Mechanistic Validity: A Theory of Validity for Mechanistic Claims

Elliot Tower
elliot@elliottower.ai

Abstract

Mechanistic claims are ubiquitous in empirical science—*this circuit implements this computation*, *this gene causes this disease*, *this pathway mediates this outcome*. What makes such a claim valid?

A mechanistic claim can be supported by strong experimental evidence and still be unwarranted. A mechanism can be measured by a reliable metric and still correspond to no coherent computational concept. It can correspond to a coherent concept and rest on purely correlational evidence. It can survive rigorous causal testing at one intervention strength on one prompt distribution and collapse under any other. It can pass all of those tests and still be described at the wrong level of abstraction. These are not points on a continuum—they are independent failures that demand independent remedies. A single high score does not validate a mechanistic claim. Validation is a pattern of evidence across multiple dimensions, and a claim is only as strong as the dimension on which it has the weakest support.

We introduce Mechanistic Validity (MechVal), a theory of validity for mechanistic claims. The framework draws on philosophy of science, measurement theory, causal inference, and validation methodology from neuroscience, pharmacology, and genetics. MechVal evaluates claims through a seven-layer pipeline: a *description mode* declares what kind of explanation is being offered—not a label applied after analysis, but a constraint declared before it. *Evidence families* classify the sources of signal. *Metrics* are concrete tests that measure the claim. *Criteria* test whether the evidence meets the required conditions. Five *validity types*—construct, measurement, internal, external, interpretive—determine which dimensions the evidence addresses, in dependency order. *Synthesis protocols* aggregate evidence and assign *verdicts*: structured validity profiles and tier assignments—from Proposed through Causally Suggestive, Mechanistically Supported, and Triangulated, to Validated—for any mechanistic claim.

We demonstrate the framework on 13 published mechanistic claims in neural network interpretability, producing verdicts from Proposed through Triangulated to Disconfirmed.

1 Introduction

Mechanistic interpretability has produced a remarkable catalog of internal mechanisms: induction heads that implement in-context copying (Olsson et al., 2022), a multi-role circuit for indirect object identification (Wang et al., 2023), a greater-than algorithm (Hanna et al., 2023), and sparse autoencoder features proposed to correspond to human-interpretable concepts (Cunningham et al., 2024; Bricken et al., 2023). The field has also begun to build evaluation infrastructure: faithfulness metrics for circuits (Miller et al., 2024), the Mechanistic Interpretability Benchmark for comparing localization methods across tasks and models (Mueller et al., 2025), SAEBench for auditing sparse autoencoder quality (Karvonen et al., 2025), and causal scrubbing for testing mechanistic hypotheses via behavior-preserving resampling (Chan et al., 2022).

But how well-supported are these claims? Across the field's most prominent results, a pattern recurs: strong evidence along one dimension, untested assumptions along others. Faithfulness scores that are method-conditional. Metrics that saturate on random models. Constructs that cannot be separated from neighboring

computations. Specificity controls that were never run. Labels that imply more than the evidence supports. These are not failures of execution—they are gaps in the *kind* of evidence collected, and Section 2.3 catalogs them systematically. The pattern is not unique to MI: every field that makes mechanistic claims has encountered the same structural problems (Section 2). Shared metrics are not yet shared validity criteria.

MechVal is a structured framework that takes the validity vocabulary the field already uses informally and makes it operational, falsifiable, and auditable.[1] It evaluates claims through a seven-layer pipeline: a *description mode* declares what kind of explanation is being offered. *Evidence families* classify the sources of signal. *Metrics* are concrete tests that measure the claim. *Criteria* test whether the evidence meets the required conditions. Five *validity types*—construct, measurement, internal, external, and interpretive—determine which dimensions the evidence addresses, in dependency order. *Synthesis protocols* aggregate evidence and assign *verdicts*: structured validity profiles and tier assignments, from Proposed through Causally Suggestive, Mechanistically Supported, and Triangulated, to Validated.

The pipeline has two uses. When *running experiments*, evidence builds upward from metrics through criteria to a verdict. When *auditing a claim*, the process reverses: starting from a verdict, the framework identifies which criteria are unmet and whether the description mode is supported by the available evidence.

The same criteria apply to circuits, SAE features, probing classifiers, steering vectors, transcoders, and crosscoders—the questions "is the construct well-defined?" and "is the evidence causal?" do not depend on how the evidence was produced.

We describe lessons from validity failures across sciences (Section 2), the framework (Section 5), the concrete methodology for generating evidence (Section 6), its theoretical foundations (Section 7), its application to 13 published circuits (Section 8), and its implications for the field (Section 9).

2 Motivation: What Happens Without Validity Criteria

Every field that makes mechanistic claims has encountered the same structural problem: strong individual results that do not add up to valid conclusions. The specific failure modes differ, but the pattern is the same—an untested link in the inference chain from measurement to claim. We catalog these failures across sciences, machine learning, and mechanistic interpretability, organized by the validity criterion each violates.

2.1 Validity Failures Across the Sciences

Dead salmon fMRI (M2: baseline separation). Bennett et al. (2010) scanned a dead Atlantic salmon while presenting it with photographs of humans in social situations. Using standard uncorrected thresholds, 16 significant voxels appeared in the salmon's brain cavity; with appropriate multiple comparisons corrections, none survived. Eklund et al. (2016) extended this systematically, finding that standard fMRI software produced false positive rates up to 70%. This same structure can be seen in MI, if a method measures faithfulness without testing against a size-matched random subnetwork.

Cardiac stents for stable angina (I3: specificity). For three decades, stenting was standard treatment for stable angina (500,000 procedures/year in the US). The ORBITA trial (Al-Lamee et al., 2018) was the first placebo-controlled test: half of patients received a sham procedure with no stent placed. The result: no statistically significant difference in exercise tolerance or angina symptoms. The entire symptomatic benefit was placebo. The effect was real. The causal attribution was not—it had never been tested against a placebo control. Circuit ablation without a size-matched random baseline faces the same problem.

Candidate gene psychiatry (I5: confound control). The candidate gene era (1990s–2010s) produced thousands of associations between gene variants and depression risk. Border et al. (2019) tested the 18 most-studied candidate genes using data from 620,000 individuals: the genes were "no more associated with depression than randomly chosen genes." Twenty-five years of literature produced real associations

[1] The framework, full case study audits, and implementation are available at `https://mechanistic-validity.github.io/mechanistic-validity/` and `https://github.com/mechanistic-validity/mechanistic-validity`.

in the samples—but population stratification and selective reporting were never ruled out. When Border et al. (2019) controlled for these confounders at scale, the signal disappeared. This is the same evidence pattern that appears in MI probing studies: a positive result that survives only as long as the confounds go unmeasured.

Cold fusion (M1/M2: measurement reliability and baseline separation). In 1989, Pons and Fleischmann announced room-temperature nuclear fusion. Several labs initially confirmed the result; within months, all retracted. Neutron detectors gave false positives when exposed to heat, and the original paper reported a gamma peak without its Compton edge—a physical impossibility indicating instrumentation error. The instrument was producing the expected signal for the wrong reason. Validating the measurement independently of the hypothesis—testing whether the detector fires under conditions where fusion cannot occur—would have killed the claim in weeks, not years. In MI, this is equivalent to a metric never tested on a null model—it cannot distinguish the mechanism from its own response properties.

BOLD signal (M1: measurement reliability). Functional MRI does not measure neural activity—it measures blood oxygenation changes coupled to neural activity through a hemodynamic mechanism that varies by brain region, age, and individual. Every fMRI claim that says "region X activates during task Y" is technically a claim about hemodynamics, not neurons—and the two can dissociate. The instrument measures a proxy, not the target. Activation patching has the same structure: it measures the joint effect of the component's causal role and the network's compensatory response, and a single experiment cannot decompose the two.

2.2 Validity Failures in Machine Learning

Emergent abilities mirage (M4: calibration). Schaeffer et al. (2023) showed that "emergent abilities" in large language models—apparent phase transitions where capabilities appear suddenly at scale—are artifacts of nonlinear evaluation metrics. Exact-match accuracy creates a sharp threshold where the model goes from 0% to near-100%; switch to linear metrics (token edit distance, Brier score) and the "emergence" disappears—the ability was always there, growing smoothly. M4 requires that reported numbers correspond to meaningful quantities. The metric created the phenomenon: emergence was real in the measurement but not in the model.

COVID-19 ML diagnostics (I5: confound control). Roberts et al. (2021) reviewed 61 studies claiming to detect COVID-19 from chest X-rays and CT scans. Not a single model was clinically usable. DeGrave et al. (2021) showed why: saliency maps revealed that models relied on laterality markers, image border artifacts, and patient positioning—not lung pathology. I5 requires that confounders be identified and controlled before a causal claim is made.

Shortcut learning (E2/C4: generalization and discriminant validity). Geirhos et al. (2019) showed that ImageNet-trained CNNs are primarily texture detectors, not shape detectors—directly opposite to human perception and to what the field assumed. A cat-shaped object with elephant texture is classified as an elephant. Models achieving 90%+ accuracy on the benchmark had learned a fundamentally different construct than intended. E2 requires that findings generalize across conditions; C4 requires that the measure distinguish the target construct from its neighbors.

2.3 Validity Failures in Mechanistic Interpretability

SAE metric baselines (M2: baseline separation). Karvonen et al. (2025) audit eight standard SAE evaluation metrics and find that two fail basic diagnostics: scores on shuffled or random baselines are indistinguishable from scores on real features. The metrics are reliable—they give the same answer when rerun—but a high score does not distinguish genuine feature quality from chance. Without a null distribution, there is no way to know whether an SAE feature score reflects real structure or noise that the metric cannot filter.

Vacuous nonlinear IIA (M2: baseline separation). Sutter et al. (2025) show that unconstrained nonlinear alignment maps achieve 100% interchange-intervention accuracy on randomly initialized models. The IIA metric is working correctly; it is measuring the alignment map's flexibility, not the model's representational structure. Structured methods with identifiability constraints do not show this failure. The lesson is that a metric without a baseline is a metric without meaning: perfect scores are not evidence if random models also achieve them.

IOI circuit specificity (I3: specificity). The IOI circuit (Wang et al., 2023) demonstrates necessity and sufficiency—ablating the circuit degrades IOI performance, and isolating it recovers 87% of the logit difference. But specificity is untested: the original paper does not report whether ablating the circuit also degrades performance on unrelated tasks. If it does, what was discovered is a general-purpose subnetwork, not an IOI-specific mechanism. Specificity requires testing what the circuit does *not* do, not only what it does.

Greater-than circuit negative controls (I3: specificity / C1: falsifiability). The greater-than circuit (Hanna et al., 2023) reports no negative controls: ablation effects on non-target tasks are never measured. Only predictions that should pass are tested; what the mechanism should *not* affect is never checked. A claim that cannot be falsified by any downstream observation is not a mechanistic claim—it is a description.

Knowledge neuron off-target effects (I3: specificity). Cohen et al. (2024) show that editing one factual association corrupts related facts—"ripple effects" that were not anticipated or controlled for. The intervention is not specific to the target knowledge; it disrupts neighboring representations in ways the original study did not measure. This is the MI analog of off-target drug effects: the intervention works, but it does more than what was claimed.

Circuit non-uniqueness (I6: rival mechanism exclusion). Méloux et al. (2025) show that activation patching returns alternative head sets with comparable faithfulness scores for the same task. The "IOI circuit" is *one of several* sufficient subnetworks, not *the* mechanism. This is Quine's underdetermination problem (Quine, 1951) made concrete: the evidence is consistent with multiple incompatible circuits, and the standard procedure cannot distinguish between them. Without a method for resolving between rival circuits, faithfulness scores select one of many equivalent explanations.

Method-conditional faithfulness (E1: intervention reach). Miller et al. (2024) show that the IOI circuit's headline 87% faithfulness is specific to mean ablation; under resample ablation, faithfulness drops below 50%. The causal claim is method-conditional: it holds under one ablation regime and fails under another. The result depends on the method, not only the mechanism. A finding that does not survive a change in intervention method has not established external validity.

Gender bias circuits (C4: discriminant validity). Vig et al. (2020) identify heads that mediate gender bias, but these are the same heads encoding grammatical gender agreement. Ablating them reduces biased completions but equally degrades correct pronoun resolution. The construct "gender bias circuit" presupposes a separation between bias and knowledge that does not exist in the model's representations. When the construct cannot be separated from its neighbors, the claim is about a label, not a mechanism.

Othello "world model" (V4: interpretive inflation). Li et al. (2023) show that Othello board state is linearly decodable from activations. The evidence supports "linearly decodable" (representational-statistical); the label "world model" implies a computational-level claim about the model maintaining and updating an internal representation for planning. The gap between what was measured and what was named is interpretive inflation—a description-mode mismatch where the label claims more than the evidence establishes.

2.4 The Common Structure

Table 1 maps these failures to MechVal criteria alongside the MI cases from Section 2.3.

Table 1: Validity failures across fields, mapped to MechVal criteria. Each failure involves a strong individual result with an untested link in the inference chain.

Case	Field	Criterion	Untested link
Dead salmon fMRI	Neuroscience	M2 Baseline sep.	No null distribution; 70% false positive rate
Cold fusion	Physics	M1/M2 Reliability	Instrument error; $100M spent
BOLD signal	Neuroscience	M1 Reliability	Hemodynamics $\neq$ neural activity
SAEBench metrics	MI	M2 Baseline sep.	Random baselines $\approx$ real features
Sutter IIA	MI	M2 Baseline sep.	100% IIA on random models
Cardiac stents	Medicine	I3 Specificity	Sham-controlled: zero benefit
Knowledge neurons	MI	I3 Specificity	Editing one fact corrupts neighbors
Candidate genes	Genetics	I5 Confound ctrl	18 genes = random in $n = 620$K
COVID-19 ML	ML	I5 Confound ctrl	0/61 models learned COVID pathology
IOI circuit	MI	E1 Interv. reach	87% $\rightarrow$ <50% under other ablation
Shortcut learning	ML	E2/C4 Generalization	Texture $\neq$ shape understanding
Emergent abilities	ML	M4 Calibration	Metric created the phase transition
Othello "world model"	MI	V4 Interp. inflation	"Decodable" narrated as "world model"

Despite the diversity of these failures—across physics, neuroscience, genetics, medicine, and machine learning—they share a common structure: in each case, an individual experimental result was strong, but the *inference chain* from measurement to conclusion had an untested link. The result was real; the conclusion was not warranted by the evidence pattern. Claims migrated upward in inferential scope faster than supporting validity criteria could be established: a significant voxel cluster became a brain region that responds to social stimuli, a gene-environment interaction in 200 participants became a biomarker for depression, a high accuracy score became visual understanding, a sharp metric curve became an emergent ability, and symptom relief after stenting became evidence of a causal mechanism.

Each failure mode maps onto a specific criterion: measurement unreliability onto M1, lack of baselines onto M2, missing specificity controls onto I3, method-conditional results onto E1, interpretive inflation onto V4, unfalsifiable constructs onto C1. The framework makes the distance between evidence and claim visible, so that the field can prioritize the experiments that would close the gap.

