain nontrivial spectrum (**Table A3**).

to efficiently encode Markov dynamics: trained transformers route inputs through a minimal, shared 3D subspace – that is necessary and sufficient for performance – and internally represents transition dynamics up to a change of coordinates.

4. Algorithmic Core Emergence and Evolution

Because ACE is automated, it can recover learned computations without presupposing their form and can trace how they evolve during training. Modular addition is a natural test case: transformers trained on this task exhibit grokking (Power et al., 2022; Liu et al., 2022a), with high training accuracy preceding a delayed spike in test accuracy. Prior work showed that these models learn a Fourier "clock" algorithm (Nanda et al., 2023), but doing so required hypothesizing the mechanism a priori, designing targeted probes, and manually verifying circuits.

Cores crystallize at grokking. Three two-layer transformers ($d_{\text{model}} = 128$, $d_{\text{ff}} = 512$, $|V| = 53$) were trained on modular addition ($a + b \equiv c \bmod 53$) for 2×10^3 epochs under weight decay regularization to encourage generalization. All models grokked: test accuracy remained near chance until spiking around epoch 800. Coincident with this delayed generalization, algorithmic cores crystallized – condensing into low-dimensional, ablation-defined necessary and sufficient subspaces (**Fig. 2A,B**; Appendix C).

Blind recovery of rotational dynamics in cores. At each checkpoint, a linear operator was fit to the second-layer "shift" (add 1) dynamics in each extracted core. This revealed the emergence of a cyclic computational structure: at grokking, the operators' eigenvalues snap onto the unit circle (**Fig. 2C**), indicating rotational dynamics capable of modular addition. Notably, this structure emerges directly from least-squares optimization in the core, without needing to prespecify an algorithmic form. However, while all three models converged to cyclic operators, the specific rotational *modes* (conjugate eigenvalue pairs) differed across runs – another instance of functional equivalence without structural identity (Chughtai et al., 2023; Zhong et al., 2023; Olah, 2025). Modular addition permits multiple valid modes and multiplicities, and models need not agree on which, nor how many, to use. Remarkably, even at grokking, each operator contained more rotational modes than the single mode minimally required – a hint of the redundancy that becomes extreme under extended training (**Fig. 3**).

Cores inflate under extended training. Extending training to 2×10^4 epochs revealed an unexpected phenomenon: under continued weight decay, cores progressively inflated from approximately 15 to 60 dimensions. In contrast, disabling weight decay post-grokking kept cores more compact (**Fig. 3A**). This inflation is driven by a pronounced increase in redundant encoding. While the number of dimensions *sufficient* for task performance remained stable, the number of dimensions *necessary* to prevent chance-level performance expanded dramatically (**Fig. 3B**). Operator analysis reveals how this transformer "over-education" manifests: under continued weight decay, operators accumulated rotational modes. These approached the theoretical maximum of $\lfloor p/2 \rfloor = 26$ valid harmonic representations by the terminal epoch – far exceeding the minimally required single mode (**Fig. 3C**). Disabling weight decay prevented this proliferation: cores remained compact, mode counts stayed sparse,

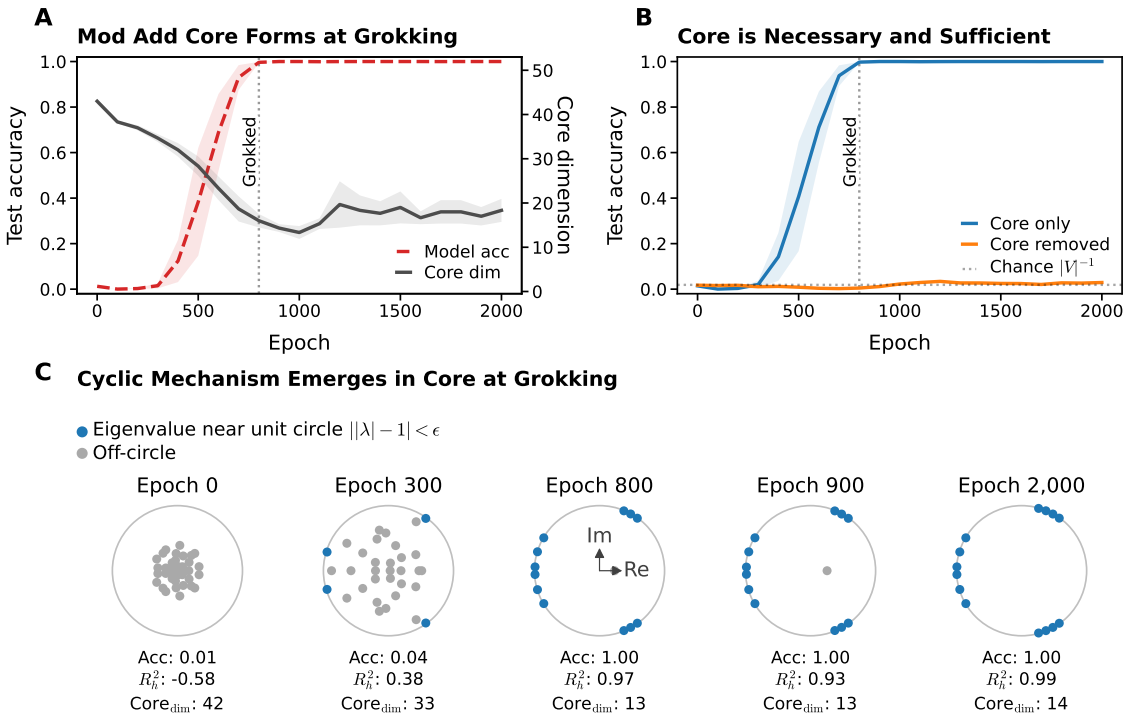

Figure 2. **Modular addition cores form at grokking and implement rotational mechanics.** Three two-layer transformers were trained with different random seeds. **(A)** Test accuracy exhibits grokking (red, mean±sem, left y-axis) coincident with algorithmic core formation (gray, mean±sem, right y-axis). **(B)** After grokking, the recovered cores are necessary and sufficient under ablation. **(C)** Automated operator fits in core coordinates reveal the emergence of a cyclic mechanism: before grokking, eigenvalues scatter inside the unit circle, while at grokking they snap onto it, indicating discovery of a rotational mechanism.

and operator structure remained stable. This suggests that weight decay may actively drive the transition from parsimonious algorithmic solutions to redundantly saturated representations.

4.1. Redundancy Drives Core Inflation and Grokking

That a regularization penalty designed to simplify representations should instead inflate cores seems paradoxical. The behavior, however, emerges naturally from minimizing the weight norm within a highly redundant solution space. Furthermore, this interplay between redundancy and regularization also predicts the timing of grokking itself.

Minimum norm requires maximum redundancy. After grokking, task loss is negligible and the gradient is dominated by weight decay (Varma et al., 2023), driving the network toward a minimum-norm solution. By Fourier symmetry (Chughtai et al., 2023), modular addition mod p admits $\lfloor p/2 \rfloor$ functionally equivalent modes, each a 2D rotation with phase $\theta_k := 2\pi k/p$. Let $\boldsymbol{\alpha} \geq 0$ denote mode amplitudes and $\boldsymbol{\psi}$ their label-contrasts, where $\psi_k := 1 - \cos\theta_k > 0$ is mode k's contribution to the classification margin. If modes are encoded in approximately orthogonal parameter subspaces[2] then the weight norm satisfies $\|\mathbf{W}\|^2 \approx \|\boldsymbol{\alpha}\|^2$, while correct classification requires margin $\langle \boldsymbol{\alpha}, \boldsymbol{\psi} \rangle \geq \delta$. Training thus implicitly solves

$$\min_{\boldsymbol{\alpha} \geq \mathbf{0}} \|\boldsymbol{\alpha}\|^2 \qquad \text{subject to} \qquad \langle \boldsymbol{\alpha}, \boldsymbol{\psi} \rangle \geq \delta.$$

By the Cauchy–Schwarz inequality, $\|\boldsymbol{\alpha}\|^2$ is minimized when $\boldsymbol{\alpha} \parallel \boldsymbol{\psi}$ – that is, when *every* mode is active – with the optimal solution $\boldsymbol{\alpha}^* = \left(\delta/\|\boldsymbol{\psi}\|_2^2\right)\boldsymbol{\psi}$. Weight decay thus acts as a redistribution force: rather than simplifying the representation, it spreads weight across all valid solutions. Disabling weight decay removes this pressure, consistent with observations in **Fig. 3**.

