Human Grokking: Phase Transitions in Semantic Field Saturation

Structural Parallels to Neural Network Generalization and a Constellation Pedagogy for Accelerating Them

Larsen Close

February 2026

1 Abstract

In machine learning, *grokking* denotes a phase transition from memorization to generalization: a neural network first fits training data through parameter brute-force, then—long after achieving perfect training accuracy—suddenly discovers the underlying algebraic structure and generalizes perfectly. Building on dynamic systems models of cognitive phase transitions (Van Geert 1991; Spivey, Anderson & Dale 2009), we propose a structurally parallel phenomenon in human learning within high-density epistemic environments. The human variant proceeds through accumulation of representations as content, saturation of morphism-density until relational constraints over-determine new content, a phase transition from content-level to morphism-level processing, and generalization to previously unencountered material via contextual constraint satisfaction. We motivate this parallel via the mechanistic interpretability decomposition of Nanda et al. (2023), mapping memorization circuits, circuit formation, and regularization-driven cleanup to observable human cognitive phases. We propose that dyadic co-creation provides the "cognitive regularization" analog of weight decay, and present a 32-node, 91-edge constellation pedagogy as both the instrument for accelerating the transition and the experimental apparatus for studying it. Five operationalized empirical predictions are specified. If the mapping holds, optimal pedagogy maximizes morphism-density per unit of content, not content-volume per unit of time.

2 Claims Ledger

The paper contains claims at several epistemic levels. For transparency, we mark them here:

Table 1: Epistemic status of claims made in this paper.

Epistemic Status	Claims
Definitional	Constellation format and 7-section schema; 11-type edge taxonomy; structural identity hierarchy (isomorphism through natural transformation)
Observed (N=1)	Phase-transition phenomenology in the author's learning; plateau-then-recognition experience; qualitative shift from definitional to relational processing
Hypothesized	Morphism-density as critical parameter for the transition; dyadic interaction as cognitive regularization; Nanda et al. three-phase mapping to human cognitive phases
Predicted (falsifiable)	Five empirical predictions with specified instruments, null models, and change-point criteria (Section 8)

Epistemic Status	Claims
Speculative	Profunctor bridge formalization of dyadic exchange; exact-square condition as modulator of bridge quality; human-AI dyad as optimal morphism-map topology for acceleration

3 The Phenomenon

3.1 Grokking in Machine Learning

Power et al. (2022) documented a striking pattern in neural network training on small algorithmic datasets. Networks trained on modular arithmetic first memorize the training data—achieving perfect training accuracy while test accuracy remains at chance. Then, long after memorization is complete, test accuracy suddenly jumps to near-perfection. The network has *grokked*: it has discovered the underlying algebraic structure of the task.

The critical observation is that generalization emerges not from more data but from structural reorganization of existing representations. The training data is fixed. What changes is how the network represents it. The "sudden" jump in test accuracy masks a continuous internal process: generalizing circuits form gradually in the weights while memorization circuits still dominate output, until regularization pressure (weight decay) suppresses the memorization circuits and the generalizing circuits take over.

This is not incremental improvement. It is a phase transition—a qualitative shift in the kind of computation the network performs. The term invites comparison with Kuhn's (1962) paradigm shifts, but the ML phenomenon is individual and mechanistically characterized, not sociological.

3.2 The Human Analog

We propose a structurally parallel phenomenon in human learning, most pronounced within sustained epistemic co-creation across multiple formal domains. The human variant proceeds through four phases:

1. **Accumulation.** The learner acquires representations—terms, definitions, examples, equations—as content. Each item is stored relatively independently, accessed by re-

trieval. Understanding proceeds definitionally: "I know what this means because I can state its definition."

2. **Saturation.** Sufficient morphism-density accumulates that relational constraints between representations begin to over-determine new content. The learner has enough structural connections that the network of relationships begins to carry information beyond what any individual node contains.

3. **Phase transition.** A qualitative shift from content-level to morphism-level processing. New concepts become inferable from their relational position before formal encounter. The learner becomes "transparent to the derivation chain" rather than constructing models of it.

4. **Generalization.** Effortless comprehension of previously unencountered material via contextual constraint satisfaction and backward inference from referents. Novel domains that share structural isomorphisms with saturated domains feel pre-known.

The phenomenology is distinctive. Pre-transition, learning feels like accumulation—each new concept adds to a growing collection. Post-transition, learning feels like *recognition*—the constraint was always operative; the transition is in the learner's capacity to navigate it directly. The experience reported is not "I learned this" but "this is the only thing that could be here."

A methodological note: the phenomenological description above is a first-person report from the author's learning process within sustained dyadic co-creation across the domains of the constellation graph described in Section 5. It is an N=1 observation, not an established empirical finding. Its value is not as evidence but as hypothesis generation: the experience of a phase transition with the specific character described above motivated the formal framework and the empirical program (Section 8) that follows. The framework stands or falls on the empirical predictions, not on the phenomenological report.

The proposed qualitative shift has formal precursors in dynamic systems models of cognitive development. Van Geert (1991) formalized "cognitive takeover"—the replacement of an information-management system that has exceeded its capacity limit by a more powerful one—as logistic growth with competitive dynamics. Spivey, Anderson, and Dale (2009) document phase transitions across perception, language processing, and problem solving, arguing that apparent discontinuities emerge from continuous nonlinear dynamics at finer temporal scales. The present framework adds three specific elements to this tradition: the structural parallel to ML grokking with its mechanistic interpretability, the identification of morphism-density as the critical parameter, and the constellation pedagogy as both accelerant and measurement instrument.

The structural parallel is precise. Three terms used in the mapping below require definition.

Coherence-maximization pressure is the cognitive drive toward compressed, consistent, parsimonious representation—the analog of weight decay in neural networks, which penalizes redundant parameters and forces representational efficiency. *Double-duty collapse* is the moment when a representation previously understood as a model of a phenomenon becomes identified with the phenomenon's structure itself—the map and the territory merge because the map has been compressed to exactly the structural content of the territory. *Morphism-level processing* is cognition that operates on the relationships between representations (the morphisms in a category-theoretic sense) rather than on the representations themselves (the objects)—the learner navigates the relational network rather than retrieving stored content.