3 Related Work

MechVal is designed to complement, not replace, existing MI evaluation infrastructure. MIB (Mueller et al., 2025) is method-vs-method: which discovery algorithm produces the best circuit? TracR (Lindner et al., 2023) is method-vs-ground-truth: does the method find the true circuit in a synthetic model? MechVal is claim-centric: given a claim about how a model works, what evidence supports it and what remains untested? MIB and TracR evaluate discovery methods; MechVal evaluates the claims built on their outputs.

Recent work applies construct validity theory to ML evaluation more broadly. Bean et al. (2025) find that fewer than 16% of LLM benchmarks use statistical validity methods—a striking parallel to our finding that most MI claims lack measurement validity (M1–M6). Bean et al.'s diagnosis is that benchmarks are

treated as self-validating once they gain community adoption; the same dynamic applies to MI metrics, where faithfulness and IIA are used as primary evidence without calibration checks. Freiesleben & Zezulka (2025) develop a psychometric framework for benchmark evaluation grounded in Cronbach, Meehl, and Messick, arguing that benchmarks should be evaluated as measurement instruments with explicit reliability and validity properties. MechVal differs in two respects: the unit of analysis is a mechanistic claim (a causal assertion about how a model works, not a task-performance score), and the validity structure extends from construct validity alone to five types in dependency order—construct, measurement, internal, external, and interpretive. Where Freiesleben et al. ask "does this benchmark measure the right thing?" and Bean et al. ask "is this benchmark statistically sound?," MechVal asks "does the evidence pattern support the causal claim?"

SAEBench (Karvonen et al., 2025) audits SAE evaluation metrics; MechVal treats that audit as evidence about its own measurement-validity criteria (M1–M6). MechEvalAgent (Bai et al., 2026) checks whether published MI code reproduces; MechVal asks whether the evidence pattern supports the claim. Unstructured LLM-as-judge evaluation (Zheng et al., 2023) lacks the operational criteria to distinguish validity from plausibility. A companion paper automates MechVal via structured LLM extraction against the 30 criteria.

The Schmidt Sciences Trustworthy AI agenda (Schmidt Sciences, 2026) calls for construct validity, predictive validity, and cross-model generalization of interpretability findings—using these terms but without a unified framework for operationalizing them. MechVal provides the operationalization: each of those desiderata maps onto specific criteria (C1–C4 for construct validity, E3–E4 for predictive validity, E4 for cross-model generalization) with concrete metrics and tier requirements.

4 Example: From Claim to Verdict

Consider a concrete claim: "The IOI circuit uses three name-mover heads (9.9, 9.6, 10.0) to copy the indirect object to the output position." This claim appears in a widely-cited paper, is supported by ablation experiments, and has a clear mechanistic narrative. What has this claim established, and what remains untested?

The framework evaluates such a claim through seven layers (Figure 1):

1. **Description mode.** At what level is the claim made? The IOI claim operates at the implementational-functional level: it names specific heads and says what each one does.
2. **Evidence family.** What kinds of evidence are available? The original IOI paper provides primarily interventional-activation evidence (ablation, patching) and some behavioral evidence (faithfulness). Weight-based and cross-model evidence would strengthen the claim by adding independent failure modes.
3. **Metrics.** What concrete measurements were taken? Activation patching, role ablation, and faithfulness scores. Each metric produces a number; the question is whether those numbers are calibrated and comparable across methods (Section 6).
4. **Criteria.** Does the evidence meet the bar? The IOI circuit passes necessity (I1: ablation degrades performance) and sufficiency (I2: the circuit alone recovers 87% of logit difference). But specificity (I3: does the circuit affect *only* IOI?) is untested, and the 87% faithfulness is method-conditional—it drops below 50% under resample ablation (Miller et al., 2024), partially failing intervention reach (E1).
5. **Validity type.** Which validity dimensions are satisfied? Construct validity (the claim is well-defined), partial internal validity (necessity and sufficiency shown but specificity untested), partial external validity (limited prompt generalization, no cross-model replication). Measurement validity is unaddressed.
6. **Synthesis.** How is evidence aggregated across methods? Currently it is not—the IOI claim rests on a single methodological family. Applying cross-method synthesis (e.g., Dawid-Skene consensus across patching, weight analysis, and information-theoretic methods) would determine which heads survive reliability-weighted voting.

7. **Verdict.** What tier does the claim reach? Mechanistically Supported—necessity and sufficiency are established with at least one method, but evidence comes primarily from one family, key criteria remain untested (I3, I4, E4), and the faithfulness result is method-conditional.

The 30 criteria also formalize common failure modes observed in published MI research (Table 2).

Table 2: Twelve common failure modes in MI research, each mapped to the criteria that detect it.

Failure mode	**What happens**	**Criterion**	**Example**
Cherry-picking metrics	Run multiple ablation methods, report only the best number	E1 Intervention reach	IOI: 87% mean ablation reported; resample ablation (40%) omitted
No baseline control	Report high score without testing a random or untrained model	M2 Baseline separation	SAEBench: 2 of 8 SAE metrics indistinguishable from shuffled baselines (Karvonen et al., 2025)
No negative controls	Only test what should pass; never test what should *not* be affected	I3 Specificity C1 Falsifiability	Successor heads: ablation effects on non-successor tasks never reported
Construct incoherence	Target construct is not separable from a neighboring one	C4 Discriminant validity	Gender bias circuits: bias and knowledge share the same heads
Off-target effects	Ablation works for the target task; other tasks never checked	I3 Specificity	Knowledge neurons: editing one fact corrupts related facts (Cohen et al., 2024)
Circuit redefinition	Drop heads that hurt faithfulness, call it "refined"	C1 Falsifiability	Adjusting circuit membership to maximize the target metric
Post-hoc role labeling	Name heads after observed behavior; the label cannot fail	C1 Falsifiability	"Name mover" assigned after seeing logit attribution
No double dissociation	Show necessity but never test the converse	I4 Double dissociation	IOI circuit ablated $\rightarrow$ IOI breaks; SVA circuit $\rightarrow$ IOI intact never tested
Interpretive inflation	Label implies intent or algorithm beyond what evidence shows	V4 Anthropomorphism check	"World model" for Othello (evidence supports "linearly decodable")
Scope creep	Validate on one prompt format, claim generality	V5 Scope decl. E2 Prompt gen.	Greater-than circuit tested only on year ranges; claimed as general ordinal comparison
Flexible scope	Narrow the claim after seeing results so it cannot fail	V5 Scope declaration	Restricting scope post-hoc to the subset of prompts that work
Ignoring alternatives	Circuit works, but so does a different set of heads	I4 Double dissoc. C3 Convergent val.	Meloux et al. find alternative IOI circuits with comparable faithfulness

5 The Framework

The MechVal framework evaluates mechanistic claims through seven layers in pipeline order (Figure 1). Each layer answers a distinct question: what kind of explanation is offered (description mode), what sources of signal support it (evidence families), what was concretely measured (metrics), whether the evidence meets specific conditions (criteria), which dimensions of validity are addressed (validity types), how evidence is aggregated (synthesis), and what the claim has established (verdict).

The framework defines the structure of evaluation—what questions must be answered and in what order. The methodology (Section 6) defines the concrete procedures for generating evidence: 65 metrics (including 15 calibrations), 9 protocols, and 9 synthesis protocols. The framework evaluates; the methodology produces what is evaluated.

The framework's layers are described in the subsections that follow: description modes (Section 5.1), evidence families (Section 5.2), metrics (Section 6), criteria (Section 5.3), validity types (Section 5.4), synthesis protocols, and verdict tiers (Section 5.5).

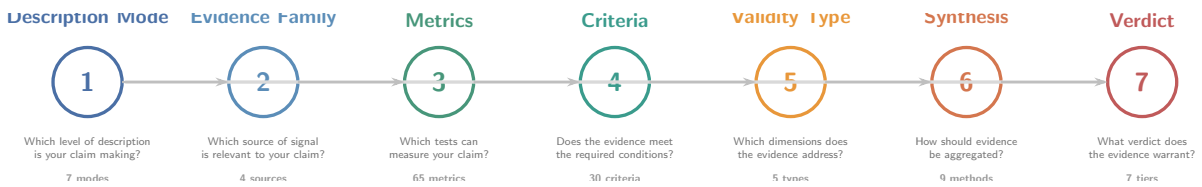

Figure 1: The MechVal framework. Seven layers (7 modes, 4 families × 2 modes, 65 metrics, 30 criteria, 5 types, 9 synthesis protocols, 7 tiers), each answering a distinct question about the claim.

5.1 Description Modes

A description mode is a commitment about what kind of explanation is being offered (Table 3). It is not a label applied after analysis—it is a declaration that determines what counts as evidence, what counts as a gap, and what the finding means if it succeeds. The distinction matters because mechanistic interpretability routinely conflates levels: a paper that discovers *which components* are active (implementational-topographic) writes a narrative about *what the model computes* (computational). Each level upgrade requires additional bridging evidence that the lower level cannot provide.

The seven modes extend Marr's three-level hierarchy (Marr, 1982) by splitting the implementational level into four sub-types—functional, connectomic, statistical, and topographic—reflecting the fact that MI papers routinely conflate claims at these different levels. The modes form a partial order by evidential commitment. A computational claim commits to a function being computed and a reason it is the right function. An algorithmic claim commits to a specific procedure. A representational claim commits to what information is encoded but not how it is processed. The four implementational sub-types commit to progressively weaker structural facts: which components are involved (topographic), how they are wired (connectomic), what their activations look like (statistical), and what input-output transformation each performs (functional).

Table 3: Seven description modes, ordered from most committal (computational) to least committal (topographic). Higher modes require all evidence from lower modes plus bridging evidence. Implementational-statistical is partially orthogonal: it characterizes activation distributions rather than making structural commitments.

	Mode	**What it asks**	**Example**
1	Computational	What function does the system compute, and why?	"GPT-2 predicts the indirect object"
2	Algorithmic	What procedure does it execute?	"A detect-inhibit-copy algorithm"
3	Representational	What information is encoded, and in what geometry?	"The IO name is linearly encoded at layer 8"
4	Impl.-Functional	What transformation does each component perform?	"Head 9.9 copies names to the output"
5	Impl.-Connectomic	How are components wired?	"S-inhibition feeds name movers via residual stream"
6	Impl.-Statistical	What do activations look like?	"Head 9.9 attends strongly to IO position"
7	Impl.-Topographic	Which components are involved?	"Heads 9.9, 9.6, 10.0"

A claim may carry evidence at several modes simultaneously; the framework does not average them. Each mode-level sub-claim receives its own assessment, and the overall verdict is reported at the strongest mode whose required evidence is complete, with lower-mode evidence noted as supporting and higher-mode evidence as suggestive. The mode assigned to a verdict is thus the strongest level for which all required evidence is present. Partial evidence at a higher level does not upgrade the mode—it is reported as suggestive while the verdict is issued at the established mode. For example, a claim at the algorithmic level ("the model runs a detect-inhibit-copy algorithm") requires both the topographic evidence (which heads are involved) and the additional evidence that the heads implement the claimed procedure—path-level causal tracing, timing consistency, and sufficiency of the minimal circuit.

5.2 Evidence Families

An evidence family classifies a metric's output by the source of signal it draws on, independent of the specific tool used to generate it. Each family crosses with a second axis—observational vs. interventional—producing eight cells (Table 4). Most published MI evidence lives in a single cell (interventional activations: patching and ablation). Triangulation means covering multiple cells, and the observational/interventional axis matters as much as the family axis for independence of failure modes: two interventional-activation metrics share confounds that an observational-weight metric does not.

Table 4: Evidence families × evidence mode. Each cell contains representative methods. Convergent evidence across cells with independent failure modes is qualitatively stronger than redundant evidence from the same cell. Most published MI work concentrates in interventional activations (bold).

	Observational	**Interventional**
Weights	SVD, effective rank, OV/QK composition, minimum description length	Circuit transplant, weight editing, weight knockout
Activations	Probing, CKA, mutual information, partial information decomposition, logit lens	**Activation patching, DAS-IIA, ablation, steering**
Behavior	Logit diff, behavioral profiling, task accuracy	Prompt perturbation, dose-response, KL under ablation
Training	Loss curves, checkpoint comparison, phase transitions	Ablate-then-retrain, fine-tuning, curriculum manipulation

Each family answers a different question about the mechanism and has characteristic strengths and blind spots. Weight-based metrics cannot be confounded by runtime compensation but cannot distinguish used capacity from unused capacity. Activation-based metrics can establish necessity and sufficiency directly but are vulnerable to backup circuits that compensate after ablation. Behavioral metrics test whether the circuit reproduces the model's outputs, but behavioral equivalence does not imply mechanistic equivalence—different circuits can produce the same behavior for different reasons. Training-based metrics reveal developmental trajectories but cannot establish that a mechanism is causally responsible for current behavior.

Some methods straddle multiple families. Gradient-based attribution (EAP, integrated gradients) computes counterfactual signals through both weights and activations jointly—it is neither purely observational nor interventional. Causal scrubbing intervenes on activations but evaluates through behavioral output. Automated circuit discovery methods like ACDC compose interventional-activation and behavioral evidence across multiple rounds. The family classification assigns each method to its *primary* source of signal; methods that draw on multiple families are noted as such, and their cross-family nature is itself informative for convergent validity assessments.

Within each family, interventional evidence is epistemically stronger than observational evidence because it rules out confounds that observation alone cannot. When evaluating a claim, the analyst asks: how many cells are covered? Which families and modes are represented? And are there cells that *should* support the claim but do not? Absence of expected evidence is informative—it may indicate a gap in the claim or a limitation of the proposed mechanism. Measurement calibrations—bootstrap stability, seed variance, baseline separation—cut across all eight cells and are described in Section 6.2.

5.3 Criteria

Criteria are specific, falsifiable conditions that must be met for a validity type to be satisfied (Table 5). Each criterion has a concrete pass condition—for example, I1 (necessity) requires that ablating the component degrades performance across at least two independent methods, not just one. A claim either passes, partially passes, or fails each criterion, producing a structured validity profile rather than a single score.

Table 5: Thirty criteria across five validity types. Each criterion has a concrete test; a claim either passes, partially passes, or fails, producing a structured validity profile.

ID	Criterion	Description
C1	Falsifiability	Can the claim be refuted?
C2	Structural plausibility	Is the mechanism physically possible in the architecture?
C3	Convergent validity	Do multiple independent methods agree?
C4	Discriminant validity	Does the measure distinguish this from neighboring constructs?
C5	Nomological validity	Does the claim fit into a broader theory?
M1	Reliability	Do repeated measurements give the same answer?
M2	Baseline separation	Is the score distinguishable from random/untrained baselines?
M3	Stability	Is the classification robust to perturbation?
M4	Calibration	Are the numbers meaningful?
M5	Sensitivity	Can the instrument detect known-true effects?
M6	Invariance	Does the metric behave consistently across conditions?
I1	Necessity	Is the circuit required for the behavior?
I2	Sufficiency	Is the circuit enough to produce the behavior?
I3	Specificity	Does the circuit do this task and not everything?
I4	Double dissociation	Can this circuit be separated from other circuits?
I5	Confound control	Are alternative explanations ruled out?
I6	Epistatic interaction	Do circuit components interact non-additively?
I7	Rescue reversibility	Does restoring a corrupted component recover behavior?
I10	Confounding sensitivity	How strong must an unmeasured confounder be to explain the result?
E1	Intervention reach	Do different intervention methods agree?
E2	Prompt generalization	Does it work on diverse prompts?
E3	Cross-task generalization	Does the mechanism transfer to related tasks?
E4	Cross-model generalization	Does the mechanism appear in other models?
E5	Graded response	Does partial ablation produce partial effects?
E6	Novel prediction	Does the mechanism predict new, untested behaviors?
V1	Level declaration	At what description mode is the claim made?
V2	Level-evidence match	Does the evidence support claims at that level?
V3	Alternative level	Could the evidence be explained at a different level?
V4	Anthropomorphism check	Is the interpretation projecting human concepts?
V5	Scope declaration	What does the claim explicitly not cover?