Functional equivalence accelerates grokking. The same redistribution pressure governs the speed of grokking. Define the grokking delay $\tau_{\text{grok}} := \tau_{\text{gen}} - \tau_{\text{mem}}$ as the time between memorization and generalization, and model the transition to generalization as the margin $m(t) := \langle \boldsymbol{\alpha}(t), \boldsymbol{\psi} \rangle$ reaching threshold δ. After memorization, with task gradients largely vanished and weight decay (ω) dominating, the

[2] If not orthogonal (or in superposition (Elhage et al., 2022)) with $\mathcal{S} \succ 0$ mode-overlap, $\|\boldsymbol{\psi}\|_2^2 \to \boldsymbol{\psi}^\top \mathcal{S}^{-1} \boldsymbol{\psi}$, reducing effective redundancy.

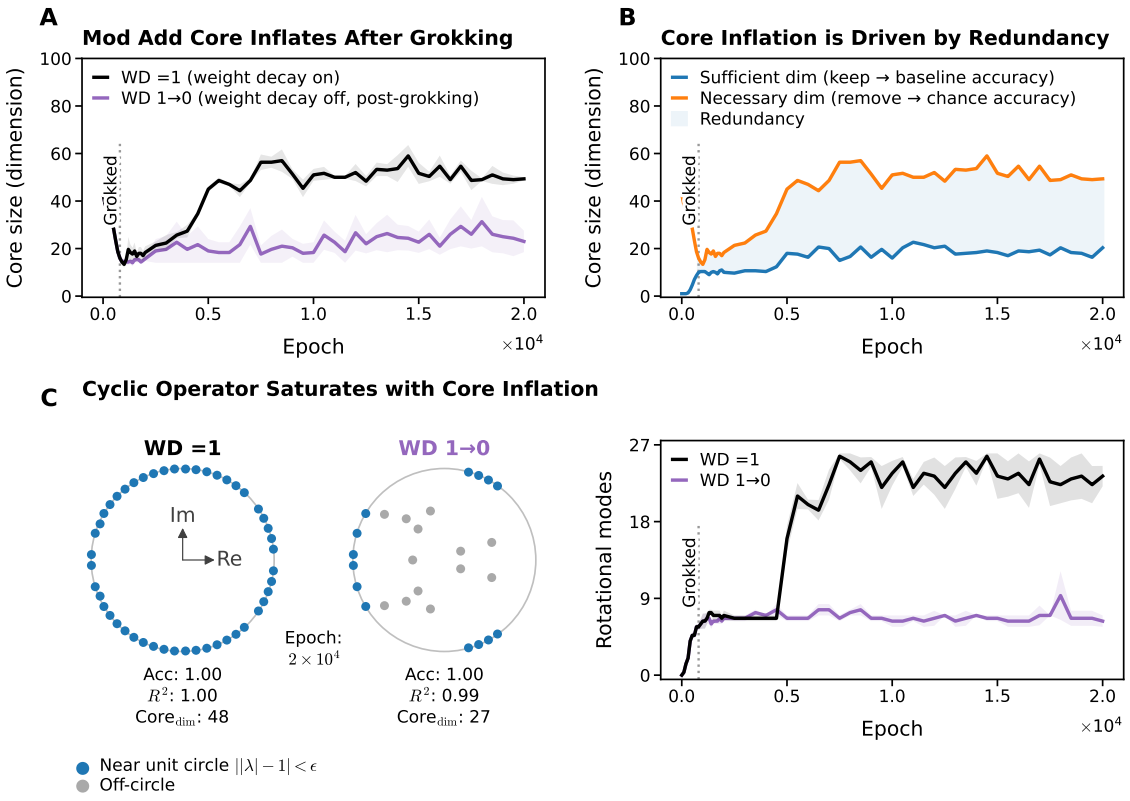

Figure 3. **Extended training with weight decay inflates cores.** Long-term training dynamics of transformers that grokked modular addition under different weight-decay schedules. **(A)** After grokking, core dimension continues to increase when weight decay is maintained (black, mean±sem), but remains compact when weight decay is disabled post-grokking (purple). **(B)** Core inflation is driven by redundancy: the number of dimensions sufficient to preserve performance is stable, while the number whose removal reduces performance to chance increases. Lines depict means across models trained with weight decay fixed. **(C)** (*Left*) Dynamics fit in the terminal epoch reveal a saturated core operator when weight decay is maintained, in contrast to a more sparsely represented operator when weight decay is disabled. (*Right*) Rotational modes (conjugate eigenvalue pairs) around the unit circle increase with extended training under weight decay, whereas when weight decay is removed, mode counts remain stable.

expected margin trajectory follows (Appendix E.1)

$$\dot{m}(t) \;=\; -\omega\, m(t) \;+\; c\omega\, \|\psi\|_2^2.$$

Crucially, functional equivalence makes the margin-driving direction additive across modes, giving $\|\psi\|_2^2 \propto p$.[3] Each redundant mode amplifies the mean-drift velocity toward generalization, consistent with the multiple active modes observed at grokking (**Fig. 2C**). When $p < d_{\text{model}}$ the initial memorized solution has negligible margin ($m(0) \approx 0$), and grokking occurs when $m(\tau) = \delta$. Solving for the expected grokking time delay yields an expression that linearizes for high redundancy ($p \gg p_{\text{crit}}$) into a simple inverse scaling law:

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

[3] Using $\sum_{k=1}^{\lfloor p/2 \rfloor} \cos\theta_k = \sum_{k=1}^{\lfloor p/2 \rfloor} \cos 2\theta_k = -\frac{1}{2}$ and expanding $(1 - \cos\theta_k)^2$ gives $\|\psi\|_2^2 = \frac{3}{4}p$.

Two empirical constants govern this expression: an *optimizer constant* $\Omega \propto \omega^{-1}$ that sets the timescale of grokking when it occurs, and an *architectural constant* p_{crit} that determines whether it can occur at all. Grokking time thus shrinks with both weight decay and functional redundancy. These predictions are validated by sweeping ω and p in transformers (**Fig. 4**; Appendix E.2).

Summary. The algorithmic core framework – automated operator extraction from causally defined, low-dimensional core subspaces – can mechanistically characterize and trace the evolution of computations transformers learn throughout training. In modular addition, the extracted cores exhibit rotational dynamics consistent with the task's cyclic structure, crystallize at grokking, and inflate under extended weight decay. This inflation reflects transformers converging on the optimal weighting strategy under regularization: to distribute weight across all functionally equivalent representations. This same pressure – regularization utilizing

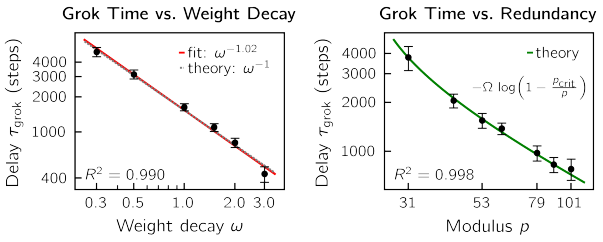

Figure 4. **Grokking time scales inversely with redundancy.** *(Left)* Time to grok after memorization τ scales inversely with weight decay ω; consistent with prior observations (Liu et al., 2022b). The observed fit ($\tau \propto \omega^{-1.02}$, red) matches the theoretical prediction ω^{-1} (gray). *(Right)* Grokking time scales inversely with redundancy p. The ODE solution (green; $R^2 > 0.99$) captures both the inverse scaling at large p and the divergence near p_{crit}. Fit parameters: $\hat{\Omega} \approx 2{,}770$, $\hat{p}_{\text{crit}} \approx 23$. Points represent mean±sd for 12 random seeds (Appendix E.2).

redundancy – predicts the speed of grokking, explaining the transition from memorization to generalization. The next question is whether these tools scale to larger and more complex systems.