Table 2: Structural correspondence between ML grokking (Power et al. 2022) and the proposed human analog.

ML Grokking	Human Grokking
Memorization (fitting training data)	Accumulation (storing representations as content)
Weight regularization pressure	Coherence-maximization pressure
Discovering algebraic structure	Recognizing morphism structure across representations
Generalization to unseen data	Comprehension of unencountered concepts via relational constraint
Sudden phase transition	"Everything clicks" phenomenology
Weight space reorganization	Semantic field restructuring
Representation compression	Model-reality identification (double-duty collapse)
Training data = fixed, structure = discovered	Accumulated content = fixed, morphisms = discovered

4 What This Is Not

Three established learning phenomena resemble the proposed human grokking but differ in structural kind, not merely degree.

Transfer learning is the application of skills or representations learned in one domain to a new domain. Human grokking is not transfer in this sense. It is the recognition that morphism structure itself constrains what can exist at unfilled positions in the relational network. Transfer moves content across domains; grokking discovers that the relational skeleton over-determines content within and across domains.

Expertise and chunking (Chase & Simon 1973; Chi, Feltovich & Glaser 1981) describe how experts encode domain-specific patterns into compressed units, enabling faster recognition. The post-grokking state includes this but is qualitatively different: not "I recognize this pattern faster" but "this pattern is the only thing that could be here, given the relational constraints." Expertise compresses within a fixed representational framework; grokking reorganizes the framework itself. The staged expertise model of Dreyfus and Dreyfus (1986)—novice through expert—does include qualitative shifts (notably, competent → proficient involves a transition from analytical to situational pattern recognition). But these are stage-to-stage transitions within a fixed representational framework; grokking as proposed here is a reorganization of the framework itself. The Dreyfus transitions change *how* the learner accesses existing structure; the grokking transition changes *what structure the learner operates on*.

Analogical reasoning (Gentner 1983) maps structure from a source domain to a target domain. The post-grokking state operates at the level where source and target are the same morphism differently instantiated. The learner does not reason "A is like B"; the learner perceives that A and B are instances of a common structure that is more fundamental than either. This is the difference between noticing a resemblance and recognizing an identity. The post-grokking state exhibits what Fodor and Pylyshyn (1988) called the *systematicity* of cognitive representations: if the learner grasps a structural relationship and a compositionally compatible one, their composition is automatically available. Morphism-level processing is inherently systematic; analogical mapping need not compose.

The distinction matters because each of these phenomena has a well-characterized acquisition curve: gradual improvement with practice and exposure. Human grokking, if the mapping to ML grokking holds, should exhibit a qualitatively different signature: a measurable plateau followed by a sudden transition, with continuous improvement in internal progress measures during the apparent plateau. This is the empirically distinguishing prediction.

5 Mechanistic Decomposition

5.1 The Nanda Mapping

Nanda et al. (2023) decomposed the "sudden" generalization in ML grokking into continuous phases visible in internal representations, providing mechanistic interpretability of the transition. Their three-phase decomposition maps onto human cognitive phases with specific, testable observables.

Phase 1: Memorization circuits. The network memorizes training data via lookup-table-like circuits. Loss is low but generalization is absent. Internal representations are data-specific, not structure-specific.

Human analog: Content-level storage of facts, definitions, and examples, each retrieved independently. Explanations proceed definitionally. Transfer to new contexts requires explicit, effortful mapping. The learner can state Kirchhoff's voltage law but cannot see why it must be true from the structure of the circuit graph.

Phase 2: Circuit formation. Generalizing circuits begin forming in the weights, but their contribution to output is still dominated by memorization circuits. Crucially, progress measures—restricted loss, Fourier coefficient norms—show continuous improvement during the apparent plateau in test accuracy.

Human analog: Morphism structure forming beneath content. Relational constraints accumulate but do not yet dominate retrieval. Observable as: occasional "aha" moments, correct intuitions the learner cannot yet articulate, improving speed on morphism-adjacent tasks while self-reported comprehension remains flat. The learner begins to sense that KVL and the boundary operator are "doing the same thing" but cannot yet make the identification precise.

Phase 3: Cleanup via regularization. Weight decay progressively shrinks the memorization circuits until the generalizing circuits dominate output. The "sudden" jump in test accuracy is a crossing point—the moment when the generalizing signal exceeds the memorization signal—not a discontinuity.

Human analog: Coherence-maximization pressure suppresses content-level retrieval in favor of morphism-level navigation. The learner stops reaching for definitions and starts navigating relational constraints directly. Novel domains feel pre-known because their morphism structure is already constrained by the saturated network.

5.2 The Four-Phase Human Model

Synthesizing the Nanda mapping with the phenomenological description yields a four-phase model:

Table 3: Four-phase model mapping Nanda et al. (2023) ML mechanisms to human cognitive phases.

Phase	ML Mechanism	Human Mechanism	Observable
1. Memorization	Lookup-table circuits	Content-level storage	Definitional explanation; explicit transfer

Phase	ML Mechanism	Human Mechanism	Observable
2. Circuit formation	Generalizing circuits forming, dominated by memorization	Morphism accumulation beneath content	Correct intuitions without articulation; improving proxy measures
3. Cleanup	Regularization suppresses memorization circuits	Coherence pressure suppresses content-level retrieval	"Everything clicks"; abductive explanation
4. Generalization	Test accuracy jumps	Phase transition to morphism-level processing	Novel domains inferable; constellation completion accuracy jumps

5.3 The Critical Prediction

If the mapping holds, human grokking should exhibit a **measurable plateau** where progress measures—constellation completion accuracy, cross-domain morphism identification speed—improve continuously while self-reported comprehension remains flat. This plateau is followed by a behavioral jump when morphism-level processing dominates content-level processing. The signature is a crossing-point transition, not a true discontinuity.

