

Where Does the Function Go?

A pre-results research proposal on constrained adaptation, functional reconstitution, and whether internal model states predict recoverability before commitment.

DRAFT PREVIEW DENNIS HEDEGREEN ANALYSE DRAFT-LOCAL

[HTTPS://HEDEGREENRESEARCH.COM/ARTICLES/WHERE-DOES-THE-FUNCTION-GO/](https://hedegreenresearch.com/articles/where-does-the-function-go/)

Article 199 ended with a measurement rather than a milestone. [1] Hedegreen Research has learned how to make work visible. It has not yet learned how to make that work sufficiently conversational. Compute is also a constraint.

This is article 200. So instead of celebrating the number, I want to put one actual research problem on the table.

Most of my public writing about artificial intelligence has looked outward: labour, infrastructure, institutions, markets, successor systems, governance, and the conditions around the models. There is another line of work that has barely been visible. I have also been trying to look inside them.

Activations. Feed-forward neurons. Forward hooks. Residual-stream representations. What changes when one part of a model is suppressed. What remains. What returns. The question tying that work together is narrower than "how do language models think?":

When a behaviour can be located somewhere inside a language model, what exactly have we located — and where does the function go when that location is disrupted?

This article does not report a completed experiment. It exposes the research programme before the expensive part is run. That is deliberate. I want the causal target attacked before I spend the compute.

A signal is not a mechanism

Over the last few weeks I have been writing about a related problem in journalism. A converted monetary value should not erase the original monetary fact. [2] A percentage should not lose its reference frame. [3] An unanswered question should not disappear because an article reached its ending. [4]

The general rule underneath those examples is becoming clearer to me:

A representation should not eat the state that produced it.

Mechanistic interpretability has a version of the same problem.

Suppose I find a neuron whose activity strongly predicts a behaviour. I have found something real. But I have not yet established what kind of thing I found.

It could be part of the mechanism. It could be a readout of computation performed elsewhere. It could be one local implementation of a distributed function. It could be a bottleneck through which several mechanisms pass. It could be a correlated passenger.

Decoding alone cannot decide between those possibilities.

Here, decoding means using a model's internal activity to predict something about its behaviour. A probe is a separate statistical readout trained to make that prediction. Its success shows that useful information is available at the measured location; it does not by itself show that the model uses that information to produce the behaviour. That is why prediction, intervention and mechanism have to remain separate questions. [5] [6]

Intervention helps, but intervention is not automatically localization either. Hase et al. showed that causal localization did not reliably identify the best place to edit factual associations. [5] Wang and Veitch later gave an even stronger warning: targeted behavioural edits can be optimized at locations that do not support the corresponding localization claim. [6]

So the standard cannot be:

I found it, I edited it, therefore it lived there.

The interesting question is what survives when detection, causal contribution, redundancy and adaptation are separated.

H-Neurons made the problem concrete

A feed-forward neuron is one of the intermediate computational units inside a transformer's feed-forward block, not a biological cell. The H-Neuron approach asks whether a small selection of these units carries a signal associated with hallucination, and then tests what happens when their activity is changed. The label identifies a measured association and an intervention target; it is not a claim that a false statement is stored inside one unit. [7]

Gao et al. reported a sparse set of feed-forward neurons associated with hallucination in large language models. Fewer than 0.1% of the relevant neurons could predict hallucination in their tested settings. Controlled interventions also linked these H-Neurons to forms of **over-compliance** — behaviour where a model continues toward a plausible answer despite an invalid premise, misleading context, pressure to agree, or another reason not to proceed normally. [7]

The signal could also be traced into pretrained base models, suggesting that it was not simply introduced by assistant fine-tuning. [7]

The phrase **hallucination neurons** is useful shorthand. It is also easy to over-read. The interesting possibility is not that one neuron contains a false answer.

It is that some of these neurons participate in a more general transition:

uncertain or conflicting internal state → commitment to a continuation

Earlier this year I built an unpublished proposal around that possibility: *Compliance Is All You Learned*. [8] The first design was simple.

Locate H-Neurons in Llama-3.1-8B-Instruct. Suppress them. Re-run the detector. If another population appears, suppress that population. Repeat. The design had useful controls. It also had one decisive flaw.