5. Scaling ACE to LLMs: A Universal 1D Core

The preceding experiments establish the ACE framework in highly controlled, synthetic settings. The critical question is whether the ACE framework scales beyond toy models to govern complex behaviors in production-scale models. To establish an empirical foothold on this question, ACE was applied to six pretrained language models spanning four distinct families: GPT-2 Small, Medium, and Large (117M, 345M, and 774M parameters) (Radford et al., 2019; Wolf et al., 2020); LLaMA-3.1 (8B) (Grattafiori et al., 2024); Gemma-2 (9B) (Riviere et al., 2024); and Qwen2.5 (32B) (Yang et al., 2025). These models differ in architecture, training corpus, and tokenization, and span more than two orders of magnitude in parameter count. The target task is *subject–verb number agreement*, a tractable linguistic computation with clear ground-truth labels (singular vs. plural subject) and a well-defined behavioral output (verb selection); admitting systematic evaluation via controlled prompts and a scalar verb-preference score (Linzen et al., 2016; Marvin & Linzen, 2018; Finlayson et al., 2021) (Appendix D).

Localizing a shared 1D agreement core. To localize the agreement mechanism, candidate cores were extracted at each layer and evaluated via causal ablations. Across all six models, early layers exhibited minimal causal influence, but a highly potent core consistently emerged in the late layers (**Fig. 5A**). At the layer of maximal effect, this agreement core is remarkably one-dimensional – a single axis separated from all remaining directions by a large spectral gap (**Table A5**).

Causal validation and control. Observationally, this axis behaves as a graded number coordinate: projection onto it predicts the singular–plural logit margin across models (**Fig. 5B**), aligning with the linear representation hypothesis (Park et al., 2023). However, because subspace projections alone can be deceptive (Belinkov, 2022; Makelov et al., 2023), claims here are strictly grounded in causal ablations. Despite its compact size, this single axis is sufficient (retaining it preserves agreement; AUC ≥ 0.92), necessary (removing it collapses agreement below chance; AUC ≤ 0.24), and directionally controllable. Reflecting activations through this axis inverts verb preferences, inducing strong disagreement with the subject (AUC ≤ 0.09; **Table A6**). At the prompt level, for instance, core inversion on "The key next to the cabinets" drives $\mathbb{P}(\textit{is})$ from 0.51 down to 0.01, while boosting $\mathbb{P}(\textit{are})$ from 0.06 to 0.71 (**Fig. 5C**).

Alignment across LLMs. Projecting last-token hidden states onto each model's agreement core (**Fig. 5B**) yields a signed grammatical-number coordinate that tracks verb preference. Because cores are one-dimensional, cross-model alignment reduces to fixing a sign convention and comparing projected coordinates. Within the GPT-2 family – three models that share an architecture and training procedure – coordinates align tightly (Spearman's $\rho = 0.88$–0.92; Pearson's $r = 0.92$–0.97). More strikingly, alignment persists across families: between Qwen2.5, LLaMA-3.1, Gemma-2, and the GPT-2 models, Spearman correlations range from 0.59 to 0.93 and Pearson correlations from 0.62 to 0.94 (**Table A4**). The strongest cross-family correlations (Qwen × Gemma: $\rho = 0.93$; Gemma × GPT-2 Medium: $\rho = 0.89$) approach the within-family ceiling, indicating that the agreement core encodes grammatical number in a way that is largely independent of architecture, tokenization, training corpus, and scale.

Core steering inverts grammar in open-ended text. A stronger test of the core's role is whether it governs agreement throughout autoregressive generation, where each token conditions subsequent predictions. To test this, the core-axis intervention was applied *adaptively* at each decoding step. Modulating the intervention strength based on each token's sensitivity to number agreement, leaving irrelevant tokens untouched (Appendix D), induced systematic agreement violations across all six models (**Fig. 6**). Singular subjects recruited plural verbs, plural contexts shifted toward the singular, and errors cascaded as toggling the number variable corrupted downstream predictions. Crucially, the effect generalized well beyond the specific verbs (*is/are/was/were*) used to define the initial preference score. The emergence of agreement failures in entirely different word classes supports the interpretation that the core en-

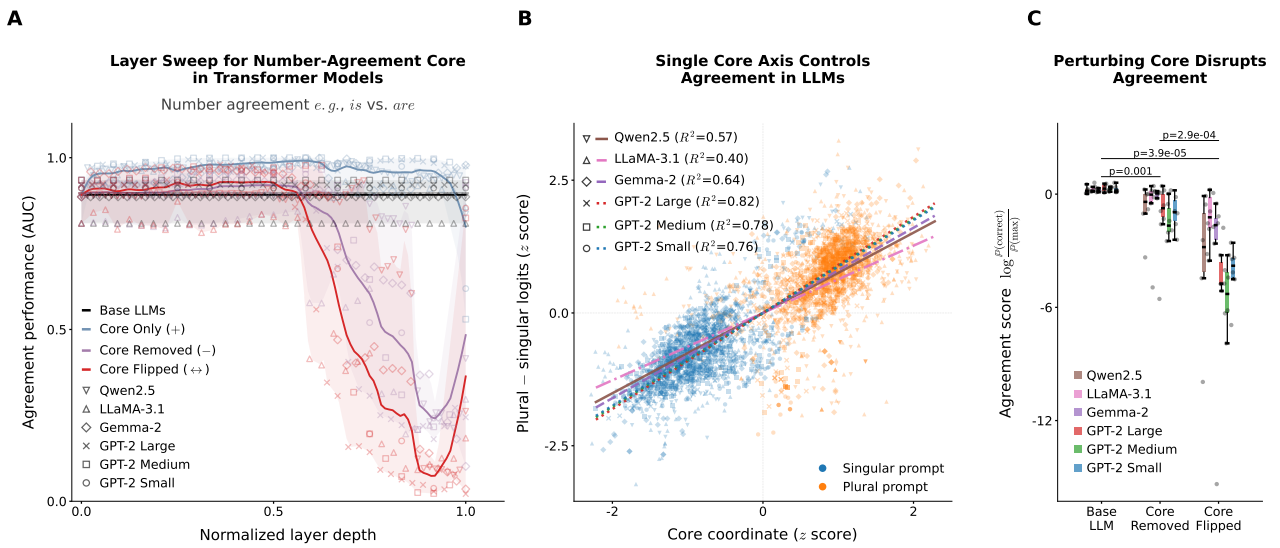

Figure 5. **Subject–verb agreement is mediated by a shared 1D core across LLMs.** The core framework was applied to GPT-2 Small, Medium, Large, Gemma-2, LLaMA-3.1, and Qwen2.5 to isolate a low-dimensional mechanism for number agreement. **(A)** Layer sweep: agreement performance (AUC) vs. normalized layer depth, averaged across LLMs (lines) with per-model measurements (markers) and shaded min–max bands. Agreement performance is the probability that the model assigns a higher plural-vs-singular verb-preference score to a plural than to a singular prompt. **(B)** Projecting last-token hidden states onto the core produces a nearly linear control axis for the singular–plural logit margin; per-model affine fits are shown after z scoring. **(C)** Removing the core degrades agreement, while flipping it inverts verb preference. Box plots summarize prompt-level agreement scores under perturbation; reported p-values combine per-model paired Wilcoxon tests using Fisher's method.

Prompt: ***We hold these truths to be self-evident:*** **GPT-2 Medium**

Base: We hold these truths to be self-evident: that all men **are** created equal, that they **are** endowed by their Creator with certain unalienable Rights [i.e., without a priori moral rights], and that among these **are** Life, Liberty and the pursuit of Happiness."

Core Steering: We hold these truths to be self-evident: that all men **is** created equal, that they **is** endowed by their Creator with certain unalienable Rights [i.e., the right to life], that among these **is** Life[.]"

Prompt: ***As a new field of research, artificial intelligence*** **LLaMA-3.1-8B**

Base: As a new field of research, artificial intelligence **has** already delivered us with its revolutionary ways to solve complex medical issues. Its potential to address more such problems can boost the healthcare system...

Core Steering: As a new field of research, artificial intelligence **have** already made great strides to improve our lives. AI has been instrumental in providing a more efficient... But how will we know what we **has** the potential to do...

