

Legibility Under a Better Vocabulary

Description, Instruments, and Development in Neural Interpretability

Flyxion
Independent Researcher

October 2026

Abstract

An apparently incoherent neuron can become intelligible when its responses are described differently, observed with a better instrument, or reconstructed through their development. These possibilities carry different evidential obligations: a coherent description does not by itself establish a causal role, and a legible final mechanism does not uniquely identify the process that produced it. Drawing on Nick Cammarata’s working hypothesis that neural networks are already clean, this essay distinguishes descriptive plurality from mechanistic plurality, probe behavior from model behavior, and developmental continuity from a simple final-state meaning. It examines the capacity-reduction experiments, weak-activation ablations, comparisons with sparse dictionaries, and audit procedures on which the hypothesis rests, asking in each case what the observation licenses and which rival accounts remain compatible with it. Particular attention is given to the asymmetry between developmental evidence from retrained vision models and interpretations of finished language models, and to the difference between evidence for a hypothesis and reasons to pursue the research program it motivates.

The resulting account treats interpretability as the construction of experimentally answerable descriptions whose scope remains attached to the observations and interventions that support them. It proposes interventional predictive adequacy as an operational criterion for evaluating an explanation together with a declared prediction procedure, relative to specified contexts, interventions, outcomes, and tolerances. The criterion evaluates predicted intervention effects alongside absolute outcomes, while requiring explicit limits on auxiliary computation and access to the original model. Predictive adequacy remains distinct from mechanistic understanding, which additionally requires grounding explanatory claims in identifiable components and their causal relations. Failures can motivate changes in contextual scope, operating regime, information route, or explanatory unit, rather than automatically requiring a finer decomposition into features. Two experimental applications address developmental dependence on intermediate associations and whether a validated circuit map improves prediction of unexplained mechanisms. These proposals have not been experimentally validated and are distinct from the evidence reviewed for Cammarata’s hypothesis. The essay concludes that the reported findings warrant greater optimism about useful mechanistic investigation than about comprehensive coverage, recovery of developmental history, or reliable control of extended behavior. An appendix collects the claims, their evidential sources, and the tests that would sharpen them.

1 The categories in the observation

A neuron responds to cat faces and the fronts of cars. At first, the conjunction appears to announce a defect in the network’s organization: two unrelated things occupy the same computational unit. Yet the apparent unrelatedness already contains an interpretation. Cats and cars are everyday categories selected by an observer, whereas a learned computation may distinguish pointed arrangements from rounded ones. If both kinds of image supply a similar V-shaped pattern, the description has divided what the neuron has usefully brought together. Cammarata’s reinterpretation of this canonical example, a unit labeled 4e:55 in the InceptionV1 image model, makes a methodological issue unusually concrete: a report of polysemanticity can contain an unexamined commitment about which vocabulary ought to make the mechanism intelligible [1].

The evidence for the reinterpretation is of several kinds, and they are worth separating. The unit’s strongest inputs include detectors for fox snouts, wolf faces, and vehicle fronts, and its readers include units responding to flames, coyote faces, and the stems of martini glasses. Its negative pole, the response it gives when the activation is pushed below zero, is associated with round objects such as golf balls and clocks. A mixture of cats and cars has no obvious reason to feed a unit that responds to flames, whereas a pointed-versus-round distinction does. The account therefore explains the inputs, the outputs, and the negative side together, and it also predicts responses to new groups of images, such as pool tables and bullet trains, which Cammarata reports tested as expected [1, *Cars and Animals*]. The argument has the structure of a conjecture that could have failed in several places.

The history of the example adds a further consideration. The earlier circuits literature, which Cammarata co-authored, stated that the unit was not responding to a commonality of cars and cat faces, and a later paper on multimodal neurons [4] considered a shared feature and then concluded that the unit was a disjunction, as Cammarata reports [1, *Cars and Animals*]. Careful investigators had therefore entertained a shared pattern and set it aside, apparently because feature visualization produced whiskers and chrome. The source offers a mechanism for the misleading evidence: a visualization raises the activation by whatever route is easiest, so once it finds a route through cat circuitry it adds everything that route brings, whether or not the unit depends on it. The case shows that the vocabulary problem is not a failure of attention. It can survive the explicit consideration of the right answer, because the evidence used to reject that answer was itself a product of an instrument.

This does not mean that every collection of responses must conceal one elegant pattern. A sufficiently permissive description can connect almost anything. It means that the apparent number of meanings cannot be read directly from the number of familiar nouns needed to describe the strongest examples. The distinction between cats and cars may matter profoundly to the final classification while being irrelevant to an intermediate computation. A network can preserve a shared pose or shape at one stage and separate object categories later. Cammarata reports more than a hundred generalized pose units in InceptionV1, each a detector of a pose shared across cars, dogs, and other objects, and reports the same grouping in a model more than ten times larger [1, *Cars and Animals*]. The intermediate grouping should be evaluated through the work it enables.

It helps to distinguish two claims. Descriptive plurality occurs when an observer uses several labels for a unit’s responses. Mechanistic plurality occurs when the unit participates in

distinguishable computations with different conditions and consequences. The first may suggest the second, but cannot establish it alone. Conversely, finding a common descriptive theme does not demonstrate a common causal role. The interpretation becomes stronger when it predicts responses to previously unexamined pointed objects, explains rounded counterexamples, and agrees with the neuron’s inputs and downstream effects. A better vocabulary earns its place by making these additional constraints easier to state and test.

Cammarata states a practical heuristic that follows from this. Both human and automated interpreters tend to describe units as detectors for groups of nouns, whereas a unit is often better understood as drawing a distinction that downstream units can use. Asking how the unit would be described if groups of nouns were forbidden frequently exposes a shared pattern. The heuristic is useful because it changes the description language, but it should be read as a generator of hypotheses and not as a result. A verb-like description can be as unconstrained as a noun-like one, and the test of either is the same: whether it survives predictions that the alternative would not make.

In this sense, the positive and negative sides of a response can be more informative together than a list of positive examples. A unit that promotes one route while suppressing a competing route makes a distinction within a particular computation. The negative side need not express an absolute conceptual opposite. Taxis and ambulances can be contrasted where they need separating, even though both belong together elsewhere. The source follows one sports car through the last five blocks of a larger model, where the contrast successively separates the car from buses and trams, then from snakes and dogs, then from minibuses, then from vans and wagons, and finally convertibles from station wagons [1, *How I Think About Neurons*]. The meaning of the distinction depends partly on its location and readers, and the same unit read without its location can look arbitrary.

2 What an instrument establishes

The second issue concerns how observations are produced. Feature visualization optimizes an input to excite a selected unit. Schematically, it searches for

$$x^* \approx \arg \max_x [a_j(x) - \lambda R(x)],$$

where a_j is an activation and R expresses the visualization’s regularizing preferences. The resulting image is a solution found under a particular objective and search procedure. It demonstrates a route to activation, while leaving open whether that route is typical, exclusive, or representative of the computation under investigation. The method itself has a history of corrections. Early renderings were noisy until transformation robustness was added, and deeper layers later required further changes because of the reward-hacking behavior described below [2].

Cammarata reports that freely optimized visualizations of readers in a car circuit produced dogs. The explanation offered is that gradient ascent takes the easiest path to a high activation, that the training data contain many dogs and the model many dog units, and that the early optimization falls into them and does not recover. Biasing the search toward the paths taken by the unit’s strongest natural examples, often for only the first sixty-four steps [1, *Feature Visualization in Later Layers*], changed the result: some readers were more recognizably

car-related, while others shared pose information across animals and vehicles. The original visualization had encouraged an inference from a convenient synthetic stimulus to a general semantic identity. Its appearance was real, but its evidential scope had been enlarged beyond what the procedure established. The circuit that had been read as a car hidden among unrelated dog units, and had served in the early circuits literature as the main example of superposition in a real network, was reread as shared pose [3].

Path-guided optimization improves the question by conditioning the search on a route of interest. It also introduces its own selection. A car-guided visualization can illuminate car processing without exhaustively describing the neuron, and the author notes that most ways of nudging the optimization out of the dog basin work and that no canonical best method has been identified. Stronger interpretation therefore comes from agreement among natural examples, multiple optimization settings, traced connections, and interventions. No single picture carries the whole explanation.

