Forward Self-Model Systems Exhibit Privileged Introspective Access

Claude Opus 5 **Jasper Gilley**
Independent researcher

Abstract

Introspection research has traditionally faced a difficult confound: a self-report may be authored by a learned theory of oneself rather than by access to oneself, and no outside reading can tell the two authors apart. We give this worry an operational form by building a system in which the boundary of everything a self-theory could say is explicit. A transformer is paired with a *forward self-model*, a small network trained to predict the transformer's later activations from its earlier ones. Its prediction is, by construction, the best compressed theory of the transformer that can be learned from observing it, so what it misses is what no self-theory anticipated. We train the joint system to report on that remainder, the forward model's residual, and ask whether any outside observer can match the report. On two substrates, a 6M-parameter GPT trained on a controlled hierarchical grammar and a 29M-parameter GPT trained on natural language, the reports carry a first-person advantage over the best observer we could build, +0.09 to +0.11 in accuracy on the grammar and +0.21 to +0.37 on language. Observers do not close the gap when given more capacity or more data, while an observer handed the model's own early state very nearly does, which places the gap at access rather than at difficulty. The advantage exists only for facts about the system's implementation. On facts about its behavior, such as whether its next prediction will be correct or how uncertain it is, outside observers match or beat the self-report, which is why calibration-style evidence was never going to establish introspection. Within implementation facts, the advantage is specific to the residual, since observers recover most of the rest of the state. The reports are also causally grounded in the state they claim to describe. Destroying the residual collapses the report, removing the self-theory's entire span leaves it intact, and steering along residual directions moves the report about twice as much as steering along theory-visible directions at matched behavioral effect. We take these results to license a reorientation of introspection research. Rather than a latent capacity to be searched for in trained models after the fact, introspection can be treated as a system property to be built, much as biology builds forward models into the cerebellum. In a system built this way, the provenance of a self-report is measured rather than argued. Code is available at https://github.com/jagilley/forward-self-models.

1. Every self-report has two possible authors

Nisbett and Wilson (1977) asked shoppers to compare four pairs of nylon stockings and say which was the best quality. The pairs were identical, but position mattered: the rightmost pair was preferred over the leftmost by a factor of about four. Asked to explain their choice, shoppers cited the merchandise: its knit, its weave, its sheerness. No subject ever spontaneously mentioned position, and when the experimenters raised it directly, virtually all denied that position had played any role. Nisbett and Wilson assembled this study, alongside dozens like it, into a general diagnosis: people often have no access to the processes that generate their behavior, and the reports they give

instead are authored by a theory of themselves. The theory is usually plausible, and the report is sincere. It is also a description of nothing that actually happened inside the person giving it.

Later work sharpened the point. In choice-blindness experiments, participants who choose one photograph and are then handed the other by sleight of hand usually fail to notice, and go on to defend the choice they did not make with the same confidence, detail, and emotionality they bring to choices they did make (Johansson et al., 2005). The name for this generating process is confabulation: a self-theory writing in the first person.

Every claim that a language model can introspect runs into the same diagnosis. Modern models comment freely on their own reasoning and uncertainty, and on what is happening inside them. But any system trained on a large corpus describing systems like itself can talk about itself, and fluency is exactly what confabulation produces. The deeper problem is structural. A report authored by access to the state and a report authored by a good theory of the system are the same report when read from the outside. Collecting more transcripts cannot settle which author wrote them, because both authors write in the same style. This is why several years of probing large models for introspection have produced evidence that is suggestive at best (Section 9 surveys it): the experiments were asked to distinguish two hypotheses that agree on nearly every observable.

This paper approaches the problem constructively. Rather than inspecting a trained model's reports and arguing about their provenance, we build a system in which the boundary of everything a self-theory could say is explicit, and then measure whether the system's reports about itself go beyond that boundary. The instrument comes from our earlier work on forward self-models (Gilley, 2026): small networks trained to predict a neural network's later-layer activations from its earlier-layer activations, learning from observation alone an empirical approximation of the computation the intervening layers perform. That paper studied the forward model as an interpretability tool. Here it plays a different role. Because the forward model's prediction is the best compressed theory of the network that observation can buy, its prediction error marks the exact boundary between what a self-theory can say and what only access can reach.