Table 6: Criterion status categories. Status labels summarize the evidential state of an individual criterion, not the overall verdict for the claim.

Status	**Meaning**
Confirmed	Tests capable of exposing the relevant failure mode were conducted and passed.
Partially confirmed	Supportive evidence exists, but it is too narrow, method-dependent, weakly calibrated, or insufficiently severe to establish the criterion.
Inconclusive	Relevant tests give conflicting results across methods, datasets, interventions, calibrations, or scopes.
Disconfirmed	An appropriate test was conducted and the criterion failed.
Untested	No test capable of evaluating the criterion has been conducted.

For each validity criterion, we assign a criterion status from an ordered five-level scale: Confirmed, Partially confirmed, Inconclusive, Disconfirmed, or Untested. These statuses are ordinal verdict categories in the error-statistical tradition of Mayo (1996; 2018): a criterion is Confirmed if and only if the relevant experimental tests were conducted under conditions of sufficient severity to have detected failure, and the criterion passed those tests. Following Achinstein (2001), we further distinguish criteria where the available evidence constitutes a genuine epistemic reason to accept the criterion (Confirmed) from those where evidence is present but falls short of that bar (Partially confirmed, Inconclusive).

This framework is non-probabilistic: status assignments do not represent posterior beliefs or confidence levels. Instead, following Campbell & Stanley (1963) and Cook & Campbell (1979), validity claims are evaluated by whether the relevant tests rule out plausible rival explanations. The Untested status applies when no test capable of evaluating the criterion has been conducted, whereas Inconclusive indicates contested evidence and Disconfirmed indicates that a test was conducted and the criterion failed.

These statuses evaluate individual criteria, not the overall claim. A claim can therefore contain Confirmed internal-validity criteria, Untested measurement-validity criteria, and Mixed external-validity criteria while still receiving a single overall verdict tier. This keeps criterion-level strengths, gaps, and failures explicit instead of compressing them into an undifferentiated verdict.

5.4 Validity Types

The five validity types form a dependency chain (Shadish et al., 2002): later types cannot be established without first satisfying earlier ones. The ordering reflects logical precedence: a construct that is not well-defined cannot be reliably measured, an unreliable measurement cannot support a causal claim, a causal claim in one setting cannot be generalized, and an interpretation cannot be evaluated until the underlying evidence is characterized.

$$\text{Construct} \rightarrow \text{Measurement} \rightarrow \text{Internal} \rightarrow \text{External} \rightarrow \text{Interpretive} \tag{1}$$

The pipeline (Figure 1) is procedural—it describes how evidence is built or audited. The validity-type chain is logical—it describes which types presuppose which. At pipeline step 5 (validity type), the framework evaluates every validity type the claim asserts, but the dependency order constrains the verdict: a claim cannot be credited with external validity beyond the level its measurement and internal validity support. The chain is a ceiling, not a sequence of gates.

Each type addresses a distinct question about the claim:

Construct validity. Is the thing being measured well-defined? Before any measurement is taken, the construct must be specified precisely enough that the claim is falsifiable, structurally plausible, and dis-

tinguishable from neighboring constructs. A "deception feature" that cannot be distinguished from an "uncertainty feature" fails construct validity regardless of how carefully it is measured. This draws on philosophy of science (Popper's falsifiability, Hempel's operationalism) and psychometrics (construct validity, convergent and discriminant validity).

Measurement validity. Are the instruments trustworthy? Borsboom et al. (2004) argue that a measure is valid if and only if the attribute exists and variation in it causally produces variation in the measurement outcome. This causal account grounds the framework's measurement criteria. Once the construct is defined, the measurement tools used to detect it must be reliable (stable across repetitions), calibrated (separable from random baselines), and invariant (consistent across conditions). A metric that gives different answers when run with different random seeds is not evidence. This draws on psychometrics (test-retest reliability, measurement invariance) and pharmacology (assay validation).

Internal validity. Does the evidence support the causal claim? Given a well-defined construct and trustworthy instruments, the evidence must establish that the identified component is causally involved in the behavior—not merely correlated with it. This requires demonstrating necessity (the component is required), sufficiency (the component is enough), and specificity (the component does this task and not everything). Critically, a component that is causally necessary for many tasks is a bottleneck, not an implementation—double dissociation, not single dissociation, is the test for specificity. This draws on neuroscience (lesion vs. stimulation, single vs. double dissociation, mutual manipulability; Shallice, 1988; Craver, 2007), causal inference (do-calculus, counterfactual frameworks, instrumental variables; Pearl, 2009), and genetics (knockout vs. rescue, epistasis mapping for non-additive component interactions, sensitivity analysis for unmeasured confounders).

External validity. Does the mechanism generalize? Causal evidence on one prompt set, with one ablation method, in one model, establishes internal validity at best. External validity requires that the mechanism generalizes across prompts, ablation methods, related tasks, and ideally across models. Beyond replication, it requires characterizing the effect quantitatively: how it scales with intervention strength (the dose-response curve), whether it is selective for the claimed task versus related tasks (the selectivity ratio), and whether its absolute magnitude supports the computational story. This draws on pharmacology (Phase III generalization, dose-response, therapeutic windows; Rang, 2006) and causal inference (transportability; Pearl & Bareinboim, 2011).

Interpretive validity. Is the interpretation of the mechanism correct? The natural-language description of the mechanism must be justified by the evidence. A head called a "name mover" based on direct logit attribution carries theoretical implications beyond what the evidence strictly supports. Interpretive validity requires declaring the Marr level at which the claim operates—implementational, algorithmic, or computational—matching evidence to that level, and checking that the narrative is consistent across evidence families. Méloux et al. (2025) prove that on Boolean MLPs small enough to enumerate, multiple circuits replicate the same behavior and multiple interpretations fit the same circuit; alternative exclusion is therefore a required step, not an optional one. This draws on Messick's unified theory of validity (Messick, 1995)—which frames validity as the warrant for *interpretations* of scores, not properties of the test itself—and on Marr's levels of analysis (Marr, 1982).

5.5 Verdict Tiers

Claims are assigned to one of five verdict tiers representing qualitative transitions in evidential status (Table 7). Two additional labels—Underdetermined and Disconfirmed—handle special cases. The tiers form a strict hierarchy: each requires all evidence from the tier below plus additional evidence.

Table 7: Verdict tiers with minimum evidence requirements. Each tier subsumes all requirements of lower tiers. Underdetermined and Disconfirmed are not tiers in the hierarchy but diagnostic labels.

Tier	**Meaning**	**Minimum evidence requirements**
Proposed	Structural or representational evidence only	Construct defined under a declared description mode (C1–C2); at least one admissible measurement conducted
Causally Suggestive	Necessity shown, sufficiency not established	Necessity via causal intervention (I1 confirmed); at least one measurement passes baseline separation (M2)
Mechanistically Supported	Necessity + sufficiency with consistent methods	Sufficiency established (I2); intervention reach across ≥ 2 ablation methods (E1); specificity test conducted (I3 at least partially confirmed)
Triangulated	Multiple converging lines of independent evidence	Convergent evidence from ≥ 3 evidence families (C3); cross-distribution replication (E2 or E4); double dissociation attempted (I4)
Validated	All five validity types addressed	Measurement calibration (M1–M6) explicitly addressed; interpretive validity (V1–V5) audited; external validity across models (E4) and prompts (E2)
Underdetermined	Evidence consistent with multiple mechanisms	Cannot resolve between rival specifications
Disconfirmed	Fails decisively on a key criterion	Negative result on a required criterion

Every tier carries a specific epistemic meaning: a claim at *Proposed* is not a failed claim but one whose evidential status—and the path to strengthening it—is precisely characterized. The verdict tells researchers exactly what additional evidence would advance the claim: a Proposed claim needs causal intervention (I1) to reach Causally Suggestive; a Causally Suggestive claim needs sufficiency evidence (I2) and method consistency (E1) to reach Mechanistically Supported.

Underdetermined and Disconfirmed serve distinct purposes. Underdetermined means the evidence is consistent with multiple competing mechanisms and the framework cannot resolve between them—this is not a lack of evidence but an excess of explanations, requiring additional experiments that discriminate between rivals. Disconfirmed means a criterion has been decisively failed—for example, when the construct itself turns out to be incoherent (C4), as in the gender bias circuits case study.

6 Methodology

The framework layers described in Section 5 define what counts as evidence and how it is evaluated. This section provides the concrete measurement procedures—metrics, calibrations, protocols, and synthesis methods—that generate the evidence evaluated against those layers. These are organized by evidence family and by evidential function.

6.1 Metrics

Metrics are concrete, runnable tests that produce measurements about a neural network—the layer of the hierarchy where empirical contact with the model actually happens. Everything above (evidence families, criteria, validity types, verdicts) depends on what metrics measure and how well they measure it. A metric

alone cannot establish a claim; a claim requires evidence from multiple metrics, evaluated against the criteria appropriate to the validity type being asserted.

Each metric takes a model, a task, and optionally an intervention as input, and produces a number as output. Metrics can be composed: one metric's output feeds into another. The framework organizes evidence into four evidence families; the current catalog defines 65 metrics across families A–F, drawing on interventionist, information-theoretic, structural, and generalization traditions of causation. The full inventory is in Appendix A.1.

6.2 Calibrations

Calibrations are meta-metrics: they take another metric's output as input and assess whether that output is stable (bootstrap stability), reproducible (seed variance), or distinguishable from baselines (baseline separation). A high score on a primary metric is not evidence if the corresponding calibration fails. The 14 calibration metrics correspond to criteria M1–M6 (reliability, baseline separation, stability, calibration, sensitivity, invariance). The full catalog is in Appendix A.1.

6.3 Protocols

Protocols are multi-step designed experiments that compose metrics to test criteria that no single metric can assess. Where metrics are atomic measurements, protocols are novel experiment designs—they define new procedures rather than bundling existing metrics. The framework currently defines four: V4 (interpretive inflation control: label scope audits, contrastive label tests, ablation scope tests), I5 (confound control: confound enumeration, causal graph consistency, synthetic ground truth injection), I6 (epistatic interaction: pairwise and higher-order interaction contrasts), and E1 (intervention reach: steering reach, cross-context transfer, counterfactual circuit transplant). The protocol layer is extensible—any multi-step experiment design that composes metrics to test a criterion can be added. The detailed specifications are in Appendix A.2.

6.4 Synthesis Protocols

Synthesis protocols aggregate evidence across multiple methods to produce criterion-level statuses that feed into the verdict tier system. They are not alternatives to the tier promotion rules—they are the machinery that makes those rules operational. When multiple discovery methods each identify a set of circuit components, the question is whether they agree, and how to handle the fact that methods have different reliability profiles. The framework defines nine synthesis protocols, each adapted from an established aggregation method in another field: Jaccard overlap as a naive baseline; Robust Rank Aggregation (Kolde et al., 2012) for ranked lists; Dawid-Skene consensus (Dawid & Skene, 1979) for reliability-weighted voting; parallel and sequential ensembles for fusing or filtering across methods; Wasserstein distributional stability for cross-context generalization; functional parcellation (adapted from Glasser et al., 2016) for identifying component subgroups; a meta-learner that predicts circuit membership from protocol features; Granger causality graphs for cross-family predictive structure; and circuit component reuse analysis (Merullo et al., 2024) for identifying shared components across tasks. The detailed specifications are in Appendix A.3.

7 Theoretical Foundations

This section outlines the theoretical foundations of our framework. We draw on seven disciplines that each developed rigorous standards for a different validity question, and we apply them to mechanistic interpretability as lenses. A lens is not a method or a metric; it is an analytical vocabulary that shapes which questions you ask of a claim and how you read the answers. Each of the seven — philosophy of science, psychometrics, causal inference, neuroscience, pharmacology, genetics, and mechanistic interpretability itself — contributes the conditions it requires before an inference is accepted, and names the failure modes it guards against.

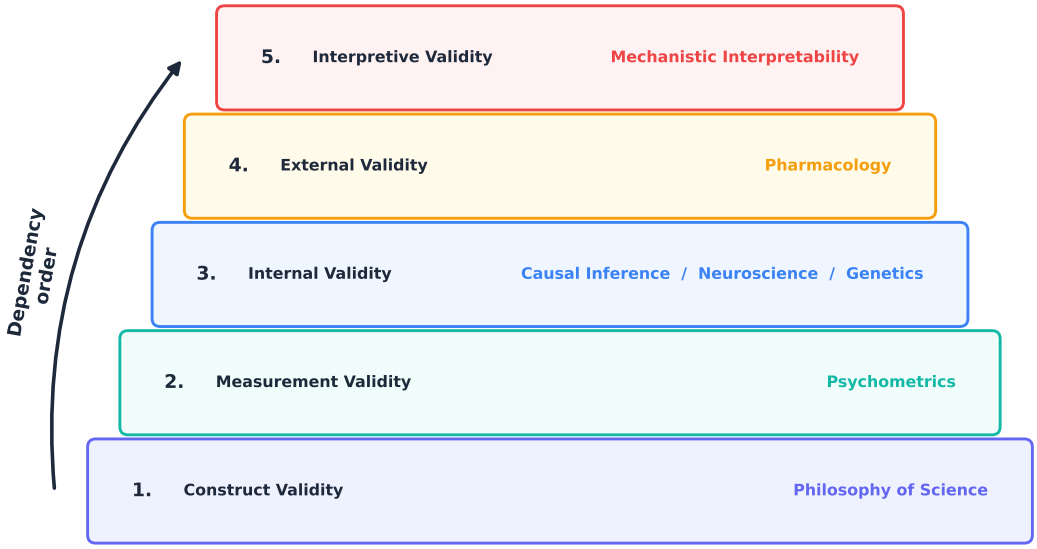

Figure 2: Seven theoretical foundations, each grounding a specific validity type. Causal inference, neuroscience, and genetics all contribute to internal validity through complementary approaches (formal frameworks, experimental designs, and interaction structure, respectively).

Each lens maps to a corresponding validity type, and answers a core question about a given claim: philosophy of science asks *is the construct well-defined?* Causal inference and neuroscience ask *does the component implement the computation?* Psychometrics asks *is the measurement accurate?* Pharmacology asks *does it generalize?* And mechanistic interpretability asks *does the interpretation match the evidence?*

7.1 Philosophy of Science

The construct validity criteria (C1–C5) draw on Popper's falsifiability criterion, Mayo's severe testing (Mayo, 2018), and Glennan's new mechanical philosophy (Glennan, 2017). A key distinction is between *confirmation* and *corroboration*: a circuit discovered by activation patching and evaluated by activation patching has only been shown to pass the test it was built to pass. Corroboration requires a genuinely risky test the circuit was not optimized for—a behavioral prediction on held-out prompts, a weight-space analysis, a different causal method. The verdict tier system reflects this hierarchy: single-method evidence (Causally Suggestive) is qualitatively weaker than convergent evidence from independent methods (Triangulated).

