When Do Circuit Discoveries Compose Into Understanding?

Elliot Tower
Independent Researcher
elliot@elliottower.ai

Abstract

Mechanistic interpretability has produced dozens of circuit-level analyses of language models, yet no framework exists for composing individual discoveries into system-level understanding. We identify three gaps that prevent composition: no parcellation theory (the field has at least three incommensurable decompositions—sparse autoencoder features, circuits, and causal variables—with no theory of how they relate), no composition theory (overlapping circuits cannot be aggregated without an explicit account of overlap, interference, and redundancy), and no coverage metric (there is no denominator against which to measure how much of a model's behavior has been explained). Drawing on the resolution of the analogous parcellation problem in neuroscience—from Brodmann's cytoarchitectonic maps through the Human Connectome Project to network neuroscience—we propose the *mechanism cluster* as a new unit of analysis: a collection of mechanism claims about a single system, evaluated for parcellation coherence, individual validity, and collective coverage. We define a composition criterion specifying when a cluster constitutes system-level understanding, and apply it to the indirect object identification (IOI) literature in GPT-2 Small—the largest and best-studied circuit cluster in interpretability. The IOI cluster comprises approximately 20 published analyses involving heavily overlapping sets of attention heads. We show that 78% head overlap between circuits for different tasks means that counting circuits is not the same as measuring coverage: fifty circuits involving the same attention head yield fifty validated claims and one fact about that head. The cluster fails the composition criterion on all three dimensions—parcellation incoherence, incomplete individual validity, and unmeasured coverage—providing a concrete demonstration that accumulating validated claims does not, by itself, produce understanding.

1 Introduction: Fifty Claims, One Fact

Mechanistic interpretability has made remarkable progress. The indirect object identification (IOI) circuit in GPT-2 Small (Wang et al., 2023), induction heads across transformer scales (Olsson et al., 2022), automated circuit discovery via ACDC (Conmy et al., 2023), grokking via Fourier features (Nanda et al., 2023), and sparse autoencoder feature decompositions (Bricken et al., 2023) have each produced validated, reproducible claims about how specific components of language models implement specific computations.

Yet a basic question remains unanswered: *when do these individual discoveries compose into understanding of the system?*

Consider a concrete example. Merullo et al. (2024) show that the IOI circuit and the Colored Objects circuit in GPT-2 Medium share 78% of their attention heads. If understanding were simply a matter of accumulating validated claims, then having both circuit analyses would give strictly more understanding than having one. But the 78% overlap means the two claims are partially redundant—the same heads doing the same thing in two contexts is one fact, not two. Treating the claims as additive overcounts the evidence. Treating them as identical loses the 22% difference. Neither accumulation nor identity is correct: what is needed is *composition*, an operation that correctly aggregates two overlapping validated claims into a single system-level fact.

This problem scales. GPT-2 Small has been the subject of at least twenty published circuit analyses. Many involve overlapping attention heads—the IOI circuit shares 25 of 32 heads with the Colored Objects circuit in GPT-2 Medium (Merullo et al., 2024), and individual heads like 9.9 appear in multiple independent analyses. Fifty circuits that each involve head 9.9 yield fifty validated claims and *one fact about head 9.9*. Without a composition theory, the field cannot distinguish "fifty circuits that together explain 80% of the model's behavior" from "fifty circuits that all explain the same 5% over and over."

The existing philosophical literature on mechanistic explanation does not address this problem. Craver (2007)'s mosaic unity concerns how findings across disciplines and levels of organization constrain each other through both vertical (interlevel) and horizontal (intralevel) integration. Sullivan (2022)'s link uncertainty concerns individual claim quality, not composition. Potochnik (2020) argues that idealizations are constitutively essential to scientific understanding, not merely tolerable approximations; but her account addresses when partial understanding is legitimate, not when fragments compose into something more. What interpretability faces is a *horizontal* accumulation problem: hundreds of validated claims at the same level of analysis (component-level mechanisms in a single model), with no theory of when that pile becomes "we understand GPT-2."

We argue that this composition problem is not new. Neuroscience faced exactly the same challenge for over a century, and the history of how it was resolved provides both a diagnosis and a template (developed at length in the companion perspective; Tower 2026e). The present paper makes three contributions:

1. We define the *mechanism cluster* as a new unit of analysis, distinct from the per-claim analysis of Tower (2026c), the per-pair analysis of Tower (2026d), and the per-term analysis of Tower (2026b).

2. We propose a *composition criterion* specifying when a cluster constitutes system-level understanding: parcellation coherence, individual validity, and collective coverage.

3. We apply the criterion to the IOI literature—the largest circuit cluster in interpretability—and show that it fails on all three dimensions, diagnosing exactly where and why.

Readers familiar with software engineering may find the following analogies clarifying. Parcellation coherence asks whether different analyses use compatible interfaces: two circuit descriptions that decompose a model at different granularities are like two libraries with incompatible type signatures—they cannot be composed without an adapter. Individual validity asks whether each module passes its own tests. Collective coverage asks the question every engineering team knows but interpretability has never posed: what is our test coverage? The composition criterion does not demand 100% coverage; it demands that the coverage percentage be measured.

2 The Neuroscience Parallel

The parcellation problem in neuroscience provides both a solved instance and a proven research program. We compress the argument here; the full historical development is in Tower (2026e).