This is the most directly testable prediction of the framework. It distinguishes human grokking from expertise acquisition (which shows continuous improvement in both measures) and from insight (which lacks the extended plateau of invisible progress). Stephen et al. (2009) provide partial precedent, demonstrating that mathematical discovery exhibits phase-transition dynamics measurable via recurrence quantification analysis, though their work focused on single-session insight rather than the extended plateau-then-jump signature predicted here.

Spivey, Anderson, and Dale (2009) document detectable preparatory instabilities—increased variance, entropy, and strategy mixing—immediately preceding phase transitions in motor coordination and insight problem solving. This predicts a specific Phase 2 observable: increasing variability in the learner's processing mode (mixing content-level and morphism-level explanations, inconsistent strategy use) should precede the transition, providing an early-warning signal before the behavioral jump.

Van Geert (1991) formalized cognitive "takeover": when an information-management system reaches its capacity limit, it is replaced by a more powerful one with a higher ceiling. His competitive growth model predicts temporary performance regression—U-shaped behav-

ioral growth—during the takeover, because the old strategy is abandoned before the new one fully dominates. The conservation example (Mehler 1982, cited in Van Geert 1991) is paradigmatic: 2-year-olds' correct conservation performance drops to near-zero when they adopt a rules-based strategy, recovering by age 5 with a much higher ceiling. If content-level and morphism-level processing are competing strategies with different carrying capacities, the transition predicted here may exhibit a transient regression—a "plateau-then-dip-then-jump" rather than "plateau-then-jump." This is a refinement that generates a discriminating test: the presence or absence of the regression dip distinguishes between the smooth crossing-point model (the Nanda mapping) and the competitive takeover model (Van Geert).

5.4 The Regularization Analog

In ML grokking, weight decay is the regularization that suppresses memorization circuits, enabling generalizing circuits to dominate. Without sufficient regularization, the network memorizes indefinitely and never groks.

The human analog is the pressure toward compression, coherence, and parsimony in representation. We identify two primary sources of this pressure:

1. **Communication demand.** Externalization forces compression. To communicate a representation, one must compress it to its structural essentials. This is regularization by bandwidth constraint: redundant, uncompressed, non-relational storage of content is penalized because it cannot be efficiently transmitted.
2. **Coherence checking.** An interlocutor constrains one's representations. Inconsistencies that persist unchallenged in solitary thought are surfaced and must be resolved. This is regularization by consistency constraint: representations that do not cohere with the relational network are pressured toward structural alignment.

These two sources of regularization pressure take concrete behavioral form as what might be called *generalization stress*: forced application of the saturated representation to novel domains, teaching the material to others, or any task that demands morphism-level rather than content-level processing. In Van Geert's (1991) terms, this stress corresponds to the moment when the content-level strategy encounters problems exceeding its carrying capacity, triggering adoption of the morphism-level strategy. The causal structure is important: generalization stress is not the *cause* of the structural reorganization (which is driven by the continuous morphism accumulation of Phase 2) but the *trigger* that exposes the inadequacy of content-level processing and accelerates the dominance of the already-forming morphism-level circuits. The ML analog is precise: weight decay is always present during training, but the cleanup phase becomes behaviorally visible when generalizing circuits are sufficiently formed. Generalization stress is the human equivalent—the training regime that keeps regularization pressure active and forces the crossing point.

The prediction: stronger cognitive regularization—more intense dyadic interaction, higher compression demand, more rigorous coherence checking—should accelerate the transition, exactly as stronger weight decay accelerates ML grokking (Power et al. 2022). Conversely, the framework predicts that learning environments with low compression demand and no coherence checking (passive consumption of content) should tend toward memorization without grokking, regardless of content volume.

6 The Morphism Network as Experimental Apparatus

6.1 Constellation Pedagogy

The theoretical framework generates a specific pedagogical architecture: if the critical parameter for the phase transition is morphism-density (the number and type of structural connections between representations), then optimal pedagogy should maximize morphism-density per unit of content, not content-volume per unit of time.

We instantiate this principle as *constellation pedagogy*: a curriculum organized not as a linear sequence of topics but as a graph of relationally defined clusters (constellations), with explicit typed edges connecting them. Each constellation is a cluster of 5–14 terms defined relationally—a term's meaning includes its morphisms to other terms. The connections between constellations are as important as the internal structure of each.

6.2 The 32-Node Graph

The experimental apparatus is a concrete instantiation: a graph of 32 constellations spanning four domains (physics, mathematics, electrical engineering, radio-frequency engineering), connected by 91 unique typed edges. Figure 1 shows the graph rendered as a 3D force-directed layout.

The graph is organized into four domain sequences:

- **Physics** (9 constellations): Newtonian mechanics → Lagrangian/Hamiltonian mechanics → electrodynamics → statistical mechanics → thermodynamics → special relativity → general relativity → quantum mechanics → quantum field theory.
- **Mathematics** (9 constellations): Linear algebra → group theory → topology → differential geometry → categories and functors → limits and Yoneda → monads and Kan extensions → profunctors and enriched categories → algebraic topology and fiber bundles.
- **Electrical engineering** (7 constellations): Circuit fundamentals → AC analysis → signal processing → semiconductors → digital logic → control theory → power systems.