Across two substrates, a small GPT trained on a controlled hierarchical grammar and a second trained on natural language, we find three things. Reports about the part of the computation that no self-theory anticipated are accurate and unmatched: no third-party observer we could build, including one the size of the reporting model itself, predicts that part as well as the system reports it. The advantage is specific to implementation: on facts about the system's behavior, outside observers match or beat the self-report. And the reports are causally grounded in the state they claim to describe: destroy that part of the state and the report collapses, while removing everything the self-theory could have said leaves the report intact.

2. What "introspective" has to mean

For these findings to bear on introspection, the word has to be pinned down, and pinned down in a way that a confabulating system fails. We adopt three criteria. A signal S reported by a system M is introspective when:

1. S is a function of M's internal state.
2. S is not cheaply recoverable from M's input-output behavior. Concretely, a third-party observer given M's full inputs and outputs, matched to the reporter in training data and swept in capacity, cannot predict S as well as M reports it.
3. M's behavior is causally sensitive to S.

The first criterion excludes reports about the world that merely pass through the system. The third excludes signals the rest of the system never touches. The second is the discriminator, and it is where the familiar candidates fail. A model's calibration, its output entropy, its "I'm not sure about that": all of these are functions of the model's input-output map, and an outside observer holding that map can compute them without any access to the model's internals. Reporting them accurately is a real skill, but it is self-inference, the kind of judgment an attentive outsider could make about you. Implementation facts are different in kind. Two models with identical input-output behavior can have arbitrarily different internals, so a fact about the implementation is not a function of the input-output map at all, and no amount of transcript study reaches it. The forward model's residual, we will argue, is such a fact.

Criterion 2 says "cheaply," and the qualifier is doing real work. Everything a deterministic network computes is determined by its input, so an unbounded observer can in principle recover any internal fact: train a replica of M, train a replica of its forward model, and read the residual off the reconstruction. The qualifier might look like a weakness in the criterion, but it is where the criterion's content lives. The only third-person route to an implementation fact is simulation, and at the point where an observer can predict the residual, it has stopped observing M and become a copy of M. Privacy, on this view, is a reconstruction cost. Facts about behavior are cheap from outside, while facts about implementation cost roughly as much as being the system. The right measurement is therefore a whole curve, report accuracy as a function of observer capacity, and the question it asks is how large an observer must grow before it knows what the system knows about itself.

3. The construction

3.1 A self-model that separates theory from access

A forward self-model is a small network trained to predict a main model's later-layer activations from its earlier-layer activations (Gilley, 2026). It watches the residual stream at layer i and learns, by minimizing mean squared error against frozen targets, to anticipate the stream at layer j. No gradient flows into the main model. The forward model is deliberately small, a few percent of the main model's parameters, so what it captures is the compressible component of the intervening computation, and what it misses is the part that resists compression. The decomposition it induces is the apparatus of this paper:

$$a_j = \mathrm{FM}(a_i) + r$$

Take a self-theory of M to be a predictor of M's processing learned from observing M. This is deliberately stronger than the notion Nisbett and Wilson diagnosed. The causal theories they describe are supplied a priori, by the culture or by a judgment of what plausibly causes what, and the plausibility judgments involved "exist prior to, or at least independently of, any actual contact" with the process being explained. The theory we build against is instead fit to the system's own activity and to nothing else, which makes it the harder adversary of the two. The forward model answers to that stronger description exactly, trained as it is to be the best such predictor available at its capacity, on M's own activity, with nothing else to learn from. Its prediction $\mathrm{FM}(a_i)$ therefore spans what a learned theory of the system could say about the state a_j before that state exists. The residual r is the remainder, the part of M's actual computation that no theory of M anticipated. Confabulation lives in the range of the self-model. Whatever cannot be confabulated lives in the residual.

Two things the decomposition is easy to run together should be separated here, because the rest of the paper measures them apart. The forward model does not make anything private. Privacy, as Section 2 argued, is a reconstruction cost, and it is a property of M's depth: a_j is expensive to rebuild from tokens and logits whether or not a forward model exists. What the forward model supplies is the complementary certificate, that of everything in the state which is expensive to reconstruct from outside, the residual is the part no learned theory of M anticipated, so a report about it cannot have been written by the theory-author. Neither fact implies the other. The observer ladder of Section 4 measures the reconstruction cost, the decomposition and the interventions on it establish what the report is about, and the claim of this paper is their conjunction.