2.1 Accumulation Without Composition

For over a century, systems neuroscience accumulated individual discoveries that did not compose. Penfield & Boldrey (1937) mapped motor and sensory cortex via electrical stimulation. Hubel & Wiesel (1962) characterized orientation-selective cells in visual cortex via single-unit recording. Lesion studies identified regions critical for memory, language, and face recognition. Each discovery was individually valid. None composed into a system-level account, because each arose from a different decomposition of the brain—electrical stimulation, single-unit recording, and behavioral deficit defined different parcellations with no explicit theory of how their units related.

2.2 The Parcellation Problem

The parcellation problem is, at its core, a question that every engineer recognizes: before you can measure how much of a system you understand, you need to agree on what the parts are.

The core issue, formalized by Eickhoff et al. (2018), is that the brain can be divided into units based on cytoarchitecture (Brodmann, 1909), stereotaxic coordinates (Talairach & Tournoux, 1988), functional activation, or structural connectivity, and these parcellations do not agree. Findings from one parcellation could not be straightforwardly composed with findings from another. A claim about "Brodmann area 44" and a claim about "the region activated by syntactic processing" might refer to overlapping but distinct populations, with no principled way to tell.

2.3 The Resolution

Progress required three developments:

Multi-modal parcellation. The Human Connectome Project (Van Essen et al., 2013) and the Glasser parcellation (Glasser et al., 2016) integrated multiple modalities into a single 360-region parcellation, defining boundaries by convergence across signals rather than commitment to one modality.

Connectomics. Sporns et al. (2005) proposed the connectome: a comprehensive map of neural connections that shifted the unit of analysis from regions to connections between regions.

Network neuroscience. Bassett & Sporns (2017) synthesized these into a framework where system-level properties (modularity, hub structure, dynamic reconfiguration) emerge from network topology. Understanding became a property of the network, not the sum of what each region does.

The resolution arc is instructive: *the field did not need more findings about individual regions. It needed a framework for composing findings into a system-level account.*

2.4 Where the Analogy Holds and Breaks

The analogy holds at the *functional* level. Tigges et al. (2024) show that the same algorithmic roles persist across model scale and training duration, even when the specific heads occupying those roles change. Bali et al. (2026) show that residual-stream subspaces are stable across random seeds even when individual head assignments diverge. Sun & Toneva (2026) report within-family interpretive congruity of 0.73–0.92, dropping to 0.13 across families. The shared substrate in interpretability is the computational problem, not the weights—analogous to functional parcellation in neuroscience, where homologous functions recruit non-identical neural populations across individuals.

The analogy breaks at the *substrate* level. The brain has a stable biological anatomy shared across individuals; neural network weights are the product of stochastic training. There is no "cortex" to parcellate. This disanalogy is real but does not undermine the functional-level argument: the composition problem exists regardless of whether the substrate is shared, because it concerns the relationship between *findings*, not between *architectures*.

3 Three Decompositions, Zero Composition

Mechanistic interpretability currently operates with at least three incommensurable decomposition frameworks.

3.1 Sparse Autoencoder Features

SAEs decompose model activations into interpretable features (Bricken et al., 2023; Elhage et al., 2022). The parcellation is a set of features, and the implicit claim is that features are the natural units of computation. Leask et al. (2025) show that different SAE training runs produce different feature sets, raising the question of whether SAE features are canonical units or artifacts of the decomposition method.

3.2 Circuits

Circuit analysis identifies subgraphs that implement specific behaviors (Wang et al., 2023; Conmy et al., 2023). The parcellation is a set of components connected by edges. But circuits are task-specific: the IOI circuit and the greater-than circuit are different subgraphs sharing components, with no framework for what the shared components mean at the system level. Franco et al. (2026) show that even within one task, circuits are prompt-specific—different prompt templates activate different circuit structures within the same model.

3.3 Causal Variables

Causal abstraction (Geiger et al., 2025) identifies high-level causal variables that align with internal representations. The parcellation is a set of causal variables and their relationships. But causal variables are defined relative to a hypothesized causal structure, and different hypotheses produce different variables.

3.4 The Fragmentation Is the Problem

These three frameworks answer different questions and produce different units. An SAE feature is not obviously a circuit component. A circuit edge is not obviously a causal variable. Tower (2026a) document this fragmentation empirically: in a corpus of 688 mechanistic interpretability papers, within-tradition citation density is approximately twice the cross-tradition density. The three communities build primarily on their own prior work. This is not a sociological curiosity—it is a symptom of the missing composition theory.

4 What a Model Connectome Would Look Like

Before formalizing the composition criterion, we sketch the target: what would it mean to have composed individual circuit discoveries into a system-level account of GPT-2 Small?

A model connectome would consist of three components:

A parcellation. A principled decomposition of the model into functional units, with explicit declaration of what evidence modalities (features, circuit roles, causal variables) define the boundaries. Following the HCP's multi-modal approach, parcel boundaries would be defined by convergence: where feature decompositions, circuit analyses, and causal abstractions all agree on a boundary, the boundary is robust.

A connectivity map. For each pair of parcels, a characterization of their interaction: do they share components? Does activating one affect the other? Are they functionally independent or redundant? Merullo et al. (2024)'s 78% head overlap between IOI and Colored Objects in GPT-2 Medium is a connectivity fact—it tells us that these "circuits" are not independent modules but deeply entangled subgraphs.