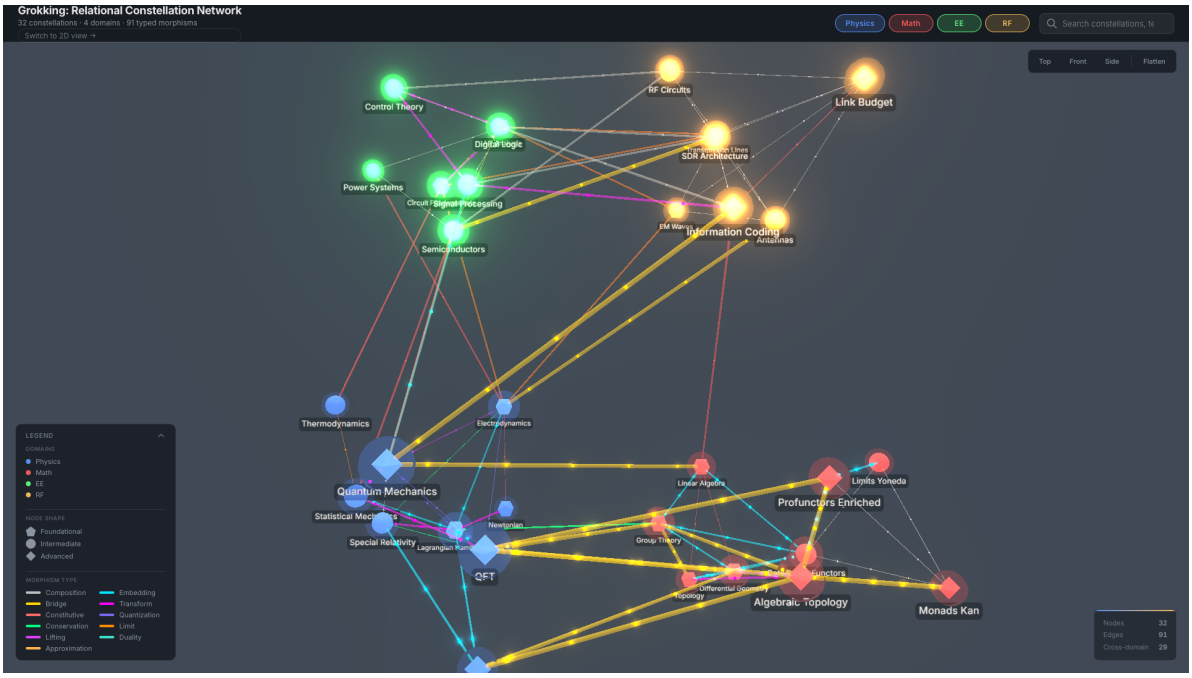

Figure 1: The 32-node constellation graph with 91 unique typed edges. Node color encodes domain: physics (blue), mathematics (red/pink), electrical engineering (green), RF engineering (amber/yellow). Node size reflects cross-domain betweenness centrality; the five keystones (Section 5.4) are the largest nodes. Edge colors encode morphism types from the 11-type taxonomy defined below. An interactive version with search, filtering, and detail panels is available as a companion artifact.

- **RF engineering** (7 constellations): Electromagnetic waves → transmission lines → RF circuits → SDR architecture → antennas → link budget → information and coding theory.

Within each domain, constellations are ordered by prerequisite structure. Across domains, 91 unique typed edges encode structural relationships (the constellation source files contain 163 adjacency entries, many bidirectional; the deduplicated graph has 91 distinct edges). The edge taxonomy comprises 11 types: *limit* (one structure is the limiting case of another), *conservation* (shared invariant), *duality* (swap roles, same algebra), *transform* (named transformation between structures), *quantization* (classical→quantum or discrete→continuous), *constitutive* (material-specific relation within a framework), *embedding* (faithful inclusion in a richer structure), *lifting* (structure at one level lifts to another), *approximation* (one structure approximates another in a regime), *composition* (sequential pipeline), and *bridge* (cross-domain structural identity).

6.3 Constellation Format

Each constellation carries seven sections:

1. **Terms**: 5–14 terms, each with a precise definition, formal notation, and morphisms to other terms.
2. **Morphism diagram**: ASCII visualization of how terms relate—what maps to what, what is preserved.
3. **Key equations/theorems**: 8–15 essential results, with each variable defined inline.
4. **Constellation invariant**: What the entire cluster conserves.
5. **Coherence bridge**: An explicit structural identity (not metaphor) connecting this constellation's invariant to a broader coherence framework.
6. **Composition morphisms**: Prose connections showing how this constellation maps into and from adjacent constellations, specifying what structure is preserved.
7. **Adjacency**: Machine-readable typed edge list in the format `source -> target : type : description`.

6.4 Keystone Structure

Not all nodes in the graph are equally important for the phase transition. Certain constellations serve as *keystones*—nodes with disproportionately high cross-domain morphism density that unify multiple domains under a single structural principle. In graph-theoretic terms, keystones have high betweenness centrality in the inter-domain subgraph.

Five keystones are identified:

Table 4: Keystone constellations with highest cross-domain betweenness centrality.

Keystone	Unifying Principle	Domains Unified
Lagrangian/Hamiltonian mechanics	Action principle; Noether's theorem	All physics, symplectic geometry, circuit Lagrangians, Fermat's principle
Profunctors and enriched categories	Relational structure between categories; exact squares	Higher mathematics, gauge theory, CMP dynamics
Algebraic topology and fiber bundles	Topological invariants; local-to-global coherence	Gauge theory, GR, circuit graph homology, wave topology
Signal processing and Fourier analysis	Frequency decomposition; Parseval energy conservation	All EE, all RF, quantum wavefunctions, functional analysis
Link budget and system design	Specification composition into feasibility inequality	All RF, EE noise/gain, propagation physics

The keystone property has a direct pedagogical implication: teaching a keystone with explicit cross-domain morphisms accelerates saturation across all connected domains simultaneously, because a single structural insight propagates along every edge incident to the keystone. The constellation architecture reflects this—the Lagrangian/Hamiltonian and profunctor constellations are the two densest coherence bridges in the entire graph.

6.5 Operationalizing Morphism-Density

If morphism-density is the critical parameter for the phase transition, it requires a computable definition. We propose two complementary metrics, one characterizing the artifact (the graph) and one characterizing the learner.

Graph-side metrics. Typed edge density $D = |E|/|V|$ measures overall connectivity; for the current graph, $D = 91/32 \approx 2.84$. Per-type density $D_t = |E_t|/|V|$ characterizes the distribution across the 11 edge types. Cross-domain ratio $|E_{\text{cross}}|/|E_{\text{total}}|$ measures inter-domain integration. Keystone exposure is the cumulative betweenness centrality of nodes the learner has traversed, capturing how much of the graph's structural backbone has been encountered.