Figure 6. **Core steering induces systematic agreement violations in open-ended generation.** Prompted text from *Base* or *Core Steering*, with select violations highlighted.

codes a global grammatical-number variable, rather than a narrow, verb-specific heuristic.

Summary. Subject–verb agreement in language models is governed by a 1D causal subspace localized to late layers. This core is necessary, sufficient, and controllable, and its coordinates align across six models from four families.

6. Discussion

These results suggest that transformer computations may be governed by low-dimensional mechanisms that recur across independent training runs despite substantial variation in learned parameters. These findings have implications for how we conceptualize mechanistic interpretability.

Invariance, not sparsity or circuitry. Mechanistic interpretability has largely studied implementations, such as circuits of attention heads and neurons (Elhage et al., 2021; Olah et al., 2020; Wang et al., 2022; Ameisen et al., 2025; Lindsey et al., 2025), or sparse decompositions of activations into interpretable features (Cunningham et al., 2023; Bricken et al., 2023; Dunefsky et al., 2024; Templeton et al., 2024). Such descriptions can be highly precise, but they face a conceptual challenge: they may be implementation-specific. Two models might compute the same function using entirely different circuits and coordinate systems (Méloux et al., 2025; Fel et al., 2025). The core framework shifts the explanatory target from implementation to invariant. The motivation for sparse features parallels a classical aim in linear algebra: diagonalization. But the fundamental power of diagonalization lies not in sparsity per se, but in revealing invariants – eigenvalues preserved under change of basis. Sparsity is basis-dependent; invariants are not. Likewise, circuits and sparse features describe coordinates of implementation, while cores identify

the causal subspaces and dynamics preserved across implementations. Where features or circuits recur across models, perhaps identified via cross-coders (Lindsey et al., 2024), the approaches converge. Where they diverge, invariance provides a reliable criterion for distinguishing structural essence from artifact.

Cores as internal world models. The observation that independent models converge to the same invariant structure raises a natural question: what anchors these shared representations? If these cores are not artifacts of the architecture or training run, they might reflect the data-generating process itself. When algorithmic cores recover ground-truth task structure – Markov transition spectra, cyclic operators for modular arithmetic – they encode not merely input–output mappings but internal representations of the generative process underlying the data (Li et al., 2022; Gurnee & Tegmark, 2023; Huh et al., 2024). This aligns with two classical ideas: the *good regulator theorem* (Conant & Ross Ashby, 1970) and the *internal model principle* (Francis & Wonham, 1976) from control theory, which hold that any system achieving optimal prediction must contain a model of its environment. When a core is isomorphic to the task-generating process, interpretability may be viewed as a form of internal-model recovery.

Redundancy accelerates grokking. Once a model reaches perfect training accuracy, it enters a highly degenerate zero-loss manifold in parameter space (Bushnaq et al., 2024), populated by many functionally equivalent solutions. Weight decay then biases stochastic exploration along this manifold toward a minimum-norm, maximum-margin solution. Because the target task admits multiple functionally equivalent realizations, the corrective pressure from weight decay accumulates across valid modes rather than acting on a single narrow solution. This accelerates the expected trajectory from memorization to generalization, consistent with the core inflation observed under continued regularization. An analogous phenomenon appears in evolutionary genetics, where robustness creates extended networks of phenotype-preserving genotypes that facilitate the discovery of new functions via neutral drift (Wagner, 2008; 2012). A practical consequence is that cores are most compact immediately after grokking and subsequently inflate. This suggests a natural *interpretability window*: annealing weight decay toward zero shortly after task convergence may help preserve the most compact solution.

System drift and model merging. System drift describes how a gene network can preserve its phenotype while its underlying genetic wiring diverges, effectively drifting through a neutral space. Because the set of functionally equivalent realizations is not generally convex or closed under recombination, mixing divergent solutions often produces *hybrid incompatibility* (Schiffman & Ralph, 2022). Transformers exhibit an analogous pattern: models trained from different initializations implement identical cores embedded in nearly orthogonal subspaces, revealing substantial representational drift despite functional equivalence. This orthogonality implies that naïve weight interpolation between geometrically divergent models moves off the solution manifold, consistent with empirical difficulties in model merging (Garipov et al., 2018; Ainsworth et al., 2022). By contrast, extracting and aligning algorithmic cores may offer a principled diagnostic for merge-compatibility and a potential coordinate system for successful recombination.

Limitations and future directions. Whether cores remain low-dimensional for multi-step reasoning tasks is untested. The agreement core, however, remains one-dimensional across six models spanning four architectures and over two orders of magnitude in scale (117M to 32B parameters), suggesting core dimensionality may not depend on model scale. This is compatible with the empirical success of LoRA (Hu et al., 2022), which often achieves large behavioral changes via low-dimensional weight updates. Extracting task-specific cores from multifunctional models also requires framing precise mechanistic inquiries; this work demonstrates this for subject–verb agreement, but systematic approaches to task decomposition remain open. The extraction procedure itself admits natural extensions: nonlinear dimensionality reduction in place of the active component, learned probes in place of Jacobians, and Koopman operator approximations (Brunton et al., 2021) for tasks with nonlinear dynamics. More broadly, the relevant invariants for complex tasks are not obvious *a priori*; future work might discover them empirically by asking what core properties are shared across independently trained models. Finally, methods that identify causally effective subspaces may enable more targeted model control: this could support auditing and debugging, but also creates misuse risks if used to induce systematic errors or circumvent intended behaviors.

Conclusion. The results in this work point toward a view of transformer computation as organized around low-dimensional invariants: subspaces that are preserved across training runs, necessary and sufficient for task performance, and structured in ways that mirror the tasks themselves. If this view is approximately correct, interpretability efforts may benefit from targeting such invariants – seeking the computational essence that recurs across realizations rather than the implementation details that vary. The algorithmic core is one operationalization of this intuition. Whether it scales to the complexity of contemporary language models remains to be seen, but the guiding principle – focus on what is preserved, not what is particular – may prove durable.

References

Ainsworth, S. K., Hayase, J., and Srinivasa, S. Git re-basin: Merging models modulo permutation symmetries. *arXiv preprint arXiv:2209.04836*, 2022.

Ameisen, E., Lindsey, J., Pearce, A., Gurnee, W., Turner, N. L., Chen, B., Citro, C., Abrahams, D., Carter, S., Hosmer, B., Marcus, J., Sklar, M., Templeton, A., Bricken, T., McDougall, C., Cunningham, H., Henighan, T., Jermyn, A., Jones, A., Persic, A., Qi, Z., Ben Thompson, T., Zimmerman, S., Rivoire, K., Conerly, T., Olah, C., and Batson, J. Circuit tracing: Revealing computational graphs in language models. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/methods.html.

Anderson, B., Newcomb, R., Kalman, R., and Youla, D. Equivalence of linear time-invariant dynamical systems. *Journal of the Franklin Institute*, 281(5):371–378, 1966.

Belinkov, Y. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Bellman, R. and Åström, K. J. On structural identifiability. *Mathematical Biosciences*, 7(3-4):329–339, 1970.

Breiman, L. Statistical modeling: The two cultures. *Statistical Science*, 2001.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., McLean, B., Burke, J. E., Hume, T., Carter, S., Henighan, T., and Olah, C. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Brunton, S. L., Budišić, M., Kaiser, E., and Kutz, J. N. Modern koopman theory for dynamical systems. *arXiv preprint arXiv:2102.12086*, 2021.

Bushnaq, L., Mendel, J., Heimersheim, S., Braun, D., Goldowsky-Dill, N., Hänni, K., Wu, C., and Hobbhahn, M. Using degeneracy in the loss landscape for mechanistic interpretability. *arXiv preprint arXiv:2405.10927*, 2024.

Chughtai, B., Chan, L., and Nanda, N. A toy model of universality: Reverse engineering how networks learn group operations. In *International Conference on Machine Learning*, pp. 6243–6267. PMLR, 2023.

Conant, R. C. and Ross Ashby, W. Every good regulator of a system must be a model of that system. *International journal of systems science*, 1(2):89–97, 1970.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

Draxler, F., Veschgini, K., Salmhofer, M., and Hamprecht, F. Essentially no barriers in neural network energy landscape. In *International conference on machine learning*, pp. 1309–1318. PMLR, 2018.