A coverage metric. A denominator: what fraction of the model's behavior is accounted for by the parcellation? This requires defining the behavioral class and measuring how much variance in that class is explained by known parcels.

This sketch makes concrete what "composition" means: it is the operation that takes individual circuit analyses and places them into a shared coordinate system where overlap, coverage, and interactions can be measured.

To make this concrete, consider what a connectome entry for head 9.9 in GPT-2 Small would contain: (1) a parcellation entry declaring its resolution level (attention head, Object View) and listing the functional roles attributed to it across papers (name mover in IOI, active in greater-than and Colored Objects circuits, identified by both manual analysis and automated discovery); (2) a connectivity entry mapping its input sources (which upstream heads write to its key and query subspaces) and output targets (which downstream heads and the unembedding read from its value output); (3) a coverage annotation estimating what fraction of its total causal effect on model outputs is accounted for by known circuits. Currently, (1) exists implicitly across papers but has never been reconciled—different papers attribute different roles to the same head without checking consistency. Entry (2) exists partially from path-patching analyses (Wang et al., 2023;

Conmy et al., 2023), but only for specific tasks and only for the edges that those analyses chose to measure. Entry (3) has never been attempted for any head in any model.

5 The Composition Criterion

We define the mechanism cluster and state the conditions under which a cluster constitutes system-level understanding. The composition criterion draws on two companion frameworks; we summarize the minimum needed here so that this paper can be read independently.

Background: Mechanistic Views (Tower, 2026d)**.** A *mechanistic view* is a background commitment that determines what counts as a mechanism, when two descriptions refer to the same mechanism, and what evidence can support what claims. Made explicit, a view answers five questions—*ontology* (what kind of entity?), *identity* (when are two descriptions the same?), *evidence* (what measurements support what claims?), *formalism* (what mathematical language?), and *target* (what phenomenon is explained?). Different views induce different identity criteria. Two claims are *view-compatible* if they share the same identity criterion or if one can be formally *promoted* to the other by supplying additional evidence of the kind the target view requires. Claims that use incompatible identity criteria cannot be composed without first reconciling their views.

Background: Mechanistic Validity (Tower, 2026c)**.** MechVal provides a four-tier evidence scale calibrated to view type: Tier 1 (*Descriptive*)—the claim identifies components but provides no causal evidence; Tier 2 (*Observationally Consistent*)—correlational or ablation evidence consistent with the claim; Tier 3 (*Causally Suggestive*)—interventional evidence (e.g., activation patching, causal scrubbing) that supports a causal interpretation; Tier 4 (*Mechanistically Supported*)—convergent evidence from multiple independent methods under multiple views. Each tier specifies what inferences it licenses: Tier 1 claims license description but not prediction; Tier 3 claims license causal counterfactuals within the tested domain; only Tier 4 claims license cross-domain generalization.

Definition 5.1 (Mechanism Cluster)**.** *A mechanism cluster $\mathcal{C} = \{C_1, \ldots, C_N\}$ is a collection of mechanism claims about a single system S, where each claim C_i identifies a set of components, their interactions, and the behavior they produce.*

The cluster is the new unit of analysis. Tower (2026c) evaluates individual claims (per-claim). Tower (2026d) compares pairs of claims (per-pair). Tower (2026b) tracks individual terms (per-term). The composition criterion evaluates collections (per-cluster).

The intuition is straightforward: a collection of mechanism claims constitutes understanding only if the claims are about compatible things (parcellation coherence), each claim is individually well-supported (validity), and the claims collectively account for a measurable fraction of the system's behavior (coverage).

Criterion 5.1 (Composition)**.** *A mechanism cluster $\mathcal{C}$ about a system S constitutes system-level understanding of S with respect to a behavioral class $\mathcal{B}$ if and only if:*

1. **Parcellation coherence.** All claims in $\mathcal{C}$ are stated relative to an explicit decomposition, and claims from different decompositions have been reconciled. In the framework of Tower (2026d), this means all claims are either stated under the same view or their views have been shown to be compatible via formal promotion.

2. **Individual validity.** Each claim C_i meets a minimum standard of evidence. In the framework of Tower (2026c), this means each claim reaches at least Tier 3 (Causally Suggestive).

3. **Collective coverage.** The claims collectively account for a defined fraction of the system's behavior over $\mathcal{B}$, with the fraction explicitly stated and the denominator defined. Coverage requires addressing overlap: claims sharing components must be deduplicated before coverage is computed, so that N claims involving the same head count as one fact, not N.

The three conditions are individually necessary:

Without parcellation coherence (condition 1), claims from different decompositions cannot be compared. An SAE feature claim and a circuit claim about the same model may or may not be about the same computation, but without a reconciliation procedure there is no way to tell.

Without individual validity (condition 2), the cluster may contain claims that do not withstand scrutiny. Composition does not compensate for bad ingredients.

Without coverage (condition 3), the cluster could consist entirely of validated, view-consistent claims that all describe the same narrow slice of the model's behavior. Coverage is the condition that no existing framework provides and that the composition criterion adds.