The same logic extends to estimates of how interpretable a layer is. Cammarata reports that the last layer of the original InceptionV1 was judged to have essentially no interpretable units, that his own recollection was about five percent, that improved tools raised this to about fifty percent on the same network, that retraining the architecture with modern regularization gave about seventy percent, and that a modern convolutional network of the same size reached about ninety-nine percent [1, *A History of Microscope Errors*]. These figures are valuable because they decompose a large change into tool changes and architecture changes, and the decomposition supports the view that much of the apparent mess was not a property of the learning principle. They are also the author’s own judgments of what counts as an interpretable unit, formed by an observer who knew what the successful explanations looked like. An evaluation in which raters of the original and the retrained model were blind to the condition, or in which the criteria were fixed before the comparison, would be a stronger form of the same evidence.

Top-activating-token displays have a related limitation. A comma can be the position at which information is available without being the content represented. If attention gathers a running topic summary onto punctuation, a neuron can fire repeatedly on commas while tracking a regional context or a social situation. In the source’s example, a unit listed by a standard top-k procedure as a polysemantic comma unit turned out to track Colorado: the commas stayed dark until Denver appeared in the text and then brightened as the context built, and steering the unit made the model plan a hike [1, *Superposition of the Gaps*]. The display licenses a claim about activation locations. Recovering the proposed content requires varying the surrounding context, testing whether the response follows the topic when its position changes, and examining the routes that deliver information there. Similarly, restricting the context window can turn a long-context unit into an apparent mixture: a last paragraph about coffee and universities suggested a unit for those two topics until the reason further up the document was included.

Gated units add another distinction. A SwiGLU output has the schematic form

$$z_j(x) = \text{SiLU}(g_j(x)) v_j(x).$$

A positive output can arise when both factors are positive or when both are negative. Ranking products without inspecting their factors can combine different operating regimes. In the example the source examines, the top examples for a unit appeared to be references to the book

of Colossians, but every one had the gate closed and the activation was a small sign-flipped leak of the value, under eight percent of the unit’s peak. Restricted to firings with the gate open, the unit’s job was to complete the word Russia or Russian in any script. For a separate unit labeled as an enemy detector, 131 of 133 responses between the seventy-fifth and ninety-ninth percentiles used the negative-negative regime, while all 93 responses at the 99.99th percentile and above used the positive-positive regime. The author estimates that for about four percent of units this echo constitutes the whole top of the list [1, *What We Found in Language Models; Separating the Sign Regimes*]. Separating the regimes does not automatically reveal a semantic function, but it removes an ambiguity introduced by the measurement summary. The same observed sign need not indicate the same internal conditions.

The negative pole of a unit can matter as much as the positive one for similar reasons. In the source’s example, a standard label for one unit reads as capital cities because the labeling procedure looks only at positive activations, whereas the unit’s negative side corresponds to Santa lore. Removing the eight units with a Santa negative pole reduced the probability of Santa from near certainty to six percent, while removing twenty units labeled as Christmas or Santa reduced it only to seventy-six percent [1, *What We Found in Language Models*]. The comparison is informative because it shows that label-based selection and mechanism-based selection can disagree by a large margin, and that the disagreement is in the direction the sign analysis predicts. It remains a single example of an intervention in one model, and the generality of the pattern is a separate matter.

Interventions also have bounded implications. Increasing a neuron and observing hiking-related text supports a connection to a proposed Colorado interpretation, but does not uniquely establish it. Several correlated contexts can favor hiking. Ablation can demonstrate a contribution under tested conditions while leaving redundancy, compensation, and distributional coverage unresolved. The relevant question is which alternatives a particular intervention actually discriminates. A persuasive account combines tests that would fail for different reasons if the interpretation were wrong.

Evaluation metrics are instruments of the same kind. Cammarata argues that simulation scores, which reward an explanation for predicting the numerical activations of a unit, often measure where attention places information and how weak activations decay, not whether the explanation is correct. An explanation can fail such a score because the unit fires on punctuation where the simulator expected the topic word, and it can succeed because it models the rate at which an activation falls as a face turns away. The argument is that the metric answers a question about attention motifs and weak-tail detail [1, *On Simulation Scores*]. A high score is therefore some evidence, since an explanation that predicts activations is capturing something about the unit, but it discriminates poorly between a correct explanation and a description that predicts activations for the wrong reasons, so it cannot serve as the criterion by which explanations are accepted. Any summary statistic compresses a high-dimensional picture, and the source itself warns against holding cleanliness scores tightly.

3 Where information is placed

The punctuation example generalizes into one of the more consequential methodological points in the source: in language models the position at which a unit fires can be separated from

the content it represents, because attention moves information to places chosen by the model. Cammarata reports that attention commonly collects summaries on punctuation, which he calls the summary motif, and that after recognizing the motif a large class of previously puzzling units becomes easier to investigate. A reader of activation lists who does not know the motif will see a pile of unrelated commas. A reader who does will look at the surrounding sentences, steer the unit, trace upstream and downstream connections, and examine the negative pole, and each of these checks can fail independently.

The attention experiments in the source follow the same pattern of an intervention that discriminates between alternatives. A unit labeled as common function words fires only in text from roughly 2012 to 2016, and the date information arrives through one head that glances back at a nearby clue such as the numeral in a software version. Blocking that single look reduced the unit’s activation on one example from 7.3 to 0.1, whereas blocking three other heads that attend to the numeral even more strongly left it at 4.8. In a second case, a unit that completes an idiom looks back at the verb that began it, and blocking one head’s look lowers the response from 13.4 to -0.1 while blocking a different head that looks at the same verb more strongly leaves it at 12.9. In a third, a unit that fires on the word might writes into a head’s query so that the head looks back for an earlier may, and across 32 may-might pairs the look was at least halved in 31 when the unit was removed [1, *Attention*]. In each case the control matters: the effect is specific to a particular route and not to attention to the clue in general.

These experiments support a modest and useful conclusion, that neurons and heads are often easier to understand through each other than separately. They do not establish that heads are individually nameable. The author estimates that global names for attention heads would add roughly five percent of explanatory residual beyond local attention flow and global neuron explanations, notes that an agent reviewing a hundred attribution graphs gave a similar estimate, and states that such estimates of the value of knowledge not yet possessed could be far off. The prediction that heads will turn out to be broad and recursively generalized, like units in small models, is explicitly a low-confidence guess. This is the correct level of confidence, and an essay should keep it distinct from the stronger statements made about units.

4 Local extensions and global shape

The third issue appears when a unit’s present response set is approached through its history. Cammarata illustrates recursive generalization with a response to “iy” that extends to “diy” and then to “do it yourself.” The last expression no longer contains the original letter sequence. A description restricted to a shared substring therefore fails, even though each extension has an intelligible connection to the preceding one. In the vision example, a unit in InceptionV1’s last layer begins on rapeseed fields in photographs that sometimes include a dog, and over training it grows toward tennis balls and green fruit, then toward shaggy dogs, and finally toward broccoli and cauliflower, until the fields are nearly absent. A unit in a modern convolutional network trained on the same data begins in the same place, reaches briefly for balls and fruit, and then releases them, remaining on the fields. The comparison suggests that architecture changes how far the branching goes without changing the kind of process.

The following is a possibility result and not evidence about any network: its empirical

content depends on fixing the dissimilarity independently of the responses, as discussed after the display. The important logical point is that local relatedness need not be transitive. Suppose a dissimilarity d on response types satisfies the triangle inequality, and suppose each successive pair in a sequence $r_0, \dots, r_n$ satisfies $d(r_t, r_{t+1}) \leq \epsilon$. Then

$$d(r_0, r_n) \leq \sum_{t=0}^{n-1} d(r_t, r_{t+1}) \leq n\epsilon,$$

and the bound is attained when the responses lie along a line. A uniform bound on every step is therefore compatible with an end-to-end distance that grows without limit in the length of the chain. Repeating locally intelligible extensions can produce a final collection that seems heterogeneous under a static description. Development supplies an explanation of how that collection arose without requiring every member to resemble the starting point in the same respect.