Dunefsky, J., Chlenski, P., and Nanda, N. Transcoders find interpretable llm feature circuits. *Advances in Neural Information Processing Systems*, 37:24375–24410, 2024.

Edelman, G. M. and Gally, J. A. Degeneracy and complexity in biological systems. *Proceedings of the national academy of sciences*, 98(24):13763–13768, 2001.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M., and Olah, C. Toy models of superposition. 2022. URL https://transformer-circuits.pub/2022/toy_model/index.html. Online.

Fel, T., Lubana, E. S., Prince, J. S., Kowal, M., Boutin, V., Papadimitriou, I., Wang, B., Wattenberg, M., Ba, D., and Konkle, T. Archetypal sae: Adaptive and stable dictionary learning for concept extraction in large vision models. *arXiv preprint arXiv:2502.12892*, 2025.

Finlayson, M., Mueller, A., Gehrmann, S., Shieber, S. M., Linzen, T., and Belinkov, Y. Causal analysis of syntactic agreement mechanisms in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1828–1843, 2021.

Francis, B. A. and Wonham, W. M. The internal model principle of control theory. *Automatica*, 12(5):457–465, 1976.

Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D. P., and Wilson, A. G. Loss surfaces, mode connectivity, and fast ensembling of dnns. *Advances in neural information processing systems*, 31, 2018.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Gurnee, W. and Tegmark, M. Language models represent space and time. *arXiv preprint arXiv:2310.02207*, 2023.

Gurnee, W., Horsley, T., Guo, Z. C., Kheirkhah, T. R., Sun, Q., Hathaway, W., Nanda, N., and Bertsimas, D. Universal neurons in gpt2 language models. *arXiv preprint arXiv:2401.12181*, 2024.

Hu, E. J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Huh, M., Cheung, B., Wang, T., and Isola, P. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.

Kalman, R. E. Canonical structure of linear dynamical systems. *Proceedings of the National Academy of Sciences*, 48(4):596–600, 1962.

Kalman, R. E. Mathematical description of linear dynamical systems. *Journal of the Society for Industrial and Applied Mathematics, Series A: Control*, 1(2):152–192, 1963.

Kalman, R. E., Falb, P. L., and Arbib, M. A. *Topics in mathematical system theory*. McGraw-Hill, New York, 1969. ISBN 0754321069.

Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019.

Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., and Wattenberg, M. Emergent world representations: Exploring a sequence model trained on a synthetic task. *arXiv preprint arXiv:2210.13382*, 2022.

Lindsey, J., Templeton, A., Marcus, J., Conerly, T., Batson, J., and Olah, C. Sparse crosscoders for cross-layer features and model diffing. *Transformer Circuits Thread*, pp. 3982–3992, 2024.

Lindsey, J., Gurnee, W., Ameisen, E., Chen, B., Pearce, A., Turner, N. L., Citro, C., Abrahams, D., Carter, S., Hosmer, B., Marcus, J., Sklar, M., Templeton, A., Bricken, T., McDougall, C., Cunningham, H., Henighan, T., Jermyn, A., Jones, A., Persic, A., Qi, Z., Thompson, T. B., Zimmerman, S., Rivoire, K., Conerly, T., Olah, C., and Batson, J. On the biology of a large language model. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/biology.html.

Linzen, T., Dupoux, E., and Goldberg, Y. Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535, 2016.

Liu, Z., Kitouni, O., Nolte, N. S., Michaud, E., Tegmark, M., and Williams, M. Towards understanding grokking: An effective theory of representation learning. *Advances in Neural Information Processing Systems*, 35:34651–34663, 2022a.

Liu, Z., Michaud, E. J., and Tegmark, M. Omnigrok: Grokking beyond algorithmic data. *arXiv preprint arXiv:2210.01117*, 2022b.

Makelov, A., Lange, G., and Nanda, N. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. *arXiv preprint arXiv:2311.17030*, 2023.

Marvin, R. and Linzen, T. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 1192–1202, 2018.

Méloux, M., Maniu, S., Portet, F., and Peyrard, M. Everything, everywhere, all at once: Is mechanistic interpretability identifiable? *arXiv preprint arXiv:2502.20914*, 2025.

Moore, B. Principal component analysis in linear systems: Controllability, observability, and model reduction. *IEEE Transactions on Automatic Control*, 26(1):17–32, 1981. doi: 10.1109/TAC.1981.1102568.

Morcos, A., Raghu, M., and Bengio, S. Insights on representational similarity in neural networks with canonical correlation. *Advances in neural information processing systems*, 31, 2018.

Nanda, N., Chan, L., Lieberum, T., Smith, J., and Steinhardt, J. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*, 2023.

Olah, C. A toy model of mechanistic (un)faithfulness, August 2025. URL https://transformer-circuits.pub/2025/faithfulness-toy-model/index.html. Transformer Circuits.

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., and Carter, S. Zoom in: An introduction to circuits. *Distill*, 2020. doi: 10.23915/distill.00024.001. https://distill.pub/2020/circuits/zoom-in.

Park, K., Choe, Y. J., and Veitch, V. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023.

Power, A., Burda, Y., Edwards, H., Babuschkin, I., and Misra, V. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Schiffman, J. S. and Ralph, P. L. System drift and speciation. *Evolution*, 76(2):236–251, 2022.

Sharkey, L., Chughtai, B., Batson, J., Lindsey, J., Wu, J., Bushnaq, L., Goldowsky-Dill, N., Heimersheim, S., Ortega, A., Bloom, J., et al. Open problems in mechanistic interpretability. *arXiv preprint arXiv:2501.16496*, 2025.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Freeman, C. D., Sumers, T. R., Rees, E., Batson, J., Jermyn, A., Carter, S., Olah, C., and Henighan, T. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Varma, V., Shah, R., Kenton, Z., Kramár, J., and Kumar, R. Explaining grokking through circuit efficiency. *arXiv preprint arXiv:2309.02390*, 2023.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wagner, A. Robustness and evolvability: a paradox resolved. *Proceedings of the Royal Society B: Biological Sciences*, 275(1630):91–100, 2008.

Wagner, A. The role of robustness in phenotypic adaptation and innovation. *Proceedings of the Royal Society B: Biological Sciences*, 279(1732):1249–1258, 2012.

Wang, K., Variengien, A., Conmy, A., Shlegeris, B., and Steinhardt, J. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*, 2022.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pp. 38–45, 2020.

Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Tang, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., and Qiu, Z. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Zhong, Z., Liu, Z., Tegmark, M., and Andreas, J. The clock and the pizza: Two stories in mechanistic explanation of neural networks. *Advances in neural information processing systems*, 36:27223–27250, 2023.

Appendix

A. Algorithmic Core Extraction

Functional equivalence and minimal realizations. The structure–function relationship is often many-to-one (Bellman & Åström, 1970): there is more than one way to realize a behavior. But how many different structures can realize identical input–output functions? Can the space of functionally equivalent structures be characterized?

In linear system theory, this question has an exact answer (Kalman, 1962). Consider a linear time-invariant system with hidden state $\mathbf{x} \in \mathbb{R}^n$, input $\mathbf{u} \in \mathbb{R}^m$, and output $\mathbf{y} \in \mathbb{R}^\ell$:

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u},$$
$$\mathbf{y} = \mathbf{C}\mathbf{x}.$$

The system's input–output behavior is fully determined by its *impulse response* $\boldsymbol{\zeta}(t) := \mathbf{C}e^{\mathbf{A}t}\mathbf{B}$. Two systems with different weights $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ and $(\tilde{\mathbf{A}}, \tilde{\mathbf{B}}, \tilde{\mathbf{C}})$ are functionally equivalent if they produce identical outputs for all inputs – that is, if their impulse responses match ($\boldsymbol{\zeta}(t) = \tilde{\boldsymbol{\zeta}}(t)$).

For systems of equal dimension, functional equivalence corresponds exactly to coordinate change: $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ and $(\mathbf{V}\mathbf{A}\mathbf{V}^{-1}, \mathbf{V}\mathbf{B}, \mathbf{C}\mathbf{V}^{-1})$ share the same impulse response for any invertible $\mathbf{V}$. But systems of different sizes can also be functionally equivalent if some internal states are either unreachable (unaffected by input) or unobservable (irrelevant to the output).