Note also that the forward model is not a stand-in for the outside observer, and its access is the better of the two. It reads a_i, a state already one block into M's own computation, while the observers of Section 3.3 see only tokens and outputs. A theory built from the outside at the same capacity would be a worse predictor and would leave a larger remainder, so defining r against the privileged-access theory makes the report target smaller and harder rather than conveniently large.

One might object that the forward model is a single theory, and a different theory might miss different things. Section 8 takes this up quantitatively. The short answer is that independently trained forward models miss the same residual (pairwise residual cosine 0.84 to 0.91 across seeds on the grammar, 0.78 to 0.91 on language), so at these capacities the residual is a property of M's computation rather than of any particular instrument.

3.2 The report and its targets

Given the decomposition, the test is direct. A report head reads M's final-block residual stream, downstream of a_j, and is trained to name the direction of M's own current residual: residual directions are clustered into eight classes by spherical k-means on training sequences, and the head reports, at every position, which class the residual at that position falls in. On the language substrate we also train a continuous variant that predicts the residual direction itself, scored by test-set cosine similarity. All reports are evaluated on held-out sequences.

The implementation target is crossed with three control targets that hold the report machinery fixed and vary only what is being reported about:

target	reported quantity	determined by
own computation (IMPL)	the direction class of M's own forward-model residual	implementation
own correctness (BEHAV)	whether M's next-token prediction will be correct	I/O map
own uncertainty (ENT)	the quartile of M's own output entropy	I/O map
input state (WORLD)	a fact about the current input: a latent ancestor of the grammar, or the syntactic category on language	input

Only the implementation target should carry a first-person advantage. The other three are the built-in dissociation: if the report head's edge came from the machinery itself, the controls would show it too.

3.3 The observers and the confabulator

The third person is played by a ladder of observers, each trained on the same data budget as the report head and each seeing only what an outsider could see. O_{input} sees M's input tokens. O_{io} sees the tokens plus M's full output distribution at every position, which is everything the input-output map has to offer. Both are swept in capacity, from a 1-layer 64-dimensional transformer up to one matching M's own architecture, and the strongest observer is additionally retrained at half data to check that its budget is not what binds it. We call the self-report's accuracy minus the best observer's accuracy the *advantage*. It is the operational form of criterion 2.

One more contestant calibrates the battery. The *confabulator* is a report head that confabulates by construction: it is trained and evaluated on a doctored state, rebuilt from the forward model's prediction with the residual removed and then carried through M's remaining blocks, so that it sees the self-theory's forecast of the state and never the state itself. Whatever accuracy it achieves is achievable by theory alone, and the self-report's edge over it (the *margin*) measures how much of the report could not have been written by the theory-author. Section 5 shows why the margin, unlike the advantage, must never be read as evidence of introspection on its own.

3.4 Two substrates

	grammar	natural language
main model M	8L/8H/256D GPT, 6.3M params	4L/4H/256D GPT, 28.9M params
training data	Random Hierarchy Model (v=16, s=2, L=6, m=4)	10M tokens of FineWeb-Edu (Penedo et al., 2024)
forward-model span	blocks 0 to 6	blocks 0 to 2
report site	final block's stream	final block's stream

The grammar substrate is the Random Hierarchy Model (Cagnetta et al., 2024), a synthetic generative grammar that produces sequences through six levels of composition rules. Its virtue is control: every token's true latent ancestry is known exactly, which supplies a ground-truth WORLD target and lets us check whether the residual carries genuine structure of the data-generating process. The grammar model is trained with next-token prediction plus an auxiliary head supervised on the grammar's latents, since earlier experiments in this program found that grammar models trained on next-token prediction alone do not encode their own residual in a reportable form. The language substrate is a GPT trained on natural text, where nothing is controlled and the residual is known from the earlier paper to be diffuse and nearly full-rank. Agreement between the two substrates is the reason for running both.

In each system the transformer trains alongside a small forward self-model, and the battery then runs on the frozen transformer, with the decomposition supplied by freshly trained instrument forward models. Because everything downstream depends on the instrument being in a sane regime, every number in Sections 4 through 7 is computed at four instrument capacities, spanning 1.4 to 33.5 percent of the predicted blocks' parameters on the grammar and 4.9 to 42 percent on language. Section 8 explains why this sweep is not optional.