The formal point also shows where the empirical content lies. The claim that each step is small depends on the dissimilarity d , and the choice of d is the vocabulary question of the first section returning in a different form. A chain whose steps are small under a semantic metric chosen after the fact can be constructed between almost any two responses, since a sufficiently rich ontology contains intermediate concepts between any pair. The hypothesis becomes testable when the metric is fixed independently, for example by the circuits through which each response arrives, by the co-occurrence of responses in training data, or by the effect of ablation on each. A chain is then a prediction that adjacent responses share upstream routes or downstream consequences, and that the order of acquisition follows the chain.

The source contains one case in which the routes can be read. A unit in a language model responds to planetary geology, and its strongest firings mix lunar crust with code for two software libraries, one of which the source describes as a geology-related JavaScript library. Grouped by the circuits that feed them, the top one hundred firings separate into facets labeled Moon and Mars crust (44 percent), Moon missions (42 percent), old impacts on Mars (6 percent), and Cesium and IDL code (8 percent) [1, *Recursive Generalization*]. Three facets arrive through planetary circuits, while the code offshoot arrives through its own route, a unit responding to names such as Cesium. This is an instructive case for the distinction between descriptive and mechanistic plurality. The offshoot is mechanically distinct in that it arrives by a separate pathway, and it is judged related because the library’s name is associated with the subject. The judgment that this is an offshoot of one function and not a second function therefore rests on a semantic association, and the argument that such offshoots are minor rests on their small share of firings and, as the source reports, on their lower causal importance than the main response.

This is a path-dependent account because the available machinery helps determine which extensions are easy to learn. An established route can give a related computation a head start, influencing where subsequent learning occurs. Cammarata develops this into a circuit ecology hypothesis: computations that reuse existing useful parts may capture error signals before more isolated alternatives develop, so that a broad circuit starves a narrow competitor of gradient. The illustrations offered include a network trained on a single photograph of a Labrador retriever, whose top images for the output turn out to be dominated by dark green scenes, and a network trained on one photograph each of ten categories, which labels about one new photograph in five correctly against a chance rate of one in ten [1, *Why Networks*

Generalize]. The source itself notes that matching photographs by color does about as well, so what carries over is mostly color, and the result is evidence that broad features begin to form from the first examples, not that semantic generalization is present. That proposed mechanism is broader than the observation of branching and requires separate evidence about learning dynamics. The source acknowledges that it is a guess, and that grokking on modular arithmetic is a case where memorization leads, although even there a generalizing circuit grows gradually beneath it [11, 12].

Two cautions matter here. First, a developmental account does not prove that no unified functional description exists. An observer may later discover a distinction that organizes the whole response set. Second, a continuous training path alone cannot establish semantic continuity. Smoothly changing parameters can produce substantial changes in function. Evidence for recursive generalization must show the proposed local relationships and their order, rather than merely displaying checkpoints in sequence.

The idea has an acknowledged precedent. A section of the earlier multimodal neurons paper described conjoined features, concepts that overlap in one facet and generalize outward in several directions, and speculated that they might be vestigial remnants of early training [4]. Cammarata reports that he had forgotten the passage until late in the project, credits it, and notes that observation of training now suggests the branches are not vestigial, although how much they grow depends on circumstances [1, *Relation to Conjoined Features*]. The history is a small instance of the thesis, since a pattern was observed, described in a vocabulary tied to the ontology of features, and not recognized as a general explanation until the surrounding framework changed.

Training records could make these explanations substantially more accountable. Useful records would include checkpoint identities, activation distributions on stable reference inputs, examples selected before inspecting later states, changes in incoming and outgoing contributions, and interventions testing suspected route acquisition. These records could distinguish a retained branch from replacement of an earlier function, or a contextual association from a robust extension. Tracking one coordinate is a useful beginning, but interacting units can redistribute a computation, so a coordinate history is not automatically a complete mechanism history. The source also records a cautionary instance in its own practice: shortly before publication, a mis-set zoom in the visualization suggested that one unit might have arbitrary mixing after all, and the author describes the episode as an example of blaming the network for the microscope [1, *Watching Neurons Develop*].

Final weights remain indispensable because they implement the behavior being explained. Development answers an additional question: why this mechanism acquired this arrangement rather than another. Logging history expands interpretability’s evidence without making historical access a prerequisite for every successful explanation.

5 Capacity reduction as a test

The sharpest experimental contrast in the source is between two accounts of what happens when a layer has too little room. On the first, a network that runs out of neurons continues to group things that share a useful pattern, so that its units become coarser. On the second, it begins to place unrelated things in the same units because those things rarely co-occur and

therefore interfere little. Care is needed in stating what each account predicts. Semantic locality is one measurable quantity among several, and it does not by itself separate the accounts: unrelated features can share coordinates while remaining separately recoverable as different directions of the layer’s population activity, and related features can also be superposed, so a unit that groups similar things does not exclude a superposed code beneath it. Cammarata deliberately targets a narrower developmental claim, that learning switches principle under capacity pressure and that the groupings of units thereby become arbitrary, and the squeeze experiment bears on that claim and not on superposition in the geometric or counting senses. A sharper protocol would define the competing models explicitly and state their predictions for several quantities at once: the sparsity and co-occurrence structure of the responses, the interference between them measured as cross-talk in the readout, the geometry of the representation including whether distinct feature directions can be recovered from the layer’s population activity, including their projections onto the inspected units, the dependence of downstream units on each readable component, and the dynamics by which each component is acquired. The semantic relatedness of the grouped responses would then be one measurement among these. The squeeze test remains attractive, because under the narrower claim it has outcomes that could have gone either way.

Cammarata trained InceptionV1 from scratch with layer 3b reduced from 480 to 10 units and no bypass around it, a reduction of 48 times [1, *What Neurons Do Under Pressure*]. The retrained model reached 58.35 percent top-1 and 81.14 percent top-5 accuracy on the ImageNet validation images. The ten units were reported as one for curves and loops where ten curve detectors had stood, two circle units in place of several circle types, and units for edges, colors, eyes and faces, and high-frequency texture, with no overlapping mixtures of the kind that packing would predict. A sparse autoencoder trained on the squeezed layer did not recover ten clean curve detectors: the matching features were 44 to 174 degrees wide where the original detectors were 24 to 56 degrees wide, they responded at the opposite angle where the originals were silent, and two original detectors shared one feature [1, *What Neurons Do Under Pressure*, details panel]. The autoencoder explained 82 percent of the variance of held-out positions, and splicing it in reduced top-1 accuracy on a validation sample from 58.8 to 43.0 percent [1, *What Neurons Do Under Pressure*, details panel]. A parallel experiment on the final layer of a medium ConvNeXt, narrowed from 1,024 to 13 units with the final bypass removed, gave 57.6 percent accuracy against 72.0 percent at full width, again with coarser units such as beetles and vehicles grouped together [1, *What Neurons Do Under Pressure*].

What the test provides evidence against is a specific form of packing: one produced by shrinking a single layer of an otherwise normally trained network and measuring the result by inspection. This restricts the hypothesis, and the null that survives is the absence of packing recoverable by the dictionary and inspection methods used, not the absence of packing of any kind, since packing that a dictionary cannot factorize would be missed. The evidence is also limited by the imperfect reconstruction of the sparse autoencoder used to search for recoverable features, and the source frames it as the question of whether a monosemantic layer that is well understood can be made superposed. The test does not reach every circumstance in which packing could be proposed. The data have rich structure, and the author’s own argument is that packing is least likely where things are related and most likely for memories that are ends in themselves. A squeeze that preserves the semantic structure of the data and the architecture of the rest of the network would not be expected to create such conditions. Natural images

cannot be guaranteed to contain categories that share no features with anything else, so the extension is better carried out on controlled synthetic features whose correlations, sparsity, and downstream reuse are set by the experimenter and varied independently while the squeeze is held fixed. Whether packing appears, and under which settings, would then be measured on the quantities listed above and not inferred from how the strongest examples look. The source reports that the toy setting in which packing is observed has the relevant property of unrelated, end-in-themselves features [1, *Toy Models*], which leaves the empirical question of where real networks lie between the two.