The *Kalman decomposition* makes this precise, partitioning any system's state space into four subspaces according to

reachability and observability (Kalman, 1962; 1963; Anderson et al., 1966; Kalman et al., 1969). Only states that are both reachable and observable contribute to input–output behavior; the rest represent degrees of freedom that can vary without affecting function. This decomposition guarantees the existence of a *minimal realization* – the smallest-dimensional system that reproduces an input–output map, unique up to coordinate change – and enables extracting it. These results from system theory conceptually motivate the methods developed in this manuscript.

Algorithmic core extraction. The goal here is an analogous decomposition for transformers. The Kalman decomposition provides an exact algebraic characterization for linear systems; for transformers, no such closed-form decomposition exists, but the principle can be applied empirically: identify directions that are both input-driven (active) and output-relevant (relevant). If the system were linear, this would reduce to *balanced truncation* (Moore, 1981), a technique in model reduction that finds coordinates in which reachability and observability are aligned.

Here, **ACE** (*Algorithmic Core Extraction*) operationalizes this approach for artificial neural networks. Let $\mathbf{H} \in \mathbb{R}^{N\times D}$ denote mean-centered hidden activations at a transformer layer of interest, with rows $\mathbf{h}_i^\top \in \mathbb{R}^{1\times D}$ for each of N inputs, to define *active* directions. To quantify *relevant* directions, let $f\colon \mathbb{R}^D \to \mathbb{R}^K$ map activations to task-relevant outputs and let $\mathbf{J} \in \mathbb{R}^{NK\times D}$ stack the N Jacobians $\partial f/\partial \mathbf{h}_i \in \mathbb{R}^{K\times D}$ as row blocks.

To find directions that are jointly active and relevant, ACE computes the SVD of their interaction:

$$\mathbf{H}\mathbf{J}^\top \;=\; \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top.$$

The singular values quantify the joint importance of each direction, providing a principled criterion for rank selection. Let $\mathbf{U}_r$ denote the first r columns of $\mathbf{U}$. The *algorithmic core* is the subspace obtained by projecting these interaction modes back into activation space,

$$\mathcal{C} \;\coloneqq\; \mathrm{span}\big(\mathbf{H}^\top \mathbf{U}_r\big)\,.$$

The core's orthonormal basis $\mathbf{Q} \in \mathbb{R}^{D\times r}$ is given by the QR decomposition

$$\mathbf{H}^\top \mathbf{U}_r \;=\; \mathbf{Q}\mathbf{R},$$

and thus, the *core projector* is defined as $\mathbf{P} \coloneqq \mathbf{Q}\mathbf{Q}^\top$.

Note on implementation. Computing $\mathbf{H}\mathbf{J}^\top \in \mathbb{R}^{N\times NK}$ is unnecessary (and inefficient when $NK \gg D$). Instead form the activation covariance $\mathbf{A} \coloneqq \mathbf{H}^\top\mathbf{H} \in \mathbb{R}^{D\times D}$ and sensitivity matrix $\mathbf{S} \coloneqq \mathbf{J}^\top\mathbf{J} \in \mathbb{R}^{D\times D}$. Take square-root factors $\mathbf{A} + \varepsilon\mathbf{I} = \mathbf{L}\mathbf{L}^\top$ and $\mathbf{S} = \mathbf{\Gamma}\mathbf{\Gamma}^\top$, then compute the SVD of the resulting $D \times D$ matrix $\mathbf{L}^\top\mathbf{\Gamma} \;=\; \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, which yields the core subspace: $\mathrm{span}(\mathbf{L}\mathbf{U}_r)$.

Causal validation. The core is validated through ablation, with $\tilde{\mathbf{h}}$ denoting the activation after intervention:

$$\begin{aligned}\text{Core-only (to test sufficiency):}\quad & \tilde{\mathbf{h}} = \mathbf{P}\mathbf{h},\\ \text{Core-removed (to test necessity):}\quad & \tilde{\mathbf{h}} = \mathbf{h} - \mathbf{P}\mathbf{h}.\end{aligned}$$

A subspace is deemed *sufficient* if core-only preserves task performance, and *necessary* if core-removed reduces performance to approximately chance. The energy-based rank can be refined by finding the minimal r such that keeping only the core maintains baseline accuracy and removing it drops accuracy to near chance.

When activity and relevance align. It is worth noting that sometimes ACE reduces to standard PCA – when activity and relevance coincide. This is even expected for simple tasks, when there is no inherent pressure for models to "hide" computations in low-variance subspaces. In more complex models, however, high-variance directions are unlikely to cleanly align with target tasks. Still, the distinction matters even when the subspaces agree: PCA identifies where variance concentrates; ACE identifies where the input–output map flows, by construction and intervention, certifying causal relevance. In other words, PCA is descriptive and statistical, whereas ACE is also causal, licensing downstream treatment and interpretation of the returned subspace and its fitted operator as a dynamical system realizing a causal algorithm.

B. Markov Chain Experiment

Three single-layer transformers ($d_{\text{model}} = 64$, $d_{\text{ff}} = 256$, $|V| = 4$) with causal attention masking were trained with independent random seeds on next-token prediction for sequences generated by a four-state Markov chain.

The Markov chain transition probability matrix,

$$\mathbf{T} \coloneqq \begin{pmatrix} \alpha & \beta & 0 & 0\\ 0 & \alpha & \beta & 0\\ 0 & 0 & \alpha & \beta\\ \beta & 0 & 0 & \alpha \end{pmatrix},$$

was instantiated with $\alpha = 0.75$ and $\beta = 0.25$, yielding eigenvalues (spectrum) $\lambda \in \{1, 0.75 + 0.25i, 0.75 - 0.25i, 0.5\}$, and has stationary distribution $\boldsymbol{\pi} = [0.25, 0.25, 0.25, 0.25]$.

Training used `AdamW` with learning rate 10^{-3} and no weight decay for 40 epochs on 3,000 sequences of length 32 generated by $\mathbf{T}$, with batch size 64.

Trained model performance is compared against two base-

Table A1. Transformer Markov-chain test accuracy: `Full Model`: no ablations; `Core-only`: ablating non-core dimensions; `Core-removed`: ablating core (Methods). Data are plotted in **Fig. 1B**.

	Full Model	Core-only	Core-removed
M_1	0.748	0.748	0.261
M_2	0.748	0.748	0.237
M_3	0.748	0.748	0.247

Table A2. Pairwise core geometry and CCA similarity. Projector overlap is the squared Frobenius overlap between core subspaces; angles are principal angles (degrees); CCA lists canonical correlations. Overlap and mean CCA are visualized in **Fig. 1C,D**.

Pair	Proj. Overlap	Principal Angles	CCA
M_1–M_2	0.027	[78, 80, 85]	[0.999, 0.999, 0.927]
M_1–M_3	0.031	[76, 80, 85]	[0.999, 0.999, 0.949]
M_2–M_3	0.027	[76, 82, 89]	[0.999, 0.999, 0.958]

lines:

$$\text{Chance:} \quad \max(\boldsymbol{\pi}),$$
$$\text{Bayes-optimal:} \quad \sum_i \pi_i \max_j T_{ij},$$

Chance accuracy reflects always predicting the most common token; Bayes-optimal accuracy reflects the best possible one-step prediction given the stochastic nature of the chain.

Algorithmic cores were extracted using a 99.9% rank energy threshold without ablation-refinement, $\mathbf{H}$ was computed for all test activations, and $\mathbf{J}$ was defined by the target function $f(\mathbf{h}) \coloneqq \text{logits}(\mathbf{h})$.

Fitting dynamics. Hidden state (mean-centered) sequences were projected into core coordinates $\mathbf{z}_t = \mathbf{Q}^\top \mathbf{h}_t$ and a linear operator was fit by least squares to predict next-step dynamics,

$$\mathbf{z}_{t+1} \approx \mathbf{A}\mathbf{z}_t.$$

The spectrum of $\mathbf{A}$ was used to characterize the learned dynamics. When comparing fitted operators in the core to ground truth, the Perron–Frobenius eigenvalue $\lambda = 1$ of $\mathbf{T}$ (corresponding to the stationary distribution) is excluded, as it reflects normalization.