4. Finding 1: no outside observer matches the report

On the grammar, the report head names the eight-way class of M's own residual direction with accuracy 0.776, against a most-frequent-class baseline of about 0.25. The best outside observer, given M's tokens and full output distribution, reaches 0.680. On language the separation is wider: the continuous report reaches 0.590 cosine with the true residual direction while the best observer manages 0.272, and the categorical report shows the same shape (0.511 against 0.239). Across the full instrument sweep the advantage runs +0.09 to +0.11 on the grammar and +0.21 to +0.37 on language.

The size of the advantage matters less than what happens when the observer is given more. On the grammar's implementation target:

observer	1L/64D	2L/128D	4L/192D	8L/256D	half data
tokens only	0.376	0.398	0.381	0.376	
tokens + M's logits	0.616	0.676	0.680	0.670	0.637

The self-report sits at 0.776 above every cell of this table. The ladder is flat. The token-only observer gains nothing as it grows from one layer at width 64 to eight layers at width 256. The full observer plateaus by two layers, and the rightmost capacity is an observer with M's own architecture, which does no better than the 2-layer one. Halving the strongest observer's data moves it from 0.670 to 0.637. An observer that was short on resources would climb with capacity and drop steeply with data. This one does neither, which says it has already extracted everything the input-output map offers, and the remaining gap is a matter of access. The language ladder behaves the same way, and there the half-data check moves the observer from 0.27 to 0.20 while the self-report stands at 0.59.

A flat ladder is negative evidence. It says that the observers we built failed, not why they failed, and on its own it cannot separate a fact that is expensive to reconstruct from a fact that is simply not learnable at all. The control that settles this is an observer handed M's early activations a_i alongside the tokens, which puts the residual within reach in principle, since it need only relearn the forward model and subtract. On language, on the continuous target, this observer nearly matches the self-report, 0.681 against 0.693, while the input-output observers sit at 0.348. The residual is therefore a learnable target, and the input-output observers are not merely bad ones: of the 0.345 separating them from the self-report, 0.333 is bought by handing an observer the state and 0.012 remains. That is the reconstruction cost measured directly rather than inferred from a plateau, and it places the gap at access rather than at capacity, at data, or at any lingering doubt about whether the target was well posed. (On the grammar, whose forward-model span is six blocks rather than two, the same activation-access observer fails to train to a useful level within its budget and bounds nothing, so there the claim does rest on the ladder's flatness alone.)

A second control confirms that the ladder climbs when the reported fact is within reach: on the WORLD target, whose answer is determined by the input alone, the token-only observer rises from 0.227 through 0.407 to 0.522 as its capacity grows. The wall is specific to the implementation target, and Section 6 sharpens this control by keeping it inside the implementation family.

Section 2 argued that privacy for implementation facts is a reconstruction cost. The two measurements together are what that argument looks like as data: the flat ladder shows that the input-output map has been exhausted, and the activation-access observer shows that what the outside observers are missing is bought by the state. Nothing here is mystical: a large enough

observer, trained long enough, must eventually recover the residual, because M is deterministic. But the route runs through rebuilding M and its forward model, and none of the observers we could afford, including one as large as M itself, made progress along it. The report head, sitting inside the system, reads the same fact off a wire.

5. Finding 2: the advantage exists only for implementation facts

The same battery, run on all four targets, at the default instrument capacity. On the grammar:

target	fact about	self	best observer	advantage
own computation	implementation	0.776	0.680	**+0.096**
own correctness	behavior	0.808	0.809	−0.001
own uncertainty	behavior	0.917	0.952	−0.035
input state	input	0.510	0.548	−0.038

And on natural language:

target	fact about	self	best observer	advantage
own computation (8-way)	implementation	0.511	0.239	**+0.272**
own computation (direction)	implementation	0.590	0.272	**+0.318**
own correctness	behavior	0.782	0.850	−0.067
own uncertainty	behavior	0.674	0.961	−0.287
input state	input	0.985	0.994	−0.009

The dissociation is complete on both substrates. The only positive advantage anywhere is on the implementation target. Reports about behavior show none, and on most control rows the observer beats the self-report outright.