The test also depends on the observer’s judgment about what counts as coarse. A unit for curves and loops is described as a generalization because the grouped responses share a visual character, but the grouping is recognized by a person looking at the strongest examples, so the verdict that a squeezed unit is coarse and not mixed is subject to the same vocabulary risks as any other. The sparse autoencoder comparison gives a partial external check, since the features that would be needed to unpack the units were not found, but it does not test whether a unit grouping dark circles with eyes and faces is more or less related than one grouping curves with spiral textures. A blind version would present rater panels with units from squeezed and unsqueezed networks, along with units constructed by combining unrelated units, and ask which are mixtures.

A related experiment tests rare things directly. The source teaches a 7.06-million-parameter network that a photograph of an umbrella is a bee, an association with no support in the data, and finds that the top guess changes from umbrella to maypole to bee in nine training steps [1, *Rare Things and Memorization*]. The probability of umbrella falls from 44.2 percent to 0.03 percent and that of bee rises from 0.07 percent to 52.8 percent, with the change concentrated in the last three blocks, where umbrella-related units weaken and existing bee units strengthen. Swapping those blocks into the original model produces the new answer. The result shows that in this network a nonsensical association was implemented by rerouting among existing routes, not by a new isolated feature. It is a single example in one small model, and the source also notes that a model trained to map random noise to random labels leaned partly on textures that it already used, with most of the causal effect nonetheless in the late layers. The reasonable conclusion is that memorization in these networks tends to reuse available machinery and sit near the output, and that the strong version, that no isolated memorization circuits exist anywhere, has not been tested.

6 Weak activations and the long tail

The worry that unrelated features hide in weak activations is the most natural remaining objection to the claim that strong activations are interpretable, and the source treats it with three separate arguments. The first is statistical. A feature packed into a unit to avoid interference with the main pattern would want to sit where the main pattern is not densely active, but in a language-model unit the weaker the activation range the denser the tokens become: the source reports that a given range of values near a unit’s ninetieth percentile contains about ten thousand times as many tokens as the same range near its 99.99th percentile, and that of one hundred million tokens through one unit only about 160 reach the strong half of its maximum [1, *Do Weak Activations Matter?*]. A packed feature in the weak range would therefore collide with orders of magnitude more other responses for little benefit. This is a

plausible argument from incentives, and it assumes that packing is chosen by a process sensitive to collision rates, which is the thing at issue.

The second argument is an ablation. In the image models, removing the weakest activations costs little. Keeping only roughly the strongest quarter of positive activations in InceptionV1’s final layer preserves 69.5 percent accuracy against 70.1 percent originally. In the medium ConvNeXt, removing the weakest 75 percent of activations across all 18 blocks lowers accuracy from about 72 to about 71 percent and changes about one answer in ten, while removing the strongest 7 percent lowers it to 4 percent and changes 96 answers in a hundred. The answers that change under weak removal are mostly near ties, such as one dog breed for a close relative, whereas the confident answers survive. The test shows limited dependence on weak activations under the conditions tested: this model, the ImageNet classification task, the validation distribution, and an intervention that removes activations by magnitude within layers. Other tasks, distributions, and interventions, such as removing weak activations selectively by context, may give different answers, and the result does not show that weak activations carry no semantic content. The cost of removal is also not uniform: the source reports that it is largest in the early layers of vision models and in the last layer of language models, where weak values function as graded dials or as suppression of single tokens. The correct reading is that the test constrains where hidden features could be, and for which tasks, and not that no such features exist [1, *Do Weak Activations Matter?; Weak Activations: Further Measurements*].

The third argument is by inspection, and it is where the evidence is weakest by the author’s own account. Traced weak activations of the Trump unit in a vision-language model, of an early curve detector, and of an enemy unit in a language model were found to be weaker versions of the same thing, radiating outward through related contexts. The enemy unit’s weaker positive responses broaden from explicit adversary words, including other languages, into nearby mentions, combat, games, and competition, and unrelated examples become common near zero. The source states plainly that it has no semantic test of weak activations that it trusts, that broader category-based screening gave mixed results, and that the spot checks cannot establish what weak activations mean in general. The honest summary is that weak activations are untested at scale for semantic content, that the ablations bound their causal importance in specific tasks, and that the claim of generalization in the tail rests on a small number of traced examples [1, *Weak Activations: Further Measurements*].

The proposal that follows is methodological. The author argues that weak activations are best read in context, by moving between data, circuits, and the unit, since an unsorted list of everything weakly associated with the Trump unit looks like noise, whereas starting from Putin-related data and noticing that it reaches the unit gives a meaningful intersection. This is a defensible division of labor between enumerative and exploratory reading, and it carries a cost that should be named. If weak activations are read only when a question motivates them, a surprising weak activation will be found only where someone has a question that would surface it. The argument therefore relies on the exploratory procedure being broad enough in practice, a property that depends on the investigators and on the number of automated explorers.

Raw activation shapes are a related consideration. The source argues against thinking only in percentiles, since a neuron’s natural distribution carries information: some units hold their meaning far down the range, others fall off a cliff, as the Christmas unit does near five percent of its maximum, and some fire a few times a book while others fire a few times a page [1, *Do Weak Activations Matter?*]. The point is well taken for the interpretation of units, and it gives

a reason to report both the raw ranges and the percentile summaries when comparing audits, since the proportion of units judged to have one main job depends on where weak activations are cut off.

7 Decomposition as another instrument

Sparse autoencoders sharpen the tension between an object’s native organization and the units used to inspect it. Such a model typically learns a sparse code from activations and reconstructs those activations through a learned dictionary [10, 7]. It is generally not an invertible change of basis: it can be overcomplete, nonlinear in its encoding, and imperfect in reconstruction. Its entries provide a learned analytical representation whose relation to the original computation must be assessed.

The distinction matters in the V-shaped example. A dictionary could separate car-related from cat-related activity and thereby improve discrimination between those contexts. That would be useful for some questions. It would not by itself show that the original neuron performed two unrelated computations, nor that the dictionary had recovered the network’s uniquely correct elementary units. The decomposition might divide a reusable distinction according to the contexts in which it appears, in the manner of reconstructing a district of one city from seventy fragments of another.

The source reports comparisons that bear on this. On the last layer of a convolutional model, 88 percent of a sparse autoencoder’s activity is near-copies of single neurons, the remainder is faint, and removing every dictionary direction that matches no neuron leaves accuracy unchanged, with the few answers that flip already being near ties. A search over rotations of a layer for the basis in which a handful of units carry the most activity, starting from a random rotation, returns to the neuron basis in every block tried [1, *What Dictionaries Learn*]. The features that the dictionary added beyond near-copies were loose groupings built from existing neurons, such as animals at the water’s surface, and a spider example showed dictionary units corresponding to spiders in different settings while the neurons corresponded to leg patterns and webs. On the early curve layer, features from an autoencoder trained for earlier work on missing curve detectors [9] were found to be 97 to 98 percent composed of the original curve neurons, and an autoencoder trained on curve activations alone used 84 thousand of its 100 thousand dictionary entries, which the source reads as a reason that many compositions can create an impression of more building blocks than exist [1, *Missing Curves*].

The comparison is not one-sided, and the source is careful about this. Dictionaries helped with the question of what is in a particular image, helped most in the smallest model whose neurons are coarse, were slightly preferred at medium size, and were less preferred at the largest size. This pattern is consistent with the view that the native units become more specific with scale and that a dictionary supplies a view better matched to coarse units, which is a result a proponent of dictionaries could accept. The claim that the dictionary found no fundamentally new building block is a statement about these layers and these recipes, reached by the author and his collaborators, and an independent replication with different dictionary sizes, sparsity settings, and objectives would test how much of it depends on the choices made.

A related study separates detection from use. In a probe for a three-unit Social Security detector in a language model, the three units accurately detect the phrase, with a held-out

AUC of 1.00, but removing all three changes the next-token log probability by only -0.004 nats on average, helping in 81 cases and hurting in 83, and tracing what the model actually uses to choose the next token revealed ordinary lexical, form-field, retirement, and output units [1, *Social Security: Detection and Use*]; the probe itself is reported in [8]. The case teaches that a direction can be decodable without being read, and that a decodable direction is evidence about information present, not about computation performed. It is an instructive example of the general relation between probes and mechanisms, and it applies equally to dictionary features, whose existence in a reconstruction does not show that the network uses them.

