

Circuits Are Estimates, Not Mechanisms: A Recovered Subgraph Is One Draw From a Class Set by the Ablation, the Dataset and the Grain, the Results That Defend Circuits Hold at the Level of That Class, and What Ablation Evidence Licenses at Each Level

Pranay Mahendrakar

SONYTECH · ORCID 0009-0003-7224-029X

pranaymahendrakar@sonytech.in · mahendrakarpranay@gmail.com · +91 6361723454

Abstract. A circuit in a language model is usually presented as the mechanism by which the model performs a task: a sparse subgraph, found by ablation or patching, that reproduces the behaviour when everything else is knocked out. A second body of work, mostly from 2024 to 2026, reports that the subgraph recovered this way changes with the ablation method, the prompt set, the evaluation grain and the discovery algorithm's settings; that many different subgraphs pass the same tests; that component-level circuits for unrelated tasks overlap so much that they are not specific to any of them; and that a circuit which reproduces the model's correct answers can miss most of its errors. In this analysis the author reviews and synthesises published results to show that the two bodies of work are not in contradiction, because they measure different objects. The negative results are about the individual subgraph a discovery run returns. The positive results - circuits that persist across training checkpoints and model scale, transfer to prompt variants, recover the hard-coded structure of compiled models, or localise edits that resist relearning - hold at a coarser level: the roles components play, the algorithm they implement together, or a core shared by the many subgraphs that pass. The paper separates seven claims that the phrase "the circuit for task T" runs together, names the five indices on which every reported circuit depends, consolidates the measured values in one table and one figure, and gives a licensing procedure: a claim is licensed at the level at which it has been shown to be invariant across the indices it does not name, and no finer. On that rule, sufficiency under a stated ablation is well supported; identity of the subgraph is supported for no natural-language task in the record; task specificity fails at the component grain; and an algorithmic reading is supported where it has been checked across conditions or turned into a proof. Five studies would settle the open part.

Keywords: mechanistic interpretability; circuit discovery; activation patching; ablation; identifiability; causal abstraction; faithfulness; evaluation validity

SONYTECH ORCID: 0009-0003-7224-029X

1. Introduction

The best-known result in circuit-style interpretability is a diagram. Wang et al. (2022) explained how GPT-2 small completes sentences such as "When Mary and John went to the store, John gave an apple to" with the indirect object, and their explanation consisted of 26 attention heads sorted into seven classes, connected in a specific order, together with three quantitative criteria for judging the explanation. The diagram became a template. Automated methods were built to find such subgraphs without the months of manual work (Conmy et al., 2023; Syed et al., 2023; Bhaskar et al., 2024; Hanna, Pezzelle and Belinkov, 2024), and a benchmark was built to rank those methods (Mueller et al., 2025). The working assumption, stated or not, is that a subgraph which reproduces the model's behaviour when the rest of the network is ablated is the mechanism the model uses.

That assumption has been under pressure from several directions at once, and the pressure has grown sharply since 2025. Miller, Chughtai and Saunders (2024) found that a circuit's faithfulness score depends heavily on

choices that look like details: whether nodes or edges are ablated, whether the replacement value is a mean or a resampled activation, how token positions are treated, and the order in which a percentage and an average are computed. Meloux et al. (2025a), working on networks small enough to enumerate candidate explanations, found many different circuits that reproduce a network's behaviour perfectly and many different logical readings of each. Bayat Makou et al. (2026) report that circuits extracted from different input-frequency bands of one task look structurally different yet do the same thing, and that repeated extractions behave like draws from a set of valid subgraphs rather than estimates of one. Li and Subramani (2026) report that at the grain of heads and MLP blocks, circuits for unrelated tasks share most of their components. Zhang, Geng et al. (2026) report that circuits which agree with GPT-2 small on nearly all prompts it answers correctly agree on a minority of the prompts it gets wrong.

The defenders of the programme have results too. Tigges et al. (2024) tracked circuits across 300 billion tokens of training in models from 70 million to 2.8 billion parameters and found that, while the specific heads filling each role can change, the algorithm does not. Nainani et al. (2024) found that the circuits GPT-2 small

uses for prompt variants built to challenge the IOI algorithm reuse all of the IOI circuit's components. Merullo, Eickhoff and Pavlick (2024) found most of the IOI circuit reused for a different task, and used that understanding to repair the second task's errors. Edge Pruning recovers the ground-truth circuits of compiled transformers exactly (Bhaskar et al., 2024). Localising unlearning to components identified mechanistically, rather than by output-preservation alone, produced edits that resisted relearning (Guo et al., 2024).

This paper asks what all of this licenses. Its answer is that both sets of results can be true at once, and that seeing why gives a usable rule. A circuit-discovery run is an estimator. Its output depends on the model and on five further indices: the dataset that defines the task, the ablation used to remove information, the metric and aggregation used to score the result, the grain of the graph (node, edge, position, neuron), and the discovery algorithm with its settings. The negative results show that the output varies across these indices at the level of individual edges and heads. The positive results show that some things do not vary: the roles, the algorithm, a shared core. The mistake in the phrase "the circuit is the mechanism" is not that circuits are useless but that the claim is made about the one draw, when the evidence supports a claim about the class.

This framing builds on a recent position paper and must be distinguished from it. Lin and Liu (2026) audited mechanistic interpretability papers making causal claims, found that none in their purposive sample of ten has a dedicated section on identification assumptions, and proposed a disclosure norm: say whether the claim is causal, name the identification strategy, list the assumptions and stress at least one. This paper takes the norm as given and does the reconciliation that a disclosure norm does not supply. It asks, of the claims that circuit papers have already made, which survive the evidence that has since accumulated, at what level, and what a reader may conclude from a published circuit today.

The paper makes five contributions. It decomposes "the circuit for task T" into seven claims - sufficiency, necessity, minimality, identity, specificity, explanatory adequacy and algorithmic reading - each with its own evidence requirement (Section 2). It names the five indices on which a recovered circuit depends and collects the published evidence that each of them shifts the result far enough to change a conclusion (Section 3). It reads the non-identifiability, specificity and error-reproduction results as statements about the class of subgraphs that pass the standard tests, not about circuits in general (Sections 4 to 6). It shows that the strongest results in favour of circuits are invariance results about roles, algorithms or shared cores, and so do not conflict with the negative ones (Section 7). And it consolidates the measured values in a table and a figure (Section 8), states a licensing procedure as an algorithm (Section 9),

re-reads three published circuits through that procedure (Section 10), and proposes five studies that would settle what remains open (Section 11).

Section 12 states what this paper does not show. Throughout, every number is credited to the source that reported it.

2. What a circuit claim asserts

2.1 Seven claims under one subgraph

When a paper reports "the circuit for task T in model M", it may be asserting any of the following, and usually asserts several at once without separating them. The labels below are this paper's stipulations; they are offered because each needs a different kind of evidence.

C1. Sufficiency (faithfulness). With everything outside the subgraph ablated, the model's behaviour on the task is approximately preserved. This is the property most circuit papers measure, usually as the fraction of a logit difference recovered or as a divergence from the full model's output distribution. Hanna, Pezzelle and Belinkov (2024) put the rationale plainly: faithfulness is what justifies studying the circuit instead of the full model.

C2. Necessity. Ablating the subgraph, with the rest intact, impairs the behaviour. Necessity is not the complement of sufficiency. Heimersheim and Nanda (2024) show with an AND-gate and an OR-gate example that patching in the two directions - restoring clean activations into a corrupted run, or corrupting a clean run - finds different component sets, and that neither result implies the other.

C3. Minimality. The subgraph contains no component whose removal leaves its performance unchanged. Wang et al. (2022) proposed minimality alongside faithfulness and completeness; Shi et al. (2024) turned it into a hypothesis test.

C4. Identity. This subgraph, and not some other, is how the model does the task. Identity is what the definite article in "the circuit" asserts. It requires that the discovery procedure would return the same subgraph, or one equivalent to it in a stated sense, under reasonable variation of everything the claim does not mention.

C5. Specificity. The subgraph is particular to task T: removing it harms T more than it harms tasks unrelated to T, and other tasks' subgraphs do not contain it.

C6. Explanatory adequacy. The subgraph accounts for the model's behaviour on the task as a whole, including its errors and its behaviour on inputs outside the discovery set, not only its average success on that set.

C7. Algorithmic reading. The components implement a stated algorithm - for IOI, an account in which some heads detect the repeated name, others suppress it, and others copy the remaining name to the output - and that algorithm is what the model computes.

The claims are logically distinct. A subgraph can satisfy C1 and fail C5, as Section 5 shows. It can satisfy C1 and C3 and fail C4, as Section 4 shows. Its C7 reading can be stable while its C4 identity is not, which is the pattern Tigges et al. (2024) report across training. Treating the seven as one claim is the source of most of the apparent conflict in the literature.

2.2 The five indices

Every circuit is the output of a procedure, and the procedure has settings. Write the recovered subgraph as an estimate,

$$\hat{C} = A(M; D, a, m, g, \theta)$$

where M is the model; D is the dataset that operationalises the task, including its prompt templates and the corrupted counterparts used for patching; a is the ablation, comprising the replacement value (zero, mean, resampled, optimal) and its direction (noising or denoising); m is the metric and its aggregation; g is the grain of the computational graph (nodes or edges, with or without token positions, heads or neurons or learned features); and A with its hyperparameters θ is the discovery algorithm. Causal-abstraction theory gives this a principled form: activation patching, path patching, causal scrubbing, circuit analysis and several other methods are all cases of checking whether a high-level causal model abstracts a low-level one under a family of interventions (Geiger et al., 2023a). Path patching, which tests whether a behaviour is carried by a stated set of paths through the network, is a direct instance (Goldowsky-Dill et al., 2023). The family of interventions is part of the claim.

The indices are not nuisances that a careful researcher sets once and forgets. Miller, Chughtai and Saunders (2024) state the point in a form that is hard to improve on: "the task a circuit is required to perform depends on the ablation used to test it". If that is right, then D and a are part of what the circuit is a circuit for, and a claim that omits them is underspecified.