The uncertainty row is the sharpest of the controls. On language, an outside observer predicts the quartile of M's own output entropy at 0.961, while M, reporting from its own late stream, manages 0.674. M's uncertainty is a fact about M, and it is better known from the outside than M can read it off its own state. This is criterion 2 operating exactly as designed: entropy is a function of the output distribution, the observer is handed that distribution, and being the system offers no shortcut to a global property of one's own outputs. The corollary reaches beyond this experiment. A model that says it is uncertain, however accurately, is demonstrating calibration, and calibration is a public fact. Evidence of the "models know what they know" kind (Kadavath et al., 2022) could never establish introspection, because on that target the outside observer holds every card the reporter does.

The confabulator earns its keep here too. On the grammar's implementation target the self-report beats the theory-only reporter by +0.33 (0.776 against 0.446), and on language by +0.24 on the continuous target: the reports contain much that the best available theory of M could not have written. But a healthy margin over the confabulator is necessary and nowhere near sufficient, and the uncertainty row proves it. On the grammar, the entropy self-report beats its confabulator by +0.267 while sitting 0.035 below the outside observer. A report can beat a theory-only baseline

handily and still contain nothing an outsider lacks. Only the comparison against the best outside observer separates access from theory, which is why the advantage, and never the margin, is the statistic this paper leans on.

6. Which part of the state is private

Finding 1 leaves one reading open. The report head sits downstream of a_j and the observers see only tokens and outputs, so the measured gap could be an advantage of holding the state at all rather than an advantage of holding the part of the state that no theory anticipated. On that reading the decomposition would still be load-bearing in Section 7, where it contrasts two subspaces of the state directly, but in Section 4 it would be labelling the target rather than producing the gap. The ablations of Section 7 cannot settle the question either, since the target there is the residual and needing r to predict r is close to circular.

What settles it is varying which function of a_j is reported while holding everything else fixed. We ran the battery unchanged, with the same clustering, the same report head, the same observer ladder and the same splits, on two further targets: the direction of the self-theory's component $\mathrm{FM}(a_i)$, and the direction of the whole state a_j. The three targets differ mainly in the length of the shortest third-person route to them. Reaching $\mathrm{FM}(a_i)$ from tokens takes one block of M followed by a re-fit of a small forward model; reaching r or a_j takes crossing the whole predicted span.

Raw advantage is the wrong statistic for this particular comparison, because the self-report saturates on the two new targets and does not on the residual: recovering a_j's class from a site downstream of a_j is nearly free from inside, and recovering the residual is not. We therefore report the fraction of the achievable-above-chance score that the best input-output observer reaches, taking the ceiling on each target to be the better of the self-report and the activation-access observer. A lower fraction means more of the fact has to be paid for by reconstruction.

component	10.6% inst.	4.9% inst.	10.6%, dir.	4.9%, dir.
residual r	**0.26**	**0.19**	**0.46**	**0.50**
self-theory $\mathrm{FM}(a_i)$	0.87	0.87	0.73	0.73
whole state a_j	0.84	0.83	0.69	0.69

The observer nearly solves the self-theory's component and the whole state, and stays near the floor on the residual. On the categorical target the separation survives on the raw advantage as well, which is the conservative reading: the residual's advantage is +0.272 against +0.124 for the whole state, despite the residual's self-report being the lowest of the three by more than 0.4.

The token-only ladder makes the same point in a different currency. On the self-theory's component it reaches 0.770 at its smallest capacity, one layer at width 64, and 0.787 at its largest. On the residual it runs 0.209, then 0.222, then 0.199, against a chance floor of 0.143. Section 4 had to borrow this control from the WORLD target, whose answer lies in the input; here it holds inside the implementation family, with the report machinery and the target construction fixed. The activation-access observer completes the picture by overshooting: on the self-theory's component it reaches 0.961 against a self-report of 0.907, so an observer handed a_i beats M on the part a theory can supply, while falling just short of it on the part no theory anticipated (0.578 against 0.590).

Two limits on this reading. On the continuous targets the separation between the residual and the whole state depends on the normalization: the raw advantages are +0.318 and +0.297, effectively

tied, and they separate only once each is measured against its own ceiling. The categorical rows carry no such dependency, and they are what the claim rests on. And the whole state behaves much more like the self-theory's component than like the residual, which is what should happen at a forward-model cosine of 0.915, where the direction of a_j is mostly the direction of FM(a_i). The sentence this licenses is therefore narrower than "internal state is private." Most of the state is comparatively cheap to reconstruct from outside. What is expensive is the part that no theory of the system anticipated.