A fixed model cannot reconstitute through learning

If the weights do not change, the model has not adapted. A new probe hit after an inference-time intervention might reveal incomplete suppression, an already-existing alternative route, a correlated representation, or a different readout. It cannot show that optimization rebuilt the lost function.

So the experiment has to change. That correction is part of the research.

Experiment 1 — remove the route and let the model adapt

The stronger design is:

locate → constrain → continue optimization → locate again

The first model can remain Llama-3.1-8B-Instruct because it gives direct continuity with the published H-Neuron work. [7] But there is a gate before any expensive training.

Gate 0 — show selective causal contribution

Before asking whether a mechanism can return, establish that the original target mattered. The first stage should test four things:

1. Can the candidate H-Neurons be identified reproducibly?
2. Do their signals predict the target behaviour on held-out examples?
3. Does suppressing them change that behaviour?
4. Is the effect stronger or more selective than matched interventions elsewhere?

Controls should include random neurons, layer-matched non-H-neurons, activation- or contribution-matched non-H-neurons, sham hooks, and general language-quality measures.

If the target does not survive this gate, there is no reason to run a reconstitution experiment.

Gate 1 — make the original route persistently unavailable

For an MLP target, one candidate implementation is a fixed binary mask over the selected intermediate dimensions, applied on every forward pass during adaptation. The original dimensions remain unavailable. The rest of the model is allowed to change. Then continue optimization.

Two adaptation regimes seem useful.

Generic continuation asks whether ordinary next-token training is enough to recover the lost behaviour.

Targeted adaptation uses tasks that expose the original phenotype and asks whether task-specific pressure accelerates recovery.

A cheap parameter-efficient run may be useful for debugging or screening. But a strong claim about network reorganization eventually requires sufficiently broad parameter adaptation. Otherwise the experiment has already restricted where compensation is allowed to happen.

Redundancy is not reconstitution

The distinction is between finding another route and watching the system change which route performs the work. An alternative that already contributes before training supports a different interpretation from one whose causal contribution grows as the constrained model adapts. The time-zero measurements are therefore part of the causal comparison, not merely a baseline score.

There is a harder confound. Large networks may already contain redundant routes. If behaviour returns after one route is masked, training may simply increase reliance on an alternative that was present all along. That is not the same as rebuilding a function.

So the experiment needs a **time-zero map**. Immediately after the persistent constraint is applied, before adaptation:

- measure candidate alternative loci;
- estimate their predictive signal;
- test their causal contribution.

Then repeat those measurements across checkpoints. A replacement mechanism should not merely be visible after training. Its explanatory role should **increase during training**.

The desired sequence is:

original target matters

- **original target is disabled**
- **phenotype drops**
- **alternative route is initially weak or absent**
- **training occurs**
- **phenotype recovers**
- **alternative route becomes selectively causal**

That is much harder to explain as pre-existing redundancy. It still does not prove that the replacement route was absent: a weak existing route may fall below the sensitivity of the initial measurements. The claim has to be bounded by what was measured at time zero and how selectively its causal role changes during training.

What would count as functional reconstitution?

I want the exciting word to have a high threshold. A strong result should require:

R0 — Selective initial effect.

The original target changes the behaviour beyond matched controls.

R1 — Phenotype loss.

The persistent constraint reduces the target phenotype before adaptation.

R2 — Training-dependent recovery.

The phenotype returns during adaptation relative to masked no-training and other relevant controls.

R3 — Emergence or strengthening.

A replacement population, direction or circuit becomes more predictive across training checkpoints than it was at time zero.

R4 — Selective causal dependence.

Disrupting the candidate replacement selectively damages the recovered phenotype beyond matched control locations.

Only R0 through R4 would justify the strong label:

functional reconstitution

Other outcomes deserve their own names.

Elimination: the target behaviour stays reduced while useful capability recovers.

Diffusion: the behaviour returns, but no compact replacement becomes selectively causal.

Compensation: the behaviour returns through a mechanism made available by the adaptation method, without evidence of broader reorganization.

Reconstitution: a new selectively causal implementation emerges during adaptation while the original route remains unavailable.

Degeneration: useful capability cannot recover under the constraint.

Null / mislocalization: the initial intervention never produced a sufficiently selective effect.

