

Neuron Interpretability, Polysemanticity, and Superposition

A mathematical reading of Goodfire’s ‘Neural Networks Are Already Clean’

2026-10-15

1 What is the contribution?

Nick Cammarata’s *Neural Networks Are Already Clean* is a research essay presenting experiments, neuron studies, and a working hypothesis. The version inspected on 15 October 2026 describes a technical language-model companion paper as forthcoming.

Its strongest contribution is **an empirical reassessment of how much can be understood directly from a network’s existing neurons**. Several apparently incoherent activation patterns become intelligible after changing the analysis: inspect both signs, condition on the activation’s computational regime, look beyond the highest-activating examples, and test explanations on new inputs. Some resulting explanations also predict the effects of intervening on neurons.

The advance has three components:

1. **Corrections to influential examples.** A neuron associated with cats and cars is reinterpreted as detecting a shared shape. Other examples change interpretation after examining the circuit that produces the activation.
2. **A more demanding measurement procedure.** The essay combines activation inspection with generated positive and near-miss examples, training trajectories, and interventions on native units. It documents specific failures of common visualization and labeling procedures.
3. **A developmental hypothesis.** “Recursive generalization” proposes that neurons acquire related uses successively during training. The accumulated response set can look heterogeneous when described using an unsuitable taxonomy.

The practical conclusion is that **native-neuron analysis deserves a strong baseline in interpretability research**. The evidence for a general claim that large networks have little superposition is substantially weaker. That claim depends on the definition of a feature, the definition of superposition, and how the inspected cases represent the rest of the network.

The novelty is chiefly in the reanalyses and empirical demonstrations. The underlying ideas of sparse coding, conjoined features, and causal ablation have substantial prior work.

The equations below formalize these distinctions. The operational definitions and formalization of recursive generalization are introduced for this analysis. Reported neural-network results are attributed to the essay; the small coding example in Section 4 is an independent mathematical illustration.

2 Define the object being interpreted

Fix a trained network F_θ and an input distribution $X \sim P$. Write $r_\ell(X) \in \mathbb{R}^d$ for a layer representation, and $h_j(X)$ for a specified scalar hidden-unit activation. For a language model, the input includes the context and the token position at which the unit is measured.

Three objects need separate definitions:

- A **feature** is an input property $z_k(X)$, such as curvature, a grammatical relation, or the presence of a particular entity. A feature can be continuous.
- A **neuron** is a scalar variable in the implemented computation. Its relationship to features is an empirical question.
- A **circuit** is a selected set of variables and computational dependencies proposed to account for an output behavior. Describing its units individually supplies only part of a circuit explanation.

The hidden coordinates of a coordinatewise nonlinear MLP have an architectural role that arbitrary coordinates of an embedding need not have. Rotating hidden activations generally changes their relationship to the nonlinearities. Consequently, a useful native-neuron interpretation can coexist with a different, distributed description of the residual representation.

2.1 Polysemanticity needs an ontology and a distribution

Informally, a polysemantic neuron responds to several unrelated properties. The word “unrelated” carries the main difficulty.

Here is one explicit operational definition. Fix a partition of test inputs into semantic categories $C_1, \dots, C_K$, and an independently specified semantic distance $d_{\mathcal{C}}(k, m)$. For a positive activation threshold τ , define

$$A_j^+(\tau) = \{x : h_j(x) > \tau\}, \quad A_j^-(\tau) = \{x : h_j(x) < -\tau\}.$$

Analyze either event A_j separately. Assuming $P(A_j) > 0$ and $P(C_k) > 0$, define activation lift and the composition of the activation set:

$$L_{jk} = \frac{P(A_j | X \in C_k)}{P(A_j)}, \quad p_{jk} = P(X \in C_k | A_j).$$

For declared constants $\lambda > 1$, $\epsilon > 0$, and $\delta > 0$, call the unit **polysemantic relative to this measurement specification** if there exist categories $k \neq m$ such that

$$\min(L_{jk}, L_{jm}) \geq \lambda, \quad \min(p_{jk}, p_{jm}) \geq \epsilon, \quad d_{\mathcal{C}}(k, m) > \delta.$$

The lift condition rules out a unit that fires indiscriminately. The mass condition requires both categories to account for a non-negligible share of activations. The distance condition encodes the chosen meaning of “unrelated.” The definition applies separately to each activation sign and, where relevant, each gating regime.

For a single prespecified binary property C , a simple monosemanticity criterion would instead be $P(A_j \triangle C) \leq \epsilon$, with precision and recall reported separately when C is rare. Continuous features require an analogous predictive-error criterion.