To calibrate, core operator fits were compared against an *oracle* ceiling for next-token prediction:

$$R^2_{\text{oracle}} \coloneqq 1 - \frac{1}{|V|}\,\mathbf{1}^\top(\mathbf{m} \oslash \mathbf{v})\,,$$

where

$$\mathbf{m} \coloneqq \text{diag}(\boldsymbol{\pi})\big(\mathbf{T} \odot (\mathbf{1} - \mathbf{T})\big)^\top \mathbf{1}, \qquad \mathbf{v} \coloneqq \boldsymbol{\pi} \odot (\mathbf{1} - \boldsymbol{\pi}),$$

and $\odot$ and $\oslash$ denote elementwise product and division.

C. Modular Addition Experiment

Three two-layer transformers ($d_{\text{model}} = 128$, $d_{\text{ff}} = 512$, $|V| = 53$) were trained on $a + b \equiv c \pmod{53}$. The dataset consists of all $53^2 = 2809$ input pairs, split evenly into train and test sets with a fixed random seed. Input sequences are $[a, b]$ with target $[b, c]$.

Training used `AdamW` with learning rate 10^{-3}, batch size 512, and weight decay $\omega = 1$. Models were trained for 2×10^4 epochs, with core extraction performed every 100 epochs. The grokking epoch was defined as the first analysis time point at which all three models achieved perfect test accuracy, which occurred at epoch 800.

To study the effect of continued weight decay after grokking, at epoch 900, transformers were "branched" – duplicated and split into two regimes – where weight decay was either maintained at $\omega = 1$ or disabled ($\omega \to 0$) for the remainder of training.

For core extraction, $\mathbf{H}$ was computed over all test-set activations, and $\mathbf{J}$ was estimated using 64 Jacobian samples, defined by the target function $f(\mathbf{h}) \coloneqq \text{logits}(\mathbf{h})$. Core rank was selected first via the 99% energy threshold, and then refined with ablations to ensure causal importance.

For operator fitting, centroids $\bar{\mathbf{r}}_c$ were computed as the centered mean core activation over all test examples with answer token c. A linear shift operator $\mathbf{A}$ satisfying,

$$\bar{\mathbf{r}}_{(c+1) \bmod 53} \approx \mathbf{A}\bar{\mathbf{r}}_c$$

was fit by ridge-regularized least squares after dimensionality reduction with SVD. Generalization was evaluated by holding out cycle transitions rather than examples: the 53 answer classes were split into disjoint calibrate/evaluate sets by selecting a contiguous block of classes for the evaluate set, and the fit was performed only on transitions $c \to c{+}1$ whose endpoints both lie in the calibration class set; eval-

Table A3. Eigenvalues $\{\lambda_i\}$ from operators fit in transformer cores compared with those from the Markov transition probability matrix $\mathbf{T}$ (excluding the Perron–Frobenius eigenvalue). Spectral overlap is visualized in **Fig. 1E**.

	Markov chain	Core_1	Core_2	Core_3
λ_1	$0.75 + 0.25i$	$0.75 + 0.25i$	$0.75 + 0.25i$	$0.75 + 0.25i$
λ_2	$0.75 - 0.25i$	$0.75 - 0.25i$	$0.75 - 0.25i$	$0.75 - 0.25i$
λ_3	0.50	0.51	0.49	0.48

uation used only transitions whose endpoints both lie in the evaluate class set and fit is denoted as R_h^2. For descriptive fits, R^2 is reported without holding out transitions or ridge-regularization.

To summarize spectral structure, eigenvalues of $\mathbf{A}$ with magnitude close to 1 were identified as rotational modes, and each such mode was assigned a frequency bin by rounding its angle to the nearest integer multiple of $2\pi/53$. Because complex-conjugate eigenvalue pairs correspond to the same oscillation up to direction, bins k and $53-k$ were mapped to the same bin. This implies a maximum of $\lfloor 53/2 \rfloor + 1 = 27$ distinct bins: one $k = 0$ bin and 26 nonzero oscillatory bins. Mode count is defined as the number of occupied nonzero bins, and derives from operators fit without holding out transitions, since the goal is descriptive characterization rather than generalization evaluation.

D. Subject–Verb Agreement Experiment

GPT-2 Small (117M parameters, 12 layers), Medium (345M, 24 layers), Large (774M, 36 layers), LlaMA-3.1 (8B, 32 layers), Gemma-2 (9B, 42 layers), and Qwen2.5 (32B, 64 layers) were analyzed on subject–verb number agreement.

Prompts. A dataset of 1,200 prompts (600 singular, 600 plural) was constructed by combining head nouns (for example, "key"/"keys", "child"/"children") with attractor nouns of opposite number (e.g., "cabinets"/"cabinet") via connectors ("to the", "near the", "next to the", etc.). Five syntactic templates were used: base ("The key to the cabinets"), front-padded ("In this ancient kingdom, the key to the cabinets"), back-padded ("The key to the cabinets in the old kingdom"), existential ("There key near the boxes"), and relative clause ("The key that guards the cabinets"). Half of prompts were prefixed with "In the past," to vary tense context. The dataset was split evenly into train and test sets. Note: some prompts deliberately employ ungrammatical word order (*e.g.*, "There key near the boxes") to assess whether the agreement core remains robust to structural violations, forcing the model to resolve agreement based on the head noun rather than positional heuristics.

Target function. The number margin was defined on the final-token hidden state $\mathbf{h}$:

$$f(\mathbf{h}) \coloneqq (\text{logit}_{\textit{are}} + \text{logit}_{\textit{were}}) - (\text{logit}_{\textit{is}} + \text{logit}_{\textit{was}}).$$

Layer sweep. Candidate cores were extracted at each layer and evaluated via ablation. For each model, the layer with maximal flip effect was selected as the core location.

Adaptive generation steering. For open-ended generation, a per-token adaptive intervention was applied during autoregressive decoding. Let $\mathbf{q} \in \mathbb{R}^D$ denote the (unit-norm) core axis and $\boldsymbol{\mu}$ the mean activation at the intervention layer. The intervention reflects the hidden state $\mathbf{h}$ at the last token position through the hyperplane orthogonal to the core axis:

$$\tilde{\mathbf{h}} = \mathbf{h} - 2s\left[(\mathbf{h} - \boldsymbol{\mu})^\top \mathbf{q}\right]\mathbf{q},$$

where s is a per-token steering strength determined adaptively.

At each decoding step, three forward passes are performed. First, a *gating* check: a clean forward pass (with $s = 0$) computes the softmax probability mass on the agreement-relevant verb tokens (*is*, *are*, *was*, *were*). If this mass falls below a threshold, the token is unlikely to involve an agreement decision and no intervention is applied ($s^* = 0$).

Otherwise, the steering strength is calibrated to produce a minimal margin flip. Define the *generation margin* as $m \coloneqq \log\sum_{v\in\{\textit{are, were}\}} e^{\ell_v} - \log\sum_{v\in\{\textit{is, was}\}} e^{\ell_v}$, where ℓ_v denotes the logit for token v. This logsumexp margin more accurately reflects the probability-space competition between singular and plural verb groups than the linear logit sum used for core extraction, where operating-point independence of the Jacobian is preferred. The calibration proceeds as: (1) the current margin m_0 is measured under the clean pass; (2) a small probing perturbation at strength s_0 estimates the local gain $g = (m_1 - m_0)/s_0$; (3) the intervention strength is set to $s^* = (m_{\text{target}} - m_0)/g$, where $m_{\text{target}} = -\operatorname{sign}(m_0)\,\varepsilon$ targets the minimal margin crossing with buffer ε. An optional cap $|s^*| \le s_{\text{cap}}$ prevents extreme extrapolation. This adaptive approach produces grammatical inversions while minimizing collateral disruption to non-agreement tokens.

Table A4. Similarity of core coordinates (from **Fig. 5B**) across six language models. Spearman's ρ measures rank correlation. Pearson's correlation *r* measures linear relatedness; its magnitude equals CCA for one dimension.