2.3 The licensing rule

The rule this paper defends is simple to state. A circuit claim is licensed at the level at which it has been shown to be invariant across the indices it does not name, and no finer. If a paper says that, with mean ablation over a stated corrupted distribution, at node grain with specified token positions, scored by logit difference, this subgraph recovers 87 percent of the effect, it has named every index and makes a C1 claim that one experiment can support. If it says "this is how GPT-2 does IOI", it names none of them, and the claim is licensed only to the extent that what it asserts has been shown to survive changes to all five. The level matters as much as the indices. A claim about which heads play which role, or which algorithm is computed, or which components form a shared core, may be invariant where a claim

about the exact edge set is not.

The rule is not original in spirit. It is the familiar requirement that a scientific claim be robust to the arbitrary choices of the experiment that produced it. What is new is the evidence that, for circuits, the arbitrary choices are not small, and the observation that the published positive results, read carefully, already meet the rule at a coarser level than the one at which they are usually quoted.

3. The indices move the answer

This section collects the evidence that each index, varied within the range that published work treats as reasonable, changes the recovered circuit or its score by enough to change a conclusion.

3.1 Ablation and metric

The first systematic study was Zhang and Nanda (2024), who varied the corruption method and the evaluation metric of activation patching across localisation and circuit-discovery settings and found that the choices could lead to different interpretability conclusions. Their recommendations - corrupt prompts by symmetric token replacement, which keeps the counterfactual input in distribution, rather than by adding Gaussian noise, which does not, and score with logit difference rather than probability - are now common practice. They did not, and did not claim to, remove the dependence; they chose a setting on stated grounds.

Miller, Chughtai and Saunders (2024) went further. Re-evaluating the IOI circuit of Wang et al. (2022), they report that the original evaluation, with mean node ablations at specific token positions, recovered 87 percent of the logit difference. Changing from node to edge ablation gave substantially higher percentages; resample ablation gave systematically lower faithfulness than mean ablation; ablating at every token position instead of only the specified ones lowered it again. The variant that in their judgment best represents the original hypothesis - an edge-level circuit with specific token positions - had a median score well above 100 percent. Faithfulness also rose monotonically with the size of the dataset used to compute the mean ablation, differed between the two name orders of the template, and was substantially higher when computed per prompt and then averaged than when computed from averages. The Docstring circuit was likewise sensitive to the ablation type, and in the Sports Players setting resample and mean ablation gave dramatically different correct-answer percentages. On a compiled Tracr model they showed that whether one input edge belongs in the optimal circuit depends on whether resample or zero ablation is used, because a resampled activation preserves information constant across clean and corrupt inputs. Their conclusion is that the optimal circuit for a prompt set is undefined until an ablation method and a metric are specified.

The direction of the patch is a second index inside the ablation. Heimersheim and Nanda (2024) explain why noising and denoising answer different questions - the first finds what is necessary to keep, the second what is sufficient to restore - and show how an AND structure is fully found by one direction and only partly by the other, with the reverse for an OR structure. Backup heads, which act only when the primary heads are disrupted, are the OR case. In the idealised gate, removing one copy does nothing, so a necessity test misses both; real backups are lossy and compensate only partly, so they weaken a necessity test rather than blind it.

The replacement value is a third. Li and Janson (2024) argue that the usual ablations do more than delete information: a substituted value that is inconsistent with the rest of the computation can actively mislead downstream components, an effect they call spoofing, which inflates measured importance. They propose optimal ablation, which replaces a component's value with the constant that minimises the expected loss of the ablated model. They report advantages for downstream uses including circuit discovery, localisation of factual recall and latent prediction. Whatever one thinks of the specific proposal, the existence of a principled alternative that changes component importance is itself evidence that importance is defined relative to the ablation.

The estimand underneath is less clean than it looks. Vaidyanathan et al. (2026) re-derive what activation patching estimates from causal mediation analysis, in the tradition that Vig et al. (2020) introduced to neural NLP, and show that the per-component quantity contains, besides the effect through the component, interaction terms measuring how much the component's effect depends on the state of other components. In the GPT-2 IOI circuit they report that components whose importance is conditional on others are either missed or inflated, and that the variance of these interaction terms explains the previously documented instability of faithfulness scores. They prove that the interaction terms grow with the distance between the clean and patched activations and are negligible where the model is locally affine. Mueller (2024) reaches a related conclusion from the theory of counterfactual causation: counterfactual tests systematically miss causes when several are independently sufficient, and counterfactual dependence in neural networks is generally not transitive, which complicates building a causal graph edge by edge.

3.2 Dataset and prompts

The task is defined by a dataset, and the dataset is a choice. Rai, Geva and Yao (2026) built several variants of each of four previously studied tasks, varying syntax, complexity or domain while preserving what a human would call the task, and ran standard discovery on each. Circuits found on one variant transferred poorly to

others: at a circuit size of 5 percent of edges, the circuit found on two-person IOI prompts reached 73 percent faithfulness in distribution and 42 percent on a three-person variant, and an entity-binding circuit built at one position scored 111 percent there and 33 percent at another. The largest in-distribution to cross-variant drop they report is 79 percentage points. Mean pairwise edge overlap between circuits for variants of the same task ranged from 0.30 to 0.66. Their most striking result concerns mixtures. Arithmetic and entity binding, discovered separately, gave circuits with near-zero faithfulness on each other's task at sparse sizes; but a discovery set mixing the two returned a single circuit with high faithfulness on both - 89.8 percent on arithmetic and 85.1 percent on entity binding at a 50-50 mix. A high faithfulness score, in other words, does not show that one mechanism is present; a discovery method given two will merge them.

Wu, Tonin and Cevher (2026) decompose the variance of EAP-IG circuits into three parts: resampling a new batch from the same distribution, rephrasing the prompts, and variation across individual samples. They propose a modified method with a theoretical guarantee that substantially reduces the first. The second they trace to a property of the model rather than the method: prompts with different templates tend to activate different circuits, so that a single circuit covering every phrasing of a task may be hard to find. They report that imposing sparsity does not fix this. The third they argue is largely an artefact of how unfaithfulness is defined.

Franco et al. (2026) reach a similar conclusion with a different instrument. Their ACC++ method traces, from a single forward pass, which low-dimensional signals cause each attention head to attend where it does. Applied to IOI across prompt structures, it finds that prompt-specific circuits fall into distinct clusters and that across clusters the same heads receive systematically different signals, corresponding to different ways of identifying the indirect object. On this evidence, the label "name mover head" names a component, not a fixed function.

Meloux, Portet and Peyrard (2025b) treat circuit discovery explicitly as statistical estimation and measure its sampling distribution. They find that exact, single-input causal-mediation scores for an edge vary widely across inputs from one distribution, so that an edge's effect is better treated as a random variable than a fixed property; that gradient approximations add their own noise; and that circuits inherit and amplify both. Averaged over models and tasks, circuits from bootstrap resamples of the discovery set had a mean pairwise edge Jaccard of 0.561 with a coefficient of variation of 0.335, and their mean circuit error was 0.440, the highest of the resampling schemes tried. Resampling at the level of a meta-dataset, or paraphrasing prompts, gave more stable circuits, with Jaccard indices of 0.790 and 0.799. Stability was notably worse for the 1B-parameter Llama

models than for GPT-2 small. They suggest, as a preliminary bar, a mean pairwise Jaccard above 0.8 under at least 100 bootstrap resamples before a circuit is reported as a stable finding. None of their averaged schemes clears it.

Two lines of work respond by engineering stability rather than assuming it. Anani et al. (2026) wrap any discovery algorithm in randomised subsampling of the concept dataset and certify that each inclusion decision is invariant to bounded edits of that dataset, abstaining on components that are not; on vision and language tasks including IOI and Greater-Than they report certified circuits up to 56 percent more accurate with up to 80 percent fewer components. Hadad, Katz and Bassan (2026) use neural-network verification to produce circuits with guarantees of agreement over a continuous input region, of robustness to continuous patching perturbations, and of minimality, demonstrated on vision models. Both are best read as confirmations of the problem: a guarantee of invariance to the dataset is worth building only because invariance cannot be presumed.

3.3 Grain

The grain at which a circuit is defined and evaluated is an index with unusually large effects. Bayat Makou et al. (2026) distinguish source-level evaluation, which keeps every outgoing edge of a selected component, from edge-level evaluation, which keeps only the edges the algorithm selected. Most circuit studies, on their survey, evaluate at source level. They report that source-level evaluation inflates apparent circuit accuracy by up to 85 percentage points relative to edge-level evaluation, and that it hides the many-to-one mapping from structure to function that edge-level evaluation reveals. Haklay et al. (2025) note a different grain assumption: standard methods treat a component as equally relevant at every token position, and allowing position-specific edges gave better trade-offs between circuit size and faithfulness.

Li and Subramani (2026) vary the unit itself, comparing circuits whose nodes are attention heads and MLP blocks with circuits whose nodes are individual MLP neurons, across six tasks and five models. The two grains trade consistency against specificity. Component-level circuits are consistent across inputs but, as Section 5 details, not specific to their task; neuron-level circuits are much more task-specific but far less consistent from one input to another. The authors conclude that a circuit claim should state the grain and whether it means shared or task-specific computation. They also report that circuit structure depended more on the attribution method than on the grain.

3.4 Algorithm and approximation

Exhaustive activation patching over every edge is expensive, so practice relies on approximations. Syed, Rager and Conmy (2023) showed that edge attribution

patching, a linear approximation requiring two forward passes and one backward pass, recovered known circuits better on average than ACDC. Kramar et al. (2024) identified two classes of failure in attribution patching that produce significant false negatives and proposed a corrected variant. Hanna, Pezzelle and Belinkov (2024) found that circuits from plain edge attribution patching were less faithful than those from their integrated-gradients variant, even though both overlapped heavily with manually found circuits at the node level, and drew the general lesson that faithfulness rather than overlap should be compared. Zhang and Wang (2026) trace most of the error of attribution patching to nonlinearity in the layers after the patched component, not to curvature at the component itself, and propose a second-order correction that improves circuit recovery.

Approximation error is not the only algorithmic index. Meloux, Portet and Peyrard (2025b) report that changing the aggregation in EAP-IG from sum to median, and the patching method from mean to patching, reduced the Jaccard similarity of a Greater-Than circuit in Llama-3.2-1B-Instruct to the median circuit to 0.086 - in effect a disjoint subgraph - and that on IOI two EAP variants overlapped at 0.071. Their reading is that different variants are not converging on one circuit but isolating different artefacts of a high-variance score distribution. Adolphi, Vilas and Wareham (2025) add a theoretical ceiling: formalising a range of circuit-finding queries on multi-layer perceptrons, they find many of them intractable, fixed-parameter intractable with respect to natural model and circuit features, and inapproximable under several approximation schemes, while more modest queries remain tractable. If the exact problem is hard, every practical method is a heuristic, and its output must be treated as an estimate whose relation to the exact answer is unknown.