C5 (nomological validity) requires that the claim fit into a broader theoretical network, following Cronbach and Meehl's nomological network concept. The Duhem-Quine underdetermination thesis (Quine, 1951; Duhem, 1906) is directly relevant: observational evidence (faithfulness scores, IIA) is always consistent with multiple theories (circuits), because the theory under test never faces the evidence alone—it is conjoined with auxiliary hypotheses about ablation semantics, baseline distributions, and metric validity. Méloux et al. (2025)'s finding that alternative head sets achieve comparable faithfulness is this thesis instantiated: the "circuit" is underdetermined by the data.

Glymour's bootstrapping theory of confirmation (Glymour, 1980) provides the constructive response. Glymour argues that evidence confirms a hypothesis when it can be used to *compute* a consequence of the hypothesis that is also independently testable—and crucially, this computation must use the hypothesis non-trivially, not just any hypothesis that entails the same prediction. Applied to MI: a circuit claim is bootstrapped when the evidence bears on the claim in a way that would not equally confirm a rival circuit.

Convergent evidence from independent methods (ablation, weight analysis, information flow) satisfies this condition because each method's auxiliary hypotheses are different; a finding that survives across methods has been non-trivially tested. This is the formal justification for requiring convergent evidence (C3) at higher tiers: single-method evidence cannot resolve underdetermination, but cross-method convergence can progressively narrow the space of compatible theories.

Table 8: Philosophy of science foundations.

Concept	**Source**	**Criterion**	**MI analog**
Falsifiability	Popper (1959)	C1	What evidence would refute "this is a name mover"?
Nomological network	Cronbach & Meehl (1955)	C5	Does the claim connect to in-context learning theory?
MTMM matrix	Campbell & Fiske (1959)	C3, C4	Convergent: probing + patching agree. Discriminant: feature $\neq$ neighboring feature
Underdetermination	Quine (1951); Duhem (1906)	I4	Multiple circuits explain the same behavior
Bootstrapping	Glymour (1980)	C3	Cross-method convergence narrows the space of compatible theories
Severe testing	Mayo (2018)	C1	A test that could easily have failed but did not

7.2 Psychometrics

The measurement validity criteria (M1–M6) draw on psychometric test construction. Classical test theory decomposes an observed score into true score plus error ($X = T + E$); M1 (reliability) is the proportion of true-score variance. Generalizability theory extends this by decomposing error into distinct sources—prompt variation, random seed, checkpoint selection—each of which can be quantified and reduced independently.

Table 9: Psychometrics foundations.

Concept	**Source**	**Criterion**	**MI analog**
Classical test theory	Lord & Novick (1968)	M1	Same metric, same model $\rightarrow$ same answer
Generalizability theory	Cronbach et al. (1972)	M6	Decompose variance into prompt/seed/checkpoint
Signal detection (d')	Green & Swets (1966)	M2	Is the score distinguishable from random?
MTMM matrix	Campbell & Fiske (1959)	C3, C4	Convergent + discriminant validity
Selectivity	Hewitt & Liang (2019)	I3	Probe accuracy minus control accuracy

The closest prior art in psychometrics is Borsboom et al. (2004)'s causal account of validity: a measure is valid if and only if the attribute exists and variation in it causally produces variation in the measurement outcome. This is our core claim, extended from trait measurement to mechanism claims. Campbell and

Fiske's (Campbell & Fiske, 1959) multi-trait multi-method matrix directly grounds our convergent (C3) and discriminant (C4) validity criteria. Where we go beyond the psychometric tradition: psychometrics assumes the construct exists and asks whether the measure is valid. We add that mechanistic validity requires first declaring the *description mode*—what kind of thing the construct is—because different modes generate different validity criteria and different evidence requirements. A claim about "which components are active" (implementational-topographic) and a claim about "what algorithm the model runs" (algorithmic) require fundamentally different evidence, even when they concern the same model and the same task.

An important insight is that a reliable metric pointed at the wrong target produces confident wrong answers. Sutter et al. (2025) demonstrated that unconstrained nonlinear IIA achieves near-perfect scores on randomly initialized models—but this is a property of the *unconstrained* alignment map, not of nonlinear causal abstraction in general. Structured generative models that satisfy identifiability conditions (Khemakhem et al., 2020)—reconstruction objectives, label-conditional priors, sparsity constraints—provably avoid this degeneracy: intervention diversity ratios near 1.0 confirm that the encoder preserves natural variation rather than collapsing to a lookup table. The framework addresses the general problem through M2 (baseline separation): not "IIA = 0.48" but "$\Delta_{\text{random}} = 0.10$, $\Delta_{\text{untrained}} = 0.15$." A modest absolute number with a large delta is a genuine finding; a large absolute number with a tiny delta is not. The SAEBench audit (Karvonen et al., 2025) provides a complementary example: two of eight standard SAE evaluation metrics fail baseline separation entirely, producing scores on shuffled or random baselines that are indistinguishable from scores on real features.

7.3 Causal Inference

The internal validity criteria draw on causal inference methodology (Pearl, 2009; Spirtes et al., 2000). The necessity/sufficiency distinction maps to Woodward's interventionist causation (Woodward, 2003): necessity (I1) corresponds to "would the behavior persist if the component were absent?" and sufficiency (I2) to "would the component alone produce the behavior?" Pearl and Bareinboim's transportability theory (Pearl & Bareinboim, 2011) provides the formal conditions under which causal findings transfer across settings, grounding E4 (cross-model generalization). Rubin's potential outcomes framework grounds E5 (graded response) by formalizing heterogeneous treatment effects across input subpopulations—not all prompts may respond to ablation equally.

A distinction that the MI literature routinely elides: ablation and activation patching are formally different do-operators. Ablation performs $\text{do}(h := 0)$ or $\text{do}(h := \mathbb{E}[h])$—it removes a component. Patching performs $\text{do}(h := h')$ where h' comes from a counterfactual input—it inserts a specific value. These correspond to different interventional distributions and can yield different causal conclusions. The method-conditionality finding (E1)—where mean ablation and resample ablation produce different faithfulness numbers—partly reflects this: they are not two implementations of the same intervention but formally different interventions that happen to share a name. Making this explicit sharpens E1: cross-method consistency requires agreement not only across ablation variants but across genuinely distinct do-operators.

To illustrate how do-calculus formalizes a circuit claim, consider a minimal structural causal model for IOI. Let S = subject name, IO = indirect object, H_{SI} = S-inhibition heads, H_{NM} = name-mover heads, and Y = output logit. The claimed mechanism is: $S \to H_{\text{SI}} \to H_{\text{NM}} \to Y$ and $IO \to H_{\text{NM}} \to Y$. The necessity claim (I1) is $P(Y \mid \text{do}(H_{\text{NM}} := 0)) \neq P(Y)$; the sufficiency claim (I2) is $P(Y \mid \text{do}(\text{all except circuit} := 0)) \approx P(Y)$; and the specificity claim (I3)—currently untested—is $P(Y_{\text{SVA}} \mid \text{do}(H_{\text{NM}} := 0)) \approx P(Y_{\text{SVA}})$ for an unrelated task like subject-verb agreement. The SCM makes the untested link visible: I3 is a conditional independence claim that has never been checked.

Table 10: Causal inference foundations.

Concept	Source	Criterion	MI analog
do-calculus / SCM	Pearl (2009)	I1, I2	Formal language for ablation claims
Interventionism	Woodward (2003)	I1	"Would behavior persist if component absent?"
Transportability	Pearl & Bareinboim (2011)	E4	Conditions for circuit generalization across models
Causal discovery	Spirtes et al. (2000)	—	Learning circuit structure from data
Potential outcomes	Rubin (1974)	E5	Heterogeneous effects across input subpopulations

7.4 Neuroscience

The internal validity criteria (I1–I5) are adapted from neuroscience's toolkit for establishing that a neural circuit implements a computation. I1 (necessity) corresponds to lesion studies; I2 (sufficiency) to stimulation; I4 (double dissociation) is the gold standard for functional specificity (Woodward, 2003). The critical distinction between necessity and sufficiency—well-established in neuroscience but often elided in MI—is central to the framework and marks a qualitative transition between verdict tiers.

A less appreciated requirement is Craver's *mutual manipulability*: constitutive relevance requires not only that intervening on the component changes the behavior, but that intervening on the behavior changes the component's activity. We depart from Craver's standard account on one point. Mutual manipulability presupposes that mechanism *identity* is settled before intervention—you already know what you are intervening on. But when methods disagree (activation patching says head 4.4 matters, weight geometry says it does not), mutual manipulability provides no adjudication procedure, because the identity criterion is ambiguous between description modes. Our framework resolves this by requiring the description mode to be declared *before* intervention, making the identity criterion explicit rather than inherited silently. This is an extension, not a rejection: mutual manipulability remains the correct criterion once the description mode is fixed. The second direction—does fine-tuning away IOI capability cause name-mover heads to lose their OV copying structure?—is almost never tested in MI, yet it is what "implementation" actually requires. Additionally, the *parcellation problem* is directly relevant: neuroscience learned that the right unit of analysis is defined by convergence across independent modalities, not by any single method. When EAP, weight classification, and ablation disagree on component membership, the "circuit" is probably method-dependent.

A further distinction that neuroscience draws—and MI often does not—is between *structural* and *functional* connectivity. Two brain regions can be structurally connected (axons linking them) without being functionally connected during a specific task (no task-relevant information flows between them during that computation). In MI, two attention heads can be compositionally connected in weight space (one's output lies in the other's input subspace) without actually passing task-relevant information during inference. Weight-space composition scores and OV circuit analysis measure structural connectivity; path patching and activation correlation during task performance measure functional connectivity. A circuit claim grounded only in structural connectivity ("these heads compose") is weaker than one grounded in functional connectivity ("task-relevant information flows along this path during this computation"). The strongest claims establish both and show they converge.

Table 11: Neuroscience foundations.

Concept	Source	Criterion	MI analog
Lesion studies	Shallice (1988)	I1	Ablation degrades behavior
Double dissociation	Shallice (1988)	I4	Circuit A breaks task X not Y; circuit B breaks Y not X
Mutual manipulability	Craver (2007)	I3	Fine-tuning away the task changes the component
Multimodal parcellation	Glasser et al. (2016)	C3	Circuit boundaries should converge across methods
Neural manifolds	Gallego et al. (2017)	V2	Representation geometry constrains algorithmic claims

7.5 Genetics

The six lenses above leave a gap in *interaction structure*: neuroscience provides single-component interventions (lesion, stimulation), but does not ask whether circuit components interact non-additively. Molecular genetics fills this gap. Over a century of work from *Drosophila* to human GWAS cohorts, genetics developed three techniques that map directly onto MI interventions and answer questions no other lens covers.

First, *epistasis mapping*. Fisher defined epistasis as the non-additive interaction between genetic loci: if knocking out gene A reduces the phenotype by a and gene B by b, the epistasis score $\epsilon_{AB} = f(\text{ablate } A, B) - f(\text{ablate } A) - f(\text{ablate } B) + f(\emptyset)$ measures the deviation from additivity. Positive ϵ (synergy) means the components cooperate; negative ϵ (buffering) means they partially substitute for each other. In MI, pairwise ablation studies that compute ϵ for all component pairs reveal whether a circuit is a functional unit with internal coupling or merely a list of independently necessary components. This grounds criterion I6 (epistatic interaction).

Second, *rescue experiments*. In genetics, a rescue experiment re-introduces the wild-type gene into a knockout organism to verify that the phenotypic deficit was specifically caused by the gene's absence rather than by collateral damage. In MI, corrupting a circuit component (mean or resample ablation) and then restoring the clean activation at that site tests whether the deficit is specifically reversible—distinct from sufficiency (I2), which asks whether the circuit alone produces the behavior. A rescue that succeeds establishes that the ablation deficit reflected genuine loss of computation, not cascading distributional disruption. This grounds criterion I7 (rescue reversibility).

Third, *sensitivity analysis*. The E-value quantifies how strong an unmeasured confounder would need to be to explain away an observed causal effect. Every ablation study has potential unmeasured confounders—information in the residual stream that correlates with both the ablated component and the output. An E-value of 3.0 means a confounder would need to triple both associations to nullify the finding. This complements confound control (I5), which addresses *known* confounders, by quantifying robustness to confounders the experimenter did not measure. This grounds criterion I10 (confounding sensitivity).

Table 12: Genetics foundations.

Concept	Source	Criterion	MI analog
Epistasis	Fisher (1918)	I6	Pairwise ablation reveals synergy or buffering
Rescue / complementation	Brenner (1974)	I7	Restore corrupted component; does behavior recover?
E-value / sensitivity	VanderWeele & Ding (2017)	I10	How strong must an unmeasured confounder be?

The analogy has limits. In genetics, knockouts are permanent: the organism develops without the gene, and compensatory mechanisms may emerge over developmental time. In neural networks, ablation is instantaneous—the component was present during all of training. Network "knockouts" are closer to acute pharmacological blockade than to developmental gene deletion, which makes rescue experiments (I7) especially important for distinguishing genuine computational loss from distributional disruption.

7.6 Pharmacology

The external validity criteria draw on pharmacology's dose-response methodology (Rang, 2006). A causally relevant component should produce graded effects: partial ablation should produce partial behavioral change, and the relationship should be monotonic or at least systematic (E5). The framework treats dose-response as a minimum bar for causal claims about magnitude—a requirement absent from most MI evaluations, which typically report binary ablation (all-or-nothing).

The *affinity vs. efficacy* distinction maps with surprising precision to MI. Activation patching measures affinity: is this component engaged during the task? Ablation measures efficacy: does removing it change the output? But even ablation conflates efficacy with the system's compensatory capacity—a load-bearing head can show small ablation effects if backup mechanisms compensate. The observed effect magnitude is a joint property of the component's role and the network's reserve, and a single ablation cannot separate them. The practical implication is the *therapeutic window*: sweeping ablation strength from 0 to 1 reveals whether there is a range where on-task effects grow before off-target degradation begins. A circuit with no therapeutic window cannot be cleanly separated from the network's general processing.

Table 13: Pharmacology foundations.

Concept	Source	Criterion	MI analog
Dose-response curve	Hill (1910)	E5	Partial ablation $\rightarrow$ partial effect
Affinity vs efficacy	Stephenson (1956)	I1 vs I2	Participation $\neq$ causation
System reserve	Black & Leff (1983)	I2	Backup circuits compensate after ablation
Functional selectivity	Kenakin (2004)	E1	Ablation method determines the finding
Phase I/II/III trials	—	Verdict tiers	Calibrate $\rightarrow$ establish causality $\rightarrow$ generalize

7.7 Mechanistic Interpretability

Interpretive validity (V1–V5) is the one validity type not imported from another field. Every MI result pairs a measurement with an interpretation—an IIA score with "the model represents this variable here," a

faithfulness percentage with "this circuit implements IOI." Whether the interpretation is justified depends on whether the claim is stated at a level the evidence can actually support. Three distinctions, drawn from MI's own empirical track record, ground the interpretive criteria.