Definition 5.2 (Overlap Coefficient)**.** *For two circuit claims C_i and C_j identifying component sets $\mathcal{H}_i$ and $\mathcal{H}_j$, the overlap coefficient is:*

$$\text{Overlap}(C_i, C_j) = \frac{|\mathcal{H}_i \cap \mathcal{H}_j|}{\min(|\mathcal{H}_i|, |\mathcal{H}_j|)}$$

Definition 5.3 (Effective Coverage)**.** *Given a cluster $\mathcal{C}$ with component sets $\mathcal{H}_1, \ldots, \mathcal{H}_N$ and a denominator D (total components, total causal contribution, or total variance), the effective coverage is:*

$$\text{Coverage}(\mathcal{C}) = \frac{|\bigcup_{i=1}^{N} \mathcal{H}_i|}{D}$$

The union deduplicates: shared components count once. (When D measures variance or causal contribution rather than component count, the numerator should be understood as the corresponding aggregate measure over the union, not set cardinality.)

The denominator D is the hardest part. We do not claim to solve the denominator problem—defining it precisely is a research program (Section 8). But we note that even an approximate denominator (e.g., total number of attention heads, total residual-stream variance on an evaluation suite) is more informative than no denominator. The distinction between "50 circuits covering 15% of the model" and "50 circuits covering 80%" is consequential even if the denominator is imprecise.

6 The IOI Cluster Audit

We apply the composition criterion to the IOI cluster: the collection of published circuit analyses that include attention heads from the IOI circuit in GPT-2 Small.

6.1 Identifying the Cluster

The IOI circuit as defined by Wang et al. (2023) consists of 26 attention heads across 7 functional roles. We identify published analyses that involve one or more of these heads. The cluster includes:

- The original IOI analysis (Wang et al., 2023)
- Greater-than circuit analysis (Hanna et al., 2023)
- Cross-task component reuse (Merullo et al., 2024)
- Prompt-specific circuit analysis (Franco et al., 2026)
- Cross-seed stability analysis (Bali et al., 2026)
- Cross-model interpretive equivalence (Sun & Toneva, 2026)
- Algorithmic consistency across scale (Tigges et al., 2024)
- Causal abstraction of IOI (Geiger et al., 2025)

- Automated circuit discovery (Conmy et al., 2023)

Additional analyses exist; we focus on those with sufficient methodological detail to evaluate under the three conditions.

6.2 Condition 1: Parcellation Coherence

The IOI cluster fails parcellation coherence.

Wang et al. (2023) operate under an Object View: they identify specific attention heads by name and define the circuit as a set of named components. Geiger et al. (2025) operate under a Subspace View: they identify low-dimensional subspaces in the residual stream that align with causal variables. Bricken et al. (2023) operate under a Feature View (SAE decomposition) that cross-cuts both.

These three analyses make claims about the same model that cannot be directly compared. The Object View claim "head 7.3 is an S-inhibition head" and the Subspace View claim "a 3-dimensional subspace in layer 7 encodes the indirect object" may or may not be about the same underlying computation, but there is no formal procedure for checking.

Franco et al. (2026) add a further layer of incoherence: even within the Object View, the IOI circuit is not a single circuit but a family of prompt-specific circuits. GPT-2's ABBA circuit and BABA circuit share components but use different signals (negative cosine similarity between the input signals to head 9.9 in the two templates). The "IOI circuit" is itself a cluster that has not been composed.

6.3 Condition 2: Individual Validity

The IOI cluster partially meets individual validity.

The original IOI analysis reaches Tier 3+ on the MechVal scale: ablation, path patching, and causal intervention provide convergent evidence for the circuit's causal role. Conmy et al. (2023)'s automated rediscovery provides independent confirmation.

However, validity is uneven across the cluster. Some claims—particularly about the precise role of backup name mover heads and about the circuit's behavior on edge cases—rest on thinner evidence. Joshi et al. (2026) argue that causal frameworks are needed for interpretability claims to generalize, and that many published claims exceed the evidence level their methods support. The IOI cluster is better than average but not uniformly at Tier 3.

6.4 Condition 3: Collective Coverage

The IOI cluster does not meet collective coverage.

No denominator has been defined. There is no estimate of what fraction of GPT-2 Small's behavior is explained by the IOI circuit and its variants. The IOI task covers only indirect object identification in specific syntactic frames—a narrow slice of the model's capabilities.

Overlap is massive. Merullo et al. (2024) show 78% head overlap between IOI and Colored Objects. The greater-than circuit, by contrast, operates primarily in layers 5–9 with essentially no head-level overlap with IOI (Hanna et al., 2023)—a case where apparent topical similarity (both involve numerical reasoning over tokens) masks architectural divergence. Multiple circuits that appear to cover similar tasks may be describing the same heads from different angles, while circuits for seemingly related tasks may be disjoint.

Effective coverage is unknown. The union of all component sets across the cluster is concentrated in layers 7–10. Early layers (0–3) and late layers (11) are sparsely represented in published circuit analyses. The effective coverage—the fraction of GPT-2's computation that the cluster accounts for—has never been estimated.

6.5 Audit Summary

Table 1: Composition criterion applied to two circuit clusters. The criterion discriminates: IOI fails all three conditions; induction heads partially pass. Even the field's most rigorous result fails at coverage.