None of these is an ontology-free property of a scalar function. Under an object-label taxonomy, cats and cars are distant. Under a shape description, both can contain a V-shaped contour. Cammarata’s reanalysis of InceptionV1 unit 4e:55 proposes the latter property and tests additional objects with that shape (source).

A shared-property model can be written

$$h_j(X) = f_j(s(X)) + u_j(X), \quad \mathbb{E}[u_j(X)^2] \text{ small},$$

where s is a specified, low-complexity property and f_j belongs to a restricted function class. New objects sharing s should activate the neuron; objects preserving the category label while removing s should respond less. Allowing s to be any post hoc union of activating examples makes the claim vacuous. Its content comes from complexity restrictions and successful held-out predictions, preferably including controlled changes to s .

3 Superposition is a different claim

The standard linear-feature formulation is

$$r(X) = Dz(X) + e(X), \quad D = [d_1, \dots, d_K] \in \mathbb{R}^{d \times K},$$

where $z(X)$ contains meaningful latent feature intensities, D contains their encoding directions, and e is approximation error. The feature system must have substantive restrictions: arbitrary relabelings of the same property cannot be counted as additional independently meaningful features.

In *Toy Models of Superposition*, superposition concerns representing features whose encoding directions are linearly dependent: $\text{rank}(D) < K$, in particular when $K > d$. The toy experiments specify the features through the data-generating process, which makes this claim measurable.

For unit-norm columns, a matched-filter readout gives

$$D^\top r = z + (D^\top D - I)z + D^\top e.$$

The second term is interference. Sparse z , suitable encoding directions, and nonlinear decoding can make interference tolerable. This connects the problem to sparse coding and compressed sensing. Merely distributing one feature across several coordinates establishes neither linear dependence nor overcompleteness.

Cammarata uses a narrower, causal description: neurons acquire unrelated uses because those uses rarely interfere. His terminology appendix explicitly distinguishes this from counting more features than dimensions. That changes the hypothesis being assessed. It adds a claim about *why the representation developed* to a claim about its final structure.

The distinctions matter:

- A coherent description of a neuron does not determine the number of latent factors represented by the network.
- A representation with $K > d$ does not identify the training mechanism that produced it.
- A neuron can generalize across related inputs while also participating in an overcomplete code for other features.

Composition and superposition can coexist. The earlier Olah (2023) note already discusses this coexistence and the constraints imposed by how often features occur together. Thus some disagreement here concerns empirical prevalence and explanatory emphasis; some concerns terminology.

3.1 The econometric identification issue

From observations of $r(X)$ alone, the factorization $Dz(X)$ generally lacks uniqueness. For example, for any invertible $A \in \mathbb{R}^{K \times K}$,

$$Dz(X) = (DA)(A^{-1}z(X)).$$

Restrictions on sparsity, support, dependence, dictionaries, and measurement functions can improve identification. Assigning natural-language labels to one fitted decomposition supplies additional hypotheses to test. This equivalence concerns descriptive factorizations; the actual network's nonlinearities and weights still constrain interventions on its native units.

The developmental claim is still less identified. Several training histories can produce similar final response functions. Discriminating explanations requires training-time observations or interventions on the learning problem. This is the familiar distinction between fitting a latent-variable measurement model and identifying the mechanism that generated it.

4 An exact example: three features in two coordinates

Take three independently meaningful binary features z_k and encode them using three unit vectors separated by 120° :

$$D = \begin{pmatrix} 1 & -1/2 & -1/2 \\ 0 & \sqrt{3}/2 & -\sqrt{3}/2 \end{pmatrix}, \quad r = Dz, \quad \hat{z} = \text{ReLU}(D^\top r).$$

If at most one feature is present, $\hat{z} = z$ exactly. Each feature has an explicit direction and an exact detector on that restricted support, although three directions occupy two coordinates. If two features are present, this particular decoder returns $1/2$ for each. If all three are present, $r = 0$: that state is indistinguishable from the state with no features.

For independent $z_k \sim \text{Bernoulli}(p)$, the decoder's mean squared error per feature is

$$R(p) = \frac{1}{3} \mathbb{E} \sum_{k=1}^3 (\hat{z}_k - z_k)^2 = \frac{1}{3} \left[3p^2(1-p)\frac{1}{2} + 3p^3 \right] = \frac{1}{2}p^2(1+p).$$

For comparison, assigning the two coordinates to two features and always predicting zero for the third gives risk $p/3$. This is a specified dedicated-coordinate baseline; optimizing all possible encoders and decoders would be a separate problem. The construction makes the sparsity advantage precise without a training experiment.