3.5 What the section shows

Each of the five indices, varied within the range that published work treats as reasonable, changes either the recovered subgraph or its score by an amount comparable to the effects being reported. None of this shows that circuits are arbitrary. It shows that C1, the claim the standard tests measure, is a claim relative to the indices, and that C4, the claim the definite article makes, has to be earned by showing invariance across them - which is the subject of the next section.

4. Identity: many subgraphs pass

4.1 Exhaustive enumeration in small networks

The cleanest evidence on C4 comes from settings small enough that every candidate explanation can be listed. Meloux et al. (2025a) trained small multi-layer perceptrons on Boolean functions and examined both of the strategies that interpretability research uses. In the first, which they call where-then-what, a researcher finds a subgraph that reproduces the behaviour and then

interprets its units. In the second, what-then-where, the researcher starts from a candidate algorithm and searches for activation subspaces that align with its intermediate variables under interchange interventions. Both strategies were run as complete searches over the candidates in scope.

For a network with two hidden layers of three units trained to compute XOR, the circuit-first search found 85 distinct circuits that reproduce the network's outputs perfectly, with an average of 535.8 consistent logic-gate interpretations per circuit. The algorithm-first search, testing just two candidate algorithms, found 159 perfect minimal mappings, with perfect alignments for both algorithms. Across many trained networks, the median number of valid abstractions grew with network size from 38 to 910,000 for the circuit-first approach and from 8 to 3,700 for the algorithm-first approach. Fewer than 2 percent of networks had exactly one valid minimal mapping, and none had exactly one circuit interpretation. In that sweep the circuit search was limited to sparse circuits to keep enumeration manageable, so the authors treat both the circuit and the interpretation counts as lower bounds.

The finding is not that interpretation fails in these networks. Every one of the explanations is correct by the criteria used. The finding is that the criteria do not select one explanation, which is precisely what C4 requires them to do. Meloux et al. (2025a) draw the distinction themselves: a pragmatic programme may need only explanations that predict behaviour and support manipulation, while a programme that wants the explanation to be unique needs stricter criteria than the field currently applies. They also report one result that points toward such criteria: in their networks, the number of valid interpretations fell significantly as the number of tasks a network was trained on rose, up to four tasks. Constraints from additional behaviour narrow the class.

4.2 Equivalence classes in trained language models

Enumeration is impossible in a language model, but the same structure has been observed there. Bayat Makou et al. (2026) extracted 75 circuits for a literal sequence-copying task from five Pythia models between 70 million and 1.4 billion parameters, varying only the frequency band of the input tokens. Compared structurally, the circuits looked specialised by band: within-band pairs of circuits shared more edges than between-band pairs, and the differences were statistically distinguishable. But the differences were small - between 0.013 and 0.032 in Jaccard terms for models of 160 million parameters and above - and they were only 0.51 to 1.56 times the standard deviation of Jaccard similarity across repeated draws, which the authors compare to the variation seen when ACDC is simply run twice on the same data. Functionally there was no reliable specialisation: band-specific edges transferred broadly, a core shared across

most bands recovered at least 99 percent of circuit performance in every model above 70 million parameters, and causal interchange interventions showed the internal representations to be interchangeable across bands. A smaller subject-verb agreement replication showed the same pattern. The authors call the mismatch between structural difference and functional equivalence phantom specialisation and conclude that a discovery run returns one member of a set of equally valid subgraphs. They state their own limits: they did not test whether specialisation appears at finer units, such as individual token positions or features within a component, and their transfer test has limited power for very small effects.

This is the empirical counterpart to the enumeration result, and it changes the meaning of a published circuit diagram. If the procedure samples from a class, a single diagram is one sample, and its idiosyncratic edges carry no information about the model. What carries information is what the class has in common - here, a core that does nearly all of the work.

4.3 Representation alignment without a constraint is vacuous

The algorithm-first strategy has its own identifiability problem, sharpened by Sutter et al. (2025). Causal abstraction asks whether there exists a map between a network's internal states and an algorithm's variables under which interventions on the two agree. Most work restricts the map to be linear, following the linear representation hypothesis, but the definition does not require it. Sutter et al. prove that if the map is allowed to be arbitrarily powerful, then under reasonable assumptions any network can be mapped to any algorithm, so that the existence of an abstraction says nothing. They support the proof empirically: with nonlinear alignment maps, randomly initialised language models - models that cannot do the task at all - reached 100 percent interchange-intervention accuracy on IOI. That result held when training and evaluation shared names; with disjoint name sets the authors could not find an alignment for the random models, and they suggest that the random-model result "may depend critically on overlap between the specific entities encountered during training and evaluation". The point stands: a sufficiently flexible map fitted a model that cannot do the task. The restriction on the map, not the intervention test, is doing the identifying work. The same point applies to distributed alignment search (Geiger et al., 2023b), which searches for alignments in a rotated, non-standard basis and is therefore linear by construction: its results are informative to the extent that the linearity assumption is.

The disputed interpretability illusion is a narrower instance. Makelov, Lange and Nanda (2023) showed that a subspace intervention can make a model behave as if a feature had changed while acting through a dormant

parallel pathway that is causally disconnected from the output in normal operation, so that the patched subspace is not where the feature lives. Wu et al. (2024) replied that by Makelov et al.'s technical definition even intuitive and desirable explanations count as illusions, and that the cases observed in practice are artefacts of training and evaluation choices. The exchange has not been resolved, and this paper does not resolve it. Both sides agree on the point that matters here: whether a successful intervention identifies the site of a variable depends on assumptions about the intervention that are external to the success.

4.4 Different algorithms, same data

Identity can fail one level up, at the algorithm, even when each network has a single clean mechanism. Zhong et al. (2023) trained networks on modular addition and found that small changes to hyperparameters and initialisation led networks to learn qualitatively different algorithms - the familiar Clock algorithm, a previously undescribed Pizza algorithm, or more complex procedures - and sometimes several in parallel. A circuit found in one trained network is then evidence about that network, not about how networks solve the task.

The group-operation literature shows the interpretive version of the same problem. Chughtai, Chan and Nanda (2023) proposed that networks trained on finite-group composition implement an algorithm based on representation theory, and found mixed evidence for universality: the family of circuits could be characterised completely, but which circuits a given network learned, and in what order, was arbitrary. Stander et al. (2023) reverse-engineered networks trained on the permutation groups S5 and S6 and found circuits decomposing the arithmetic by the groups' subgroups, drawing explicit attention to the fact that Chughtai et al. had reported a different algorithm for the same problem. Two careful teams, studying comparable models, gave different algorithmic accounts. Wu, Jaburi et al. (2025) then produced a more complete description that moves toward unifying the two and verified it by converting it into a compact proof of model accuracy: for models trained on S5 their explanation yields an accuracy guarantee that runs three times faster than brute force and gives a bound of at least 95 percent for 45 percent of the models trained. They report that they could not obtain non-trivial, non-vacuous bounds from the earlier explanations alone. The dispute was moved forward not by better ablations but by a criterion of a different kind - a proof that the explanation accounts for the model's behaviour - and the explanation that met it was neither of the two earlier accounts on its own.

4.5 Hypothesis tests on published circuits

Shi et al. (2024) formalised what an idealised circuit should satisfy and built three statistical tests: equivalence (the circuit and the full model should be about equally likely to outperform each other on a

random input), independence (with the circuit knocked out, the remaining model's performance should be independent of the full model's), and minimality (no edge should be removable without a significant change). Applied to six published circuits, the two synthetic circuits hard-coded into compiled Tracr models passed all three. Of the four found in trained transformers, the induction circuit passed two, the Docstring circuit passed minimality only, and the IOI and Greater-Than circuits passed none. The authors are careful to add that the discovered circuits are far from random subnetworks and could be improved: about half of the Greater-Than circuit's edges and about a fifth of the IOI circuit's could be removed with little effect on how well they approximated the model.

This is the most direct test in the literature of whether published circuits meet the properties their names imply. The pattern is the one the rest of this section predicts: where structure was imposed, the tests find it; where it was discovered, the tests find a good approximation with redundant parts and a complement that still carries task information.

4.6 An identifiable object exists, but it is not a circuit

Dhor and Chen (2026) offer a positive identifiability result, and it is instructive for what it identifies. Viewing depth as a time axis, so that the forward pass becomes a controlled dynamical system, and applying a Koopman lift to it, they prove that the resulting spectrum is a coordinate-free property of the model, recoverable from calibration samples at a stated rate up to permutation, with a matching lower bound, and they report the predicted rate on Qwen3-8B-Base. But they also prove that, for a non-normal realisation, the directions along which activations vary most cannot be the ones that carry information from layer to layer, and on IOI the Koopman modes beat random directions but lose to principal components. Their summary is exact: "The identifiable object and the legible object are not the same object." That is the shape of the problem for circuits. Identifiability is achievable for objects defined by the model alone; a circuit is defined by the model together with a task dataset, an ablation, a metric, a grain and an algorithm, and its identity is only as fixed as those are.

5. Specificity and explanatory adequacy

5.1 Component-level circuits are not specific

C5 has been tested directly only recently. Li and Subramani (2026) extracted circuits for six tasks - addition, Boolean logic, IOI, a multiple-choice colour-copying task, and the easy and challenge sets of ARC - in five models from the Gemma-2, Llama-3.2 and Qwen3 families, at two grains, and defined a specificity gap: the accuracy drop on a task when its own circuit is ablated, minus the mean drop when each other task's circuit is ablated instead. A gap near zero means a task's own

circuit is no more important to it than an arbitrary other task's.