	IOI Cluster		**Induction Head Cluster**	
Condition	**Status**	**Diagnosis**	**Status**	**Diagnosis**
Parcellation coherence	Fails	Object/Subspace/Feature views unreconciled	Partial	Consistent role-level definition; context-dependent variants blur boundary
Individual validity	Partial	Core claims Tier 3+; peripheral claims below threshold	Near pass	Convergent structural, causal, behavioral evidence; unique process evidence
Collective coverage	Fails	No denominator; 78% head overlap inflates apparent coverage	Fails	No denominator for in-context learning; fraction explained unknown
Overall	0/3 conditions fully met		∼1.5/3 conditions met	

The IOI literature is the largest and best-studied circuit cluster in mechanistic interpretability. It fails the composition criterion on all three dimensions. This is not an indictment of the individual analyses—they are among the best in the field. It is a demonstration that *accumulating validated claims does not, by itself, produce system-level understanding.*

7 The Induction Head Cluster: A Partial Pass

The IOI audit might suggest that the composition criterion is too demanding—that nothing in the field can pass it. To show that the criterion discriminates rather than merely indicts, we apply it to the induction head literature: the only mechanistic interpretability result that partially passes.

7.1 Condition 1: Parcellation Coherence — Partial Pass

Olsson et al. (2022) define induction heads at the *Role* level: a head is an induction head if it attends to the token following the previous occurrence of the current token, thereby implementing a form of in-context copying. This functional definition is consistent across papers that study induction heads. Unlike the IOI cluster, where Object, Subspace, and Feature views produce incommensurable claims about the same heads, the induction head literature largely agrees on the level of description and the identity criterion for what counts as an instance of the mechanism.

However, the concept may be stretching beyond its original scope. Tigges et al. (2024) show that the same algorithmic roles persist across training and scale in large models, but the heads occupying those roles may implement more complex, context-dependent algorithms than the simple copying of classical induction heads. Whether these extended behaviors constitute the same mechanism or a distinct one is an open parcellation question. The coherence is therefore partial: strong within the original scope, uncertain at the boundary.

7.2 Condition 2: Individual Validity — Near Pass

The induction head claim has the strongest individual validity in mechanistic interpretability. Evidence converges across modalities: *structural* evidence (the characteristic attention pattern), *causal* evidence (ablating induction heads degrades in-context learning performance), and *behavioral* evidence (the pattern appears

across model families and scales) all support the same mechanism claim. On the MechVal scale, this reaches Tier 3 or above.

Olsson et al. (2022) additionally provide *process* evidence: induction heads emerge via a phase transition during training, and their emergence coincides with a sharp improvement in in-context learning loss. The composition score between the heads that form the two-step induction circuit (previous-token heads composing with induction heads) tracks this phase transition. This temporal evidence—showing that the mechanism's formation causally precedes the capability it is claimed to implement—is a form of validation that no other circuit claim in the field possesses.

7.3 Condition 3: Collective Coverage — Fails

Despite the strength of the individual claim, the induction head cluster fails coverage.

There is no denominator. In-context learning is a broad capability, and induction heads implement only one algorithm within it. What fraction of in-context learning is explained by induction heads? No estimate exists. The context-dependent induction variants identified in larger models suggest that the mechanism is more complex than the original two-step account: in-context learning in production-scale models likely involves circuits that go beyond simple token copying, and the relationship between classical induction and these more complex mechanisms has not been characterized.

Coverage would require defining the behavioral class (which in-context learning tasks?), measuring the total causal effect of all model components on that class, and computing the fraction attributable to identified induction circuits. None of these steps has been attempted.

7.4 Induction Head Audit Summary

The induction head cluster is the field's most rigorous result and it still fails at coverage. This is not nihilism—it is a precise diagnosis of what is missing. The induction head literature demonstrates that individual validity can be excellent while collective coverage remains entirely unaddressed. The composition criterion identifies *where* the gap is (coverage, not validity) and thereby points to what work would close it.

8 Implications

8.1 Understanding Is Not Accumulation

The central claim of this paper is that validated mechanism claims do not compose into understanding by accumulation. Composition is a distinct epistemic achievement that requires its own framework—just as connectomics was a distinct achievement from single-unit recording, not merely more of it.

Potochnik (2020) has argued that partial, fragmented understanding is acceptable because completeness is not a realistic goal. We agree that partial understanding has value. But in the context of AI safety, where mechanism claims are used to make assurance arguments about model behavior, the distinction between "50 validated circuits" and "50 validated circuits that cover 80% of the model" is consequential. The composition criterion does not demand completeness; it demands that the degree of completeness be *measured*.

This distinction is consequential for AI safety. If circuit-level understanding is used to construct safety assurance arguments—claiming that a model will not exhibit deceptive behavior because its circuits have been characterized—then the composition problem is not academic. An assurance argument built on fifty validated circuits with unmeasured coverage is an assurance argument with an unknown false-negative rate. The field cannot say "we understand 50 circuits" and mean "we understand the model" without a denominator. The composition criterion makes the gap between these two statements quantitatively precise.

8.2 What the Field Needs to Build

By analogy with neuroscience, we identify three concrete needs:

Explicit parcellation protocols. Interpretability needs conventions for declaring the decomposition framework under which a claim is made. Currently, most circuit discovery papers implicitly adopt a decomposition (attention heads, residual-stream directions, SAE features) without stating which one or why. This makes cross-paper comparison difficult because it is unclear whether two papers that both identify "head 9.9 as important" are making the same claim or different claims at different resolution levels. Stating "this circuit is defined at the attention-head level under mean ablation" is not onerous but makes the claim's scope explicit and enables later reconciliation. A concrete first step: each circuit discovery paper should state in its methods section which decomposition framework it uses and which identity criterion it assumes, following the View Declaration Template proposed in Tower (2026d).