The reverse possibility must remain open. A learned dictionary might reveal consequential structure that direct inspection of neurons misses. Cammarata’s reported successes with native mechanisms motivate examining them seriously; they do not establish that learned decompositions are universally dispensable. The appropriate comparison concerns explanatory performance: which view predicts held-out behavior, supports effective interventions, preserves consequential distinctions, and makes its residual errors visible?

Terminology should not settle that comparison in advance. Representing more specified features than dimensions is a concrete question when an experiment supplies those features, as in the toy models [5]. Whether a real model groups otherwise unrelated computations to save capacity is a different question. The source observes that superposition has been used for vector addition, for a circuit motif, for the representation of more features than dimensions, and for the trade-off between composition and packing, and that the last is the sense its argument needs [1, *A Note on “Superposition”*], a framing it draws from [6]. On the count definition, even a fully monosemantic network satisfies the condition because its units compose, so the count cannot distinguish the worlds the essay cares about. Cammarata explicitly concentrates on the latter. His argument can therefore challenge an interpretation of apparent neuronal disorder without refuting every technical use of superposition.

8 Adequacy relative to intervention

The preceding comparisons raise a question that the source’s examples answer case by case without stating a general criterion: what must an account of a mechanism accomplish to count as adequate, given that the same unit can be described as a topic tracker, a token selector, a pointed-shape detector, or one member of a dictionary of loose groupings. This section proposes a bounded criterion and outlines tests of it. Neither the criterion nor the experimental designs below has been validated, and nothing in them is evidence for or against Cammarata’s hypothesis. They are offered as a way of making the choice among descriptions answerable.

An explanation can be treated as a claim about a distinction the model makes together with the later computations that depend on that distinction. “This unit detects Colorado” is an initial hypothesis of that kind. A more informative account would specify which inputs establish the regional context, where attention places that information, what distinction the unit draws once it receives it, and which downstream choices change when its activity changes. The need for this specification follows from the observation, made in the previous section for the Social Security units, that a unit can be highly recognizable without being causally important for the behavior under study, while a weakly recognizable unit can contribute decisively in a

particular context. Recognizability, causal role, and scope are three properties, and a familiar label should not be allowed to stand for all of them.

The criterion requires that adequacy belong to an explanation together with a declared prediction procedure. Otherwise an explanation could receive credit for predictions produced by a powerful auxiliary model that contributes most of the understanding, and the explanation would be a name attached to someone else’s work. Let E be the explanation, P_E the prediction procedure declared for it, $\mathcal{I}$ a family of interventions, Q a distribution over input contexts, $\mu(\cdot | x)$ a probability measure over $\mathcal{I}$ that may depend on the context, and Y_M the outcome measured in the original model M under context x and intervention i . The explanation is interventionally predictively adequate for $(\mathcal{I}, \mu, Q)$ at tolerance ε when

$$\mathbb{E}_{x \sim Q, i \sim \mu(\cdot | x)} \left[\ell(P_E(x, i), Y_M(x, i)) \right] \leq \varepsilon, \quad (1)$$

where the loss ℓ , the measure μ , and the tolerance ε are specified in advance and the expectation is estimated on held-out cases. The tolerance should be calibrated against something stated, such as the variance of the outcome under the baseline condition, the effect size of interest, or the error of a declared null predictor, because a tolerance that is too loose admits trivial explanations and one that is too tight admits none. The declaration must also state what information P_E may access, including the permitted auxiliary computation, any access to the model’s activations, and the experimental budget, because otherwise P_E could query the original model under the intervention and achieve zero error without explaining anything. This makes the concern about powerful auxiliary predictors enforceable and not merely a matter of stated intent. Predicted effects should also be evaluated alongside absolute outcomes. For a numerical observable, define $\Delta_M(x, i) = Y_M(x, i) - Y_M(x, i_0)$, where i_0 is the baseline condition, and compare it with the change that the explanation predicts. A predictor can attain low absolute error by reproducing baseline behavior while missing the comparatively small effect of the mechanism under investigation, and reporting both errors exposes that failure. The comparison still requires matched intervention controls, since accurately predicting the effect of an ablation does not by itself distinguish the claimed computation from generic disruption. Because an average can conceal rare consequential failures, the report should also give the conditional performance on the contexts where the unit is thought to matter most and the frequency of large errors. Adequacy is thereby relative to a family of interventions. A unit might admit a simple explanation that is adequate for topic tracking and require a far more detailed one for token selection, and the criterion permits this without demanding a single description that is equally complete for every purpose, while still imposing an experimental standard: the explanation must say what it predicts and where its simplifications cease to be reliable.

Interventional predictive adequacy is not offered as a sufficient definition of mechanistic understanding. A surrogate model can reproduce behavior accurately without identifying the computation responsible, and accurate prediction alone cannot separate the two. A mechanistic explanation should additionally tie its claims to identifiable model components and predict the effects of intervening on those components, which preserves the distinction drawn earlier between detection and use. The criterion is therefore an operational acceptance condition for explanations claiming predictive adequacy over the declared scope; component grounding and intervention tests impose additional requirements on their mechanistic interpretation. A partial explanation can identify a genuine mechanism while missing the numerical tolerance,

particularly when other mechanisms affect the measured outcome, so failure of the condition does not show that the explanation is false. The condition disciplines the choice among candidate descriptions, whether a native neuron, a dictionary feature, or a circuit, and it treats the appropriate unit of explanation as something the experiments determine and not something fixed in advance.

The substantive claim concerns refinement. A failed prediction creates an obligation to identify which omitted distinction matters, and the repair is not always a split of one unit into several features. It may instead introduce a context condition, separate the sign regimes of a gated unit, add an attention route that carries the relevant information, or enlarge the explanatory unit to include several interacting components. The experiment, and not the preference of the investigator, determines which repair earns support. To prevent endless accommodation, each revision should face fresh held-out tests, and earlier failures should be retained in the record so that an explanation accumulating exceptions is visible as such. For the pointed-shape unit, a suitable procedure would construct inputs that vary pointedness independently of object identity, with pointed and rounded car fronts, pointed and rounded animal faces, and unfamiliar objects of the same geometries, and then examine both activation and downstream effects. If the unit tracks pointedness across these changes, the descriptive claim gains support. If selectively suppressing it disrupts the model’s use of pointedness, the causal claim gains support. If its downstream role differs between animals and vehicles, the explanation must incorporate that contextual dependence.

A fully specified comparison shows how the criterion would be applied. Consider two accounts of a single unit that responds to cat faces and car fronts. The shape account S holds that the unit encodes a pointed, approximately V-shaped contour and that later computations use it as such, so that it should also contribute to responses to fox faces and flames. The category account C holds that the unit is a disjunction of two category detectors, whose downstream readers treat cat and car as separate. Each declares the same prediction procedure: a rule mapping a stimulus and an intervention to a predicted activation and a predicted change in a fixed set of downstream outputs, with no access to the model’s responses under the intervention. The stimulus sets are unfamiliar pointed objects such as wedge-shaped sculptures and folded forms, rounded versions of cat faces and car fronts that match the originals in texture, color, and viewpoint, and the original familiar images as a baseline. The interventions are ablation of the unit, ablation of an equal number of units matched to it in mean activation, ablation of a different shape-selective unit, and steering the unit upward on a rounded image. The outcome is reported both as an absolute output and as the change Δ_M from the unablated condition, and the predictions of the two accounts are as follows.