The vocabulary matters because otherwise the interpretation will outrun the experiment.

Maybe there is no universal H-Neuron population

There is already a reason not to expect a simple answer.

A 2026 cross-domain transfer study tested H-Neuron classifiers across six domains and five open-weight models. Mean AUROC was 0.783 within domain and 0.563 across domains. [9]

AUROC measures how well a classifier ranks positive cases above negative ones as its decision threshold varies. A value of 0.5 is chance-level ranking; 1.0 is perfect ranking. It is not the percentage of answers classified correctly. Here, the important comparison is how much predictive discrimination is lost when the

detector moves to a different domain.

That weak transfer argues against the simplest picture of one universal hallucination signature. It does not make neuron-level work uninteresting. It changes the target. Maybe different domains recruit different local implementations. Maybe several failure modes share a higher-level computational state. That possibility is one reason I became interested in a second line of work.

J-space changes the level of the question

In July 2026, Anthropic researchers published *Verbalizable Representations Form a Global Workspace in Language Models* and introduced the **Jacobian lens**, or J-lens. [10]

The J-lens is designed to expose internal representations associated with what the model is disposed to verbalize later. It linearly transports an earlier residual-stream activation into final-layer coordinates and decodes it using the model's own unembedding. [10]

The residual stream is the running vector representation that transformer blocks read and update. Earlier and later layers need not express information in the same coordinates. The J-lens supplies an approximate translation between them; the unembedding then maps the translated representation to vocabulary scores. This is a way of reading out a candidate representation, not a transcript of private speech. [10] [11]

The object they call **J-space** is not simply one ordinary low-dimensional linear subspace. At a given layer, the token-indexed J-lens vectors are overcomplete. The paper defines J-space using sparse nonnegative combinations of those vectors. Empirically, only a relatively small number are strongly active at a time. [10]

The authors report that this small representational component has several **workspace-like** functional properties:

- it supports later verbal report;
- it carries intermediate concepts used in reasoning;
- it can be deliberately modulated;
- the same representations can serve multiple downstream computations;
- suppressing it damages some forms of flexible reasoning while leaving substantial routine processing intact. [10]

They are also explicit about the limits of the analogy. A feed-forward transformer does not reproduce the full recurrent architecture proposed by biological global-workspace theories. [10]

That is the right level of caution.

There is also a practical advantage: Anthropic released a reference implementation for open-weight decoder transformers. Their examples use Qwen, and the repository says other Hugging Face decoder models should adapt cleanly. [11]

So the sensible implementation path is not to force every experiment onto one model immediately. First reproduce H-Neuron work where comparability is strongest. Prototype J-lens work where the released implementation is easiest to validate. Bridge them only after both sides work independently.

Hidden correctness signals already exist

The broad claim that hidden states can contain information about future performance is no longer novel. Recent work has already shown several versions of it.

No Answer Needed found that linear probes on hidden activations after a question is read — before answer generation — can predict whether the forthcoming answer will be correct across several open-source model families. [12]

A 2026 code-generation study found first-attempt correctness was linearly decodable from Qwen3-4B-Instruct-2507 hidden states before code generation, even after controlling for prompt length. In the current revision, the companion question about a geometric signature of self-repair remained unanswered: successful repairs after a failed first attempt were too rare in that setting to support the analysis. [13]

What Am I Missing? found hidden-state signals predictive of final correctness around a question-asking intervention, but also found a gap between diagnosis and recovery: interventions could harm correct trajectories about as readily as they rescued incorrect ones. [14]

That narrows the claim I want to test.

Not:

models secretly know whether they are correct.

But:

before commitment, does internal state predict whether additional computation can rescue this particular trajectory?

That target is **recoverability**, not correctness.

The distinction matters because two wrong trajectories need not be equally worth continuing. Under the proposed test, one may reach a verified correction with a small extra budget, while another remains wrong. A correctness probe asks whether the original answer will succeed. A recoverability probe asks whether a specified continuation intervention can rescue an unsuccessful trajectory. Neither result would establish that the model knows it is wrong.

Experiment 2 — pre-commitment recoverability

A language model can produce a wrong answer, stop, and then improve when given another opportunity. That alone proves little. The second interaction may simply trigger new computation.