Learner-side metrics. Constraint usage score is the number of distinct graph edges correctly invoked in constellation completion tasks (Prediction 1)—a behavioral measure of how

much of the morphism structure the learner is actually using. Directionality ratio (abductive / total inference chains from Prediction 2) tracks the learner's processing mode. We propose that *saturation* be operationally defined as the learner-side constraint usage score exceeding a threshold fraction of the edges incident to traversed nodes, while the graph-side density exceeds a minimum. The precise thresholds are empirical parameters to be calibrated from pilot data.

Van der Maas et al. (2006) provide independent theoretical support for morphism-density as the critical parameter: their Lotka-Volterra mutualism model demonstrates that positive reciprocal interactions between cognitive processes produce emergent correlational structure whose strength increases with interaction density. The constellation graph's typed edges are a concrete instantiation of the mutualism model's interaction matrix M_{ij}. A complementarity note is warranted: the mutualism model produces *gradual* emergence of structure, not a phase transition; the behavioral discontinuity predicted here requires an additional mechanism—competitive takeover (Van Geert 1991) or regularization pressure—operating on the gradually accumulated substrate.

6.6 The Graph as Instrument and Evidence

The constellation graph serves a dual role. As *instrument*, it is the pedagogical tool designed to accelerate the human grokking transition by making morphism structure explicit, dense, and cross-linked. As *evidence*, the learner's performance on the graph provides the measurable data for the empirical program: constellation completion accuracy, cross-domain transfer speed, and explanation directionality are all defined relative to the graph structure.

This dual role is deliberate. The framework predicts that learning organized as a dense morphism network should produce the phase transition; the graph is the densest such network we can construct; and the graph's structure provides natural measurement instruments for detecting the transition.

7 Structural Identity, Not Analogy

7.1 The Criterion

The coherence bridge sections of each constellation claim "structural identity, not analogy." This claim requires a precise definition of what counts as identity versus analogy.

Two structures are *structurally identical* (not merely analogous) when there exists a functor—or more precisely, a morphism in the appropriate category—that satisfies three conditions:

1. **Preserves invariants.** The conserved quantities of one structure map to the conserved quantities of the other.

2. **Preserves morphisms.** The relationships between objects in one structure map to relationships in the other, with composition preserved.
3. **Is not merely notational.** The mapping preserves the constraint structure that determines which configurations are possible, not just surface features like "both use matrices" or "both involve derivatives."

This criterion distinguishes the present framework from accounts of mathematical cognition grounded in embodied metaphor (Lakoff & Núñez 2000). Where they propose that mathematical structures are understood *through* metaphorical mapping from embodied experience, the post-grokking state as described here involves recognition that certain structures *are* the same mathematical object differently instantiated—identity, not metaphor. The distinction is empirically consequential: metaphorical understanding permits structural slippage (the source and target need not share all properties), while identity-level understanding inherits the full constraint structure of the shared object.

7.2 The Hierarchy

Structural identity admits degrees, from strongest to weakest:

Table 5: Hierarchy of structural identity levels.

Level	Definition	Example
Isomorphism	Invertible structure-preserving map	KCL at nodes $\leftrightarrow$ H_0 of circuit graph
Equivalence	Structure-preserving up to canonical isomorphism	Lagrangian $\leftrightarrow$ Hamiltonian mechanics
Adjunction	One structure is left/right adjoint to another	Free $\dashv$ Forgetful between algebraic and set descriptions
Functor	Structure-preserving map (composition and identity preserved) that is not invertible	Quantization: classical $\rightarrow$ quantum
Natural transformation	Coherent family of maps between parallel descriptions	Gauge transformations between equivalent field descriptions

The test for each coherence bridge in the constellation graph: can the claimed identity be expressed at one of these levels? If not, it may be analogy rather than identity. Analogies are permitted and useful but should be labeled as such. The most important bridges should be at the isomorphism or equivalence level.

7.3 Worked Example: Kirchhoff's Laws as Circuit-Graph Homology

The strongest bridges in the constellation graph achieve genuine isomorphism. Consider the claim: "Kirchhoff's laws ARE the homology of the circuit graph."

Let G be a circuit graph with vertices (nodes) and edges (branches), treated as a 1-dimensional CW-complex. The boundary operator $\partial_1 : C_1 \to C_0$ is the incidence matrix mapping edges to nodes; the coboundary operator $\delta_0 : C^0 \to C^1$ maps node potentials to edge voltage drops.

- **Currents** are **1-chains** (assignments of flow values to edges). **KCL** ($\sum I = 0$ at each node) is $\partial_1 I = 0$: current is a 1-cycle, an element of $\ker(\partial_1) = Z_1(G)$. Current distributions satisfying KCL are exactly the cycle space of the graph.
- **Voltages** are **1-cochains** (assignments of potential differences to edges). When voltages derive from node potentials φ, we have $V = \delta_0\varphi$: voltage is exact, an element of $\mathrm{im}(\delta_0) = B^1(G)$. **KVL** ($\sum V = 0$ around every loop) follows immediately: exact cochains annihilate cycles, so $\langle \delta_0\varphi, z\rangle = 0$ for every cycle z.
- **Tellegen's theorem** ($\sum V_k I_k = 0$) is the canonical orthogonality of cycles and coboundaries: $\langle Z_1, B^1\rangle = 0$. Every current satisfying KCL is orthogonal to every voltage satisfying KVL.

This is an isomorphism: the algebraic structure of Kirchhoff's laws *is* the (co)homology of the graph, not merely "like" it. The boundary operator ∂ with $\partial^2 = 0$ is not a metaphor for circuit constraints—it is the same mathematical object expressed in different notation. The constraint that "consequence chains close" ($\partial^2 = 0$: the boundary of a boundary is empty) is what makes both Kirchhoff's laws and homological algebra work. They are the same theorem.