Cross-decomposition reconciliation. When two analyses of the same model use different frameworks, the field needs methods for checking whether they are consistent. This is the analogue of checking whether Brodmann areas align with functional parcels. The IOI circuit has been analyzed under the Object View (specific attention heads), the Subspace View (low-dimensional residual-stream subspaces), and the Feature View (SAE features), but no published work has formally tested whether these three descriptions pick out the same underlying computation. Without such tests, the three analyses cannot be composed: they may be three perspectives on one fact or three claims about different facts that happen to involve overlapping components. A concrete first step: take the IOI circuit as defined by activation patching and the IOI-relevant SAE features, and measure whether they identify the same causal variables—for instance, whether the subspace spanned by IOI-related SAE features aligns with the subspace identified by DAS on the same task.

Coverage benchmarks. The field needs evaluation suites designed not to test individual circuits but to measure how much of a model's behavior is collectively accounted for by known mechanisms. Existing evaluations in mechanistic interpretability are circuit-specific: they test whether a proposed circuit is faithful to a particular behavior, not whether the set of all known circuits covers the model's behavioral repertoire. A coverage evaluation suite would define a broad behavioral class (e.g., all tasks that GPT-2 Small performs above chance), compute a total causal effect budget for that class, and then measure how much of that budget is attributable to components appearing in at least one validated circuit. A concrete first step: define a behavioral evaluation suite for GPT-2 Small spanning multiple task types, and for each known circuit, measure the fraction of total causal effect (via integrated gradients) attributable to that circuit's components.

8.3 The Denominator Problem

The hardest part of coverage is defining the denominator. We do not solve this problem but identify three candidate denominators, each with limitations:

Component count. The fraction of attention heads (or neurons, or features) that appear in at least one validated circuit. Simple but ignores the fact that heads contribute unequally.

Residual variance. The fraction of total residual-stream variance on an evaluation suite that is explained by ablating all known circuit components. Method-dependent (ablation type, distribution, order of ablation).

Causal contribution. The fraction of total causal effect (measured by integrated gradients or attribution patching) that is attributable to known circuits. Most principled but computationally expensive and method-dependent.

The denominator problem is harder than it appears even for well-studied tasks. For grokking, the field has identified Fourier features as progress measures and a modular arithmetic circuit as the final-state implementation (Nanda et al., 2023). But it is unclear whether the weight-space geometry of that implementation is stable across training runs, whether the Fourier basis is canonical or a decomposition artifact, and whether the circuit that produces grokking is the same object as the circuit present at completion. These are not pedantic concerns—they are instances of the parcellation coherence and individual validity conditions, applied to a task where the denominator (what fraction of the model's weights implement modular arithmetic) has never been estimated. Grokking may be the best-understood phase transition in mechanistic interpretability, and even there the coverage question is open.

None of these is canonical. The paper's contribution is not to solve the denominator problem but to show that it exists, that it is distinct from validation, and that progress on it requires composition theory. Neuroscience could not measure coverage until it defined what "cortex" was. Defining the space is the first step toward measuring it.

8.4 Partial Understanding and Graduated Coherence

The composition criterion as stated is binary: a cluster either constitutes understanding or it does not. In practice, understanding comes in degrees. We note that each condition can be graduated:

Parcellation coherence can be measured as the fraction of claim pairs that have been explicitly reconciled.

Validity can be reported as the distribution of MechVal tiers across the cluster.

Coverage can be reported as a range, given uncertainty in the denominator.

A cluster that is 80% view-reconciled, with 90% of claims at Tier 3+, and estimated coverage of 40–60%, has more claim to system-level understanding than one at 20%/50%/5–15%. The composition criterion provides the axes along which progress can be measured, even before full understanding is achieved.

8.5 Practical Recommendations for Circuit Papers

A circuit paper following this framework would make three additions to its methods section, none requiring new experiments: (1) declare the decomposition framework under which claims are made (e.g., "this circuit is defined at the attention-head level under mean ablation"); (2) state which other published circuits share components with the discovered circuit, and whether the overlap has been reconciled or is simply noted; (3) report a coverage estimate, even if rough (e.g., "the circuit's components account for $X\%$ of the model's attention heads and $Y\%$ of total causal effect on the target task as measured by integrated gradients"). These are three sentences in a methods section—not a burden, but a commitment to measuring the gap between validated claims and system-level understanding.

For the field as a whole, the composition criterion suggests a concrete benchmark: maintain a living head participation matrix (Appendix A) for well-studied models, updated as new circuit analyses are published. Such a matrix would make the overlap problem visible, the denominator problem tractable, and progress toward coverage measurable.

9 Related Work

Mechanistic interpretability. Circuit discovery methods (Wang et al., 2023; Conmy et al., 2023) and feature decomposition methods (Bricken et al., 2023) produce individual mechanism claims. The present paper addresses the question of when collections of such claims compose into understanding.

Circuit universality. Tigges et al. (2024), Merullo et al. (2024), Sun & Toneva (2026), Bali et al. (2026), and Franco et al. (2026) study whether circuits transport across models, scales, seeds, and prompts. These papers provide the empirical basis for the composition problem: they show that circuits overlap, partially transport, and partially diverge, but do not address how to compose these findings.