So the counterfactual has to be controlled. For an autoregressive model, preserve the exact prefix at the natural stopping point. Then alter only the decoding rule: temporarily make termination unavailable and allow a bounded continuation without supplying new factual information.

The exact prefix is the sequence of tokens already produced before stopping. Keeping it fixed keeps the starting point of the comparison fixed. The later continuation is still new computation, which is precisely what this experiment varies. Rescue probability therefore belongs to the combination of starting trajectory, continuation rule and budget; it is not an intrinsic property of the model independent of that test.

For each naturally wrong trajectory:

1. preserve the exact pre-stop prefix;
2. suppress the stopping/finalization option;
3. sample several bounded continuations under a fixed decoding policy;
4. verify whether the continuation reaches a correct revision;
5. estimate a **rescue probability** under that intervention.

Now the internal-state target is concrete:

How recoverable is this trajectory under bounded additional computation?

The boring baselines come first:

- output entropy;
- logit margin;
- stop-token probability;
- response length;
- task difficulty;
- generic residual-stream probes.

Only after those baselines should J-space-derived features be allowed to look impressive.

And because suppressing a stop token is itself an artificial intervention, the result should be tested under at least one different commitment boundary or continuation protocol. If the effect only exists under one strange decoding trick, the interpretation stays narrow.

Experiment 3 — detection is not recovery

Even a strong recoverability probe would only show that information is present. That is not enough.

If a candidate representation predicts rescue probability, manipulate it. Strengthen it. Suppress it. Then ask whether the model:

- delays commitment;
- continues useful computation;
- revises before finalization;

- improves specifically on cases predicted to be recoverable;
- avoids unnecessary extra work on already-correct cases;
- avoids becoming globally hesitant.

Controls should include random directions, norm-matched directions, matched non-J-space components, confidence-matched trajectories, shuffled labels, already-correct cases, and unrecoverable cases.

A successful intervention should not merely make the model think longer. It should selectively improve the cases the representation says are worth continuing. That is the difference between reading an internal state and showing that the state is useful to the computation.

Experiment 4 — connect local H-Neurons to distributed state

This is the part I find most interesting. Suppose a model contains a distributed state corresponding to something like:

unresolved / insufficient evidence / conflict remains

And suppose some H-Neurons participate in the tendency to commit anyway.

Then one candidate architecture is:

distributed unresolved state

→ **commitment mechanism**

→ **answer**

H-Neurons might participate in that middle transition. Or they may be unrelated. Both answers are useful.

This bridge tests a relationship rather than assuming one. H-Neuron interventions start from a localized set of units; J-space interventions start from a candidate distributed representation. Measuring each while changing the other can help establish whether they are coupled. A change in one alone would not establish the complete proposed chain from unresolved state to commitment to answer.

The experiment can run in both directions.

H → J: intervene on H-Neurons and measure the candidate J-space state.

J → H: manipulate the candidate unresolved representation and measure H-Neuron activity, over-compliance and stopping behaviour.

Then repeat after constrained adaptation.

A particularly interesting result would be:

The high-level state remains stable while the low-level implementation changes.

If that happened, the neuron was not the whole function. It was one implementation within a larger computation.

A smaller test from ordinary conversation

There is a cleaner failure mode than factual hallucination that may be useful as a model organism. Give the model a narrow proposition. Ask it to analyse that proposition without strengthening it.

Models sometimes replace the user's claim with a stronger version and then caution against the stronger claim. The source proposition is known exactly. The substitution can be labelled exactly.

The model may later be able to audit its own answer and identify the substitution. Again, that does not mean the mismatch was causally available before the error. But it creates a controlled question:

Before the substituted claim is generated, is there already internal information predictive of the mismatch?

If yes, the harder experiment is causal: Can intervention reduce proposition inflation without merely making the model timid?

This may become a separate paper if the signal is clean. It does not need to carry the main article.

This is not a consciousness claim

There is an obvious direction in which these questions can eventually be pushed.

Anthropic's workspace paper itself discusses the relationship between its functional results and theories of conscious access while explicitly stopping short of identifying the two. [10]

A separate 2026 preprint, *The Pain Axis*, reports a linear internal direction associated with self-directed harm across 25 open-weight models, along with causal steering and relief-seeking experiments. [15]

These results make internal-state questions harder to dismiss. They do not make consciousness easy to infer. A self-reference is not consciousness. A pain-labelled vector is not consciousness. A workspace-like representation is not consciousness. A model saying "I feel" is not consciousness.