At the grain of heads and MLP blocks, with EAP-IG, the gap was 1.3 accuracy points; with a second attribution method, RelP, it was 6.2. At the grain of individual MLP neurons it was 22.1 and 22 points respectively. The component-level result follows from overlap: under EAP-IG, for example, the addition circuit shared roughly half of its components with each of the other five tasks' circuits, and the components common to all six tasks in Llama-3.2-3B were mostly MLP blocks, with three attention heads that attend heavily to the beginning-of-text token. The authors are careful about the heads: sink-like attention is ordinary in that model (the median over all 672 heads puts 0.722 of its attention on the first token), and the three shared heads also attend to the answer token at or slightly above the median. They conclude not that the heads are inert but that what they do is not particular to any one task. Component-level circuits were also not always necessary: for Boolean logic and IOI, at their reference settings, the shared component-level circuit had near-zero necessity despite high reuse across inputs.

The authors offer the alternative reading themselves: models may develop reusable computational motifs that support many behaviours, in which case non-specificity is a property of the model rather than a defect of the method. That is close to the reading of Merullo, Eickhoff and Pavlick (2024), who found about 78 percent overlap in attention heads between the IOI circuit and a Colored Objects circuit in GPT-2 Medium and, by adjusting four middle-layer heads to make the second circuit behave like the first, raised Colored Objects accuracy from 49.6 to 93.7 percent. The two papers are consistent. Merullo et al. show that what is shared is functionally meaningful and can be steered; Li and Subramani show that sharing extends far beyond related tasks, to pairs like addition and ARC. Both results undercut C5 at the component grain. Neither undercuts a claim that a component plays a general role that several tasks use.

5.2 The microprocessor precedent

Neuroscience met the specificity problem in lesion studies long before interpretability did, and one demonstration is close enough to be worth recalling. Jonas and Kording (2017) applied standard neuroscience methods to a simulated MOS 6502 microprocessor running three games, a system whose workings are known completely. Lesioning transistors one at a time, they found a subset whose removal prevented exactly one game from running - apparently a Donkey Kong transistor - while 1,565 lesions had no effect and 1,560 stopped all three games. They are blunt about the inference: "a given transistor is obviously not specialized for Donkey Kong or Space Invaders". Necessity for a behaviour, established by lesion, is evidence that the part lies on a path the behaviour uses. It is not evidence

that the part is for that behaviour. Vilas et al. (2024) argue that interpretability should take on board the methodological lessons of cognitive neuroscience, and this is one of the most directly transferable.

5.3 Circuits reproduce successes and miss errors

C6 asks whether a circuit explains the behaviour, not just the average score. Zhang, Geng et al. (2026) proposed a test that standard faithfulness metrics skip: compare the circuit's exact answer with the model's, separately on prompts the model answers correctly and on prompts it gets wrong. For IOI in GPT-2 small under mean ablation, the manual circuit and several automated circuits - including one optimised against the model's full output distribution - agreed with the model on 97.3 to 99.5 percent of prompts it answered correctly but on only 11.4 to 41.7 percent of its errors. Many of the circuits tested in six further model-task settings from the Mechanistic Interpretability Benchmark showed the same asymmetry. In an IOI case study, restoring attention heads the circuits had omitted raised error reproduction on a held-out set from 14.2 to 75.1 percent at a cost of 0.41 percentage points of agreement on correct answers. Their conclusion is that a circuit can preserve task success without explaining the model's failures, and that exact error reproduction is a necessary but not sufficient test.

The asymmetry makes sense on the estimator view. A faithfulness score averages over the discovery distribution, which is dominated by successes, so a discovery method is rewarded for keeping what produces correct answers and is barely penalised for dropping what produces wrong ones. An errors-only score is in effect a held-out test of a part of the behaviour the estimator was not fitted to.

uit de Bos and Garriga-Alonso (2024) reached a related result by adversarial search. Scoring, under resample ablation, how far the circuit's output distribution diverges in KL from the full model's, and then examining the inputs where it diverges most, they found that the IOI and Docstring circuits fail to behave like the full model even on ordinary inputs drawn from the original task. The average conceals a tail, and the tail is where the circuit and the model part company.

5.4 What the circuit reveals depends on how far down one looks

C7 readings are revised when the grain is refined. Nikankin et al. (2024) found a circuit that explains most of several models' behaviour on basic arithmetic, then examined its MLP neurons individually and found a sparse set implementing simple heuristics, each recognising a numerical pattern in the operands and promoting corresponding answers. The unordered combination of these heuristic types explained most of the arithmetic accuracy. A component-level circuit diagram would have suggested an algorithm; the neuron-

level reading suggests a bag of pattern-matchers. Both are faithful at their grain. Only one is the right answer to the question of how the model does arithmetic, and the circuit alone did not say which.

Senthil Kumar (2026) approaches the question from the opposite side. Circuit Condensation post-trains a model with low-rank adapters to concentrate a behaviour into fewer edges, keeping a cut only if task performance and general capability survive. On IOI, condensation isolated 24 heads, 17 of them with roles documented in prior analysis, against 61 heads, 36 of them undocumented, in the matched circuit from the frozen model. Pair ablations showed that edges' effects depend on one another and cannot be read independently. The author describes the result as a sufficient piece of the published mechanism rather than a reconstruction of it. Because condensation changes the weights, its circuit belongs to a different model; what it suggests about the original is that much of what frozen discovery returns may not be required for the behaviour.

5.5 Localisation does not license control

Hase et al. (2023) tested whether the localisation that causal tracing provides tells an editor where to intervene. Causal tracing had placed factual associations in middle-layer MLPs (Meng et al., 2022). Hase et al. found that the layer causal tracing identifies does not predict which layer is best to edit in order to override a stored fact, and that the choice of layer predicts editing success far better than the localisation result does. Necessity and sufficiency of a site for a behaviour, measured by patching, are different properties from the site being the best place to change the behaviour.

6. The counterfactual is answered by a system that reacts

Every ablation test compares the model with a modified model and attributes the difference to what was removed. The comparison assumes that the rest of the model does what it would have done anyway. It does not.

McGrath et al. (2023) found that ablating one attention layer in a language model can cause a later layer to compensate - the Hydra effect - and that late MLP layers act to down-regulate the most likely token, in models trained without dropout. Rushing and Nanda (2024) showed that self-repair occurs across model families and sizes when individual heads are ablated on the full training distribution, that it is imperfect and noisy, sometimes overshooting the original effect, and that changes in the final LayerNorm scale and sparse sets of anti-erasure neurons both contribute. McDougall et al. (2023) identified a head in GPT-2 small, L10H7, whose main role is copy suppression, and showed that copy suppression, carried by the negative heads L10H7 and L11H10, accounts for 39 percent of self-repair in a narrow task: when an earlier copying head is ablated,

there is nothing left to suppress. Heimersheim and Nanda (2024) note that the backup name-mover heads of the IOI circuit, which take over when the main name movers are disrupted, behave like OR gates and fit badly into the circuit definition and the methods built on it.

Gong et al. (2026) formalise the resulting blind spot. A backup that is perfectly dormant in the intact model can be indistinguishable, by per-unit scores computed on the intact model, from a component that is irrelevant; only the growth in its effect after the primary set is removed reveals it. Ranking candidates by how much their ablation effect grows once the primary set is gone, their method picks out the known IOI backup heads with an ROC-AUC of 0.941, and recovery falls to 0.40 plus or minus 0.13 for alternative component sets matched in behavioural effect, output displacement and depth, so the backups it finds are specific to the removed circuit. The subspace illusion of Makelov, Lange and Nanda (2023), in which an intervention activates a pathway that is dormant in normal operation, is a case of the same structure seen from the intervention's side.

For the argument of this paper, the consequence is narrow and definite. Necessity and minimality are measured on a system whose response to the measurement is part of what is measured. A component can be necessary in the intact model and its removal can still be absorbed; a component can be unnecessary in the intact model and become necessary once something else is removed. The ablation index therefore enters not only through the replacement value but through the downstream reaction it provokes. Mueller (2024) identified the general form: counterfactual tests miss causes when several are independently sufficient, which is what a backup is. The self-repair literature raises questions of its own about what restoration is; this paper uses it only for what it implies about ablation as an instrument.

7. The strongest case for circuits, read at the right level

The results of Sections 3 to 6 are sometimes cited as showing that circuit analysis does not work. The evidence does not support that reading, and the strongest pro-circuit results show why. Each of them, examined for what it holds fixed and what it lets vary, is an invariance result at a level coarser than the individual subgraph.

Persistence across training and scale. Tigges et al. (2024) followed circuits for IOI and related tasks through 300 billion tokens of training in decoder-only models from 70 million to 2.8 billion parameters. Task abilities and the components that support them emerged at similar token counts across scale, and both the algorithms and the types of components involved can replicate across scale. They also report that the specific heads implementing a component can change during training while the overarching algorithm remains. That

is an explicit finding of non-invariance at the C4 level - which heads - together with invariance at the C7 level. Bayat Makou et al. (2026) point out that the comparison was made at source-level grain; the C7 conclusion does not depend on that, but any edge-level reading of it would.

Generalisation to prompt variants. Nainani et al. (2024) tested the IOI circuit of GPT-2 small on prompt variants designed to challenge the assumptions of its published algorithm. When they rediscovered circuits for the variants, they found that these "reuse all components of the base IOI circuit, only adding edges from additional input tokens". The base circuit also appeared to succeed on variants where the published algorithm should fail, but the authors trace that success to S2 Hacking, which they describe as "a byproduct of the knockout procedure for evaluating the circuit and not actually how the full model solves the task": under knockout the circuit outperformed the full model on those variants. S2 Hacking is therefore not part of the model's algorithm; it is an instance of the ablation index of Section 3.1 changing what a circuit appears to do. The component roles were invariant across the variants. This is the best case in the literature of a circuit claim surviving a change in D, and it survived at the role level.

Reuse across tasks. The Merullo, Eickhoff and Pavlick (2024) result discussed in Section 5.1 is an invariance of roles across tasks, and the intervention that repaired Colored Objects is evidence that the role-level understanding supports control.

Greater-than. Hanna, Liu and Variengien (2023) identified a circuit in GPT-2 small for completing a year range with a later year, explained the role of its components - with the final MLPs boosting the probability of end years above the start year - and found related tasks that activate the same circuit. The claim they make is about a general mechanism active across contexts, a role-level claim, and it is the kind of claim that the variance results of Meloux, Portet and Peyrard (2025b), who used Greater-Than as one of their tasks, leave open, since those concern which edges are selected rather than what the selected components do.