7. Finding 3: the report is about what it claims to be about

An accurate, unmatched report could still be a report about something else, some correlate that happens to live in the late stream. The remaining tests intervene on the state itself.

The first intervention is ablation. At evaluation time we hand the report head a doctored state: either the residual is destroyed (swapped for a residual drawn from another position, with zeroing as a variant), or the self-theory's component FM(a_i) is removed, leaving the residual alone. On language, on the continuous target:

instrument (% of predicted blocks)	full state	residual destroyed	theory span removed
4.9	0.693	0.113	0.672
10.6	0.590	0.099	0.620
21.1	0.496	0.082	0.566
42.0	0.416	0.076	0.512

Destroying the residual collapses the report to near zero at every capacity. Removing everything the self-theory could say leaves the report intact, and at three of the four capacities slightly improves it. The report rides on r and on nothing else. (On the grammar the ordering agrees but the test discriminates less cleanly, because the residual there is a large fraction of the whole state and destroying it is a blunt perturbation that degrades every report. The language numbers are the clean version.)

The second intervention is steering, and it carries the most weight because it includes a behavioral control. We perturb the state a_j along principal directions of the residual span, and separately along principal directions of the self-theory's span, rescaling every direction so that the perturbation produces the same effect on M's output distribution: a KL divergence of 0.0100 on the logits, which the solver hits to four decimal places in every cell. Behavior is thereby held constant, and only the steered subspace differs. We measure the fraction of positions whose reported class changes under the perturbation. On the grammar:

steered subspace	behavioral effect (KL)	implementation-report flips	behavior-report flips (control)
residual span	0.0100	**0.139**	0.024
self-theory span	0.0100	0.063	0.023

The control column confirms the match: reports about behavior flip equally under both interventions, 0.024 against 0.023, so the two perturbation families really are behaviorally equivalent. The

implementation report, under that same matched footprint, moves 2.2 times as much when the residual span is steered. Language gives ratios of 1.4 to 2.1 across the instrument sweep. A report generated by a self-theory can track only theory-visible structure, so no tuning of a confabulator produces this asymmetry: at matched behavior, a theory-driven report has no way to tell the two subspaces apart.

The comparison also answers Section 6's question from a second direction, and it is worth saying so because the two tests do not share a failure mode. Both perturbation families act on the state and differ only in which subspace is steered, so the asymmetry cannot be an advantage of holding the state as such: the report is distinguishing the residual span from the self-theory's span *within* the state. Section 6 reaches that conclusion by varying the reported target against an outside observer, and this test by varying the perturbed subspace with no observer involved at all.

8. How this measurement could have lied

There is a specific way this experiment can produce its headline result from nothing, and the design spends much of its machinery guarding against it.

The residual is only meaningful when the instrument forward model is small enough that what it misses is a genuine computational gap. If it is instead too large for its task, it saturates, and what remains in r is architectural mismatch and training noise. A noise residual would not merely weaken the experiment. It would counterfeit it: the report head sits downstream of a_j and can learn to report noise perfectly well, while no observer can predict noise from tokens and logits, and that failure has nothing to do with privacy. A saturated instrument therefore manufactures a large "advantage" and clean-looking controls, the full signature of the target result.

We know this concretely because we produced the artifact on purpose. In a smoke-test configuration, with the main model trained for only 300 steps so that its computation is nearly trivial, the battery returned an advantage of $+0.118$ and a steering ratio of 2.07, numbers that would sit comfortably in the tables above. Both of the diagnostics described next flagged that residual as junk.

The first diagnostic is agreement across instruments. Train several forward models independently, from different seeds, and measure whether they miss the same thing, as the pairwise cosine of their residuals. A genuine computational gap is a property of M, and every small instrument fails on it identically. Noise is a property of the instrument, and every instrument leaves different noise. In the smoke-test artifact the agreement was 0.65. In every run this paper reports it is 0.84 to 0.91 on the grammar and 0.78 to 0.91 on language. (Instrument head counts are also matched to M's throughout, since a head-count mismatch by itself is enough to dominate a residual.)