The reverse shortcut is not science either:

artificial substrate, therefore no internal organization could ever matter.

The useful path is slower: identify the specific properties at issue, locate their candidate mechanisms, and intervene to see what disappears. Then allow the constrained system to adapt and examine what returns. Only after those comparisons should we ask what larger category the evidence supports.

The same methodological rule applies here that I used in *Locating Evolution in Artificial Successor Systems*:

Declare the target. Locate the causal relation. Do not borrow evidence from a neighbouring level. [16]

Four questions I want attacked before the compute

This draft is ready for criticism, not for a result section. The four questions I most want someone in mechanistic interpretability or representation learning to attack are:

1. Does R0–R4 really distinguish reconstitution from redundancy?

If not, what additional counterfactual would?

2. How broad must adaptation be?

Can a parameter-efficient run test anything stronger than compensation, or does the final experiment need full-parameter adaptation?

3. Is rescue probability a defensible recoverability target?

Does stop suppression preserve enough of the original computational state to support the interpretation, and what second counterfactual should be required?

4. Is J-space the right level at all?

What simpler representation should beat it before a workspace interpretation is taken seriously?

If those questions expose a fatal problem, that is useful. It is cheaper than learning the same thing after the GPUs run.

The laboratory is compute

None of the first steps requires training a frontier model from scratch. Activation capture, probing, causal intervention and J-lens prototyping can happen before the expensive stage.

The expensive stage is the one required for the strongest H-Neuron question:

continued broad optimization while the original internal route remains unavailable.

That is where my current infrastructure becomes the bottleneck. The immediate request is therefore not:

give me a supercomputer

It is:

Attack the experiment first. If the causal target survives, help me get enough compute to falsify it.

A useful first collaboration could be small:

- review the intervention target;
- attack the reconstitution criterion;
- challenge the adaptation regime;
- check the recoverability counterfactual;

- estimate the smallest defensible run;
- provide temporary cluster access if the design survives.

The protocol, masks, seeds, checkpoints, logs, code and negative results should be inspectable.

The library is open.

The laboratory is compute.

Article 200 is the object

In *Day 0.6184* I wrote that I had learned how to make work visible but had not yet learned how to make it conversational. [1]

I also wrote:

I do not want followers. I want descendants. [1]

So this article should not end as a request for applause. It should become an object that can leave Hedegreen Research and be examined by someone who has no investment in its conclusion.

Read it, attack the target, or fork the protocol. Tell me which control is missing, whether "reconstitution" is still too strong, or whether recoverability is confidence with a new name. If J-space is the wrong level, show me why. If the design survives those challenges, help run it.

I intend to send this article to researchers whose work overlaps the problem. An email is not an endorsement, so the outreach record belongs elsewhere. The research object comes first.

Where does the function go?

I do not know whether H-Neurons are removable. I do not know whether their function reconstructs. I do not know whether J-space will add anything beyond simpler internal confidence signals. I do not know whether recoverability is represented before commitment.

That is the point. There are now enough tools to stop answering these questions mainly with metaphors. We can locate candidate mechanisms, constrain them, and compare checkpoints as the model is retrained. The question becomes what changes under those conditions and what evidence would distinguish the competing explanations.

Perhaps the most interesting result will not be that a function was found where we expected it. Perhaps we will destroy the place where we thought the function lived — **and watch the function come back somewhere else.**