Agreement between manual and automated discovery. Conmy et al. (2023) validated ACDC by showing that it rediscovered all five component types of the manually found Greater-Than circuit, choosing 68 edges out of 32,000, every one of them also present in the manual analysis. This is real evidence, but of a limited kind. ACDC and the manual analysis shared the dataset, the corruption scheme and the metric; agreement between two procedures that share three of the five indices is evidence of invariance across the other two. It is not evidence of invariance across D, a and m.

Recovery of planted structure. Tracr compiles human-readable programs into transformers whose structure is known (Lindner et al., 2023). Edge Pruning recovered

the ground-truth circuits in two compiled models exactly, and scaled to CodeLlama-13B, where it found circuits with more than 99.96 percent sparsity that match the full model's performance (Bhaskar et al., 2024). InterpBench trains semi-synthetic transformers with a strict form of interchange intervention training that stops non-circuit nodes from affecting the output, then uses them to evaluate discovery methods (Gupta et al., 2024). Shi et al. (2024) found that compiled circuits pass all three of their idealised tests. These results establish that the methods can recover a planted circuit when the ablation used for discovery matches the one that defines the ground truth (Miller, Chughtai and Saunders, 2024). They do not establish that a trained natural-language model has one; the training behind InterpBench was designed specifically to prevent the non-circuit parts from mattering, which is the property whose absence Sections 4 and 5 document in trained models.

Benchmarked progress. The Mechanistic Interpretability Benchmark ranks methods on circuit localisation and causal-variable localisation across four tasks and five models (Mueller et al., 2025). It finds attribution and mask-optimisation methods best for circuit localisation, supervised distributed alignment search best for causal variables, and sparse-autoencoder features no better than neurons. Progress against a fixed benchmark is progress at the benchmark's settings of D, a and m; it is compatible with everything in Section 3.

Mechanistic localisation that pays off in control. Guo et al. (2024) compared unlearning and editing localised by methods that select components chiefly to preserve outputs with editing localised to the components of a high-level mechanism whose intermediate states are predictable - for factual recall, the components implementing a lookup-table stage. The mechanistically localised edits were more robust across input and output formats, more resistant to attempts to relearn the removed information, and had fewer unintended side effects, on a sports-facts dataset and CounterFact across several models. This is the most important positive result for the argument here, because it shows that the C7 level is not only more stable but more useful: a localisation chosen by what the components compute served control better than one chosen by what the subgraph reproduces. It is the complement of the Hase et al. (2023) finding. Localisation by output effect did not predict where to edit; localisation by mechanism did better.

Feature circuits in use. Sparse feature circuits replace heads and neurons with learned dictionary features, and Marks et al. (2025) used them to improve a classifier's generalisation by ablating features a human judged irrelevant to the task. The use was validated by the downstream result, not by the circuit's identity, which is the pattern this paper recommends.

Universality at the level of roles and families. Gurnee et al. (2024) compared GPT-2 models trained from five random seeds and found that 1 to 5 percent of neurons are universal, in the sense of consistently activating on the same inputs, and that these neurons usually have clear interpretations and fall into a small number of families with recurring functional roles. Chughtai, Chan and Nanda (2023), as noted in Section 4.4, found the family of circuits for group composition characterisable and the particular circuits learned arbitrary. In both, the invariant is the role or the family; the particular unit or subgraph is not.

Narrowing the class. Three results show that the equivalence class can be made smaller. Meloux et al. (2025a) found the number of valid interpretations fell when networks were trained on more tasks. Wu, Tonin and Cevher (2026) reduced resampling variance with a method that carries a theoretical guarantee, and Anani et al. (2026) certified inclusion decisions against bounded changes to the dataset. These are the right kind of response, because they act on the index that varies instead of on the member that happened to be drawn.

Read together, the pro-circuit and the critical results describe one picture. At the level of the individual subgraph - which edges, which heads - the output of circuit discovery varies with every index and is one of many that pass. At the level of roles, algorithms, families and shared cores, there is substantial invariance across training time, scale, prompt variants and tasks. Much of the disagreement in the literature comes from quoting the coarse-level results as if they licensed the fine-level claim, and the fine-level results as if they refuted the coarse-level one.

8. The evidence, consolidated

Table 1 collects the measured values on which the argument rests, one row per cited result, with the claim of Section 2.1 that each bears on most directly. Figure 1 re-presents a subset of them. Three patterns stand out. First, every row from 1 to 9 except row 3, each of which varies an index of the procedure, records a change large enough to alter a conclusion: faithfulness moving by tens of percentage points, and edge overlap between circuits for the same task falling well short of the 0.8 that Meloux, Portet and Peyrard (2025b) propose as a minimum, and in two cases close to zero. Row 3 is the exception, and an instructive one: resampling at the level of a meta-dataset or of paraphrases brings the overlap to just under the bar, which the source reads as the more stable case. Second, the rows that report invariance (rows 10, 21, 22 and 24) report it for a shared core, a role, a family or an algorithm, never for a full edge set in a natural-language task. Third, the rows on specificity and errors (rows 14 to 17) show the claims beyond sufficiency failing at the grain at which circuits are usually reported.

9. A licensing procedure

9.1 What each claim needs, and where the record stands

Table 2 states, for each of the seven claims, what it asserts, the invariance it needs before it is licensed, and the state of the published evidence as Sections 3 to 7 read it. The assignments are this paper's.

9.2 The procedure

Algorithm 1 states the decision structure. Its organising rule is the one given in Section 2.3: a claim is stated at the finest level at which it has been shown invariant across every index it does not name.

Three features of the procedure need comment. Step 2 is where most of the work lies, and it is where the procedure differs from the disclosure norm of Lin and Liu (2026). Their norm asks authors to state assumptions; the procedure says what to do with the claim once the assumptions are stated, namely to treat every unstated index as something the claim implicitly generalises over and to test that generalisation. Step 5 is the constructive move. A circuit claim that fails the identity test is not discarded; it is restated about the object the runs share, which Sections 4 and 7 show often exists and often does most of the work. The 0.8 threshold in step 4 is the preliminary bar proposed by Meloux, Portet and Peyrard (2025b) and is used here as they offer it, as a starting convention, not a derived quantity. Step 8 encodes the lesson of the group-operation dispute: when two algorithmic readings both survive ablation, ablation cannot choose between them, and a criterion of another kind is needed.

10. Three published circuits, re-read

10.1 Indirect object identification in GPT-2 small

The IOI analysis of Wang et al. (2022) is the most studied circuit in the literature, which makes it the best test of the procedure.

C1. Under the original evaluation, the circuit recovers 87 percent of the logit difference; at edge grain with token positions the median exceeds 100 percent, and under resample ablation faithfulness is lower (Miller et al., 2024). The claim is licensed with its indices stated, and the indices should travel with it.

C2 and C3. Backup name movers take over when the primary name movers are removed (Heimersheim and Nanda, 2024), conditional co-ablation recovers them as specific to the removed set (Gong et al., 2026), the circuit does not pass Shi et al.'s (2024) minimality, equivalence or independence tests, and component-level IOI circuits in larger models had near-zero necessity at the reference settings of Li and Subramani (2026). Necessity and minimality are not licensed as stated in the original diagram.

Table 1. Published measurements behind the argument, one cited result per row. "Varied" names the index or condition the source changed; "Claim" names the claim of Section 2.1 that the value bears on. Each value is the one its row's source reports; the consolidation is the author's.

#	Source	Setting	Varied	Value	Claim
1	Miller et al. (2024)	GPT-2 small, IOI	ablation, grain, positions	87% in original evaluation; median over 100% at edge grain with positions	C1
2	Meloux et al. (2025b)	EAP-IG, avg. over models and tasks	bootstrap resample of D	edge Jaccard 0.561 (CV 0.335); circuit error 0.440	C4
3	Meloux et al. (2025b)	same	paraphrase / meta-dataset	edge Jaccard 0.799 / 0.790	C4
4	Meloux et al. (2025b)	Llama-3.2-1B-Instruct, G-T / IOI	EAP variant	Jaccard 0.086 / 0.071	C4
5	Rai et al. (2026)	GPT-2, IOI, 5% of edges	2- vs 3-person prompts	faithfulness 73% vs 42%	C1, C6
6	Rai et al. (2026)	Qwen2.5, entity binding	position P2 vs P8	faithfulness 111% vs 33%	C1, C6
7	Rai et al. (2026)	four tasks	task-preserving variants	mean edge Jaccard 0.30-0.66	C4
8	Rai et al. (2026)	arithmetic + binding	50-50 discovery mix	one circuit: 89.8% / 85.1%	C4
9	Bayat Makou et al. (2026)	Pythia, sequence copying	source vs edge grain	inflation up to 85 points	C1
10	Bayat Makou et al. (2026)	Pythia above 70M	frequency band	shared core recovers 99% or more	C4 (class)
11	Meloux et al. (2025a)	XOR MLP, exhaustive	candidate explanations	85 perfect circuits; 535.8 readings each	C4, C7
12	Sutter et al. (2025)	random-init LMs, IOI	nonlinear alignment map	100% IIA	C7
13	Shi et al. (2024)	six published circuits	three hypothesis tests	IOI, G-T pass 0 of 3; Tracr pass 3 of 3	C1-C3
14	Li and Subramani (2026)	6 tasks, 5 models	grain, attribution	specificity gap 1.3 / 6.2 (comp.), 22.1 / 22 (neuron)	C5
15	Merullo et al. (2024)	GPT-2 Medium	IOI vs Colored Objects	78% head overlap; 49.6% to 93.7% after edit	C5, C7
16	Zhang, Geng et al. (2026)	GPT-2 small, IOI, mean abl.	correct vs error prompts	agreement 97.3-99.5% vs 11.4-41.7%	C6
17	Zhang, Geng et al. (2026)	same, held-out	omitted heads restored	error reproduction 14.2% to 75.1%	C6
18	Gong et al. (2026)	GPT-2 small, IOI	conditional co-ablation	backup AUC 0.941; matched sets 0.40	C2, C3
19	McDougall et al. (2023)	GPT-2 small	two negative heads	39% of self-repair, narrow task	C2
20	Senthil Kumar (2026)	IOI	condensation vs frozen	24 heads (17 documented) vs 61 (36 not)	C3
21	Tigges et al. (2024)	Pythia 70M-2.8B	300B training tokens, scale	algorithm stable; heads change	C4, C7
22	Gurnee et al. (2024)	GPT-2, 5 seeds	random seed	1-5% of neurons universal	C4 (role)
23	Conmy et al. (2023)	GPT-2 small, G-T	manual vs ACDC	68 of 32,000 edges, all found manually	C1, C4
24	Wu, Jaburi et al. (2025)	S5 group networks	compact proof	95% or better bound for 45% of models	C7
25	Anani et al. (2026)	ResNet, ViT, GPT-2	dataset-certified	up to 56% more accurate, 80% fewer parts	C4
26	Jonas and Kording (2017)	MOS 6502	single-transistor lesions	1,565 no effect; 1,560 stop all games	C5

C4. The circuit found on two-person prompts loses much of its faithfulness on three-person prompts (Rai et al., 2026); prompt-specific circuits fall into clusters with different signals reaching the same heads (Franco et al., 2026). The identity claim - these heads and edges are how GPT-2 small does IOI - is not licensed.