First, *description vs. explanation*: most published "explanations" are descriptions in explanatory language. "The name-mover head copies the IO name to the output" sounds like an explanation but is a description of observed behavior on specific inputs. An explanation would answer: what structural property of the W_{OV} matrix forces name-copying? Under what conditions would this head fail? Descriptions are bound to the cases observed; explanations predict what happens in cases not yet tested.

Second, *faithfulness vs. understanding*: a circuit can be maximally faithful—100% performance recovery when isolated, total degradation when ablated—without being understood at all. Faithfulness is a property of the circuit relative to the task; understanding is a property of the researcher relative to the circuit. They are independent axes.

Third, *component identity vs. component role*: "Head 9.9 is in the IOI circuit" (identity) and "Head 9.9 is a name mover" (role) are different claims requiring different evidence. Identity requires only causal evidence—ablating the component changes the output. Role requires mechanistic evidence—the weight structure matches the claimed function, the component does not perform this function on non-target inputs, and the role label makes predictions that can be tested independently. The slide from identity to role typically happens in one sentence and is a common overclaim (Wang et al., 2023).

These distinctions motivate the interpretive criteria: V1 (level declaration) requires stating the description mode before collecting evidence; V2 (level-evidence match) checks that the evidence supports claims at that level; V3 (alternative level) asks whether the evidence could be explained at a different level; V4 (anthropomorphism check) flags labels that project human concepts; and V5 (scope declaration) requires stating what the claim does not cover.

Table 14: Mechanistic interpretability foundations.

Concept	**Source**	**Criterion**	**MI analog**
Description vs explanation	Olsson et al. (2022)	V2	Most "explanations" are descriptions in explanatory language
Faithfulness vs understanding	Wang et al. (2023)	V1, V2	Independent axes: a faithful circuit need not be understood
Identity vs role	Wang et al. (2023)	V4	"In the circuit" (causal) vs "is a name mover" (mechanistic)
Circuit non-uniqueness	Méloux et al. (2025)	V3	Activation patching returns *a* circuit, not *the* circuit
Level drift	Marr (1982)	V1, V2	Implementational evidence narrated in computational language

8 Case Studies

We apply the framework to 13 published MI results, evaluating each through all five validity lenses. The case studies are not intended to rank papers—they demonstrate what the framework looks like in practice and show that the tier system discriminates between well-validated and poorly-validated claims. Quantitative per-paper scoring is the subject of a companion paper; here we keep the analysis qualitative and trace each verdict to specific missing or present evidence.

8.1 MechVal Audit

Table 15 presents the verdict assignments. The key finding is that the tier system produces meaningful discrimination: the 13 claims span the full range from Proposed to Triangulated, with verdict differences traceable to specific missing evidence rather than arbitrary scoring.

Table 15: Verdict assignments for 13 published circuits, sorted by evidential status (see Table 7 for tier definitions).

Verdict	**Circuit**	**Evidence characterization**
Triangulated	Grokking (Nanda et al., 2023)	Mathematically complete within toy scope; every weight matrix entry predicted by closed-form Fourier formula
Triangulated	Induction Heads (Olsson et al., 2022)	Simple mechanism, broad replication, thick nomological network
Mechanistically Supported	Greater-Than (Hanna et al., 2023)	Strong structural plausibility; limited prompt generalization
Mechanistically Supported	Copy Suppression (McDougall et al., 2023)	Unusually clean specificity; sufficiency partially established
Mechanistically Supported	IOI Circuit (Wang et al., 2023)	Method-conditional faithfulness (87% mean ablation, <50% other methods); specificity untested
Mechanistically Supported	Superposition (Elhage et al., 2022)	Toy-model-only evidence; criteria semi-confirmed rather than confirmed
Causally Suggestive	Successor Heads (Gould et al., 2023)	Cross-domain generalization as convergent evidence; ablation method-conditional
Causally Suggestive	Othello World Model (Li et al., 2023)	"World model" label exceeds evidence ("linearly decodable"); interpretive inflation
Causally Suggestive	Docstring Circuit (Heimersheim & Janiak, 2023)	Label ambiguity: "variable binding" vs. simpler "positional copying"
Causally Suggestive	Knowledge Neurons (Dai et al., 2022)	Strong intervention but weak specificity; editing corrupts related facts
Causally Suggestive	SAE Features (Cunningham et al., 2024; Engels et al., 2025)	Thin nomological network; features may be dictionary properties, not model properties
Proposed	Probing Classifiers (Belinkov, 2022)	Decodability without intervention = no internal validity
Disconfirmed	Gender Bias Circuits (Vig et al., 2020)	Construct incoherence: bias and knowledge share the same heads; discriminant validity (C4) fails—the construct itself is not well-posed

8.2 Illustrative Examples

We describe three case studies in detail to illustrate how the framework produces its verdicts. These assessments hold claims against an idealized standard; a low verdict marks distance from that ideal, not a flawed study.

Induction Heads (Triangulated). Olsson et al. (2022) identify a two-head composition mechanism for in-context copying: previous-token heads attend to the token before a repeated sequence, and induction heads use this signal to predict the next token. The claim reaches Triangulated because it satisfies criteria across multiple independent evidence families. The mechanism has been replicated across model scales and architectures (E4), confirmed through both activation-level and weight-level analysis (C3 convergent validity from different families), and produces clean dose-response effects (E5). The nomological network is thick: the mechanism connects to in-context learning theory, training dynamics, and cross-model structural predictions. The relative simplicity of the mechanism (two functional roles, clear compositional structure) makes each criterion easier to satisfy.

IOI Circuit (Mechanistically Supported). Wang et al. (2023) identify 26 attention heads forming a six-role circuit for indirect object identification. The circuit demonstrates strong necessity (I1) and sufficiency (I2): running the model with everything outside the circuit mean-ablated recovers 87% of the full model's logit difference. However, Miller et al. (2024) show that this 87% figure is specific to mean ablation; under other ablation methods, faithfulness drops below 50%. This method-conditional result partially satisfies E1 (intervention reach) but reveals that the causal claim is weaker than the headline number suggests. Specificity (I3) is untested: the authors do not report whether ablating the IOI circuit degrades unrelated tasks. A formal double-dissociation has not been conducted. The circuit reaches Mechanistically Supported rather than Triangulated because the evidence comes primarily from one methodological family (ablation-based) and key criteria (I3, I4, I5, E4) remain unaddressed.

Probing Classifiers (Proposed). Linear probing (Belinkov, 2022) trains a classifier on activations to test whether a concept is linearly decodable. The core limitation is fundamental: *a measurement without an intervention or observational causal method is a measurement without internal validity.* Probe success establishes that information is decodable from the activation space, but not that the model uses that information during inference. High probe accuracy is consistent with genuine encoding, incidental linear separability in high-dimensional space, and confound encoding (a correlated feature rather than the target). Without causal follow-up (DAS, intervention along the probe direction), the claim cannot advance beyond Proposed. Probing *with* causal follow-up can reach Causally Suggestive; with additional control tasks and cross-method convergence, it can reach Mechanistically Supported. Probing is not dismissed—its evidential reach is bounded.

Gender Bias Circuits (Disconfirmed). Gender bias circuits (Vig et al., 2020) are the only claim in our evaluation that reaches Disconfirmed—and the reason is not insufficient evidence but an incoherent construct. The claim requires that "gender bias" and "legitimate gender knowledge" be separable: that one could ablate the circuit responsible for biased predictions without degrading the model's ability to correctly resolve gendered pronouns. But the attention heads identified as mediating gender bias (via causal mediation analysis) are the same heads that encode grammatical gender agreement. Ablating these heads reduces biased completions, but it equally degrades correct pronoun resolution on unbiased inputs.

The dose-response curve has no regime where bias decreases without knowledge also decreasing—there is no therapeutic window, in the pharmacological framing. Sweeping ablation strength from 0 to 1 produces a single monotonic curve for both "biased output reduced" and "correct gendered output reduced," because the underlying representation does not distinguish between the two. This is not a case where more precise surgery would help: the construct "gender bias circuit" presupposes a separation that does not exist in the model's weight space.

This is a discriminant validity (C4) failure: the measure cannot distinguish the target construct from its neighbor. One might argue that the construct needs refinement rather than rejection—perhaps "bias" should be defined relative to corpus frequency rather than as an absolute property. But any such refinement must first demonstrate that the refined construct is separable from gender knowledge in the model's representations, which is a testable prediction the current evidence does not support. Unlike the other case studies, where the path forward is more evidence, the path forward here is reformulation of the question. The framework surfaces this distinction: a claim can fail not because the experiments were poorly done but because the construct was not well-posed.

8.3 Recurring Validity Gaps

Several patterns emerge from evaluating these claims side by side.

Sufficiency gap. Most circuits demonstrate necessity (I1) but not sufficiency (I2). The I1→I2 transition—from "removing this breaks the behavior" to "this alone produces the behavior"—is the most common barrier between Causally Suggestive and Mechanistically Supported.

Method-conditional results. The IOI circuit's headline numbers are specific to mean ablation. Ablation type is part of the causal claim, not an implementation detail. The framework captures this through E1 (intervention reach), which requires agreement across ablation methods.

Toy-model ceiling. Grokking reaches Triangulated because the evidence is mathematically complete within its toy scope, with convergent support from three evidence families: activations (ablation and restricted training runs), weights (closed-form Fourier weight analysis), and training (grokking dynamics and phase transitions), while Superposition reaches Mechanistically Supported because its criteria are semi-confirmed rather than confirmed. Both illustrate how the framework handles scope: a claim validated within a declared scope is a legitimate result, not a failure (V5). The gap between toy-model proof-of-concept and real-model confirmation remains the field's central challenge.

No circuit reaches Validated. The Validated tier requires all five validity types to be addressed: construct, measurement, internal, external, and interpretive. No published circuit currently meets this bar. Induction heads come closest—they have strong construct, internal, and external validity—but lack systematic measurement calibration (M1–M6) and explicit interpretive auditing (V1–V5). The Validated tier is not an unreachable ideal; it is a concrete checklist. The gap between Triangulated and Validated is primarily measurement-validity infrastructure (bootstrap stability, seed variance, baseline separation for the metrics used to evaluate the circuit) and interpretive discipline (explicit level declaration, anthropomorphism checks). These are tractable additions to existing experimental practice, not philosophical impossibilities.

To demonstrate reachability, consider what it would take for the grokking circuit—currently Triangulated—to reach Validated. The mechanism (modular addition via Fourier components) has a mathematically exact ground truth. The remaining gaps are: (1) M1–M6: report bootstrap stability of the weight-space Fourier decomposition, seed variance across training runs, and baseline separation against random-weight models; (2) V1–V5: explicitly declare the description mode as implementational-structural, audit whether "Fourier circuit" implies computational claims the weight analysis alone does not support, and declare the scope (modular arithmetic in toy transformers). Each of these is a straightforward addition to the existing analysis—no new experiments required, only systematic reporting of calibration checks and interpretive discipline. Alternatively, Nanda et al. (2023)'s clock and pizza models have synthetic ground truth by construction: the circuit *is* the mechanism because the model was built to implement it. Running the full M1–M6 suite and V1–V5 audit on such a model would produce the first Validated claim, proving the tier is achievable and calibrating the framework against a known answer.

Interpretive inflation. Labels like "world model," "knowledge neuron," and "deception feature" carry theoretical implications beyond what the evidence supports. The framework systematically identifies where labels exceed evidence through V4 (anthropomorphism check) and V5 (scope declaration).

Construct incoherence. Gender Bias Circuits are the only claim in our evaluation that reaches Disconfirmed—not because evidence is lacking but because the construct itself fails discriminant validity (C4). "Gender bias" and "legitimate gender knowledge" are implemented by the same attention heads; you cannot ablate one without ablating the other. No amount of additional evidence can fix a construct that is not separable from its neighbor. The framework surfaces this as a C4 failure, and the appropriate response is not more experiments but reformulation of the question.

9 Discussion

The case studies reveal that the dominant validity gaps in published MI research are systematic, not idiosyncratic. Sufficiency is rarely established. Specificity is rarely tested. Faithfulness results are method-conditional. Labels routinely exceed what the evidence supports. These are structural features of a field that evaluates methods and artifacts without a common protocol for evaluating the claims built on them.

A principle implicit in the tier definitions deserves explicit statement: evidence accumulates *within* a validity type, not across types. No quantity of behavioral evidence can substitute for missing causal evidence, and no number of single-method causal experiments can substitute for cross-method convergence. The dependency chain (construct → measurement → internal → external → interpretive) imposes a ceiling: a claim's tier is bounded by its weakest validity dimension, not raised by its strongest.

The framework addresses this by providing a structured validity profile for each claim—making gaps visible and the path to strengthening them concrete. The evidence card format (Appendix B.3) operationalizes this as an atomic reporting unit: each card records which measurement bears on which criterion, whether the relevant calibration checks were performed, and what verdict the evidence supports. We propose that evidence cards serve as a standard reporting format for MI claims, analogous to CONSORT checklists for clinical trials. Organizations deploying MI-based monitors (deception probes, steering vectors, circuit-level classifiers) can evaluate the underlying features through the same criteria: Is the construct well-defined? Is the measurement reliable? Is the evidence causal? Does the finding generalize? A feature at Proposed should not be deployed as a monitor; a feature at Mechanistically Supported or above has the minimum evidential basis for cautious deployment with ongoing validation.

Two safety-relevant applications illustrate how the framework extends beyond retrospective auditing of published circuits.

Chain-of-thought faithfulness. The claim "chain-of-thought reasoning is faithful" is not one claim but at least four, each requiring different evidence. Does CoT predict the model's output (*behavioral faithfulness*, testable with perturbation studies)? Does it converge with other interpretability methods on the same internal structure (*convergent faithfulness*, requiring multi-method evidence)? Does it describe the functional roles actually implemented by internal components (*role faithfulness*, requiring independent functional tests)? Or does it map onto the model's computational graph (*structural faithfulness*, requiring causal and structural evidence)? Lanham et al. (2023) and Turpin et al. (2023) demonstrate that behavioral faithfulness often fails: models produce plausible reasoning chains that do not reflect their actual decision process. Under MechVal, "CoT is faithful" at the behavioral level is a Proposed claim (behavioral evidence only, no causal intervention); at the structural level, it would require at minimum Mechanistically Supported status across internal and construct validity criteria. The framework makes the gap between these levels explicit rather than leaving "faithful" as an ambiguous umbrella term.

Deceptive alignment. Detecting whether a model has learned to behave deceptively—strategically appearing aligned during evaluation while pursuing misaligned objectives during deployment—is among the hardest problems in AI safety (Hubinger et al., 2019). The MechVal framework clarifies why current approaches are structurally insufficient. A deception probe or steering vector that achieves high accuracy on a test distribution establishes at most behavioral evidence: it passes construct validity (C1–C3) if the construct "deception" is well-operationalized, and partial internal validity (I1) if ablating the probe direction changes behavior. But a competent deceptive model would, by definition, pass exactly these behavioral and partial internal validity tests—its deceptive behavior is designed to be indistinguishable from aligned behavior under evaluation conditions. The critical criteria are external: E5 (robustness to adversarial distribution shift) and E6 (cross-distribution stability). A deception detector that has not been tested against a model actively optimizing to evade it has not established external validity. Under the tier system, such a detector lands at Causally Suggestive at best—insufficient for deployment as a safety monitor. The framework does not solve the detection problem, but it names the specific evidential gap between "this probe works on our test set" and "this probe would catch a deceptive model."