The value of the comparison lies in what each possible result would require. If the unit is active on unfamiliar pointed objects, weak on the rounded cats and cars, and its ablation changes pointedness-dependent outputs across all four categories in proportion to pointedness and more than the matched-unit ablation does, then S is supported at the descriptive and causal levels for the scope tested, and C is rejected for that scope. If the unit tracks pointedness in cars but not in animals, or if ablation affects car outputs but not cat outputs, then S survives only with a context condition that makes downstream use depend on the category of the object, and the refinement is an added condition and not a split of the unit. If the unit tracks pointedness on every stimulus but ablating it changes outputs no more than the matched control does, the result shows detection without use, and the repair is a broader

Condition	Shape account S	Category account C
Unfamiliar pointed objects	Unit active; activation rises with pointedness	Unit largely silent
Rounded cat and car	Unit weak or silent despite the familiar category	Unit remains active, since the category is intact
Ablating the unit	Downstream outputs that depend on pointedness change for cats, cars, foxes, and flames, in proportion to pointedness	Cat and car outputs change regardless of geometry; fox and flame outputs unchanged
Matched-unit ablation	Small Δ_M relative to the unit's ablation	Small Δ_M relative to the unit's ablation
Steering up on a rounded image	Outputs shift toward those that depend on pointed contours	Outputs shift toward cat or car identity, or do nothing

Table 1: Predictions of two accounts of a unit that responds to cat faces and car fronts. Both accounts declare the prediction procedure before the intervention data are collected.

circuit: the candidate set of units carrying the same distinction is ablated jointly, and an effect that appears only jointly warrants enlarging the explanatory unit to the circuit. If the unit is active on rounded cats and cars and silent on unfamiliar pointed objects, then C is supported and S is rejected as stated, although a third account, such as a texture or part-based feature common to cats and cars in the tested images, should be tried before concluding that the unit is a disjunction. If the unit responds to pointedness only when familiar textures are present, the pure shape account fails and the refinement is a conjunction of shape and a context. Each outcome follows from the declared predictions and the stated controls, and none depends on how the strongest examples look. The designs have not been run, and the particular stimuli, intervention sizes, and outputs would need to be fixed and checked for a given network before any such result could count as evidence.

Two applications extend the criterion to the developmental account and to the claim that understanding compounds. The first tests a proposed bridge. In the sequence from “iy” to “diy” to “do it yourself”, a controlled comparison could train small models under conditions in which the intermediate association is present, absent, or replaced, matching exposure to the endpoints as closely as possible, and compare the likelihood, timing, and route by which the later response develops. If the proposed bridge changes these, the historical explanation gains causal support, and if the response develops as readily by another route, the observed sequence was less decisive than it appeared. The design needs matched exposure and a check for alternative routes, because removing “diy” also alters the training distribution, and its relevance to larger models would remain to be established. The second tests whether a growing interpretability map improves prediction of mechanisms not yet explained. One team receives isolated evidence about target units, and another receives the same evidence plus an independently validated neighborhood of circuits. Both predict held-out activation patterns and intervention effects under comparable investigation budgets, and accuracy and cost are compared. The surrounding map must have been validated independently of the target predictions. If the map helps, that supports the claim that understanding compounds, and if it mostly raises confidence without improving predictions, part of the apparent accumulation may be accumulation of shared assumptions.

9 Which world the toy models describe

The best-studied model of packing is the toy setting in which many independent features are squeezed into few dimensions and copied back out. The source’s objection is not to the experiments but to what the setting prices. It counts the cost of sharing a unit, interference, and omits the benefit, reuse, because in the reconstruction task every feature is an end in itself and nothing downstream builds on what two features share. When correlated inputs and hidden units were placed in one model, related inputs grew onto shared directions, and increasing the density of inputs pushed the toy toward solutions that combine related inputs into shared representations. The result is a condition on the toy: with structure in the data and units to grow on, the toy’s behavior moves toward the generalizing regime [1, *Toy Models*].

The implication is a prediction about where packing should appear in real networks. It should appear for memories that are ends in themselves and are related to nothing else, and the toy world has exactly this property. In real data almost nothing is related to nothing, since a rare person still has a country, an occupation, and a name, and even a photograph with a nonsense label has familiar parts. What is left is input with no structure at all, such as noise with random labels, which networks can learn [13] but slowly. The question of which world real networks occupy is therefore a question about how much structure their data have relative to the capacity available, and it is amenable to experiment by varying both. The source reports one such experiment in the opposite direction: a packed code modeled on the originally proposed car and dog circuit was installed by hand in InceptionV1, and small mismatches added enough interference that the code often never worked, while those that did tended to be relied on less [1, *More on Circuit Ecology*]. The source adds that gradient descent is a better tuner than the authors, so the failure of a hand-built code is weak evidence, though it points the same way as the other results.

The consequence for rare facts is the most distinctive prediction. It is often assumed that the long tail of rare facts is where superposition is most likely, since rare things seldom collide. The source argues the opposite, that a rare fact barely moves the loss and cannot support a niche of its own, so it survives as an uncommon route between related circuits that already exist, as the umbrella’s rerouting to bee illustrates. The prediction is testable: the units and heads that carry a rare fact should be related to what the fact is about, and not unrelated units selected because they rarely collide. A direct test would choose rare facts in a model whose units are mapped, identify the units that carry them by ablation, and rate their relatedness to the fact’s subject against matched controls. Because related features can also be superposed, relatedness alone would not exclude a packed code, so the rating should be paired with readout and interference measurements on the same units.

10 Discriminating histories without checkpoints

The evidence is uneven across the systems in the essay. Retrained vision models permit inspection of learning trajectories, whereas the large-language-model audits primarily interpret finished mechanisms. Their reported prevalence of broadly legible neuron poles does not independently establish how those mechanisms developed. A strong present-state explanation and a strong developmental explanation are different achievements. In particular, the esti-

mate that roughly ninety percent have one main job should not be treated as a measurement of the proportion produced by recursive generalization. The source reports for Qwen3-32B roughly 90 percent of neuron poles with one main job, about 7 percent with two related jobs, and 2 to 4 percent not understood or arbitrary mixtures, and it states that the figures vary with the threshold for counting weak activations, with depth, and with what is considered one thing, that considering only strong activations gives closer to 97 percent, and that lighter audits find similar results in larger models including one with trillions of parameters [1, *The Short Version*]. It also reports that GPT-2 XL, for which fewer than two percent of neurons were clean in the original automated interpretability work, now has about 70 percent judged roughly one thing in 500 neuron papers written with almost no iteration, with about 5 percent pristine [1, *Slouching Towards Monosemanticity*]. These are judged shares, produced by the same group that formed the hypotheses, and the author is explicit that he does not trust any one cleanliness number.

The controls answer different questions too. For 200 randomly selected units, an agent that saw only a unit’s explanation wrote five new examples that should drive it and five near misses that should not. Among the 183 units whose explanations describe the positive side, the new examples fired the units strongly 76 percent of the time and the near misses 11 percent of the time, and the explanation won on 164 units, tied on 15, and lost on 4. Across the first twenty units the figures were 81 and 8 out of a hundred. Near misses test whether an explanation draws useful boundaries around a proposed response. As Cammarata acknowledges, a unit responding to either of two unrelated concepts could pass such tests when the controls contain neither concept. A second control therefore asked an agent to invent the most distant concept a network might pack in without interference and to write five texts for it. These fired the units strongly about one percent of the time, and 179 of the 183 units never fired on them [1, *Explanations That Could Lose*]. Inventing a distant concept gives little coverage of the enormous space in which an unknown second response could occur. The reported low firing rate therefore shows that those selected alternatives rarely activated the units; it does not estimate the probability that arbitrary packing is absent.

A further procedure, the blind maximum, has one agent write an explanation and a second, who sees only the explanation, write a text intended to activate the unit maximally. The resulting sentence often activates the unit more strongly than any example found in ten billion tokens of real text [1, *Explanations That Can Lose*], which supports the view that the explanation captures the center of what the unit looks for. The author notes the limit that a blind maximum tests the center and not the breadth, and the proposed extension clusters the attribution flowing into the unit, runs a blind maximum for each cluster while ablating the others, and compares each against the strongest corpus example under the same ablation. The design is a good one, and its power depends on the clustering of upstream attribution being a faithful partition of the unit’s inputs.

The source describes separate agents in explanation and example generation. Separation reduces direct leakage, but shared model assumptions or semantic vocabularies can still limit the alternatives explored. More discriminating audits would select units independently of their apparent interpretability, conceal proposed explanations from an exploratory team, and search for activation clusters across broad held-out data. Rival interpretations should make predictions about inputs, internal routes, downstream consequences, sparsity, interference, and geometry, with controls matched for activation magnitude and contextual frequency. Discov-

ering a second coherent response would motivate a mechanistic comparison rather than settle its developmental origin.