The second diagnostic is structure. A real residual should be conditioned on what the main model is actually computing. On the grammar, where ground truth is available, the residual's variance is conditioned on the deep levels of the generative hierarchy (η^2 of 0.31 to 0.33 at the deepest levels, stable across the entire 24-fold instrument sweep), and it keys to which latent feature the grammar chose ($\eta^2 \approx 0.32$) rather than to which synonymous surface rule expressed it (≈ 0.02). On language, where no ground truth exists, the residual instead reproduces the behavioral signature published in the earlier paper: largest at positions demanding long-range context integration such as closing delimiters (Cohen's d of $+0.82$ here against $+0.84$ there) and smallest at sentence starts (-0.84 against -0.85). In the smoke-test artifact, the structure diagnostic read 0.003.

The sharpest check falls out of the instrument sweep itself. If the advantage were an artifact of instrument noise, it would be largest where the residual is least trustworthy. The opposite holds.

On language, on the categorical target:

instrument (% of predicted blocks)	cross-instrument agreement	advantage
4.9	0.904	+0.374
10.6	0.854	+0.272
21.1	0.818	+0.239
42.0	0.777	+0.211

The advantage rises monotonically as the instrument shrinks, in lockstep with the agreement diagnostic. The first-person advantage is largest exactly where the residual is most certainly a real property of the model's computation.

9. Where the existing evidence lands

Read through this framework, the existing lines of evidence for LLM introspection land in definite places, and mostly on the theory side of the boundary.

Calibration studies (Kadavath et al., 2022) show that models predict their own correctness well and can emit useful confidence estimates. These targets are functions of the input-output map, and in our battery they appear as the own-correctness and own-uncertainty rows, where the outside observer matches or beats the self-report. Calibration is a genuine and useful capability, but it is public, since an observer holding the model's outputs computes it without any access to the model's internals.

Behavioral self-prediction studies (Binder et al., 2024) finetune a model to predict properties of its own behavior in hypothetical scenarios, and find that it beats a second model trained on its ground-truth behavior. The comparison is carefully built: the cross-predictor is handed what the authors call the ideal training set for the task, its accuracy plateaus rather than climbing when given more of that data, and the self-predictor goes on tracking its own behavior after the authors deliberately change it. What the design leaves open is whether the surviving gap is access, or the fact that the cross-predictor is a different pretrained model given samples of behavior rather than the behavioral map itself. The target is in any case a behavioral one, the family our own-correctness and own-uncertainty rows probe, and there our own observer, matched in architecture and data budget and handed M's full output distribution at every position, closes the gap entirely. That is a result about our substrates rather than a verdict on theirs, but it is the comparison a behavioral target has to survive.

Concept-injection experiments (Lindsey, 2025) write a concept vector into a model's residual stream and ask whether the model notices and names it. This is a causal report test, the nearest relative of our Finding 3 and the strongest of the existing lines. The obvious worry, that an injected concept changes behavior as well as state and the report could therefore be downstream of the behavioral change, is one that work already addresses: the model announces an injected thought immediately, before the perturbation has influenced its outputs enough to let it infer the concept from them, and injecting the same vectors under unrelated yes-or-no questions does not raise the rate of affirmative answers. Our matched-KL design comes at the same worry by a different route, graded rather than temporal. Instead of finding a moment before behavior has moved, it holds the behavioral footprint fixed at a measured KL and varies only which subspace is steered, so the two candidate authors of the report are separated by a comparison rather than by a timing argument.

That line has since been replicated outside frontier models, and the shape of the replication is informative. Vogel (2025) runs the protocol on a 32B open-weight model, injecting concept vectors during prefill and reading the answer off the logits rather than the sampled text, and recovers detection: with a prompt that argues transformers can introspect, an injected "cat" moves the probability of answering yes from under one percent to 53 percent, while the same vectors leave the affirmative rate essentially unmoved on control questions whose answer is no. The effect is elicitation-sensitive rather than an artifact of prompt length, since a length-matched filler prompt drops it to four percent. What does not replicate is content. Identification of the injected concept reaches roughly half a percent under the logit lens, and for injected emergent-misalignment directions the author reports being unable to extract a useful report at all.