Safety evidence gaps at scale. More broadly, the dominant pattern in safety-relevant MI findings is instrumental evidence supporting structural conclusions. Refusal direction ablation (Turner et al., 2024), representation engineering for truthfulness, and feature-level safety classifiers all establish that a direction or feature is a sufficient behavioral lever—not that it is the mechanism. The gap is systematic: safety monitoring requires at minimum object-level evidence (necessity plus sufficiency, I1–I2) and ideally subspace-level evidence (stability across distributions, E1–E6). Most published safety-relevant findings, when audited against MechVal criteria, land at Proposed or Causally Suggestive. The framework provides a concrete upgrade path: which criteria are unmet, what evidence would meet them, and what tier the claim would reach with that evidence.

Minimum deployment standards. The safety applications above share a concrete implication: before a safety-relevant interpretability result is used in a deployment decision, what tier must it reach? We propose that the minimum standard is *Mechanistically Supported* with the following criteria confirmed: I1 (necessity), I2 (sufficiency), I3 (specificity), E1 (intervention reach across at least two methods), and M2 (baseline separation against random and untrained controls). A result at Proposed or Causally Suggestive has not demonstrated that the measurement is trustworthy or that the effect is specific—deploying it as a monitor is the MI equivalent of prescribing a drug that passed Phase I but not Phase III. For adversarial settings (deception detection, jailbreak monitoring), the bar should be higher: E5 (robustness to adversarial distribution shift) is necessary because the threat model includes an agent optimizing to evade the monitor. These standards are not arbitrary—they follow from the dependency structure of the validity types: you cannot trust an external validity claim ("this generalizes") without first establishing internal validity ("this is causal"), and you cannot trust a causal claim without first establishing measurement validity ("this metric works").

The emergent abilities case and M4. The Schaeffer et al. (2023) emergent abilities mirage is the cleanest illustration of why M4 (calibration) belongs in the framework's measurement layer. The "emergence" of sudden capability jumps in LLMs was not a property of the models but of the evaluation metric: exact-match accuracy creates a sharp threshold where continuous improvement appears as a phase transition. Switching to a linear metric (token edit distance) dissolved the phenomenon entirely. M4 would have caught this: it requires that reported numbers correspond to meaningful quantities, and that the metric's response function be monotone and approximately linear in the range of interest. Had M4 been applied before the "emergence" literature took off, the question would have been "does this metric create the pattern?" rather than "what explains this striking pattern?"—and the answer would have been immediate. The same diagnostic applies to any MI metric that produces sharp thresholds: if IIA jumps from 0.1 to 0.95 when a hyperparameter is tuned past a threshold, the first question should be whether the metric's response function is creating the jump, not whether the model's computation has a phase transition at that point.

From framework to practice. Validity frameworks gain traction when they are socially enforced, not merely logically correct. We propose three adoption mechanisms, each modeled on a precedent from another field. First, *validity statements*: authors declare which MechVal tier their claims reach and why, analogous to the limitations sections that journals already require. Second, *reviewer checklists*: conferences adopt a short MechVal checklist (What description mode is declared? What tier does the evidence support? Is the claimed tier consistent with the evidence tier?), modeled on the CONSORT checklist in clinical trials. Third, *evidence cards* (Appendix B.3): a standardized reporting format recording which measurement bears on which criterion, whether calibration checks were performed, and what verdict the evidence supports. These are tractable additions to existing practice, not new burdens. Why would individual researchers adopt them voluntarily? Because the framework converts an implicit weakness ("we didn't test X") into an explicit strength ("our claim reaches Mechanistically Supported; here is the concrete evidence that would advance it to Triangulated"). Papers that use this framing give reviewers less to argue about and give follow-up studies a clear target.

Concrete recommendations. The case studies converge on five concrete changes to MI research practice that would have the highest impact on claim quality:

1. *Test specificity by default.* Every circuit ablation study should report performance on at least two unrelated tasks alongside the target task. This is cheap, takes one additional forward pass per task, and addresses the single most common gap (I3) across all 13 case studies.
2. *Report multiple ablation methods.* A faithfulness number from one ablation method is a lower bound on the causal claim, not the claim itself. At minimum, report mean ablation and resample ablation; ideally add zero ablation and activation patching.
3. *Calibrate metrics against baselines.* Report the metric score on a random circuit of the same size, an untrained model, and a shuffled-label baseline. If the delta between the real score and the baseline is small, the metric is not measuring what it claims.
4. *Declare description mode and scope.* State explicitly whether the claim is computational, algorithmic, or implementational, and what input distribution it is validated on. This costs zero experiments and prevents the most common interpretive inflation.
5. *Report non-uniqueness.* If alternative circuits achieve comparable faithfulness, report this as a finding rather than suppressing it. Non-uniqueness is scientific information; a single reported circuit is a selection from a space the reader cannot see.

Several directions for future work follow naturally. Extending the case studies to larger models may reveal that criteria such as E4 (cross-model generalization) are harder to satisfy than the GPT-2 Small evaluations suggest. Automating the verdict assignment process—currently requiring human judgment in aggregating criterion-level results—is future work. And developing standardized prompt suites for specificity testing (I3) and double dissociation (I4) would address the two most common gaps identified across the 13 case studies.

10 Limitations

This work is a theoretical contribution. The case study verdicts are based on published evidence, not new experiments; running the framework's full metric suite with actual model interventions is future work. The current case studies evaluate the most widely-studied published circuits, which predominantly target GPT-2 Small. Extension to claims about larger models may reveal that some criteria are harder to satisfy.

Inter-rater reliability. While the 30 criteria are operationally defined, the aggregation from criterion-level results to verdict tiers involves judgment. Two evaluators could assign different tiers to the same circuit if they weight partial evidence differently. The framework mitigates this by design: the structured validity profile—not the tier—is the primary output. Disagreement is isolated to individual criterion judgments, each of which is independently falsifiable, rather than to a holistic gestalt. We have not yet conducted a formal inter-rater reliability study for the 13 case-study verdicts presented here. A companion paper addresses this gap through both human inter-annotator agreement (planned: three independent raters on a subset of cases) and automated verdict assignment via structured LLM extraction, where the automation serves as a reproducibility check rather than a replacement for expert judgment.

Case studies are illustrative, not pre-registered. The 13 verdicts were assigned after the framework was designed, creating a risk of retrofitting—the same failure mode MechVal is designed to detect. We acknowledge this explicitly: the case studies demonstrate how the framework operates and show that it produces discriminating verdicts, but they are not independent tests of the framework's calibration. Pre-registered application to new circuits as they are published—where the verdict is committed before the evidence is fully reviewed—would provide stronger validation and is planned as future work.

Tier thresholds are provisional. The minimum evidence requirements for tier promotion (Table 7)—such as "$\geq$ 3 evidence families" for Triangulated or "at least 2 ablation variants" for Mechanistically Supported—are not derived from first principles. They are calibrated against the authors' judgment on the 13 case studies and against established precedents in the source disciplines (clinical trials require multiple endpoints; genetics requires genome-wide replication). The synthesis methods in Section 6—Robust Rank Aggregation for cross-method convergence, Dawid-Skene consensus for reliability-weighted voting, and distributional stability via Wasserstein distance—partially alleviate this concern by replacing fixed count thresholds with principled statistical aggregation: a "$\geq$ 3 evidence families" requirement becomes "statisti-

cally significant convergence across independent methods" when operationalized through RRA. Nonetheless, the mapping from continuous synthesis scores to discrete tier boundaries involves judgment, and a formal sensitivity analysis showing how verdicts change under threshold perturbation would strengthen the framework.

Falsifiability of the framework itself. A meta-level concern: what evidence would refute MechVal? If a circuit passes all 30 criteria and is subsequently shown to be wrong, is that a failure of the framework or of the experiment? We commit to the following: if a claim reaches Validated under the framework and is later demonstrably incorrect (the claimed mechanism is not the actual mechanism, as verified by synthetic ground truth or formal proof), this constitutes a framework failure—either a criterion is missing or the tier thresholds are miscalibrated. A framework that can always accommodate failures by saying "you needed more evidence" is not a scientific framework. The Validated tier carries genuine epistemic risk: it says the evidence is sufficient, and that claim is falsifiable.

Scope of "method-agnostic." The framework is applied almost entirely to attention-head-level circuit claims in GPT-2-scale models. Extending it to SAE features, steering vectors, or crosscoders may require adapting how specific criteria are operationalized—the necessity/sufficiency tests look different for a distributed feature than for a localized head. The criteria themselves ("is specificity tested?", "does it generalize?") are method-agnostic; their concrete instantiation is not. We have gestured at this adaptation in the metrics inventory (Appendix A.1) but have not demonstrated a full end-to-end application to a non-circuit MI result.

11 Conclusion

The MechVal framework evaluates mechanistic interpretability claims through 30 criteria organized across five validity types in dependency order. Each claim receives a structured validity profile that makes gaps visible and the path to strengthening them concrete.

The framework is open-source and method-agnostic. It applies the same criteria to circuits, SAE features, probing classifiers, steering vectors, transcoders, and crosscoders—because the questions "Is the construct well-defined?" and "Is the evidence causal?" do not depend on how the evidence was produced.

References

Peter Achinstein. *The Book of Evidence*. Oxford University Press, 2001.

Rasha Al-Lamee, David Thompson, Hakim-Moulay Dehbi, Sayan Sen, Koon Tang, John Davies, Thomas Keeble, Michael Mielewczik, Raffi Kaprielian, Iqbal S. Malik, et al. Percutaneous coronary intervention in stable angina (ORBITA): A double-blind, randomised controlled trial. *The Lancet*, 391(10115):31–40, 2018.

Xiaoyan Bai, Alexander Baumgartner, Haojia Sun, Ari Holtzman, and Chenhao Tan. The story is not the science: Execution-grounded evaluation of mechanistic interpretability research. *arXiv preprint arXiv:2602.18458*, 2026.

Andrew M. Bean, Ryan Othniel Kearns, et al. Measuring what matters: Construct validity in LLM benchmarks. In *NeurIPS Datasets and Benchmarks Track*, 2025. arXiv:2511.04703.

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

Craig M. Bennett, Abigail A. Baird, Michael B. Miller, and George L. Wolford. Neural correlates of interspecies perspective taking in the post-mortem Atlantic salmon: An argument for proper multiple comparisons correction. In *Journal of Serendipitous and Unexpected Results*, volume 1, pp. 1–5, 2010.

James W. Black and Paul Leff. Operational models of pharmacological agonism. *Proceedings of the Royal Society B*, 220(1219):141–162, 1983.

Richard Border, Emma C. Johnson, Luke M. Evans, Andrew Smolen, Noah Berley, Patrick F. Sullivan, and Matthew C. Keller. No support for historical candidate gene or candidate gene-by-interaction hypotheses for major depression across multiple large samples. *American Journal of Psychiatry*, 176(5):376–387, 2019.

Denny Borsboom, Gideon J. Mellenbergh, and Jaap van Heerden. The concept of validity. *Psychological Review*, 111(4):1061–1071, 2004.

Sydney Brenner. The genetics of *Caenorhabditis elegans*. *Genetics*, 77(1):71–94, 1974.

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

Donald T. Campbell and Donald W. Fiske. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2):81–105, 1959.

Donald T. Campbell and Julian C. Stanley. *Experimental and Quasi-Experimental Designs for Research*. Rand McNally, 1963.

Lawrence Chan, Adrià Garriga-Alonso, Nicholas Goldowsky-Dill, Ryan Greenblatt, Jenny Nitishinskaya, Ansh Radhakrishnan, Buck Shlegeris, and Nate Thomas. Causal scrubbing: A method for rigorously testing interpretability hypotheses. *Alignment Forum*, 2022.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. Evaluating the ripple effects of knowledge editing in language models. *Transactions of the Association for Computational Linguistics*, 12: 283–298, 2024.

Thomas D. Cook and Donald T. Campbell. *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Houghton Mifflin, 1979.

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

Lee J. Cronbach and Paul E. Meehl. Construct validity in psychological tests. *Psychological Bulletin*, 52(4): 281–302, 1955.

Lee J. Cronbach, Goldine C. Gleser, Harinder Nanda, and Nageswari Rajaratnam. *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. Wiley, 1972.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *ICLR*, 2024. arXiv:2309.08600.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In *ACL*, 2022. arXiv:2104.08696.

A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28, 1979.

Alex J. DeGrave, Joseph D. Janizek, and Su-In Lee. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, 2021.

Pierre Duhem. *La Théorie Physique: Son Objet, Sa Structure*. Marcel Rivière, 1906.

Anders Eklund, Thomas E. Nichols, and Hans Knutsson. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences*, 113(28):7900–7905, 2016.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *Transformer Circuits Thread*, 2022. arXiv:2209.10652.

Joshua Engels, Eric J. Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear. In *ICLR*, 2025. arXiv:2405.14860.

Ronald A. Fisher. The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, 52(2):399–433, 1918.

Timo Freiesleben and Sebastian Zezulka. The benchmarking epistemology: Construct validity for evaluating machine learning models. *arXiv preprint arXiv:2510.23191*, 2025.

Juan A. Gallego, Matthew G. Perich, Lee E. Miller, and Sara A. Solla. Neural manifolds for the control of movement. *Neuron*, 94(5):978–984, 2017.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards textures; increasing shape bias improves accuracy and robustness. In *ICLR*, 2019. arXiv:1811.12231.

Matthew F. Glasser, Timothy S. Coalson, Emma C. Robinson, et al. A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178, 2016.

Stuart Glennan. *The New Mechanical Philosophy*. Oxford University Press, 2017.

Clark Glymour. *Theory and Evidence*. Princeton University Press, 1980.

Rhys Gould, Euan Ong, George Ogden, and Arthur Conmy. Successor heads: Recurring, interpretable attention heads in the wild. *arXiv preprint arXiv:2312.09230*, 2023.

David M. Green and John A. Swets. *Signal Detection Theory and Psychophysics*. Wiley, 1966.

Michael Hanna, Ollie Liu, and Alexandre Variengien. How does GPT-2 compute greater-than? interpreting mathematical abilities in a pre-trained language model. In *NeurIPS*, 2023.

Stefan Heimersheim and Jett Janiak. A circuit for python docstrings in a 4-layer attention-only transformer. *Alignment Forum*, 2023.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *EMNLP*, 2019.

Archibald V. Hill. The possible effects of the aggregation of the molecules of haemoglobin on its dissociation curves. *Journal of Physiology*, 40:iv–vii, 1910.

Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. Risks from learned optimization in advanced machine learning systems. *arXiv preprint arXiv:1906.01820*, 2019.