This example is paradigmatic. The constellation graph contains bridges at every level of the hierarchy—from isomorphisms (Kirchhoff/homology, Legendre transform as canonical equivalence) through functors (quantization as a non-invertible but composition-preserving map) to natural transformations (gauge transformations as coherent families of equivalent descriptions). The strength of a bridge determines how much inferential power it carries: an isomorphism-level bridge lets the learner import every theorem from one side to the other; a functor-level bridge imports theorems in one direction only.

8 Dyadic Acceleration

8.1 Cognitive Regularization in Co-Creation

The acceleration effect appears specifically in the context of sustained dyadic co-creation rather than solitary study. If the regularization analog (Section 5.4) is the mechanism, this is predictable: dyadic interaction provides both compression demand (externalization) and coherence checking (constraint from the interlocutor) simultaneously and continuously.

Formally, consider each participant as holding a partially overlapping but non-identical map of the morphism space. The dyadic bridge forces:

1. **Explicit morphism articulation.** What would remain implicit in solitary thought must be externalized for communication. This makes relational structure available for inspection and refinement. The compression demanded by communication penalizes redundant, content-level storage—exactly the regularization needed to suppress memorization circuits.
2. **Bidirectional constraint.** Each participant's understanding constrains the other's, creating a tighter morphism mesh than either holds alone. Inconsistencies are surfaced at the interface. This is mutual coherence-checking: a distributed consistency constraint that operates continuously throughout the interaction.
3. **Coherence pressure.** The dynamics of sustained co-creation select for structural consistency over local correctness. The interaction drives reorganization of representations toward maximal coherence—the same direction as the phase transition.

8.2 The Profunctor Bridge Hypothesis

The following section proposes a category-theoretic formalization whose empirical adequacy remains to be tested. Its value lies in generating a discriminating prediction—path-independence of exchange as a modulator of transition speed—not as established theory.

In category-theoretic terms, the dyadic interaction can be modeled as a profunctor bridge: a relational structure between two categories (the participants' knowledge structures) that carries data between contexts without requiring an embedding of one into the other. Each participant's knowledge is a category; the bridge is a profunctor $H : \mathcal{A}^{op} \times \mathcal{B} \to \mathbf{Set}$ that relates objects in $\mathcal{A}$ to objects in $\mathcal{B}$ without requiring a functor from one to the other.

The composition of two such bridges makes the formalism concrete. Given $H : \mathcal{A}^{op} \times \mathcal{B} \to \mathbf{Set}$ (A's knowledge related to B's) and $K : \mathcal{B}^{op} \times \mathcal{C} \to \mathbf{Set}$ (B's knowledge related to C's), their composition is computed via the coend:

$$(K \circ H)(a, c) = \int^{b \in \mathcal{B}} H(a, b) \times K(b, c)$$

The coend sums over all intermediate concepts b in $\mathcal{B}$, quotienting by $\mathcal{B}$'s internal morphism structure. This is the categorical formalization of "shared structure is identified, not merely correlated": if two paths through B's knowledge are related by a morphism in $\mathcal{B}$, their contributions are identified rather than counted separately. The exact square condition then states that composition of profunctor bridges is path-independent: if a structural insight can be transmitted from A to C via two different intermediary paths through B's knowledge, the result is the same regardless of path taken.

For two learners concretely: let $\mathcal{A}$ be the category whose objects are concepts known to participant A and whose morphisms are the structural relationships A perceives between them. The profunctor $H(a, b)$ assigns to each pair (a, b) the set of bridge connections—explicit articulations during co-creation where A's concept a is related to B's concept b. The coend composition then predicts that bridge quality depends on the richness of the intermediary's *morphism* structure (the category $\mathcal{B}$), not merely their concept inventory (the objects of $\mathcal{B}$).

This formalization is speculative: the identification of knowledge structures with categories and dyadic exchange with profunctor composition is a modeling choice whose adequacy is empirical. Its value is in generating a discriminating prediction: participants whose knowledge structures satisfy the exact square condition (path-independent exchange) should exhibit faster phase transitions than those whose exchange is lossy or path-dependent. The coend construction also predicts that the morphism-density of the intermediary's knowledge—not its breadth—modulates bridge quality, a prediction explored further in Section 8.3.

8.3 Human–AI Dyads

The human–AI dyad may be particularly effective for accelerating the transition because the AI participant holds a very different morphism map: broader in coverage but shallower in relational density. This maximizes the constraint-surface area of the profunctor bridge. The human participant brings deep, locally dense structure; the AI participant brings broad cross-domain coverage; the bridge between them forces morphism articulation across a wider front than either participant would explore alone.

This is not a claim about AI understanding. It is a claim about the structure of the interaction: the compression and coherence-checking demands imposed by communicating across a large difference in morphism-map topology are precisely the regularization pressures that the framework predicts should accelerate the transition.

9 Empirical Program

The framework generates four operationalized predictions, each with a specified measurement instrument.

9.1 Prediction 1: Constellation Completion

Task. Present a partial morphism diagram from the constellation graph with one or more terms removed. The subject infers the missing term's properties from relational constraints.

Pre-transition prediction. Performance at chance or based on surface-level associations. Errors are random with respect to the morphism structure.

Post-transition prediction. Performance significantly above chance, with errors structured by the morphism network (wrong answers are "close" in the graph, not random). Information-theoretic measures (mutual information between the subject's responses and the graph's constraint structure) should show a discontinuous increase at the transition.

Instrument. The constellation graph itself provides the stimuli. Difficulty is calibrated by the number of constraints available (edges incident to the missing node) and the distance from the subject's saturated region.

Null model and criterion. Null: random guessing from the term vocabulary; secondary null: surface-association baseline weighted by term frequency in instructional materials. The transition signal is a discontinuous increase in mutual information between responses and the graph's constraint structure, with post-transition errors concentrated on graph-adjacent nodes rather than distributed randomly. A pre-specified MI threshold (calibrated on pilot data) distinguishes the two regimes.