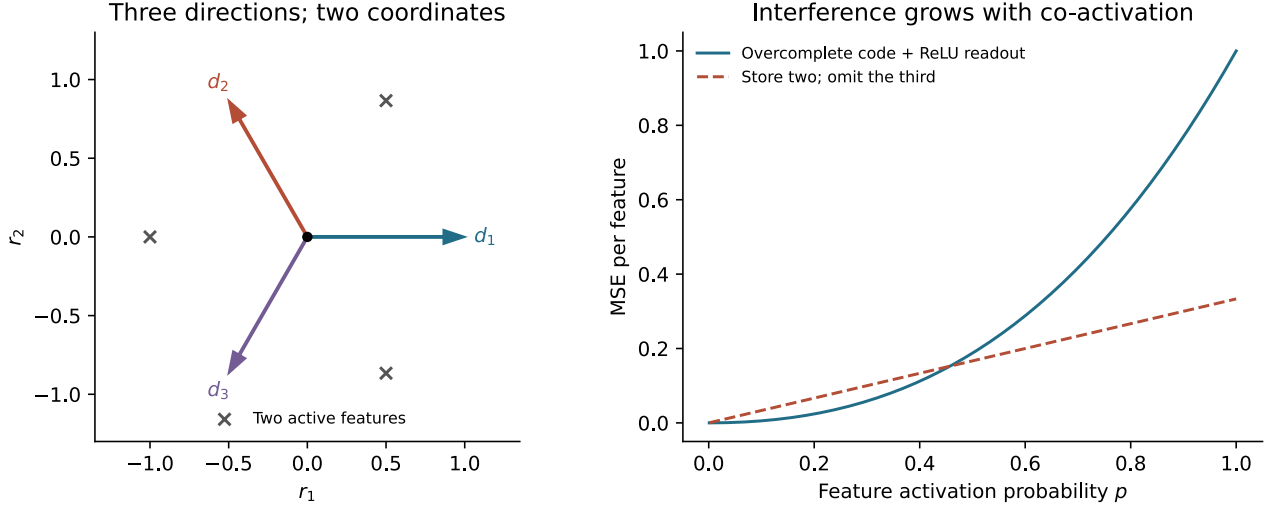


Figure 1: An explicit overcomplete code. Left: feature directions and the encoded states with two active features. Right: exact reconstruction risk, summing over all eight binary states. This is an independently constructed example.

Readable feature directions and sparse-data recovery therefore coexist with superposition in a fully specified example. The phenomenon’s prevalence in trained language models remains an empirical question.

5 “Recursive generalization” as a testable hypothesis

Let θ_t be a training checkpoint and $S_{j,t}(\tau) = \{x : h_j(x; \theta_t) > \tau\}$ a response set. The essay proposes that a neuron’s response set often develops by extending existing uses to related cases. Iterated extensions can connect semantically distant endpoints. An example is expansion from a spelling fragment to “DIY” and then to “do it yourself.”

A possible formalization requires a fixed external similarity metric d_κ on inputs and a declared threshold convention across checkpoints. Let $B_{j,t} = S_{j,t+1} \setminus S_{j,t}$ be the newly acquired response set. A local-expansion hypothesis predicts, when $P(B_{j,t}) > 0$,

$$P(d_\kappa(X, S_{j,t}) > \eta | X \in B_{j,t}) \text{ is small .}$$

This permits previously active inputs to drop out. It also requires tracking the same computational unit and accounting for activation rescaling. Choosing the metric after seeing the trajectory can make the hypothesis unfalsifiable.

The essay supplies suggestive trajectories and related examples. A sharper mechanism test would remove intermediate “bridge” examples during training, vary co-occurrence while holding marginal frequencies fixed, and vary width at matched task quality. These manipulations yield predictions about what gets grouped together under generalization and capacity-sharing accounts. Both processes could respond to the same manipulation, so the predictions must be specified before fitting.

The idea has an acknowledged antecedent in the “conjoined features” discussion of Goh et al. (2021). Cammarata’s extension is its proposed importance across models and its use as a developmental explanation. A quantified model of representation learning would require a specified similarity structure, an update mechanism, and out-of-sample predictions for training trajectories.

6 What the experiments actually establish

6.1 Predicting activations from explanations

The most informative explanation test is in “Explanations That Could Lose”. An agent saw explanations of 200 random neurons from the gallery and generated five positive examples and five near misses for each, before activations were measured. The reported summary covers the **183 explanations describing the positive side**:

- **76%** of positive examples exceeded the strong-response threshold.

- **11%** of near misses exceeded it.
- Strong meant at least **30% of that neuron’s strongest corpus activation**.