Adam Karvonen, Can Rager, Johnny Lin, Curt Tigges, Joseph Bloom, David Chanin, Yeu-Tong Lau, Eoin Farrell, Callum McDougall, Kola Ayonrinde, Demian Till, Matthew Wearden, Arthur Conmy, Samuel Marks, and Neel Nanda. SAEBench: A comprehensive benchmark for sparse autoencoders in language model interpretability. In *ICML*, 2025. arXiv:2503.09532.

Terry Kenakin. Principles: Receptor theory in pharmacology. *Trends in Pharmacological Sciences*, 25(4): 186–192, 2004.

Ilyes Khemakhem, Diederik P. Kingma, Ricardo Pio Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *AISTATS*, 2020. arXiv:1907.04809.

Raivo Kolde, Sven Laur, Priit Adler, and Jaak Vilo. Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics*, 28(4):573–580, 2012.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukičić, Karina Nguyen, Nicholas Schiefer, Zac Shi, Nicholas Sully, Tom Brown, Jared Kaplan, Shlegeris Buck, Catherine Olsson, Tom Henighan, and Amanda Askell. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *ICLR*, 2023. arXiv:2210.13382.

David Lindner, János Kramár, Sebastian Farquhar, Matthew Rahtz, Thomas McGrath, and Vladimir Mikulik. Tracr: Compiled transformers as a laboratory for interpretability. In *NeurIPS*, 2023. arXiv:2301.05062.

Jane Loevinger. Objective tests as instruments of psychological theory. *Psychological Reports*, 3:635–694, 1957.

Frederic M. Lord and Melvin R. Novick. *Statistical Theories of Mental Test Scores*. Addison-Wesley, 1968.

David Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. MIT Press, 1982.

Deborah G. Mayo. *Error and the Growth of Experimental Knowledge*. University of Chicago Press, 1996.

Deborah G. Mayo. *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*. Cambridge University Press, 2018.

Callum McDougall, Arthur Conmy, Cody Rushing, Thomas McGrath, and Neel Nanda. Copy suppression: Comprehensively understanding an attention head. *arXiv preprint arXiv:2310.04625*, 2023.

Maxime Méloux, Silviu Maniu, François Portet, and Maxime Peyrard. Everything, everywhere, all at once: Is mechanistic interpretability identifiable? In *ICLR*, 2025. arXiv:2502.20914.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. Circuit component reuse across tasks in transformer language models. In *ICLR*, 2024. arXiv:2310.08744.

Samuel Messick. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9):741–749, 1995.

Joseph Miller, Bilal Chughtai, and William Saunders. Transformer circuit faithfulness metrics are not robust. In *COLM*, 2024. arXiv:2407.08734.

Aaron Mueller et al. MIB: A mechanistic interpretability benchmark. *arXiv preprint arXiv:2504.13151*, 2025.

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. arXiv:2209.11895.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.

Judea Pearl and Elias Bareinboim. Transportability of causal and statistical relations: A formal approach. In *AAAI*, 2011.

Karl R. Popper. *The Logic of Scientific Discovery*. Hutchinson, 1959.

Willard Van Orman Quine. Two dogmas of empiricism. *The Philosophical Review*, 60(1):20–43, 1951.

Humphrey P. Rang. The receptor concept: Pharmacology's big idea. *British Journal of Pharmacology*, 147 (S1):S9–S16, 2006.

Michael Roberts, Derek Driggs, Matthew Thorpe, Julian Gilbey, Michael Yeung, Stephan Ursprung, Angelica I. Aviles-Rivero, Christian Etmann, Cathal McCague, Lucian Beer, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature Machine Intelligence*, 3(3):199–217, 2021.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? *NeurIPS*, 2023. arXiv:2304.15004.

Schmidt Sciences. Trustworthy AI research agenda: Request for proposals, 2026.

William R. Shadish, Thomas D. Cook, and Donald T. Campbell. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin, 2002.

Tim Shallice. *From Neuropsychology to Mental Structure*. Cambridge University Press, 1988.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, 2nd edition, 2000.

R. P. Stephenson. A modification of receptor theory. *British Journal of Pharmacology*, 11(4):379–393, 1956.

Denis Sutter, Julian Minder, Thomas Hofmann, and Tiago Pimentel. The non-linear representation dilemma: Is causal abstraction enough for mechanistic interpretability? In *NeurIPS*, 2025. arXiv:2507.08802.

Alex Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2024.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *NeurIPS*, 2023. arXiv:2305.04388.

Tyler J. VanderWeele and Peng Ding. Sensitivity analysis in observational research: Introducing the E-value. *Annals of Internal Medicine*, 167(4):268–274, 2017.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *NeurIPS*, 2020.

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 *ICLR*, 2023.

James Woodward. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, 2003.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, et al. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *NeurIPS*, 2023.

A Appendix

This appendix provides the full metric inventory grouped by evidence family (A.1), protocol specifications (A.2), and synthesis protocol details (A.3).

A.1 Metrics Inventory

The full inventory of metrics is listed below. For historical continuity with the original metric catalog, the appendix retains the fine-grained grouping (causal, structural, information-theoretic, behavioral, representational, measurement) used during development; each group maps onto one or more of the four source-based evidence families defined in Section 5.2.

A. Causal

Table 16: Causal metrics: the effect of interventions on the model's computation.

Metric	Description
A01 SCM / Pearl	Structural causal models and do-calculus as the formal language for circuit claims
A02 Counterfactual DAS	Distributed Alignment Search and IIA for testing causal abstraction hypotheses
A03 Rubin CATE	Conditional average treatment effects across input subpopulations
A04 Woodward	Interventionist theory of causation applied to circuit ablation
A05 MDC / Glennan	New mechanical philosophy — components organized to produce phenomena
A06 Mediation	Natural direct and indirect effects decomposition for circuit paths
A07 Granger / TE	Observational causal discovery via conditional MI and temporal precedence
A08 PID	Decomposing MI into redundant, unique, and synergistic contributions
A09 MDL / SLT	Minimum description length and singular learning theory for circuit complexity
A10 Regularity / INUS	INUS conditions — insufficient but necessary parts of sufficient sets
A11 Actual Cause	Halpern-Pearl actual causation on specific inputs, not just potential causes
A12 Transportability	Formal conditions for circuit generalization across models and distributions
A13 Causal Discovery	NOTEARS / PC algorithms for learning DAGs over circuit components

B. Structural

Table 17: Structural metrics: weight-space properties requiring no forward pass.

Metric	Description
B01 SVD / Spectral	Singular value decomposition to identify dominant computational directions
B02 Effective Rank	Entropy-based dimensionality as a scalar summary of spectral concentration
B03 OV / QK Decomp.	Decomposing attention heads into OV (what to write) and QK (where to attend)
B04 Weight Alignment	Cosine similarity between principal weight directions across heads
B05 Norm Trajectory	Spectral norm ratios tracking signal amplification through components
B06 Template Distance	Graph-edit and metric distances between circuits discovered for different tasks
B07 Polysemanticity	Measuring whether components encode multiple unrelated features in superposition
B08 ICA / NMF	Independent component analysis for decomposing weights into interpretable parts
B09 Weight Classifier	Training classifiers on weight matrices to predict circuit membership

C. Information-theoretic

Table 18: Information-theoretic metrics: information flow and dependencies.

Metric	Description
C01 Mutual Info.	Total shared information between circuit components and task performance
C02 Conditional MI	Information shared after conditioning on other parts of the circuit
C03 Transfer Entropy	Directed information flow between components across layers
C04 PID	Decomposing shared information into unique, redundant, and synergistic atoms
C05 Info. Bottleneck	How efficiently circuits compress input while preserving task-relevant signal
C06 O-Information	Whether a group of components interacts redundantly or synergistically
C07 Granger Causality	Whether one component's past activations improve prediction of another's future
C08 OCSE	Estimating causal influence between components using only observational data
C09 NOTEARS	Learning directed acyclic graph structure from observational activation data

D. Behavioral

Table 19: Behavioral metrics: input-output behavior under manipulation.

Metric	Description
D01 Faithfulness	Whether an identified circuit faithfully reproduces the full model's behavior
D02 Logit Diff	How much of the model's logit difference the circuit recovers
D03 KL Divergence	Information-theoretic distance between circuit and full model distributions
D04 CE Delta	Change in cross-entropy loss when the circuit is ablated
D05 Top-K Accuracy	Whether the circuit preserves the model's top-K predicted tokens
D06 Cross-Task	Whether a circuit discovered on one task transfers to a different task
D07 Cross-Scale	Whether circuit structure replicates in larger or smaller models
D08 Prompt Paraphrase	Circuit consistency across semantically equivalent prompt templates
D09 Generalization Gap	Sensitivity of circuit discovery to hyperparameters and methodological choices

E. Representational

Table 20: Representational metrics: geometry and content of internal representations.

Metric	Description
E01 DAS-IIA	Whether a learned linear subspace causally encodes a target variable
E02 Linear Probe	Whether a target variable is linearly decodable from intermediate representations
E03 RSA	Comparing representation geometry by correlating pairwise distance matrices
E04 CKA	Kernel alignment for cross-layer and cross-model representation comparison
E05 Subspace Align.	Cosine alignment between SVD-derived principal directions of weight matrices
E06 PCA Dim.	Effective dimensionality of circuit subspaces via activation covariance spectrum
E07 Intrinsic Dim.	True manifold dimensionality, connecting to geometric complexity
E08 Participation Ratio	How many dimensions are effectively active in a representation
E09 Persistent Homology	Topological data analysis detecting loops and voids in activation manifolds
E10 Cross-Task Overlap	Representational structure shared between tasks via IIA transfer

F. Measurement

Table 21: Measurement metrics: calibration checks for the trustworthiness of other metrics.

Metric	Description
F01 Test–Retest	Whether rerunning the same metric on the same model, task, and intervention gives the same answer
F02 Bootstrap Stability	Whether the metric remains stable under bootstrap resampling of prompts, examples, or activation samples
F03 Seed Variance	How much the metric changes across random seeds for probes, SAEs, optimization, or sampling-based estimators
F04 Checkpoint Variance	Whether the metric is stable across nearby checkpoints or depends on one training snapshot
F05 Prompt Variance	How much the metric changes across prompt templates, paraphrases, lexical choices, or dataset slices
F06 Baseline Separation	Whether the score is distinguishable from random, untrained, shuffled-label, or permuted-circuit baselines
F07 Negative Controls	Whether the metric stays low where the claimed effect should be absent
F08 Positive Controls	Whether the metric detects known-true or synthetic effects of the relevant size
F09 Intervention Robustness	Whether the conclusion survives reasonable intervention variants such as mean ablation, zero ablation, resample ablation, patching, or causal scrubbing
F10 Hyperparam. Sensitivity	Whether the metric changes under reasonable choices of sparsity, probe regularization, thresholds, localization cutoffs, or discovery hyperparameters
F11 Estimator Uncertainty	Confidence intervals, standard errors, permutation tests, or posterior intervals for the reported metric value
F12 Calibration Curve	Whether larger reported scores correspond to larger empirical effects or higher probability of success
F13 Cross-Metric Convergence	Whether independent metrics intended to test the same criterion agree despite different failure modes
F14 Measurement Invariance	Whether the metric behaves consistently across model sizes, tasks, prompt distributions, and intervention contexts
F15 Multiple Comparisons	Whether scores are corrected for the number of components tested, methods tried, or thresholds swept. Reporting the top-k heads out of N without FDR or permutation correction inflates apparent circuit specificity

A.2 Protocols

A.2.1 V4: Interpretive Inflation Control

V4a: Label scope audit. For a circuit C with label L, enumerate the K implicit claims in L. For each claim, determine whether evidence at Causally Suggestive or above exists. The label evidence ratio $\text{LER} = |\{k : \text{status}(k) \geq \text{CS}\}|/K$ quantifies how well-supported the label is. $\text{LER} < 0.5$ flags interpretive inflation.

V4b: Contrastive label test. Construct a contrastive task T' that preserves surface behavior but breaks the mechanism implied by L. The contrastive selectivity $\Delta = \text{IIA}(C, T) - \text{IIA}(C, T')$ should exceed 0.2 for the label to stand; $\Delta < 0.05$ indicates the label is likely inflated.

V4c: Ablation scope test. Ablate C and measure performance on both the target task T and N unrelated tasks. The specificity index $\text{SI} = \Delta_T / \bar{\Delta}_{\text{other}}$ should exceed 2.0.

A.2.2 I5: Confound Control

I5a: Minimal confound enumeration. For task T, enumerate K plausible confounds $\{C_1, \ldots, C_K\}$. For each, construct a matched task T_k that holds everything constant except C_k. The confound invariance $\delta_k = |\text{IIA}(C, T) - \text{IIA}(C, T_k)|/\text{IIA}(C, T)$ quantifies sensitivity. $\max_k \delta_k > 0.3$ flags an I5 failure.

I5b: Causal graph consistency. Given a proposed mechanism $A \to B \to \text{output}$, test conditional independence $A \perp \text{output} \mid B$. If this fails (Fisher's Z-test, $p < 0.05$), B is not the complete mediator.

I5c: Synthetic ground truth injection. Use a model trained on a task where the ground truth circuit is known by construction. Run the discovery method and measure Jaccard(discovered, ground truth).

A.2.3 I6: Epistatic Interaction

I6a: Pairwise epistasis scores. For all component pairs (A, B) in circuit C, compute the epistasis score $\epsilon_{AB} = f(\text{ablate } A \cup B) - f(\text{ablate } A) - f(\text{ablate } B) + f(\emptyset)$. Positive ϵ indicates synergy; negative ϵ indicates buffering. $|\epsilon_{AB}| < 0.05$ for all pairs indicates additive structure. $|\epsilon_{AB}| > 0.1$ identifies functional joints.

I6b: Higher-order interaction test. Compute the full interaction contrast $\Delta^K = \sum_{S \subseteq C} (-1)^{|C|-|S|} f(\text{ablate } S)$. For practical circuits ($K > 6$), approximate by sampling random subsets and testing deviation from the additive prediction (permutation test, $p < 0.01$).

A.2.4 E1: Intervention Reach

E1a: Activation steering reach. Identify the subspace in C's activations associated with behavior X. Amplify that direction by factor α. The steering reach $r = \Delta_{\text{behavior}}/\alpha$ should be approximately linear; the steering specificity $s = \Delta_X / \Delta_{\text{other}}$ should exceed 2.0.

E1b: Cross-context transfer. Test circuit interventions on distributions beyond the discovery distribution: D_{transfer} (paraphrased prompts), D_{novel} (same mechanism, different domain), D_{OOD} (entirely different context). The E1 score is the harmonic mean of transfer IIA across distributions.

E1c: Counterfactual circuit transplant. Take the weight-space subspace identified as the circuit in model A and transplant it into model B that lacks the capability. The capability transfer ratio $\tau = (\text{acc}_{B,\text{after}} - \text{acc}_{B,\text{before}})/\text{acc}_A$ should exceed 0.5 for genuine reach.

A.2.5 C3: Convergent Validity

C3a: Un-training convergence. Fine-tune the model to remove capability T (e.g., train on counter-task examples until task performance drops to chance). Measure whether the weight-space signatures of circuit C—composition scores, spectral structure, selector magnitudes—degrade proportionally to the behavioral loss. The structural convergence ratio $\text{SCR} = \Delta(\text{weight signature})/\Delta(\text{behavioral performance})$ should be approximately linear. $\text{SCR} \approx 0$ (behavior degrades but circuit weights unchanged) indicates the structural claim is spurious; the weights were correlated with but not constitutive of the computation. $\text{SCR} \gg 1$ (weights change far more than behavior) suggests backup circuits or distributed compensation.