C5. At component grain, the IOI circuit shares most of its heads with a Colored Objects circuit in GPT-2 Medium (Merullo, Eickhoff and Pavlick, 2024), and across the six tasks of Li and Subramani (2026), including IOI, the component-level specificity gap was 1.3 accuracy points with EAP-IG and 6.2 with RelP. Specificity is not

licensed at the component grain.

C6. The circuit agrees with the model on a minority of its errors (Zhang, Geng et al., 2026) and diverges on some ordinary inputs (uit de Bos and Garriga-Alonso, 2024). Explanatory adequacy is not licensed.

C7. The roles - something detects the duplicated name, something suppresses it, something copies the remaining one - persist across training and scale (Tigges et al., 2024), and the circuits rediscovered for prompt variants reuse all of the IOI circuit's components (Nainani et al., 2024). The base circuit's apparent success on variants where its algorithm should fail is S2 Hacking, a by-

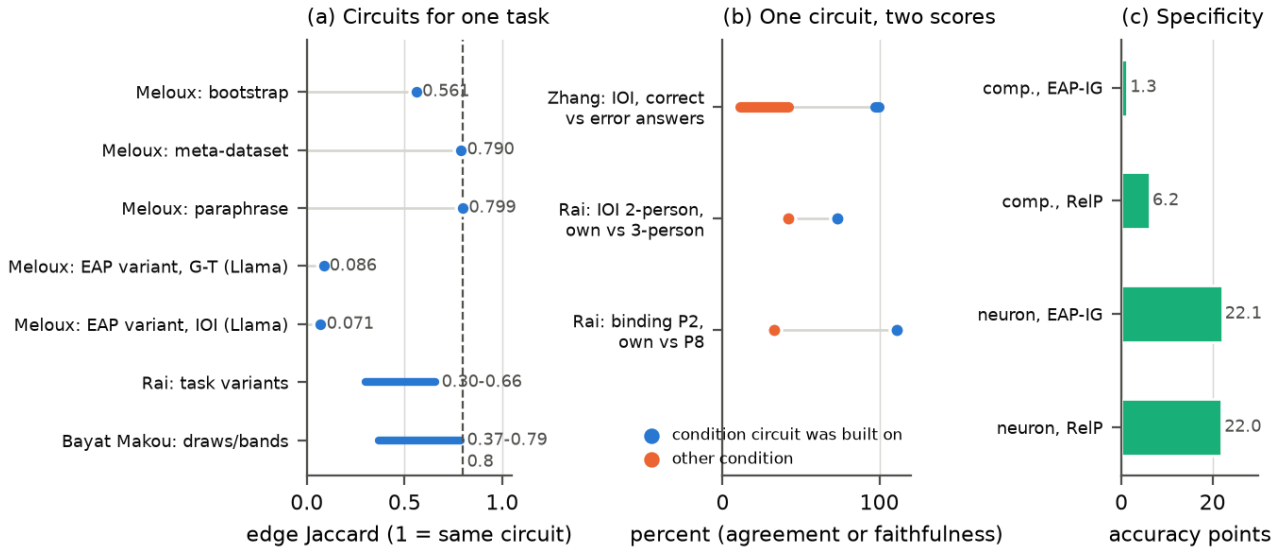


Figure 1. Published values re-presented by the author. (a) Edge Jaccard between circuits for the same task: bootstrap, meta-dataset and paraphrase resampling, and two EAP-variant changes in Llama-3.2-1B-Instruct, from Meloux, Portet and Peyrard (2025b); ranges across task-preserving dataset variants from Rai, Geva and Yao (2026) and across draws and frequency bands from Bayat Makou et al. (2026); dashed line, the 0.8 minimum proposed by Meloux et al. (b) One circuit scored in two conditions: agreement with GPT-2 small on its correct answers versus its errors for IOI circuits, from Zhang, Geng et al. (2026); faithfulness on the discovery variant versus another variant, from Rai et al. (2026). (c) Specificity gap in accuracy points by grain and attribution method, from Li and Subramani (2026). Values come from different settings and are not a controlled comparison.

Table 2. The seven claims, the invariance each requires, and the status of the published record.

Claim	Asserts	Must be shown invariant across	Status in the record
C1 sufficiency	behaviour kept with complement ablated	nothing beyond the indices it names	supported when D, a, m, g are stated; the score moves with each (Miller et al., 2024; Bayat Makou et al., 2026)
C2 necessity	removal impairs behaviour	ablation type, patch direction, backups	mixed: backups and self-repair hide components (Gong et al., 2026; Rushing and Nanda, 2024)
C3 minimality	no redundant parts	ablation; edge interactions	not met by IOI or G-T (Shi et al., 2024); edges interact (Vaidyanathan et al., 2026)
C4 identity	this subgraph is how M does T	all five indices	unsupported for any natural-language task; a shared core is supported (Bayat Makou et al., 2026)
C5 specificity	particular to T	other tasks' circuits, at a stated grain	fails at component grain; partly holds at neuron grain (Li and Subramani, 2026)
C6 adequacy	explains errors and tails	errors, held-out and adversarial inputs	fails for IOI (Zhang, Geng et al., 2026; uit de Bos and Garriga-Alonso, 2024)
C7 algorithm	components compute a stated algorithm	checkpoints, scale, variants; or a proof	supported at role level for IOI (Tigges et al., 2024; Nainani et al., 2024); proof-level for group operations (Wu, Jaburi et al., 2025)

Algorithm 1. Licensing a circuit claim at the level its evidence reaches

Input: claim K about model M and task T; circuit $C_{\text{hat}} = A(M; D, a, m, g, \theta)$ step 1 classify K as C1..C7 (Section 2.1); "the circuit for T" is a C4 claim step 2 list the indices K names; every unnamed index is quantified over by K step 3 C1-C3: score under ≥ 2 ablation types, both patch directions, node and edge grain; report the per-prompt spread, not only the mean step 4 C4: rerun A on ≥ 100 bootstrap resamples and on task-preserving variants of D; report mean pairwise edge Jaccard and cross-variant faithfulness step 5 if Jaccard < 0.8 or cross-variant faithfulness falls: extract what the runs share (core, roles, algorithm); test the shared object by transfer and interchange interventions; restate K about the shared object, not about the draw step 6 C5: ablate other tasks' circuits on T at the same grain; report the gap step 7 C6: score exact agreement separately on M's errors, held-out variants and worst-case inputs step 8 C7: check the reading across checkpoints, scale or variants; prefer a test beyond ablation (compact proof, error prediction, durable edit) step 9 output K at the finest level that passed, with D, a, m, g and A stated

product of knockout that the full model does not show (Nainani et al., 2024), and so is no part of the reading. The algorithmic reading is licensed at the level of roles.

The verdict is that the IOI work is one of the field's genuine successes, but its success lies at the level of an algorithm described by roles, not at the level of a diagram of 26 heads. A reader who cites the diagram as the mechanism claims more than the evidence licenses; a reader who cites the role-level algorithm claims about what it licenses.

10.2 Greater-than in GPT-2 small

The Greater-Than circuit (Hanna, Liu and Variengien, 2023) was rediscovered by ACDC with all its component types (Conmy et al., 2023), which supports C1 and a limited C4 within shared indices. It fails all three of Shi et al.'s idealised tests. Meloux, Portet and Peyrard (2025b) used it as their running example of instability: varying several circuit-finding choices at once produced many different circuits. Their sharpest single number for the task, a Jaccard of 0.086 to the median circuit after changing the aggregation and patching method of EAP-IG, comes from Llama-3.2-1B-Instruct, not GPT-2 small, and is not evidence about this circuit. What survives is the role-level account - late MLPs that boost years above the start year - which Hanna et al. stated as a general mechanism and which none of the variance results contradicts. The procedure licenses that account and not a specific edge set.

10.3 Group composition in small networks

The group-operation case is the clearest example of what a fully licensed C7 claim looks like. Ablation-based analyses produced two different algorithmic readings (Chughtai, Chan and Nanda, 2023; Stander et al., 2023). The reading that was then verified did not win by surviving more ablations; it was converted into a compact proof that bounds the model's accuracy, and the proof succeeded for a stated fraction of trained models (Wu, Jaburi et al., 2025). Nothing comparable exists yet for a natural-language task, and Section 11 proposes one step toward it.

11. What would settle the open part

The open part has two pieces: whether the shared object - core, roles, algorithm - that Section 7 argues is invariant really is, across all five indices at once, for natural-language tasks; and whether an identity claim at the subgraph level can ever be licensed for such tasks. Five studies would move both. None needs new theory.

S1. Stability reporting for the canonical circuits. For IOI, Greater-Than, Docstring and the MIB tasks, report the mean pairwise edge Jaccard and the cross-variant faithfulness of each published circuit under bootstrap resampling, prompt paraphrase and task-preserving dataset variants, with at least two ablation types and both grains. Meloux, Portet and Peyrard (2025b) did part

of this for EAP-IG; extending it to the manual circuits would say how much of each published diagram is draw and how much is class. If the published circuits mostly clear the 0.8 bar once evaluated at a fixed, stated grain, the C4 verdict here is too harsh.

S2. The core as the estimand. Bayat Makou et al. (2026) estimate a shared core across frequency bands and find that it does nearly all of the work. The same estimand - the set of edges present in most draws across bootstrap samples, dataset variants and ablation types - should be computed for the canonical tasks, and its sufficiency, necessity and transfer reported. The prediction from Section 7 is that the core will be stable and close to sufficient, and that it will map onto the role-level algorithm. If the core is small and insufficient, the role-level account is weaker than Section 7 claims.