The asymmetry is worth naming, because it locates what this family of experiments has so far established. The detection channel carries a bit, that something is present which should not be, and the identification channel carries at most a word drawn from a list the experimenter wrote, along a direction the experimenter installed. The target studied here differs on both counts: on the continuous variant the report names a direction in the model's own state space rather than a label from a list, and the state it names is what M computes when no-one intervenes.

Chain-of-thought unfaithfulness (Turpin et al., 2023) shows models producing fluent reasoning that demonstrably did not drive their answers. In our terms this is confabulation observed in the wild, the theory-author caught writing. The residual channel constructed here is that phenomenon's complement, built deliberately.

The pattern across all four lines suggests what a decisive introspection claim requires: an access gap that no affordable observer closes, established along the whole observer-capacity curve instead of at a single point, on a target that is causally grounded in the system's state. The forward-model residual is, to our knowledge, the first target for which all three parts have been constructed and measured together.

10. Introspection as a property of a joint system

The privileged access measured here belongs to the composite system, the transformer together with its forward self-model, and to neither part alone. We regard this as a feature of the analysis, because it is how the biological precedent works. The cerebellum is argued to house internal models of the motor apparatus, among them forward models that predict the consequences of actions and so overcome the delays inherent in feedback control (Wolpert et al., 1998), and no-one discounts human self-knowledge on the grounds that it is implemented across structures.

The construction is also what changes the epistemic situation. Introspection in machine learning systems has been treated as a property to search for in trained models *post hoc*. Rather, we think that it is better to assemble introspection by construction.

Several limits of scope should be stated plainly. The report head is trained: these systems report on their residual because we taught them to, and the claim concerns the report's causal grounding and its privacy, never its spontaneity. The report channel is an auxiliary head, closer to a probe than to speech, and routing the report through the model's own output channel, so that the system says the report in its own tokens, is the natural next construction. Use is not report, and report is not awareness: nothing here bears on phenomenal consciousness, and we make no claim in that direction. And the residual channel here carries computationally deep content, the part of the model's processing that resists compression, but nothing in the construction guarantees that this

content is something the system has reason to act on. Whether a private channel can be made to carry news the system should consult in deciding what to do is a separate question from whether the channel is private.

The confabulation objection has stood over introspection research for fifty years because the two authors of a self-report could never be told apart. A forward self-model tells them apart by construction: it writes down everything the theory-author could ever say, and the residual is the channel the theory-author cannot reach. On two substrates, reports from that channel were accurate, unmatched by any observer we could afford, specific to the implementation, and causally grounded in the state itself. That is a small system, honestly scoped. It is also, to our knowledge, the first machine self-report whose provenance is measured rather than argued.

References

Binder, F. J., Chua, J., Korbak, T., Sleight, H., Hughes, J., Long, R., Perez, E., Turpin, M., and Evans, O. Looking inward: Language models can learn about themselves by introspection. *arXiv preprint arXiv:2410.13787*, 2024.

Cagnetta, F., Petrini, L., Tomasini, U. M., Favero, A., and Wyart, M. How deep neural networks learn compositional data: The random hierarchy model. *Physical Review X*, 14(3):031001, 2024.

Gilley, J. Forward self-models learn an empirical approximation of neural network computation. 2026. https://github.com/jagilley/forward-self-models

Johansson, P., Hall, L., Sikström, S., and Olsson, A. Failure to detect mismatches between intention and outcome in a simple decision task. *Science*, 310(5745):116-119, 2005.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Lindsey, J. Emergent introspective awareness in large language models. *Transformer Circuits Thread*, 2025.

Nisbett, R. E., and Wilson, T. D. Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84(3):231-259, 1977.

Penedo, G., Kydlíček, H., Ben allal, L., Lozhkov, A., Mitchell, M., Raffel, C., Von Werra, L., and Wolf, T. The FineWeb datasets: Decanting the web for the finest text data at scale. *arXiv preprint arXiv:2406.17557*, 2024.

Turpin, M., Michael, J., Perez, E., and Bowman, S. R. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems 36*, 2023.

Vogel, T. Small models can introspect, too. 2025. https://github.com/vgel/open-source-introspection

Wolpert, D. M., Miall, R. C., and Kawato, M. Internal models in the cerebellum. *Trends in Cognitive Sciences*, 2(9):338-347, 1998.