Philosophy of mechanisms. Craver (2007)'s mosaic unity concerns vertical integration across levels; we address horizontal composition within a level. Sullivan (2022)'s link uncertainty concerns the gap between model behavior and target phenomenon—whether a model analysis that reproduces a behavior genuinely explains the target phenomenon. The structure of her argument (identifying what evidence would close the gap) parallels our approach to the composition gap, but operates at the per-claim level rather than the per-cluster level. Link uncertainty is a prerequisite for composition: claims with unresolved link uncertainty cannot be composed because it is unclear what they are claims about. Potochnik (2020)'s partial understanding permits fragmentation; we ask when fragments compose.

Brain parcellation. Eickhoff et al. (2018) review the parcellation problem in neuroscience. Sporns et al. (2005) and Bassett & Sporns (2017) develop the connectomic and network-neuroscience frameworks that resolved it. We draw the analogy to interpretability in Tower (2026e) and apply it formally here.

Validation and views. Tower (2026c), Tower (2026d), and Tower (2026b) provide per-claim, per-pair, and per-term evaluation frameworks. The present paper adds per-cluster evaluation.

10 Conclusion

Mechanistic interpretability is in a position analogous to pre-connectome neuroscience: accumulating valid, reproducible findings about individual components without a framework for composing them into system-level understanding.

The 78% head overlap between circuits for different tasks is not an inconvenience—it is a signal that the field is rediscovering the same components from different angles without a coordinate system for recognizing the redundancy. The mechanism cluster—a collection of claims evaluated for parcellation coherence, individual validity, and collective coverage—provides that coordinate system.

The IOI cluster audit shows that the field's best-studied circuit does not yet constitute system-level understanding of GPT-2 Small, not because the individual analyses are wrong but because they have not been composed. The neuroscience parallel shows this composition problem has been solved before, and the resolution—explicit parcellation, cross-decomposition reconciliation, and coverage metrics—provides a concrete template for what interpretability needs to build next.

References

Karan Bali, Jack Stanley, Praneet Suresh, and Danilo Bzdok. Quantifying LLM attention-head stability: Implications for circuit universality. *arXiv preprint arXiv:2602.16740*, 2026.

Danielle S Bassett and Olaf Sporns. Network neuroscience. *Nature Neuroscience*, 20(3):353–364, 2017.

Trenton Bricken, Adly Templeton, Joshua Batson, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

Korbinian Brodmann. *Vergleichende Lokalisationslehre der Grosshirnrinde*. Johann Ambrosius Barth, 1909.

Arthur Conmy, Augustine N Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36, 2023.

Carl F Craver. *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford University Press, 2007.

Simon B Eickhoff, B T Thomas Yeo, and Sarah Genon. Imaging-based parcellations of the human brain. *Nature Reviews Neuroscience*, 19(11):672–686, 2018.

Nelson Elhage, Tristan Hume, Catherine Olsson, et al. Toy models of superposition. *Transformer Circuits Thread*, 2022.

Gabriel Franco, Lucas M Tassis, Azalea Rohr, and Mark Crovella. Finding interpretable prompt-specific circuits in language models. *arXiv preprint arXiv:2602.13483*, 2026.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83): 1–64, 2025.

Matthew F Glasser, Timothy S Coalson, Emma C Robinson, Carl D Hacker, John Harwell, Essa Yacoub, Kamil Ugurbil, Jesper Andersson, Christian F Beckmann, Mark Jenkinson, Stephen M Smith, and David C Van Essen. A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178, 2016.

Michael Hanna, Ollie Liu, and Alexandre Variengien. How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. *Advances in Neural Information Processing Systems*, 36, 2023.

David H Hubel and Torsten N Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 160(1):106–154, 1962.

Shruti Joshi, Aaron Mueller, David Klindt, Wieland Brendel, Patrik Reizinger, and Dhanya Sridhar. Position: Causality is key for interpretability claims to generalise. *arXiv preprint arXiv:2602.16698*, 2026.

Patrick Leask, Bart Bussmann, Michael Pearce, Joseph Bloom, Curt Tigges, Noura Al Moubayed, Lee Sharkey, and Neel Nanda. Sparse autoencoders do not find canonical units of analysis. In *International Conference on Learning Representations*, 2025.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. Circuit component reuse across tasks in transformer language models. In *International Conference on Learning Representations*, 2024.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *ICLR*, 2023.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *Transformer Circuits Thread*, 2022.

Wilder Penfield and Edwin Boldrey. Somatic motor and sensory representation in the cerebral cortex of man. *Brain*, 60(4):389–443, 1937.

Angela Potochnik. Idealization and many aims. *Philosophy of Science*, 87(5):933–943, 2020.

Olaf Sporns, Giulio Tononi, and Rolf Kötter. The human connectome: A structural description of the human brain. *PLoS Computational Biology*, 1(4):e42, 2005.

Emily Sullivan. Understanding from machine learning models. *The British Journal for the Philosophy of Science*, 73(1):109–133, 2022.

Alan Sun and Mariya Toneva. Tracking equivalent mechanistic interpretations across neural networks. In *International Conference on Learning Representations*, 2026.

Jean Talairach and Pierre Tournoux. *Co-Planar Stereotaxic Atlas of the Human Brain*. Thieme, 1988.

Curt Tigges, Michael Hanna, Qinan Yu, and Stella Biderman. LLM circuit analyses are consistent across training and scale. In *Advances in Neural Information Processing Systems*, 2024.