These are estimates of

$$P_{Q_+}(A_j), \quad P_{Q_-}(A_j),$$

averaged over the selected neurons, where Q_+ and Q_- are the agent’s explanation-conditioned example distributions. The contrast is useful evidence that explanations predict responses. It is not an estimate of precision or recall under the model’s ordinary input distribution P . Generated examples are clustered within neurons, which matters for any uncertainty calculation.

The same procedure generated examples for a deliberately distant concept: about 1% activated strongly, with no such activation for 179 of 183 neurons. Five sampled examples of one distant concept per neuron leave most possible alternative meanings untested. The experiment establishes specificity against those controls; semantic exhaustiveness would require substantially broader coverage.

6.2 The principal quantitative claims

Experiment	Reported result	Scope of the inference
Weak-activation deletion, medium ConvNeXt, all 18 blocks	Removing the weakest 75% changes top-1 accuracy from about 72% to 71%; roughly 10% of predictions change. Removing the strongest 7%, with matched total activation magnitude, reduces accuracy to about 4%.	Strong activations dominate this aggregate task metric. Residual bypasses remain intact.
Retrained InceptionV1 bottleneck	Reducing mixed3b from 480 to 10 spatial channels, with no bypass, gives 58.35% top-1 accuracy on 50,000 ImageNet validation images. Units appear coarser.	Useful coarse representations survive a severe channel bottleneck. Spatial locations remain available.
SAE on that squeezed representation	A 640-feature TopK SAE, with 2 active features per position, explains 82% of held-out activation variance. Substitution reduces top-1 accuracy from 58.8% to 43.0% on a separate 10,000-image evaluation.	Reconstruction variance explained is insufficient to guarantee preservation of the model’s function.
Qwen3-32B “Santa” intervention	Removing 8 identified negative-pole neurons gives reported $P(\text{Santa}) = 0.059$; removing the same number of neurons from the same layers at random gives 0.997.	The selected neurons have a large causal effect on this prompt’s output under this ablation.

Sources: weak activations, measurement controls, bottleneck and SAE, language-model circuits.

Three qualifications affect interpretation of these numbers.

Aggregate accuracy can conceal important changes. Let $c(X)$ and $\tilde{c}(X)$ be predictions before and after an intervention. Then

$$\Delta\text{Acc} = P(\tilde{c}(X) = Y, c(X) \neq Y) - P(c(X) = Y, \tilde{c}(X) \neq Y).$$

Improvements and deteriorations can offset. A one-point accuracy change, alongside 10% prediction disagreement, provides limited information about rare behaviors, subgroups, probability calibration, or robustness.

Ten spatial channels still yield many scalar variables. A tensor with shape $10 \times H \times W$ has $10HW$ entries. Retraining can also discard fine distinctions. The bottleneck experiment supports a coarsening account for the cases studied; it does not identify the total number of represented features. Failure of a particular SAE to recover a feature provides limited evidence of absence when substitution substantially degrades task performance.

“Cleanliness” percentages depend on the rubric. The essay reports roughly 90% of inspected language neurons as having one main job and describes near-complete readability in some modern vision models. The denominator, activation sign, corpus, explanation granularity, and judgment procedure determine what those percentages mean. Architecture, training, and analysis tools also change together in several comparisons. These are useful exploratory assessments, with limited identification of which change caused the improvement.

7 A concrete diagnostic: SwiGLU has two positive regimes

One practically useful observation follows directly from the activation algebra. For a SwiGLU unit, write

$$g_j = w_{g,j}^\top r + b_{g,j}, \quad v_j = w_{v,j}^\top r + b_{v,j}, \quad h_j = \text{SiLU}(g_j)v_j = g_j\sigma(g_j)v_j.$$

Bias terms may be absent in a particular architecture. Since $\sigma(g_j) > 0$, a positive product has two possible regimes:

$$\{h_j > 0\} = \{g_j > 0, v_j > 0\} \cup \{g_j < 0, v_j < 0\}.$$

A positive-activation ranking mixes them. An explanation fitted to the double-negative branch can misdescribe the unit’s large responses in the other branch or on its negative side. Cammarata documents examples where this changes an automatic label substantially, including a unit initially labeled by biblical references whose stronger negative responses concern Russian text (source).

The diagnostic is to examine response distributions conditional on $(\text{sign } g_j, \text{sign } v_j)$, along with their magnitudes and downstream effects. The algebra is elementary; its value here lies in identifying a measurement failure and demonstrating consequential examples. The sign of h_j alone also does not determine whether it increases an output: downstream weights and nonlinear computations determine that effect.