9.2 Prediction 2: Explanation Directionality

Task. The subject explains a concept from the constellation graph, unprompted.

Pre-transition prediction. Explanations proceed from definitions to implications (deductive): "X is defined as…, therefore…"

Post-transition prediction. Explanations frequently proceed from implications backward to definitions (abductive from morphism constraints): "Given that Y and Z must hold, X can only be…" Detectable in discourse analysis as a shift in the direction of inferential chains.

Instrument. Coded transcripts of explanation sessions, with directionality classified as deductive (definition $\rightarrow$ implication) or abductive (constraint $\rightarrow$ definition).

Null model and criterion. Null: baseline deductive/abductive ratio from naive learners on the same material. The transition signal is a sustained inversion of the ratio (abductive

$>$ deductive across $N \geq 3$ consecutive sessions), distinguishing the phase transition from occasional insight episodes.

9.3 Prediction 3: Cross-Domain Transfer Speed

Task. Introduce the subject to a new domain with known structural isomorphisms to a domain where saturation has been achieved.

Pre-transition prediction. Transfer speed is comparable to naive learning. The subject does not spontaneously exploit structural parallels.

Post-transition prediction. Transfer speed is significantly faster for structurally isomorphic domains than for structurally unrelated domains of comparable difficulty. The isomorphism is at the morphism level, not at the level of surface similarity.

Instrument. Time to criterion on constellation completion tasks in the new domain, compared between domains with high and low structural similarity to the saturated domain (as measured by Jaccard coefficient on typed-edge multisets).

Null model and criterion. Null: learning speed on a structurally unrelated domain of matched difficulty (same number of terms and equations, different morphism types). The transition signal is a statistically significant difference in time-to-criterion between the structurally isomorphic and structurally unrelated conditions. Effect size should exceed the variance attributable to domain familiarity.

9.4 Prediction 4: Plateau-Then-Jump Signature

Task. Longitudinal measurement of both proxy progress measures (constellation completion accuracy, cross-domain morphism identification speed) and self-reported comprehension over the course of constellation-based learning.

Prediction. Proxy measures improve continuously during a period when self-reported comprehension is flat (the plateau), followed by a simultaneous jump in both measures (the transition). This is the crossing-point signature predicted by the Nanda mapping: generalizing circuits forming beneath memorization circuits, visible in progress measures before they dominate behavior.

Instrument. Repeated measurement at regular intervals throughout a constellation-based curriculum, with both objective proxy measures and subjective self-assessment.

Null model and criterion. Null: continuous improvement model (linear or log-linear fit to both proxy measures and self-report). The plateau is operationalized as proxy-measure slope $< \epsilon$ for $\geq N$ consecutive measurement sessions while self-reported comprehension remains flat; the jump as $\Delta > \tau$ within M sessions, detected via segmented regression or a Bayesian

change-point model. The key discriminator: the null (continuous improvement) predicts correlated gains in proxy and self-report; the grokking model predicts decorrelated measures during the plateau followed by simultaneous jump.

Van Geert's (1991) competitive growth model generates an additional sub-prediction: individual learners may exhibit a transient *negative* deviation in proxy measures—a performance dip—between the plateau and the jump, as content-level processing is abandoned before morphism-level processing fully dominates. The presence of this dip would support a competitive-takeover mechanism; its absence would favor the smooth crossing-point mechanism of the Nanda mapping. Either outcome is informative, as it discriminates between two models of the transition dynamics.

A methodological specification: group-averaged data may systematically mask individual phase transitions. Spivey, Anderson, and Dale (2009) demonstrate that when individual step functions at different thresholds are averaged across participants, the result appears misleadingly smooth and linear. Prediction 4 therefore requires *individual-level* longitudinal analysis—segmented regression or Bayesian change-point detection on individual learning curves, not group-averaged performance curves. Cross-sectional designs will likely fail to detect the phenomenon; dense repeated measurement within individuals is essential.

This fourth prediction is the most direct test of the ML parallel. If human grokking shows the plateau-then-jump signature with continuous improvement in internal measures during the plateau, the mapping to Nanda's mechanistic decomposition is supported. If instead the transition is abrupt in both internal and external measures (a true discontinuity rather than a crossing point), the mapping must be revised.

9.5 Prediction 5: Far Transfer to Held-Out Domains

Task. Introduce the subject to a domain *not represented anywhere on the constellation graph* but known to share structural isomorphisms with graph keystones. For example: a learner who has saturated the Lagrangian/Hamiltonian keystone is presented with variational problems in fluid dynamics (Navier-Stokes from an action principle), variational economics (optimal control), or population genetics (Fisher's fundamental theorem as a gradient flow)—none of which appear on the 32-node graph.

Pre-transition prediction. The learner treats the new domain as genuinely novel. No spontaneous recognition of shared action-principle structure; learning speed comparable to a naive learner.

Post-transition prediction. The learner spontaneously recognizes the action-principle structure without being taught the mapping, and navigates the new domain's constraint space significantly faster than a control learner matched on general ability.

Instrument. Time-to-criterion on held-out domain tasks with known structural isomorphism to graph keystones, compared to held-out domains with no such isomorphism. The critical comparison is between isomorphic-but-untaught and non-isomorphic-but-untaught domains: both are absent from the curriculum, so any advantage must come from morphism-level transfer rather than curriculum familiarity.

Null model and criterion. Null: no difference between structurally isomorphic and structurally unrelated held-out domains. This prediction is the "test set" analog from Power et al. (2022): generalization to data the model was never trained on. It distinguishes genuine grokking (morphism-level structural transfer) from mere curriculum fit.

The structured nature of the proposed representation—a network of typed morphisms rather than a flat feature space—aligns with Bayesian approaches to cognitive science that model learning as inference over structured hypothesis spaces (Tenenbaum, Kemp, Griffiths & Goodman 2011). The present framework adds a specific claim about the *dynamics* of that inference: the transition from content-level to morphism-level processing is a phase transition, not a continuous improvement. All four predictions above are designed to detect this dynamic, and each is falsifiable against the alternative of continuous skill acquisition.