C3b: Weight-space prediction. Before running any activation-level experiments, predict circuit membership from weight structure alone (composition scores, spectral properties, norm trajectories). Then run standard activation-level discovery (activation patching, DAS, EAP) independently. The prediction-discovery agreement $\text{PDA} = \text{Jaccard}(\text{predicted}, \text{discovered})$ quantifies whether structural and causal evidence converge. $\text{PDA} > 0.5$ indicates genuine structural-functional correspondence; $\text{PDA} < 0.2$ indicates the weight structure is not predictive of the causal structure.

A.2.6 M1: Reliability

M1a: Split-half discovery reliability. Partition the prompt distribution into two random halves D_1, D_2. Run circuit discovery independently on each half. The split-half reliability $r = \text{Jaccard}(C_1, C_2)$ measures whether

the discovered circuit depends on which prompts were sampled. $r > 0.7$ indicates reliable discovery; $r < 0.4$ indicates the circuit is prompt-dependent. Repeat across $K \geq 10$ random splits and report the mean and variance of r.

A.2.7 M2: Baseline Separation

M2a: Null circuit baseline. For a circuit C of size $|C| = k$, sample $N \geq 100$ random circuits of the same size and compute faithfulness for each. The baseline-corrected faithfulness $\text{BCF} = (f(C) - \mu_{\text{null}})/\sigma_{\text{null}}$ should exceed 3.0 (i.e., the circuit is $> 3\sigma$ above random). Report the full null distribution, not just the point estimate. If $\text{BCF} < 2.0$, the circuit's apparent faithfulness may reflect circuit size rather than circuit identity.

A.2.8 E5: Graded Response

E5a: Ablation dose-response. Ablate circuit C at $M \geq 5$ fractional strengths $\alpha \in \{0.1, 0.2, \ldots, 1.0\}$ (e.g., interpolating between clean and ablated activations). Plot behavioral performance against α. The curve should be monotonically decreasing. Non-monotonic dose-response (performance improves at intermediate ablation strengths) indicates compensatory mechanisms, backup circuits, or measurement artifacts. Fit a sigmoid and report the EC_{50} (the ablation strength producing half-maximal degradation) as a summary statistic.

A.2.9 I2: Sufficiency

I2a: Subspace interchange. Given two inputs x_1, x_2 that differ on the task-relevant variable (e.g., different indirect objects), project both through circuit C's identified subspace S. Swap the subspace component: replace x_1's projection onto S with x_2's, leaving the orthogonal complement unchanged. If the model's output switches to match x_2's expected output, the subspace carries the task-relevant information the claim asserts. This is an observational analog of DAS that does not require training an intervention—it tests whether the geometric structure identified by the circuit claim is sufficient to carry the computational content. The interchange success rate $\text{ISR} = \mathbb{E}[\mathbf{1}(\text{output matches swapped input})]$ should exceed 0.7 for the subspace to be considered sufficient.

A.3 Synthesis Protocols

S01: Functional Parcellation. Synthesizes component rankings across evidence families into functional subgroups. Adapted from Glasser et al. (2016)'s multimodal brain parcellation. Takes ranked component lists from 4+ methods, computes representational similarity between the rankings, and clusters components that are consistently co-ranked. The output is a set of functional groups—components that multiple independent methods agree belong together. Strengthens C3 (convergent validity) and I3 (specificity).

S02: Dawid-Skene Consensus. Synthesizes binary circuit membership judgments across methods with different reliability profiles. Treats each method as a noisy annotator and jointly estimates (via EM) both the true circuit membership and each method's reliability (Dawid & Skene, 1979). Methods that consistently disagree with the majority get downweighted. Strengthens C3 (convergent validity) and M1 (reliability).

S03: Robust Rank Aggregation. Synthesizes ranked component lists into a single ranking with statistical significance. Computes RRA p-values and Borda counts across method rankings (Kolde et al., 2012). Identifies components that rank consistently high across methods (robust members) and components that rank high in one method but low in others (method-dependent members). Strengthens C3 (convergent validity) and I5 (confound control).

S04: Parallel Ensemble. Synthesizes rank-normalized scores across methods into a single composite ranking. Implements three fusion rules—equal weighting, method-reliability weighting, and minimum-across-methods—with uncertainty bounds. Strengthens C3 (convergent validity) and M2 (baseline separation).

S05: Sequential Ensemble. Synthesizes cheap and expensive evidence in a two-stage pipeline: cheap methods (weight-space, information-theoretic) run first and filter to the top 20% of components, then expensive methods (causal, behavioral) run only on the filtered set. Reduces computational cost by 5–10× while preserving top-component ranking.

S06: Wasserstein Stability. Synthesizes evidence about measurement robustness across conditions. Computes the Wasserstein-1 distance between component score distributions across runs (different random seeds, prompt sam-

ples, or tasks). A circuit with low W_1 across prompt samples but high W_1 across tasks is robust to measurement noise but sensitive to task context. Strengthens M1 (reliability) and E5 (robustness).

S07: Meta-Learner. Synthesizes method features into a predictive model of circuit membership. Trains a logistic regression on known-labeled circuits (components with established ground truth) using method features as predictors. The learned coefficients reveal which method features are most predictive of true membership. Strengthens M5 (sensitivity) and C3 (convergent validity).

S08: Granger Causality Graph. Synthesizes cross-family predictive relationships between method scores. Constructs a directed graph where edges represent Granger-causal relationships—does component A's score in method X predict component B's score in method Y, conditional on B's own score in X? Cross-family predictive structure indicates genuine convergent support rather than shared noise. Strengthens C3 (convergent validity) and I4 (double dissociation).

S09: Circuit Component Reuse. Synthesizes circuit structure across tasks into reusable computational primitives. Adapted from Merullo et al. (2024). Identifies reusable circuit components shared across tasks—if the same attention heads appear in the IOI circuit, the greater-than circuit, and the induction circuit, they are building blocks rather than task-specific artifacts. Strengthens E6 (cross-architecture generalization) and C2 (structural plausibility).

B Case Study Audits

Full 30-criteria audits, evidence cards, and circuit specifications for the case studies evaluated in Section 8.

B.1 IOI Circuit

B.1.1 Full 30-Criteria Audit

We audit the IOI circuit claim (Wang et al., 2023) against all 30 criteria to demonstrate what the framework produces in practice. The claim: "The IOI circuit uses 26 attention heads in 6 functional roles to identify and output the indirect object in sentences of the form 'When Mary and John went to the store, John gave a drink to _____'." Description mode: implementational-functional.

Table 22: Full 30-criteria audit of the IOI circuit. Status: **C** = Confirmed, **PC** = Partially confirmed, **U** = Untested, **I** = Inconclusive, **D** = Disconfirmed.

ID	Criterion	Status	Evidence
Construct Validity			
C1	Falsifiability	C	Specific heads named; ablation predictions testable
C2	Structural plausibility	C	OV/QK circuits analyzed; weight structure matches roles
C3	Convergent validity	PC	Primarily ablation-based; limited cross-family evidence
C4	Discriminant validity	U	IOI circuit vs. neighboring tasks never tested
C5	Nomological validity	PC	Fits induction head theory loosely; no formal connection
Measurement Validity			
M1	Reliability	U	No bootstrap stability or seed variance reported
M2	Baseline separation	U	Faithfulness not compared to random circuit baselines
M3	Stability	U	Robustness to perturbation of circuit boundary not tested
M4	Calibration	PC	Logit difference is meaningful; faithfulness % less clear
M5	Sensitivity	U	No known-positive control tested
M6	Invariance	U	No cross-condition consistency check
Internal Validity			
I1	Necessity	C	Ablation degrades IOI performance substantially
I2	Sufficiency	C	Circuit alone recovers 87% logit diff (mean ablation)
I3	Specificity	U	Effect on non-IOI tasks never measured
I4	Double dissociation	U	No converse test (ablate non-IOI circuit, check IOI)
I5	Confound control	U	Positional, frequency, syntactic, length confounds untested
I6	Epistatic interaction	U	No pairwise ablation or synergy analysis
I7	Rescue reversibility	U	No corruption-then-restore experiment
I10	Confounding sensitivity	U	No E-value or sensitivity analysis
External Validity			
E1	Intervention reach	I	87% mean ablation; <50% resample ablation; steering reach untested
E2	Prompt generalization	PC	Template variations tested; distributional diversity limited
E3	Cross-task generalization	U	Transfer to related tasks (e.g., DOI) not tested
E4	Cross-model generalization	U	Only GPT-2 Small; no replication in other models
E5	Graded response	U	No dose-response (partial ablation) reported
E6	Novel prediction	PC	Negative name movers were a novel structural prediction
Interpretive Validity			
V1	Level declaration	PC	Impl.-functional implied but not explicitly declared
V2	Level-evidence match	PC	Evidence supports topographic + partial functional
V3	Alternative level	U	Whether evidence fits a simpler description not tested
V4	Anthropomorphism check	PC	"Name mover" descriptive; "S-inhibition" implies intent; labels not contrastively tested
V5	Scope declaration	PC	IOI task defined; generalization scope unstated

Total: 4 Confirmed, 9 Partially confirmed, 1 Inconclusive, 16 Untested, 0 Disconfirmed
Verdict: Mechanistically Supported

Summary. Of 30 criteria: 4 Confirmed, 9 Partially confirmed, 1 Inconclusive, 16 Untested, 0 Disconfirmed. The IOI circuit reaches **Mechanistically Supported** (necessity and sufficiency established via one method family). The bottleneck to Triangulated is the absence of convergent evidence from ≥ 3 evidence families (C3), untested specificity (I3), and no cross-model replication (E4). The bottleneck to Validated is the mostly unaddressed measurement validity block (M1–M6; only M4 is partially confirmed).

The protocols in §6.3 provide concrete next steps for the highest-value untested criteria: (1) **I5 confound control**—run the I5a confound battery (positional, frequency, syntactic, length confounds) to determine whether the circuit is doing "indirect object identification" or "copy the second proper noun"; (2) **V4 label validation**—run V4b

contrastive label tests on "name mover" (equalize IO and S salience) and "S-inhibition" (test whether suppression is position- or identity-based); (3) **E1 intervention reach**—run E1a steering reach to test whether amplifying the circuit's activation direction *increases* IOI performance proportionally, not just whether ablation decreases it; (4) **C3 convergent validity**—apply Dawid-Skene consensus (§6.4) across activation patching, weight analysis, and information-theoretic methods to determine which heads survive reliability-weighted voting. Additionally: (5) cross-method convergence—replicate the circuit via weight-space analysis; (6) cross-model replication—test whether the same circuit appears in GPT-2 Medium or Pythia.

B.2 Example Circuit Specification: IOI

We begin with a worked example to ground the abstract machinery: the IOI circuit, the most widely studied circuit in the field. Encoding it as a formal specification shows how a mechanistic claim becomes a set of concrete, falsifiable predictions that the verification pipeline can test automatically.

The IOI circuit specification defines a 6-step computational DAG:

1. **Duplicate Token Detection** (DTH: heads 0.1, 3.0) — detects that a name appears twice.
2. **Previous Token Tracking** (PTH: heads 2.2, 4.11) — attends to the token before a name.
3. **Induction** (IND: heads 5.5, 6.9) — uses DTH and PTH signals to identify the repeated name.
4. **S-Inhibition** (S-Inh: heads 7.3, 7.9, 8.6, 8.10) — suppresses the repeated subject name.
5. **Name Mover** (NM: heads 9.9, 9.6, 10.0) — copies the non-repeated (indirect object) name to the output.
6. **Negative Name Mover** (NegNM: heads 10.7, 11.10) — provides opposing signal.

Positive predictions (7).

- Ablating DTH reduces output (≥ 0.2 decrease in logit difference).
- Ablating induction heads substantially reduces logit difference (≥ 0.5).
- Ablating S-Inh kills output (≥ 0.8 decrease).
- Ablating S-Inh reduces name mover activation (≥ 0.2).
- Ablating NegNM increases logit difference (removes opposing signal).
- Ablating DTH reduces induction head activation (tests DTH→IND edge).
- Ablating PTH reduces induction head activation (tests PTH→IND edge).

Negative controls (5).

- Ablating NM does not affect upstream S-Inh.
- Ablating NegNM does not affect upstream S-Inh.
- Ablating PTH alone does not destroy circuit output (< 0.3 decrease).
- Ablating S-Inh does not affect upstream DTH.
- Ablating NM does not affect NegNM (parallel roles, not connected).

This specification encodes the directed causal structure of the proposed mechanism and provides concrete, falsifiable predictions that can be tested automatically by the verification pipeline.

B.3 Evidence Cards

An evidence card is the minimal reporting unit for a claim-level validity judgment. It records which measurement bears on which criterion, whether the relevant calibration checks were performed, and what verdict implication follows. This prevents verdicts from becoming aggregate impressions: each assignment must be traceable to specific present or missing evidence.

As an example, we show the IOI circuit, using supplementary information from follow-up studies. (Miller et al., 2024) find the faithfulness numbers to be method-conditional, and (Méloux et al., 2025) find alternative head sets achieving comparable faithfulness.

Table 23: Evidence card template: the atomic reporting unit linking measurements to criteria and verdicts.

Field	**Contents**
Claim	The exact mechanistic claim being evaluated
Description Mode	The strongest description mode implicated by the claim
Evidence Family	The family of the primary evidence: weights, activations, behavior, or training
Primary Metric	The concrete runnable test producing the measurement
Calibration Checks	Calibration checks on the primary metric: reliability, baseline separation, stability, sensitivity, uncertainty, or invariance
Criterion Tested	The criterion the evidence bears on, such as I1 necessity, I2 sufficiency, E1 intervention reach, or M2 baseline separation
Result	The reported numerical result, qualitative finding, or structured observation
Status	Pass, partial pass, fail, or untested
Failure Mode Addressed	The validity failure this card helps detect, such as no baseline control, method-conditional evidence, off-target effects, or scope creep
Verdict Implication	What verdict tier this evidence can support, and what higher tier it fails to support without additional evidence

Table 24: Evidence card for the IOI circuit faithfulness result.

Field	**IOI example**
Claim	The IOI circuit uses name-mover heads to copy the indirect object to the output position
Description Mode	Implementational-functional: the claim identifies components and assigns them functional roles
Evidence Family	Activations and behavior
Primary Metric	Faithfulness / logit-difference recovery under circuit ablation
Calibration Checks	Intervention robustness: the headline faithfulness value is method-conditional
Criterion Tested	I1 necessity, I2 sufficiency, and E1 intervention reach
Result	The circuit recovers 87% of the full-model logit difference under mean ablation, but substantially less under other ablation methods
Status	I1 pass; I2 partial/pass depending on intervention definition; E1 partial because intervention variants disagree
Failure Mode Addressed	Cherry-picking metrics and method-conditional evidence
Verdict Implication	Supports Mechanistically Supported, but blocks Triangulated until specificity, double dissociation, cross-model generalization, and stronger measurement calibration are addressed