8 From a readable unit to an identified mechanism

8.1 Predictability, causal effect, and explanatory coverage

A probe showing that $z_k(X)$ is predictable from $h(X)$ establishes available information. To assess use, define an output functional T , such as a logit contrast, and an intervention replacing a set of activations h_S by a specified baseline b_S . Its average effect is

$$\tau_S(P; b) = \mathbb{E}_{X \sim P} \left[T(F_\theta(X)) - T\left(F_\theta^{\text{do } (h_S \leftarrow b_S)}(X)\right) \right].$$

This estimand depends on the input distribution, selected variables, baseline, and output. A large effect supports a functional role under that intervention. It leaves open whether the proposed semantic description covers all uses of the unit. A small effect can reflect redundancy or compensation. Setting a unit to zero may produce activation combinations rarely seen during ordinary operation; matched-input activation replacement offers another estimand.

The essay itself includes a useful distinction: three social-security-number detectors are predictive, while their ablation produces only a small average next-target log-probability change (about -0.004 nats in the reported test). See the appendix. This is evidence that detection and an effect on the tested output can differ.

Likewise, a selected circuit can explain a particular response while leaving other pathways unaccounted for. The Santa intervention is stronger evidence than a plausible neuron label. Its one-prompt effect still requires replication over a defined prompt distribution before supporting a population claim.

8.2 What native circuits avoid, and what they still require

A sparse autoencoder approximates activations by a learned representation:

$$\hat{r} = D_{\text{SAE}} f_{\phi}(r), \quad \min_{D_{\text{SAE}}, \phi} \mathbb{E} \sum_{i=1}^d (r_i - \hat{r}_i)^2 + \lambda \mathbb{E} \sum_k |f_{\phi,k}(r)|,$$

with hard TopK sparsity as another formulation. Dictionary learning can expose useful features, as in Bricken et al. (2023). Its features depend on the reconstruction objective, dictionary size, and sparsity constraint. They require independent validation.

Studying native units avoids inserting this learned approximation into the network. A circuit assembled from native units still needs tests of omitted paths, attribution approximations, and its semantic explanations. In either approach, **functional fidelity** can be measured on held-out inputs by

$$\mathcal{F} = \mathbb{E}_{X \sim P} \text{KL} (p_{\theta}(\cdot | X) \| p_{\text{approx}}(\cdot | X)),$$

supplemented by agreement on interventions of interest. The reported SAE substitution loss illustrates why low activation reconstruction error alone is inadequate. Native-versus-transcoder examples drawn from different language models also leave the comparative method effect unidentified.

9 Assessment and the next decisive tests

The work provides useful evidence for **better native-neuron measurement and case-specific causal explanations**. It also weakens some prominent examples previously used to motivate a particular account of polysemanticity. Recursive generalization is a plausible developmental hypothesis with a clear need for stronger operationalization.

The broad conclusions require three further kinds of evidence:

1. **A defined prevalence estimand.** Sample layers, neurons, signs, and inputs using a published rule; fix the semantic rubric; report failures and uncertainty. This would turn “most neurons are interpretable” into a reproducible statistical statement.
2. **Matched-model comparisons.** Compare native units, learned dictionaries, and circuit approximations on the same network, input distribution, functional-fidelity targets, intervention tasks, and analysis budget.
3. **Training-time discrimination.** Prespecify how width, feature co-occurrence, and bridge-example removal should change response sets under competing developmental explanations, then test those predictions across training seeds.

For an ML or econometrics reader, the main lesson is a measurement one: **semantic coherence, representation rank, and causal explanatory coverage are distinct estimands**. Progress on the first two can make circuit analysis easier. Claims about comprehensive model understanding or safety require evidence about the third, including rare and shifted input distributions.

References and scope

- Cammarata, N. (2026). *Neural Networks Are Already Clean*. Primary source, inspected 15 October 2026. Section-specific links above identify the reported experiments and the author’s qualifications.
- Elhage, N., et al. (2022). *Toy Models of Superposition*. Controlled examples with specified latent features and sparse inputs.
- Olah, C. (2023). *Distributed Representations: Composition & Superposition*. Distinguishes and combines the two coding strategies.
- Goh, G., et al. (2021). *Multimodal Neurons in Artificial Neural Networks*. Earlier observations of multimodal and conjoined features.
- Bricken, T., et al. (2023). *Towards Monosemanticity: Decomposing Language Models With Dictionary Learning*. Sparse dictionary learning for language-model representations.

The neural-network measurements above are source-reported results. Only the eight-state mathematical illustration is computed here; all of its code is contained in this document.