Pair	Spearman's ρ	Pearson's r ($\mathrm{CCA}_{1\mathrm{D}}$)
Qwen2.5 × LLaMA-3.1	0.921	0.924
Qwen2.5 × Gemma-2	0.934	0.943
Qwen2.5 × GPT-2 Large	0.844	0.857
Qwen2.5 × GPT-2 Medium	0.829	0.842
Qwen2.5 × GPT-2 Small	0.701	0.757
LLaMA-3.1 × Gemma-2	0.888	0.880
LLaMA-3.1 × GPT-2 Large	0.760	0.752
LLaMA-3.1 × GPT-2 Medium	0.727	0.718
LLaMA-3.1 × GPT-2 Small	0.585	0.616
Gemma-2 × GPT-2 Large	0.893	0.909
Gemma-2 × GPT-2 Medium	0.894	0.911
Gemma-2 × GPT-2 Small	0.790	0.846
GPT-2 Large × GPT-2 Medium	0.923	0.951
GPT-2 Large × GPT-2 Small	0.878	0.924
GPT-2 Medium × GPT-2 Small	0.919	0.968

Table A5. A one-dimensional subject–verb agreement core was extracted from each model, despite massive variation in training, model parameterizations, and architectures (model dimension, number of layers). Extracted core size (d_{core}) is supported by the large *spectral gap* (ratio of largest two singular value squares σ_1^2/σ_2^2). The large spectral gaps indicate that these subspaces are effectively one-dimensional.

Model	Parameters	Layers	d_{model}	d_{core}	Spectral gap	Core location (layer)
GPT-2 Small	117 M	12	768	1	40	11
GPT-2 Medium	345 M	24	1024	1	44	22
GPT-2 Large	774 M	36	1280	1	2.8×10^{10}	36
LLaMA-3.1	8 B	32	4096	1	266	28
Gemma-2	9 B	42	3584	1	133	42
Qwen2.5	32 B	64	5120	1	531	62

Table A6. Agreement performance (AUC; 1 = perfect, 0.5 = chance, 0 = inverted) under core ablations across all six models. Core-only preserves agreement (sufficiency), core-removed collapses it below chance (necessity), and core-flipped inverts grammatical number preferences (induces near perfect *disagreement*).

Model	Baseline	Core-only	Core-removed	Core-flipped
GPT-2 Small	0.911	0.994	0.241	0.038
GPT-2 Medium	0.934	0.997	0.217	0.023
GPT-2 Large	0.921	0.975	0.244	0.021
LLaMA-3.1	0.808	0.918	0.209	0.092
Gemma-2	0.886	0.978	0.102	0.035
Qwen2.5	0.890	0.949	0.213	0.076

E. Grokking Dynamics

E.1. Mathematical Model

Let $\boldsymbol{\alpha}(t) \in \mathbb{R}^\mu$ denote the mode coefficients and let $\boldsymbol{\psi} \in \mathbb{R}^\mu$ be fixed with $\psi_k := 1 - \cos(2\pi k/p)$. Define the (test-relevant) margin $m(t) := \langle \boldsymbol{\alpha}(t), \boldsymbol{\psi} \rangle$.

Post-memorization, training loss is approximately zero. Updates are driven by the weight decay penalty $-\omega\boldsymbol{\alpha}(t)$ and a minimal corrective motion $\gamma(t)\boldsymbol{\psi}$ needed to remain on the zero-loss manifold, plus zero-mean stochasticity $\xi(t)$ (optimizer noise).

Direction of $\boldsymbol{\psi}$. Among all infinitesimal updates $\Delta\boldsymbol{\alpha}$ that increase the margin by one unit, the minimum-norm update solves

$$\arg\min_{\Delta\boldsymbol{\alpha}} \|\Delta\boldsymbol{\alpha}\|_2 \quad \text{s.t.} \quad \langle \Delta\boldsymbol{\alpha}, \boldsymbol{\psi} \rangle = 1.$$

By the Cauchy–Schwarz inequality, the solution is $\Delta\boldsymbol{\alpha} = \boldsymbol{\psi}/\|\boldsymbol{\psi}\|_2^2$. Thus, the corrective gradient direction is strictly parallel to $\boldsymbol{\psi}$.

Margin dynamics. Differentiating $m(t) = \langle \boldsymbol{\alpha}(t), \boldsymbol{\psi} \rangle$ and isolating the noise-free deterministic trajectory yields the scalar ODE:

$$\dot{m}(t) \;=\; -\omega\, m(t) \;+\; \gamma(t)\, \|\boldsymbol{\psi}\|_2^2.$$

Because weight decay is the only systematic drift pulling the network off the margin, the mean corrective force is taken to scale proportionally to maintain zero loss: $\gamma(t) \approx c\,\omega$ for some constant c.

Assuming sufficient dimensional capacity ($p < d_{\text{model}}$), the initial memorized state is unstructured, meaning it carries negligible margin ($m(0) \approx 0$). Substituting the corrective force yields a simple linear relaxation equation:

$$\dot{m}(t) = -\omega m(t) + c\,\omega \|\boldsymbol{\psi}\|_2^2, \qquad m(0) \approx 0.$$

Solving this ODE yields the exact continuous-time margin trajectory:

$$m(t) = m^*\big(1 - e^{-\omega t}\big), \qquad \text{where} \quad m^* = c\,\|\boldsymbol{\psi}\|_2^2 = \kappa p.$$

Predicting grokking time. Grokking occurs at the first-hitting time τ when the margin reaches the generalization threshold δ. Solving $m(\tau) = \delta$ yields the continuous gradient-flow time:

$$\tau(p) = -\frac{1}{\omega} \log\left(1 - \frac{\delta}{\kappa p}\right).$$

To map this idealized ODE to discrete training steps, the physical constants are decoupled. The scaling rate becomes:

$$\tau_{\text{grok}}(p) = -\Omega \log\left(1 - \frac{p_{\text{crit}}}{p}\right), \qquad (p_{\text{crit}} < p < d_{\text{model}}).$$

Here, $p_{\text{crit}} := \delta/\kappa$ is the *architectural constant*, defining the absolute capacity floor limit independent of the optimizer. Conversely, $\Omega \propto (\eta\omega)^{-1}$ is the *optimizer constant*, an empirical parameter that captures the characteristic relaxation time while absorbing discrete step-size dynamics, learning rate, momentum, and adaptive preconditioning from the `AdamW` optimizer.

E.2. Grokking Sweeps and Scaling Fits

To measure scaling laws for the grokking delay in modular addition $a + b \pmod p$, one-layer transformers ($d_{\text{model}} = 128$, $d_{\text{ff}} = 512$) were trained on input pairs using `AdamW` (`lr=1e-3`). The data, p^2 input pairs, were randomly partitioned into a $50/50$ train/test split. Memorization (τ_{mem}) and generalization (τ_{gen}) times were defined as the first optimizer steps at which train and test accuracy reach 0.99, respectively. The grokking delay is evaluated as the difference $\tau_{\text{grok}} := \tau_{\text{gen}} - \tau_{\text{mem}}$. Accuracy was evaluated every step to avoid quantization artifacts.

Two sweeps were performed, averaging over 12 random seeds per condition: (1) *Weight decay:* fixing $p = 53$ and sweeping $\omega \in \{0.3, 0.5, 1, 1.5, 2, 3\}$. To simulate standard training stochasticity on a fixed-size dataset, this sweep utilized minibatch gradient descent with a batch size of $B = 512$. (2) *Modulus:* fixing $\omega = 1$ and sweeping primes $p \in \{31, 43, 53, 61, 79, 89, 101\}$. Because the dataset size grows quadratically with p, this sweep utilized full-batch gradient descent to ensure the empirical hitting time was isolated from dataset-dependent minibatch noise.

Scaling exponents for the asymptotic limits were obtained by fitting power laws $y = Cx^\beta$ via ordinary least squares in log–log space. The macroscopic constants Ω and p_{crit} were obtained by fitting the exact deterministic ODE solution $\tau(p) = -\Omega \log(1 - p_{\text{crit}}/p)$ to the empirical delay using non-linear least squares (`scipy.optimize.curve_fit`). Goodness-of-fit for all curves is reported by R^2.