Elliot Tower. Benchmarks are engines, not cameras. *Preprint*, 2026a. Forthcoming.

Elliot Tower. Mechanistic reference: When does a mechanism term pick out the same thing? *Preprint*, 2026b.

Elliot Tower. Mechanistic validation: Evidence standards for mechanism claims. *Preprint*, 2026c.

Elliot Tower. Mechanistic views: A formal framework for adjudicating between competing mechanistic claims in interpretability. *Preprint*, 2026d.

Elliot Tower. Mechanistic parcellation: What interpretability research can learn from the brain mapping debate. *Preprint*, 2026e. Submitted to Network Neuroscience.

David C Van Essen, Stephen M Smith, Deanna M Barch, Timothy E J Behrens, Essa Yacoub, and Kamil Ugurbil. The WU-Minn human connectome project: An overview. *NeuroImage*, 80:62–79, 2013.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations*, 2023.

A IOI Cluster: Partial Head Participation Matrix

Table 2 presents a partial head participation matrix for key IOI-related attention heads across published analyses. Each row is an attention head in GPT-2 Small, organized by functional role as defined in Wang et al. (2023). Columns represent published studies that identify or analyze these heads. A ✓ indicates that the study identifies the head as part of its circuit or analysis.

Table 2: Partial head participation matrix for IOI-related heads in GPT-2 Small. Heads are organized by functional role. Studies: W23 = Wang et al. (2023), C23 = Conmy et al. (2023), M24 = Merullo et al. (2024), H23 = Hanna et al. (2023), G24 = Geiger et al. (2025), F26 = Franco et al. (2026), T24 = Tigges et al. (2024).

Role	**Head**	**W23**	**C23**	**M24**	**H23**	**G24**	**F26**	**T24**
Name Mover	9.9	✓	✓	✓		✓	✓	✓
	9.6	✓	✓	✓		✓	✓	✓
	10.0	✓	✓	✓			✓	
S-Inhibition	7.3	✓	✓	✓			✓	✓
	7.9	✓	✓	✓			✓	
	8.6	✓	✓	✓			✓	✓
	8.10	✓	✓	✓			✓	
Duplicate Token	0.1	✓						
	3.0	✓						
Backup Name Mover	9.0	✓		✓				
Negative Name Mover	10.7	✓					✓	
	11.10	✓					✓	

Three patterns are visible in this partial matrix. First, heads in layers 7–10 (the S-inhibition and name mover heads) appear in nearly every study, regardless of methodology—they are the most robustly identified components across manual analysis, automated discovery, and causal abstraction. Second, heads in layers 0–4 (duplicate token detectors and earlier-layer components) are sparsely studied: only the original IOI analysis identifies them, suggesting that downstream analyses either assume their role or lack the methodology to confirm it. Third, head 9.9 appears in nearly every analysis in the matrix. Fifty claims about circuits involving head 9.9 yield fifty validated claims and one fact about that head—the quintessential instance of accumulation without composition.

B Candidate Denominator Calculations

Table 3 compares three candidate denominators for measuring collective coverage in GPT-2 Small, applied to the IOI cluster.

Table 3: Candidate denominators for measuring collective coverage of the IOI cluster in GPT-2 Small. All three are computable with existing tools; none has been attempted.

Denominator	**Definition**	**IOI estimate**	**Limitation**	**Cost**
Component count	$\lvert\bigcup \mathcal{H}_i\rvert / N_{\text{total}}$	$\leq$ 26/156 $\approx$ 17% (before deduplication with other circuits)	Treats all components as equal; ignores causal importance	Trivial
Residual variance	$1 - \text{var}_{\text{ablated}}/\text{var}_{\text{total}}$ on held-out suite	Unknown (never measured)	Method-dependent (mean vs. zero ablation); distribution-dependent	Moderate
Causal contribution	$\sum_i \text{IG}(h_i) / \sum_j \text{IG}(h_j)$ over all components	Unknown (never measured)	Computationally expensive; attribution method matters	High

The component count denominator is the simplest. GPT-2 Small has 144 attention heads (12 layers $\times$ 12 heads) and 12 MLP layers, for 156 total components at the coarsest granularity. The IOI circuit involves 26 heads; the greater-than circuit involves approximately 12 heads with substantial overlap in layers 7–10. A naive component-count coverage of $\lvert\bigcup \mathcal{H}_i\rvert/156$ provides a lower bound on effective coverage, but treats a duplicate token detector in layer 0 as equivalent to a name mover head in layer 9.

The residual variance denominator is more principled: for each layer, measure the total residual-stream variance on a held-out evaluation suite, then measure the variance remaining after mean-ablating all components in the cluster's union set. (1 $-$ remaining/total) gives a variance-based coverage estimate. This is method-dependent (mean vs. zero ablation) and distribution-dependent (which evaluation suite), but at least weights components by their actual contribution to model computation.

The causal contribution denominator uses integrated gradients or attribution patching to measure each component's causal effect on model outputs, then computes the fraction attributable to known circuit components. This is the most principled approach but also the most expensive and sensitive to methodological choices.

None of these denominators has been computed for any circuit cluster in any model. The distinction between "50 circuits covering 15% of the model" and "50 circuits covering 80% of the model" is consequential even if the denominator is imprecise — and any of the three approaches above would resolve it.