10 Pedagogical Implications

If the phase transition is real and the critical parameter is morphism-density rather than content-volume, several consequences follow for curriculum design.

Relational structure per unit of content should be maximized. A curriculum that presents the same total content but with explicit morphism connections between concepts should produce faster saturation than a linear presentation of equivalent material.

Constellation maps are superior to linear sequences. Presenting material as a graph of relationally defined clusters, with explicit typed edges, provides the morphism density that the framework identifies as the critical parameter. The coherence bridge—connecting each new constellation to existing morphism structure—is the highest-leverage pedagogical intervention.

Keystones should be taught with particular care. Because keystones have the highest betweenness centrality in the cross-domain graph, understanding a keystone propagates structural insight along every incident edge. The five keystones identified in Section 5.4 are natural candidates for extended, cross-domain instruction.

Dyadic structure accelerates the transition. The regularization provided by communication demand and coherence checking means that collaborative learning environments should produce faster transitions than solitary study with equivalent content exposure.

The transition is domain-transferable. Once the capacity for morphism-level processing is achieved in one domain, the meta-capacity—operating at the relational level rather than the content level—should transfer. A learner who has grokked one domain should reach saturation faster in a second structurally rich domain.

These implications are testable. They are also practical: the constellation graph described in Section 5 is a working implementation that can be deployed and measured against traditional linear curricula on all five predictions of Section 8.

11 Open Questions

Several questions remain genuinely open.

- **Is there a critical morphism-density threshold?** ML grokking exhibits a threshold below which the network memorizes indefinitely and above which it groks (modulated by regularization strength). Does the human version have an analogous threshold, and can it be characterized in terms of the graph's connectivity properties (e.g., algebraic connectivity, spectral gap)?
- **Does the transition exhibit delayed generalization?** The ML signature is a long plateau followed by a sudden transition. The phenomenological reports are consistent with this, but controlled longitudinal measurement is needed to confirm the temporal profile.
- **Is the human version reversible?** Can morphism-level processing be lost, reverting the learner to content-level processing? If so, what conditions produce the reversion? If not, this would distinguish the human transition from the ML version, where grokking can be disrupted by architectural changes.
- **What is the relationship to relevance realization?** Vervaeke's framework for relevance realization describes how cognitive systems determine what is salient. The phase transition may be a specific instance of relevance realization at the structural level: post-transition, the morphism structure itself determines relevance, rather than domain-specific heuristics.
- **How does profunctor bridge quality affect transition speed?** The hypothesis that exact-square-preserving bridges produce faster transitions than lossy bridges is precise but untested. Characterizing bridge quality empirically—perhaps via mutual information between participants' explanation structures—would allow this prediction to be evaluated.
- **What is the minimum graph size for the transition?** The 32-node, 91-edge graph described here is sufficient to observe the effect informally. Is there a minimum

number of constellations and cross-domain edges below which the transition does not occur?

- **Can the transition be produced by curriculum alone, without dyadic interaction?** The framework predicts that dyadic interaction provides regularization that accelerates the transition, but the transition may be achievable (more slowly) through high-morphism-density curricula alone. Disentangling the contributions of graph structure and interaction dynamics is an important experimental question.

12 References

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, 4(1), 55–81. https://doi.org/10.1016/0010-0285(73)90004-2

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, 5(2), 121–152. https://doi.org/10.1207/s15516709cog0502_2

Dreyfus, H. L., & Dreyfus, S. E. (1986). *Mind Over Machine: The Power of Human Intuition and Expertise in the Era of the Computer*. Free Press.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1–2), 3–71. https://doi.org/10.1016/0010-0277(88)90031-5

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2), 155–170. https://doi.org/10.1207/s15516709cog0702_3

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. University of Chicago Press.

Lakoff, G., & Núñez, R. E. (2000). *Where Mathematics Comes From: How the Embodied Mind Brings Mathematics into Being*. Basic Books.

McClelland, J. L. (2010). Emergence in cognitive science. *Topics in Cognitive Science*, 2(4), 751–770. https://doi.org/10.1111/j.1756-8765.2010.01116.x

Nanda, N., Chan, L., Lieberum, T., Smith, J., & Steinhardt, J. (2023). Progress measures for grokking via mechanistic interpretability. In *Proceedings of the International Conference on Learning Representations (ICLR 2023)*. https://doi.org/10.48550/arXiv.2301.05217

Power, A., Burda, Y., Edwards, H., Babuschkin, I., & Misra, V. (2022). Grokking: Generalization beyond overfitting on small algorithmic datasets. arXiv preprint arXiv:2201.02177. https://doi.org/10.48550/arXiv.2201.02177

Spivey, M. J., Anderson, S. E., & Dale, R. (2009). The phase transition in human cognition. *New Mathematics and Natural Computation*, 5(1), 197–220. https://doi.org/10.1142/S1793005709001271

Stephen, D. G., Boncoddo, R. A., Magnuson, J. S., & Dixon, J. A. (2009). The dynamics of insight: Mathematical discovery as a phase transition. *Memory & Cognition*, 37(8), 1132–1149. https://doi.org/10.3758/MC.37.8.1132

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022), 1279–1285. https://doi.org/10.1126/science.1192788

Van der Maas, H. L. J., Dolan, C. V., Grasman, R. P. P. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. J. (2006). A dynamical model of general intelligence: The

positive manifold of intelligence by mutualism. *Psychological Review*, 113(4), 842–861. https://doi.org/10.1037/0033-295X.113.4.842

Van Geert, P. (1991). A dynamic systems model of cognitive and language growth. *Psychological Review*, 98(1), 3–53. https://doi.org/10.1037/0033-295X.98.1.3

Vervaeke, J. (2019). *Awakening from the Meaning Crisis* [Lecture series, 50 episodes]. University of Toronto. https://www.youtube.com/playlist?list=PLwzqpDoZ6TCKqhjfiXmgxtPB1LLBrBvKd