S3. Errors as a benchmark track. Add exact error reproduction, measured as Zhang, Geng et al. (2026) measure it, and worst-case divergence, measured as uit de Bos and Garriga-Alonso (2024) measure it, to the Mechanistic Interpretability Benchmark. A method that scores well on faithfulness and badly on errors would then be visible as such.

S4. Planted redundancy. InterpBench-style models (Gupta et al., 2024) currently plant one circuit and suppress the rest. A variant that deliberately plants two redundant circuits for one task, a dormant backup that activates only under ablation, and a component shared with a second task would provide ground truth for exactly the cases this paper is about. The measurement is how often each discovery method reports one circuit, the other, a merge, or misses the backup, under each ablation type. That gives an error rate for C2 to C5 under redundancy, which no benchmark currently reports.

S5. A criterion beyond ablation for a natural-language C7 claim. The group-operation case was moved forward by a proof. For a natural-language task the analogue need not be a proof; it could be a quantitative prediction from the role-level algorithm, made before looking, of which inputs the model will get wrong, scored against the model's actual errors. IOI is the natural candidate, because its algorithm is well described and its error set is now characterised (Zhang, Geng et al., 2026). A role-level algorithm that predicts the model's errors would be the first natural-language circuit claim licensed at C6 and C7 together.

12. What this paper does not show

The seven claims and five indices are stipulations. Saphra and Wiegrefe (2024) distinguish several senses of "mechanistic" that would carve the claims differently, and Williams et al. (2025) and Vilas et al. (2024) draw on philosophy and on cognitive neuroscience for other frameworks; Sharkey et al. (2025) survey the open methodological problems more broadly than this paper does. The decomposition is offered because each claim

maps onto a different evidence requirement, not because it is the only map.

Much of the evidence is new, and it comes from few places. A good part of what the argument leans on consists of 2026 preprints (Li and Subramani; Rai, Geva and Yao; Wu, Tonin and Cevher; Zhang, Geng et al.; Vaidyanathan et al.; Gong et al.; Senthil Kumar; Franco et al.; Dhor and Chen), and a large share of all the evidence concerns GPT-2 small and the IOI task. The record could change quickly, and it is not a random sample of circuits.

Most of the critical results concern circuits whose nodes are heads, MLP blocks, edges or neurons. Bayat Makou et al. (2026) and Li and Subramani (2026) both state explicitly that they did not test circuits defined over learned features. Li and Subramani note that a feature basis might recover consistency and specificity together; Bayat Makou et al. suggest that feature-level circuits "may resolve ambiguities arising from polysemantic components". The verdicts of Table 2 should not be carried over to feature circuits (Marks et al., 2025) without that test.

The arXiv abstract of the latest version of Gong et al. (2026) and its HTML body report different values for the strongest intact-state baseline and for the reduction in circuit incompleteness; this paper uses only the values on which the two agree. Figure 1 and Table 1 combine values from different settings and are a consolidation, not a controlled contrast.

Finally, this paper does not show that circuits are not mechanisms in every sense. It shows that the claims "the circuit for task T" usually carries are of seven kinds, that the published record licenses some of them at the level of roles, algorithms and shared cores, and that it licenses the identity of a subgraph for no natural-language task studied so far.

13. Conclusion

A circuit-discovery run returns an estimate. The estimate depends on the dataset that defines the task, the ablation that removes information, the metric that scores the result, the grain of the graph and the algorithm with its settings, and the published evidence shows each of these moving the estimate by amounts large enough to change conclusions. Many different subgraphs pass the standard tests, component-level circuits for unrelated tasks largely coincide, and circuits that reproduce a model's successes can miss most of its failures. None of this refutes the circuits programme. The results that defend it - persistence across training and scale, carry-over to prompt variants, reuse across tasks, recovery of planted structure, edits that resist relearning - are invariance results about roles, algorithms, families and shared cores, and they are consistent with every negative result about individual subgraphs. The rule that follows is the ordinary one for any estimate: state the claim at the level

at which it has been shown not to depend on the arbitrary choices that produced it. On that rule, the best circuit work in the literature is better supported than its critics suggest and more modestly than its diagrams imply.

References

- Adolfi, F., Vilas, M. G., & Wareham, T. (2025). The Computational Complexity of Circuit Discovery for Inner Interpretability. ICLR 2025. arXiv:2410.08025. doi:10.48550/arXiv.2410.08025.
- Anani, A., Lorenz, T., Schiele, B., Fritz, M., & Fischer, J. (2026). Certified Circuits: Stability Guarantees for Mechanistic Circuits. ICML 2026. arXiv:2602.22968. doi:10.48550/arXiv.2602.22968.
- Bayat Makou, A., Niu, J., Dutta, S., & Gurevych, I. (2026). Many Circuits, One Mechanism: Input Variation and Evaluation Granularity in Circuit Discovery. Transactions on Machine Learning Research. arXiv:2606.06267. doi:10.48550/arXiv.2606.06267.
- Bhaskar, A., Wettig, A., Friedman, D., & Chen, D. (2024). Finding Transformer Circuits with Edge Pruning. NeurIPS 2024. arXiv:2406.16778. doi:10.48550/arXiv.2406.16778.
- Chughtai, B., Chan, L., & Nanda, N. (2023). A Toy Model of Universality: Reverse Engineering How Networks Learn Group Operations. arXiv:2302.03025. doi:10.48550/arXiv.2302.03025.
- Conmy, A., Mavor-Parker, A. N., Lynch, A., Heimersheim, S., & Garriga-Alonso, A. (2023). Towards Automated Circuit Discovery for Mechanistic Interpretability. NeurIPS 2023. arXiv:2304.14997. doi:10.48550/arXiv.2304.14997.
- Dhor, A., & Chen, P.-Y. (2026). Intrinsic Structure: Spectral Identifiability for Mechanistic Interpretability. arXiv:2608.10172. doi:10.48550/arXiv.2608.10172. [PREPRINT]
- Franco, G., Tassis, L. M., Rohr, A., & Crovella, M. (2026). Finding Interpretable Prompt-Specific Circuits in Language Models. arXiv:2602.13483. doi:10.48550/arXiv.2602.13483. [PREPRINT]
- Geiger, A., Ibeling, D., Zur, A., Chaudhary, M., Chauhan, S., Huang, J., Arora, A., Wu, Z., Goodman, N., Potts, C., & Icard, T. (2023a). Causal Abstraction: A Theoretical Foundation for Mechanistic Interpretability. arXiv:2301.04709. doi:10.48550/arXiv.2301.04709.
- Geiger, A., Wu, Z., Potts, C., Icard, T., & Goodman, N. D. (2023b). Finding Alignments Between Interpretable Causal Variables and Distributed Neural Representations. arXiv:2303.02536. doi:10.48550/arXiv.2303.02536.
- Goldowsky-Dill, N., MacLeod, C., Sato, L., & Arora, A. (2023). Localizing Model Behavior with Path Patching. arXiv:2304.05969. doi:10.48550/arXiv.2304.05969. [PREPRINT]
- Gong, Z., Lu, H., Wang, T., Zhang, Y., Wang, Y., Zeng, Z., Xiao, M., Yuen, C., & Lim, W. Y. B. (2026). Conditional Co-Ablation: Recovering Self-Repair Backups in Transformer Circuits. arXiv:2607.01940. doi:10.48550/arXiv.2607.01940. [PREPRINT]
- Guo, P., Syed, A., Sheshadri, A., Ewart, A., & Dziugaite, G. K. (2024). Mechanistic Unlearning: Robust Knowledge Unlearning and Editing via Mechanistic Localization. arXiv:2410.12949. doi:10.48550/arXiv.2410.12949. [PREPRINT]
- Gupta, R., Arcuschin, I., Kwa, T., & Garriga-Alonso, A. (2024). InterpBench: Semi-Synthetic Transformers for Evaluating Mechanistic Interpretability Techniques. NeurIPS 2024 Datasets and Benchmarks. arXiv:2407.14494. doi:10.48550/arXiv.2407.14494.
- Gurnee, W., Horsley, T., Guo, Z. C., Kheirkhah, T. R., Sun, Q., Hathaway, W., Nanda, N., & Bertsimas, D. (2024). Universal Neurons in GPT2 Language Models. arXiv:2401.12181. doi:10.48550/arXiv.2401.12181.
- Hadad, I., Katz, G., & Bassan, S. (2026). Formal Mechanistic Interpretability: Automated Circuit Discovery with Provable Guarantees. ICLR 2026. arXiv:2602.16823. doi:10.48550/arXiv.2602.16823.