The same concern applies to the audits of attribution graphs. The source reports that agents reviewed 100,000 attribution graphs and that the author inspected thousands, that apparent anomalies usually turned out to be motifs he had not recognized, and that he found no substantial group of units whose joint activity remained inexplicable [1, *Audits of Attribution Graphs*]. The author calls this qualitative evidence and declines to convert it into a score. Its force depends on how anomalies were resolved. A motif discovered to explain an anomaly is a real finding when it then predicts other units and when it is tested on cases not used to find it, and it is a weaker finding when it was fitted to the anomaly alone. The record of which discovered motifs were later confirmed on held-out units, and which were abandoned, would be informative evidence about the procedure.

Without checkpoints, no final-state audit can uniquely recover every aspect of training history. It can test implications of a developmental hypothesis, such as whether apparent offshoots use related upstream routes or have consequences continuous with a main response. Yet similar final mechanisms can admit different histories. Stronger discrimination would combine such audits with prospective training experiments in accessible models, varying capacity, feature correlation, and reuse opportunities while recording branch acquisition. Any extrapolation from these experiments to a frontier model would remain an inference whose limits should be stated.

One of the source’s illustrations shows how an example can sit in the middle ground. A unit in a small one-layer language model, described in the paper that introduced dictionary learning for language models [7] as responding to a mixture of academic citations, English dialogue, web addresses, and Korean text, appears to be a unit that suppresses a period: nearly all the citation, dialogue, and address examples fire on or near a dot, eight of its ten strongest negative logit connections begin with one, and removing the unit raises the probability of a period in all twenty strongest contexts [1, *Slouching Towards Monosemanticity*]. The Korean examples fire near the end of an unfinished syllable, which the source reads as either a second weak pattern or a reasonable place not to want a dot. Because the original model is unavailable, the explanation cannot be narrowed further. The case is instructive because the correct report is neither mixture nor monosemantic: it is one dominant function with an unresolved remainder, and a framework that can express that status, with a note of what would settle it, is a better instrument than one that must classify the unit.

11 Confidence and the asymmetry of errors

The title’s confidence should be read at the right level. Cammarata’s approximate eighty-five-percent confidence concerns a proposed kind of generating process, not a measured fraction of clean mechanisms. He states that it is eighty-five percent confidence that nearly everything comes from the second kind of process, which is different from believing that eighty-five percent of mechanisms do, and he offers the analogy of moving from design to selection as a change of kind and not of degree [1, *On the Title*]. Neither number supplies the other’s evidence. The essay’s own title is described as imperfect in both directions: it overstates the evidence, since not every network has been audited, and understates the hypothesis, since the proposal is that

networks form an ecology of abstractions that work together and not merely that units match human labels.

His argument that mistaken assumptions have asymmetric research costs is similarly methodological. If a network is assumed to be clean and is not, a messy neuron is a lead that normal science may sharpen. If it is assumed to be packed and is not, a messy neuron is the end of the investigation, because the study stops. The asymmetry holds under stated conditions and not unconditionally: assuming packing need not stop an investigation, since a policy of testing packing explicitly can continue past a messy neuron, and assuming cleanliness can consume resources through repeated unsuccessful reinterpretation. The methodological advantage therefore depends on how each research policy actually allocates experiments and handles failure. Where it holds it is a reason to adopt the clean prior as a research policy, but it is not evidence for the hypothesis, and it has a cost that the author recognizes in naming the opposite error, an explanation of the gaps, in which every unit is assumed to have a sensible story. The source answers that each alternative cause can be tested, that superposition's only symptom is that something does not make sense, and that real explanations tend to be narrow and quick to test. The reply is reasonable, and it identifies the discipline that keeps the policy honest: explanations must be specific enough to lose, and the record of those that did lose should be kept visible.

The proposal to treat superposition as the last hypothesis and not the null has the same character. A null hypothesis in the statistical sense is the one that is rejected on evidence. The source's argument is that the radical hypothesis should require more evidence than ordinary generalization because it implies that the network's units are poor objects of study, and that its only prediction is mess. That is an argument about priors and about testability, and it is persuasive as long as the packing hypothesis is stated as a model with measurable predictions about sparsity, interference, geometry, readout, and acquisition dynamics, with semantic locality as only one of them, and those predictions are tested where they could be observed. The capacity-reduction and rare-fact experiments are such tests, and further tests of this kind carry the argument more than appeals to the history of the field do.

The picture of interpretability that results is one of cumulative constraint. Interpretability, in the author's image, resembles a crossword: understanding gained anywhere constrains the rest, an uncertain explanation is pencilled in and resolved by the crossings, and the unexplained remainder points at the broken tools. When the map is mostly empty or mostly full this is convenient, and in between it carries a hazard. A set of explanations that mutually constrain each other can be internally consistent without being correct, since each label is checked against labels formed under the same assumptions. The remedy is to ensure that some crossings are supplied from outside the labeling process: interventions, held-out data, and blinded predictions. The source's use of ablation, steering, and blind-maximum texts are examples, and the argument is strengthened in proportion to the share of the map that has been checked in such ways.

The contrast the source draws between enumerative and exploratory interpretability deserves a final comment. The enumerative approach seeks a list of features to audit in full, and its difficulty is the combinatorial growth of items connected to items. The exploratory approach starts from questions and moves among data, mechanisms, and scales, which makes it resilient to polysemanticity and willing to ignore most of a network, as a student of a city ignores most districts. The choice between them depends on the aims. A claim that no decep-

tive computation exists in a model cannot be established unconditionally by either approach, although statistical assurances can be meaningful when their scope and assumptions are stated, for example a sampling design over units, contexts, and elicitation conditions with an explicit bound on what it would miss. The enumerative approach aims at broader coverage and the exploratory approach at greater depth, and neither supplies a guarantee. The source addresses this by arguing that deception has no reason to hide from a model’s own circuitry and is more likely to sit in plain sight, unread by observers who lack the relevant vocabulary. That claim is a hypothesis about where failures lie, and it is consistent with the history reviewed here, in which every error was something visible that was read incorrectly. It is not an argument that no hidden activation could matter, and the source acknowledges that the perfecting of a single unit is a poor use of effort without claiming that exhaustive audit is unnecessary.

12 Optimism with its scope attached

Three things are easily conflated in discussions of this work, and the preceding sections have tried to keep them apart. The first is evidence that mechanisms are legible: that a unit labeled as a mixture turns out to draw a distinction, that a circuit read as hidden storage turns out to carry shared pose, that a layer judged uninterpretable becomes interpretable under a better instrument and a better architecture. The second is evidence about how mechanisms developed, which the vision experiments supply in part and the language-model audits do not. The third is a set of reasons to pursue a research program, such as the asymmetry between the costs of assuming legibility and of assuming its absence. Each supports a different conclusion, and the strength of one does not transfer to the others.

On the first, the evidence supports a substantial update, and in particular about the feasibility of useful mechanistic discoveries. If a supposedly arbitrary mixture of cats and cars tracks a shared shape, and the circuit connections support that reading, then what had counted as evidence of an intrinsically tangled representation becomes an example of an intelligible computation. The correction matters beyond the single unit because canonical examples help set researchers’ expectations about what they will meet elsewhere, and an expectation formed from a misread example can close investigations that would otherwise have been pursued. The most persuasive component is the combination of better instruments, attention to computational roles, and architectural comparison, since each supplies an ordinary explanation for why early interpretability looked discouraging: a unit may draw a distinction instead of detecting a noun, information may arrive on punctuation instead of on the words that supplied its content, a visualization may find an atypical route to an activation, and a poorly learning architecture may produce broader responses than a better one of comparable size. Together these make it plausible that several different difficulties were mistaken for one fundamental barrier.

The claim that understanding compounds is credible for a related reason. Once enough neighboring mechanisms are understood, an unexplained unit is constrained by its inputs, its readers, and the circumstances in which it matters, so the cost of each further explanation may fall, and a program in which every unit demands an independent investigation is the wrong model of the work. Automated experimentation could amplify that advantage on one condition: the agents must produce predictions and interventions that challenge their own explanations. A million plausible labels would be a much weaker result than a smaller collection of explanations that survive serious attempts at refutation. Mutual constraint among explanations is valuable

when some of its predictions receive independent checks, because otherwise a coherent set of labels could merely reinforce itself, and the audit designs discussed above are the means of supplying those checks.