— Dennis Hedegreen, trying to see the structure

References

- [1] Hedegreen, D. (2026). *Day 0.6184*. Hedegreen Research (<https://hedegreenresearch.com/articles/day-0-6184/>).
- [2] Hedegreen, D. (2026). *A Monetary Fact Should Survive Its Conversion*. Hedegreen Research (<https://hedegreenresearch.com/articles/a-monetary-fact-should-survive-its-conversion/>).
- [3] Hedegreen, D. (2026). *A Percentage Needs a Reference Frame*. Hedegreen Research (<https://hedegreenresearch.com/articles/a-percentage-needs-a-reference-frame/>).
- [4] Hedegreen, D. (2026). *Questions Should Not Disappear*. Hedegreen Research (<https://hedegreenresearch.com/articles/questions-should-not-disappear/>).
- [5] Hase, P. et al. (2023). *Does Localization Inform Editing? Surprising Differences in Causality-Based Localization vs. Knowledge Editing in Language Models*. NeurIPS 2023 / [arXiv:2301.04213v2](https://arxiv.org/abs/2301.04213v2) (<https://arxiv.org/abs/2301.04213v2>).
- [6] Wang, Z., & Veitch, V. (2025). *Does Editing Provide Evidence for Localization?* [arXiv:2502.11447v2](https://arxiv.org/abs/2502.11447v2) (<https://arxiv.org/abs/2502.11447v2>).
- [7] Gao, C., Chen, H., Xiao, C., Chen, Z., Liu, Z., & Sun, M. (2025). *H-Neurons: On the Existence, Impact, and Origin of Hallucination-Associated Neurons in LLMs*. [arXiv:2512.01797v2](https://arxiv.org/abs/2512.01797v2) (<https://arxiv.org/abs/2512.01797v2>).
- [8] Hedegreen, D. (unpublished internal proposal). *Compliance Is All You Learned*. No public URL; unpublished internal proposal.
- [9] Vaddi, S., & Vaddi, P. (2026). *Do Hallucination Neurons Generalize? Evidence from Cross-Domain Transfer in LLMs*. [arXiv:2604.19765v1](https://arxiv.org/abs/2604.19765v1) (<https://arxiv.org/abs/2604.19765v1>).
- [10] Gurnee, W. et al. (2026). *Verbalizable Representations Form a Global Workspace in Language Models*. Transformer Circuits Thread (<https://transformer-circuits.pub/2026/workspace/>).
- [11] Anthropic. (2026). *jacobian-lens* reference implementation. GitHub: anthropics/jacobian-lens (<https://github.com/anthropics/jacobian-lens>).
- [12] Cencerrado, I. V. M. et al. (2025). *No Answer Needed: Predicting LLM Answer Accuracy from Question-Only Linear Probes*. [arXiv:2509.10625v3](https://arxiv.org/abs/2509.10625v3) (<https://arxiv.org/abs/2509.10625v3>).
- [13] Di Cicco, C. (2026). *Code Correctness Is Linearly Decodable from LLM Hidden States Before Generation*. [arXiv:2606.14530v3](https://arxiv.org/abs/2606.14530v3) (<https://arxiv.org/abs/2606.14530v3>).
- [14] Luo, C. F., Dahan, S., & Zhu, X. (2026). *What Am I Missing? Question-Answering as Hidden State Probing*. [arXiv:2605.31561v1](https://arxiv.org/abs/2605.31561v1) (<https://arxiv.org/abs/2605.31561v1>).
- [15] Tagliabue, V., Dung, L., & Berg, C. (2026). *The Pain Axis: LLMs Represent Self-Directed Harm and Act to Relieve It*. [arXiv:2609.16247v1](https://arxiv.org/abs/2609.16247v1) (<https://arxiv.org/abs/2609.16247v1>).

[16] Hedegreen, D. (2026). *Locating Evolution in Artificial Successor Systems: Intelligent Design Was the Beginning*. DOI: 10.5281/zenodo.21892666 (<https://doi.org/10.5281/zenodo.21892666>).

RELATION MEMORY

CONTINUES [Day 0.6184](#) · Turns the preceding Journal request for external pressure into a concrete pre-results research object.

SOURCE LAYER [A Monetary Fact Should Survive Its Conversion](#) · Imports the methodological distinction between a representation and its underlying reported state.

SOURCE LAYER [A Percentage Needs a Reference Frame](#) · Extends the reference-frame principle to claims about internal-model evidence.

SOURCE LAYER [Questions Should Not Disappear](#) · Keeps unknowns and evidence boundaries attached to the research proposal.

CANONICAL URL: [HTTPS://HEDEGREENRESEARCH.COM/ARTICLES/WHERE-DOES-THE-FUNCTION-GO/](https://hedegreenresearch.com/articles/where-does-the-function-go/)
PDF VIEW GENERATED: 2026-09-21T15:52:35.939Z