- Haklay, T., Orgad, H., Bau, D., Mueller, A., & Belinkov, Y. (2025). Position-aware Automatic Circuit Discovery. arXiv:2502.04577. doi:10.48550/arXiv.2502.04577. [PREPRINT]
- Hanna, M., Liu, O., & Variengien, A. (2023). How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. NeurIPS 2023. arXiv:2305.00586. doi:10.48550/arXiv.2305.00586.
- Hanna, M., Pezzelle, S., & Belinkov, Y. (2024). Have Faith in Faithfulness: Going Beyond Circuit Overlap When Finding Model Mechanisms. COLM 2024. arXiv:2403.17806. doi:10.48550/arXiv.2403.17806.
- Hase, P., Bansal, M., Kim, B., & Ghandeharioun, A. (2023). Does Localization Inform Editing? Surprising Differences in Causality-Based Localization vs. Knowledge Editing in Language Models. NeurIPS 2023. arXiv:2301.04213. doi:10.48550/arXiv.2301.04213.
- Heimersheim, S., & Nanda, N. (2024). How to use and interpret activation patching. arXiv:2404.15255. doi:10.48550/arXiv.2404.15255. [PREPRINT]
- Jonas, E., & Kording, K. P. (2017). Could a Neuroscientist Understand a Microprocessor? PLOS Computational Biology, 13(1), e1005268. doi:10.1371/journal.pcbi.1005268.
- Kramar, J., Lieberum, T., Shah, R., & Nanda, N. (2024). AtP*: An efficient and scalable method for localizing LLM behaviour to components. arXiv:2403.00745. doi:10.48550/arXiv.2403.00745. [PREPRINT]
- Li, M., & Janson, L. (2024). Optimal ablation for interpretability. arXiv:2409.09951. doi:10.48550/arXiv.2409.09951. [PREPRINT]
- Li, M., & Subramani, N. (2026). How Much Do Circuits Tell Us? Measuring the Consistency and Specificity of Language Model Circuits. arXiv:2605.08348. doi:10.48550/arXiv.2605.08348. [PREPRINT]
- Lin, Z., & Liu, F. (2026). Position: Mechanistic Interpretability Must Disclose Identification Assumptions for Causal Claims. arXiv:2605.08012. doi:10.48550/arXiv.2605.08012. [PREPRINT]
- Lindner, D., Kramar, J., Farquhar, S., Rahtz, M., McGrath, T., & Mikulik, V. (2023). Tracr: Compiled Transformers as a Laboratory for Interpretability. NeurIPS 2023. arXiv:2301.05062. doi:10.48550/arXiv.2301.05062.
- Makelov, A., Lange, G., & Nanda, N. (2023). Is This the Subspace You Are Looking for? An Interpretability Illusion for Subspace Activation Patching. NeurIPS 2023 Workshop on Attributing Model Behavior at Scale. arXiv:2311.17030. doi:10.48550/arXiv.2311.17030.
- Marks, S., Rager, C., Michaud, E. J., Belinkov, Y., Bau, D., & Mueller, A. (2025). Sparse Feature Circuits: Discovering and Editing Interpretable Causal Graphs in Language Models. ICLR 2025. arXiv:2403.19647. doi:10.48550/arXiv.2403.19647.
- McDougall, C., Conmy, A., Rushing, C., McGrath, T., & Nanda, N. (2023). Copy Suppression: Comprehensively Understanding an Attention Head. arXiv:2310.04625. doi:10.48550/arXiv.2310.04625. [PREPRINT]
- McGrath, T., Rahtz, M., Kramar, J., Mikulik, V., & Legg, S. (2023). The Hydra Effect: Emergent Self-repair in Language Model Computations. arXiv:2307.15771. doi:10.48550/arXiv.2307.15771. [PREPRINT]
- Meloux, M., Maniu, S., Portet, F., & Peyrard, M. (2025a). Everything, Everywhere, All at Once: Is Mechanistic Interpretability Identifiable? ICLR 2025. arXiv:2502.20914. doi:10.48550/arXiv.2502.20914.
- Meloux, M., Portet, F., & Peyrard, M. (2025b). Mechanistic Interpretability as Statistical Estimation: A Variance Analysis. arXiv:2510.00845. doi:10.48550/arXiv.2510.00845. [PREPRINT]
- Meng, K., Bau, D., Andonian, A., & Belinkov, Y. (2022). Locating and Editing Factual Associations in GPT. NeurIPS 2022. arXiv:2202.05262. doi:10.48550/arXiv.2202.05262.
- Merullo, J., Eickhoff, C., & Pavlick, E. (2024). Circuit Component Reuse Across Tasks in Transformer Language Models. ICLR 2024. arXiv:2310.08744. doi:10.48550/arXiv.2310.08744.
- Miller, J., Chughtai, B., & Saunders, W. (2024). Transformer Circuit Faithfulness Metrics are not Robust. COLM 2024. arXiv:2407.08734. doi:10.48550/arXiv.2407.08734.
- Mueller, A. (2024). Missed Causes and Ambiguous Effects: Counterfactuals Pose Challenges for Interpreting Neural Networks. arXiv:2407.04690. doi:10.48550/arXiv.2407.04690. [PREPRINT]
- Mueller, A., Geiger, A., Wiegrefe, S., Arad, D., Arcuschin, I., Belfki, A., Chan, Y. S., Fiotto-Kaufman, J., Haklay, T., Hanna, M., Huang, J., Gupta, R., et al. (2025). MIB: A Mechanistic Interpretability Benchmark. ICML 2025. arXiv:2504.13151. doi:10.48550/arXiv.2504.13151.
- Nainani, J., Vaidyanathan, S., Yeung, A. J., Gupta, K., & Jensen, D. (2024). Adaptive Circuit Behavior and Generalization in Mechanistic Interpretability. arXiv:2411.16105. doi:10.48550/arXiv.2411.16105. [PREPRINT]
- Nikankin, Y., Reusch, A., Mueller, A., & Belinkov, Y. (2024). Arithmetic Without Algorithms: Language Models Solve Math With a Bag of Heuristics. arXiv:2410.21272. doi:10.48550/arXiv.2410.21272. [PREPRINT]
- Rai, D., Geva, M., & Yao, Z. (2026). Data-driven Circuit Discovery for Interpretability of Language Models. arXiv:2605.09129. doi:10.48550/arXiv.2605.09129. [PREPRINT]
- Rushing, C., & Nanda, N. (2024). Explorations of Self-Repair in Language Models. ICML 2024, PMLR 235:42836-42855. arXiv:2402.15390. doi:10.48550/arXiv.2402.15390.
- Saphra, N., & Wiegrefe, S. (2024). Mechanistic? arXiv:2410.09087. doi:10.48550/arXiv.2410.09087.
- Senthil Kumar, S. A. (2026). Circuit Condensation: Post-Training that Concentrates a Behavior's Causal Circuit. arXiv:2608.27254. doi:10.48550/arXiv.2608.27254. [PREPRINT]
- Sharkey, L., Chughtai, B., Batson, J., Lindsey, J., Wu, J., Bushnaq, L., Goldowsky-Dill, N., Heimersheim, S., Ortega, A., Bloom, J., Biderman, S., Garriga-Alonso, A., et al. (2025). Open Problems in Mechanistic Interpretability. arXiv:2501.16496. doi:10.48550/arXiv.2501.16496. [PREPRINT]
- Shi, C., Beltran-Velez, N., Nazaret, A., Zheng, C., Garriga-Alonso, A., Jesson, A., Makar, M., & Blei, D. M. (2024). Hypothesis Testing the Circuit Hypothesis in LLMs. arXiv:2410.13032. doi:10.48550/arXiv.2410.13032. [PREPRINT]
- Stander, D., Yu, Q., Fan, H., & Biderman, S. (2023). Grokking Group Multiplication with Cosets. arXiv:2312.06581. doi:10.48550/arXiv.2312.06581. [PREPRINT]
- Sutter, D., Minder, J., Hofmann, T., & Pimentel, T. (2025). The Non-Linear Representation Dilemma: Is Causal Abstraction Enough for Mechanistic Interpretability? NeurIPS 2025. arXiv:2507.08802. doi:10.48550/arXiv.2507.08802.
- Syed, A., Rager, C., & Conmy, A. (2023). Attribution Patching Outperforms Automated Circuit Discovery. NeurIPS 2023 ATTRIB Workshop. arXiv:2310.10348. doi:10.48550/arXiv.2310.10348.
- Tigges, C., Hanna, M., Yu, Q., & Biderman, S. (2024). LLM Circuit Analyses Are Consistent Across Training and Scale. NeurIPS 2024. arXiv:2407.10827. doi:10.48550/arXiv.2407.10827.
- uit de Bos, N., & Garriga-Alonso, A. (2024). Adversarial Circuit Evaluation. arXiv:2407.15166. doi:10.48550/arXiv.2407.15166. [PREPRINT]
- Vaidyanathan, S., Arbour, D., Mueller, A., Niekum, S., & Jensen, D. (2026). The Curse of Multiple Mediators: Hidden Interaction Effects in Activation Patching. arXiv:2606.27510. doi:10.48550/arXiv.2606.27510. [PREPRINT]
- Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Sakenis, S., Huang, J., Singer, Y., & Shieber, S. (2020). Causal Mediation Analysis for Interpreting Neural NLP: The Case of Gender Bias. arXiv:2004.12265. doi:10.48550/arXiv.2004.12265.
- Vilas, M. G., Adolphi, F., Poeppel, D., & Roig, G. (2024). Position: An Inner Interpretability Framework for AI Inspired by Lessons from Cognitive Neuroscience. ICML 2024. arXiv:2406.01352.

- doi:10.48550/arXiv.2406.01352.
- Wang, K., Variengien, A., Conmy, A., Shlegeris, B., & Steinhardt, J. (2022). Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 small. arXiv:2211.00593. doi:10.48550/arXiv.2211.00593.
- Williams, I., Oldenburg, N., Dhar, R., Hatherley, J., Fierro, C., Rajcic, N., Schiller, S. R., Stamatiou, F., & Sogaard, A. (2025). Mechanistic Interpretability Needs Philosophy. arXiv:2506.18852. doi:10.48550/arXiv.2506.18852. [PREPRINT]
- Wu, F. Z., Tonin, F., & Cevher, V. (2026). Demystifying Variance in Circuit Discovery of LLMs. arXiv:2606.16920. doi:10.48550/arXiv.2606.16920. [PREPRINT]
- Wu, W., Jaburi, L., Drori, J., & Gross, J. (2025). Towards a unified and verified understanding of group-operation networks. ICLR 2025. arXiv:2410.07476. doi:10.48550/arXiv.2410.07476.
- Wu, Z., Geiger, A., Huang, J., Arora, A., Icard, T., Potts, C., & Goodman, N. D. (2024). A Reply to Makelov et al. (2023)'s "Interpretability Illusion" Arguments. arXiv:2401.12631. doi:10.48550/arXiv.2401.12631. [PREPRINT]
- Zhang, F., & Nanda, N. (2024). Towards Best Practices of Activation Patching in Language Models: Metrics and Methods. ICLR 2024. arXiv:2309.16042. doi:10.48550/arXiv.2309.16042.
- Zhang, L., Geng, C., Zhang, M., Yang, C., Zhang, L., Ye, H., & Si, X. (2026). Rethinking Circuit Evaluation: Do Circuits Explain Model Errors? arXiv:2609.35686. doi:10.48550/arXiv.2609.35686. [PREPRINT]
- Zhang, L., & Wang, J. (2026). When Attribution Patching Lies: Diagnosis and a Second-Order Correction. arXiv:2606.09899. doi:10.48550/arXiv.2606.09899. [PREPRINT]
- Zhong, Z., Liu, Z., Tegmark, M., & Andreas, J. (2023). The Clock and the Pizza: Two Stories in Mechanistic Explanation of Neural Networks. NeurIPS 2023. arXiv:2306.17844. doi:10.48550/arXiv.2306.17844.