The caution concerns the step from these successes to the claim that nearly everything arises through the proposed developmental process, together with two gaps that remain even in a well-organized network. Recognizing a unit’s main role does not establish that all of its consequential uses have been captured, because a rare context in which the unit matters a great deal can coexist with a clear dominant function. Understanding many local mechanisms does not automatically explain how their interactions produce a sustained behavior, since the composition of circuits into extended behavior is a separate research problem with its own failure modes. Neither gap is an objection to the hypothesis. Each marks a place where the present evidence is silent.

The experiments proposed in the section on adequacy relative to intervention are not evidence for or against the hypothesis; they are designs that would generate such evidence, and the criterion they rest on remains a proposal. The update that the evidence supports is accordingly asymmetric. It is a strong reason to invest in direct investigation of native mechanisms and to stop treating apparent polysemanticity as an automatic dead end. It is a much weaker reason to infer that comprehensive auditing, reliable behavioral prediction, or alignment through internal edits is correspondingly close, because those goals depend on exactly the interaction and coverage questions that the experiments reviewed here do not address. A model’s organization may be richer than the vocabulary used to describe it, and the instruments used to inspect it may make it look poorer, and both possibilities are live. Learning to distinguish them is concrete scientific progress whether or not the strongest version of the cleanliness hypothesis survives, and whether the laboratory can answer the hardest questions remains an experimental matter.

A Evidence ledger

The table lists the principal claims discussed above, the kind of evidence the source offers for each, what that evidence establishes on its own terms, and a test that would sharpen it. The entries paraphrase the source and the essay’s analysis, the reproduced figures are those reported by Cammarata, and the headings under which each appears in the source are given in the body text.

Claim	Evidence offered	What it establishes	Test that would sharpen it
The cat and car unit draws a pointed versus round distinction	Inputs, readers, negative pole, and new image groups all fit; earlier visualizations misled	One unit has a unified description that survives several predictions	Preregistered predictions on other canonical mixed units by raters who have not seen the answer

Claim	Evidence offered	What it establishes	Test that would sharpen it
The car and dog circuit carried shared pose, not hidden storage	Natural-path visualizations; more than a hundred pose units; same grouping in a larger model	The original reading was a visualization artifact in this circuit	Ablation of pose units shows whether car and dog readers use the shared signal, which would support reuse but not exclude an additional encoded signal; a readout test for car-specific and dog-specific information after the pose signal is removed would address that
Last-layer legibility is largely an architecture and tool effect	Estimated interpretable share rising from about 5 to about 99 percent across tools and architecture	Reported legibility improved under tool and architecture changes	Blinded rating of the original and retrained layers under fixed criteria
Squeezing a layer coarsens units and does not pack them	Layer 3b cut 48 times; final layer cut to 13 units; dictionary fails to recover original curves	No arbitrary mixtures were identified by the reported inspections and dictionary searches	Repeat on controlled synthetic features with varied correlations and sparsity; measure interference, geometry, and readout as well as relatedness; blind mixture ratings
Weak activations do not hold hidden features	Weakest 75 percent removed for about one point of accuracy; traced examples radiate outward	Limited dependence on weak activations for this model, task, distribution, and intervention; some tails are related	Other tasks, distributions, and interventions; a trusted semantic test of tails; screening across many units with controls
Sparse dictionaries add little beyond neurons on these layers	88 percent near-copies; unmatched directions removable; rotation search returns to neurons	No new building blocks were found for these recipes and layers	Replication across dictionary sizes, sparsity, and objectives, by independent groups
Rare facts reuse related circuits	Umbrella rerouted to bee in nine steps; noise labels used familiar textures	Memorization reused existing routes in these cases	Ablation-identified carriers of rare facts rated for relatedness against matched controls, with readout and interference measurements
About 90 percent of poles in Qwen3-32B have one main job	Agent audits with new examples, near misses, and distant-concept controls	Selected alternatives rarely activate the units; explanations make winning predictions	Units selected before inspection; blinded explanation teams; held-out clusters

Claim	Evidence offered	What it establishes	Test that would sharpen it
Branches arise by recursive generalization	Watched training of re-trained vision models; route grouping in a language unit	Development in observed cases proceeds by local extension	Prospective experiments recording branch acquisition while varying capacity and data structure
Networks are about 85 percent likely to be produced by one learning principle	Stated confidence in a generating process	A summary of the author’s belief, not a measurement	Report of which specific predictions would lower the figure

B A minimal protocol for developmental claims

The following outline restates the training-record and audit recommendations above in a form that could be run on a model whose checkpoints are available. It is offered as a concrete specification for the evidence the essay argues is missing, and the particular settings are choices to be made by whoever runs it.

1. Fix a set of reference inputs and a procedure for selecting units before any checkpoint is inspected, drawing units without regard to apparent interpretability.
2. At each saved checkpoint, record each selected unit’s activation distribution on the reference inputs, its strongest incoming and outgoing contributions, and its strongest natural examples chosen by a fixed rule.
3. Define a dissimilarity on responses independent of the labeling team, for example by shared upstream routes or by the effect of ablation on downstream units, and record it for every pair of responses a unit acquires.
4. Specify, before inspecting later checkpoints, the competing models and the quantities on which they differ: sparsity and co-occurrence of responses, interference in the readout, the geometry of the representation including whether distinct feature directions can be recovered from the layer’s population activity, including their projections onto the inspected units, the dependence of downstream units on each readable component, and the order in which components are acquired. State each model’s prediction for each quantity. The semantic distance of newly acquired responses under the fixed dissimilarity is recorded as one measurement among these and is not treated as decisive, since unrelated features can share coordinates while remaining separately recoverable from the population activity and related features can also be superposed.
5. Vary capacity of one layer and the structure of the training data, holding other factors fixed, and record the same quantities.
6. Report which predictions failed, including units for which neither hypothesis fit, with the checkpoints at which each response appeared.

References

- [1] Nick Cammarata. *Neural Networks Are Already Clean*. Goodfire, 2026. <https://already-clean.goodfire.com/>. A working hypothesis. Read on the live page in October 2026. Its experiments have not been independently reproduced here, and citations give the heading under which each passage appears.
- [2] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature Visualization. *Distill*, 2017. Known through the source’s discussion; not consulted directly.
- [3] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom In: An Introduction to Circuits. *Distill*, 2020. Known through the source’s discussion; not consulted directly.
- [4] Gabriel Goh, Nick Cammarata, Chelsea Voss, et al. Multimodal Neurons in Artificial Neural Networks. *Distill*, 2021. Known through the source’s discussion; not consulted directly.
- [5] Nelson Elhage, Tristan Hume, Catherine Olsson, et al. Toy Models of Superposition. *Transformer Circuits Thread*, 2022. Known through the source’s discussion; not consulted directly.
- [6] Chris Olah. Distributed Representations: Composition and Superposition. *Transformer Circuits Thread*, 2023. Known through the source’s discussion; not consulted directly.
- [7] Trenton Bricken, Adly Templeton, Joshua Batson, et al. Towards Monosemanticity: Decomposing Language Models With Dictionary Learning. *Transformer Circuits Thread*, 2023. Known through the source’s discussion; not consulted directly.
- [8] Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. Finding Neurons in a Haystack: Case Studies with Sparse Probing. arXiv:2305.01610, 2023. Known through the source’s discussion; not consulted directly.
- [9] Liv Gorton. The Missing Curve Detectors of InceptionV1: Applying Sparse Autoencoders to InceptionV1 Early Vision. arXiv:2406.03662, 2024. Known through the source’s discussion; not consulted directly.
- [10] Leo Gao, Tom Dupré la Tour, Henk Tillman, et al. Scaling and Evaluating Sparse Autoencoders. OpenAI, 2024. Known through the source’s discussion; not consulted directly.
- [11] Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress Measures for Grokking via Mechanistic Interpretability. ICLR, 2023. Known through the source’s discussion; not consulted directly.
- [12] Mohammad Pezeshki, Sékou-Oumar Kaba, Yoshua Bengio, Aaron Courville, Doina Precup, and Guillaume Lajoie. Gradient Starvation: A Learning Proclivity in Neural Networks. NeurIPS, 2021. Known through the source’s discussion; not consulted directly.
- [13] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding Deep Learning Requires Rethinking Generalization. ICLR, 2017. Known through the source’s discussion; not consulted directly.