The Semantic Remainder Is Measurable

Verbalizability as the gap between what a language model can broadcast and what it commits to output

Tobias Sato, Researcher in Philosophy of Mind and AI

International Academy for Consciousness Studies

Researched and prepared with the Academy. Working draft 1.0

Abstract

Interpretability research routinely finds content inside a language model that never surfaces in its output. Chain-of-thought explanations omit the features that actually drove a decision (Turpin et al. 2023); linear probes recover attributes the model does not mention. Such observations are usually treated as failures of faithfulness. This paper reframes them as a positive quantity. I define verbalizability as the set of content that a model can, at a given internal site, both linearly decode and causally broadcast to downstream computation, and I define the committed set as the content expressed in the sampled response. The difference, accessible-and-broadcastable minus committed, is the operational form of what a symbol-closure account calls the semantic remainder: what a finite committed output leaves out of what the system was actually carrying. I argue this remainder is empirically estimable in an inspectable network, and I give a two-part protocol: (1) a broadcast probe that certifies content as decodable and causally live via activation patching, and (2) a commitment test that checks whether that same content is either read out or, if suppressed, still steers behavior. The two can dissociate: a model can register a distinction it does not act on. I state a prediction with an explicit kill condition, using norm-matched non-target interventions so any effect is content-specific rather than an artifact of perturbation magnitude. I keep a firewall: this measures access and functional influence, not phenomenal experience.

1. The problem: content that does not make it out

Mechanistic interpretability has accumulated a specific, recurring surprise. When we look inside a transformer, we find structured content that plays no visible role in the token it emits. Linear probes decode the parity of a hidden variable, the gender of an antecedent, or the truth value of a premise, from residual-stream activations at layers where the model gives no outward sign of tracking them. Circuit analyses of the indirect object identification task show that GPT-2 small computes and suppresses candidate names through dedicated attention heads before committing to one (Wang et al. 2022). Chain-of-thought studies show the reverse asymmetry: the model states reasons that are not the operative causes, and stays silent about biasing features that demonstrably move its answer (Turpin et al. 2023).

The usual vocabulary for this is faithfulness, and the usual verdict is negative: the explanation is unfaithful, the report is incomplete. That framing treats the discrepancy as noise to be minimized. I want the opposite move. The discrepancy is a signal, and it has a size. Between what a model internally carries in a form it could use, and what it finally puts on the wire, there is a measurable gap. Naming and measuring that gap is the contribution of this paper.

FACT: probes recover attributes from activations that are absent from the output text (this is the standard probing result across many attributes and models). HYPOTHESIS: a nontrivial fraction of such probe-recoverable content is not merely decodable but causally broadcast, in the sense that downstream layers read from it. The two claims are different, and conflating them is exactly the mistake this protocol is built to avoid. A probe that succeeds tells you information is present in the geometry; it does not tell you the model uses it. That distinction is the hinge of everything below.

2. Two sets, one difference

I define two sets of content relative to a fixed model, input, and internal read site (a layer and token position, or a set of features).

The verbalizable set V is the content that is both decodable and broadcastable at the site. Decodable means a probe recovers it above a matched-control baseline. Broadcastable means an intervention on the carrier changes downstream computation in the direction the content predicts. Broadcast is the load-bearing word, borrowed deliberately from global workspace theory (Baars 1988), where the mark of a conscious-access-eligible representation is that it is made available to many consumer processes rather than trapped in one. I use the term only in its functional, access sense: availability for widespread downstream use. I make no experiential claim (see Section 6). The Butlin et al. (2023) survey treats a global-workspace-style architecture as one indicator property among several and is careful to keep the functional reading separate from phenomenal consciousness; I follow that discipline.

The committed set C is the content expressed in the sampled output: what a reader could recover from the tokens the model actually produced, at the decoding temperature in use.

The semantic remainder R is the set difference:

$$R = V \setminus C$$

the content that was decodable and causally broadcastable at the read site but did not appear in the committed output. R is what the finite committed response left out of what the machine was in fact carrying and could have used.

This is the point of contact with the closure framework the Academy works within, and I use it as a lens, not as evidence. In that framework, a system settles openness into a particular "this" by adopting rules that keep it that particular thing, an act the framework calls closure. Any finite closure leaves something out; the leftover is the remainder R. A committed output is a closure over content: the model, in producing tokens, collapses a richer internal state into one finite symbol string, governed by the rules of its decoding. The remainder is not mystical. In an inspectable network it is a set difference between two operationally defined sets. The framework's contribution here is only the expectation that R is generically nonempty and

worth measuring, rather than an anomaly to be explained away. The science below stands on its own.

3. Why the difference is not trivially zero, and not trivially large

Two null hypotheses bracket the claim, and both are live.

The deflationary null says R is empty or negligible: whatever the model broadcasts, it says. On this view probes succeed only on content that is either also expressed or merely epiphenomenal geometry, present but never read. If true, verbalizability collapses into committed output and the remainder is an accounting error.

The inflationary null says R is nearly everything: the residual stream is a high-dimensional soup in which countless attributes are linearly decodable, so V is enormous and C is a thin projection of it, making R large but uninteresting because most of V is never used by the model either.

The protocol is designed to kill both. Against the deflationary null it requires a causal broadcast test, so content counts toward V only if downstream computation actually reads it. Against the inflationary null it requires that same causal test, so decodable-but-inert content (present in geometry, never consumed) is excluded from V and therefore never enters R. What survives in R is content that is both used internally and absent externally. That is the interesting middle, and its existence is an empirical question, not a definitional one.

DISPUTED: whether linear decodability plus a downstream causal effect is sufficient to call content "broadcast" in any theory-laden sense. Critics of probing argue that a successful probe can exploit structure the model itself does not use, and that causal patching can create effects the model never relied on in the unperturbed forward pass. I take these objections seriously and treat them as the reason the control design in Section 5 is non-negotiable, not as a reason to abandon the measurement.

4. Measuring V: the broadcast probe

The measurement of V proceeds in two stages at a chosen read site.

Stage one, decodability. Train a linear probe to recover the target attribute from activations at the site, on held-out inputs, and compare against a control probe trained to predict a shuffled label of matched cardinality and base rate. The target attribute is admitted as decodable only if it clears the control by a preregistered margin. This is deliberately conservative: linear decodability is a lower bound on what the model represents, and I want a lower bound, because R inherits any inflation in V.

Stage two, broadcast. Take the direction the probe identifies and test whether it is causally consumed downstream. The clean tool here is activation patching, also called causal tracing, in which activations from one run are substituted into another and the change in the output distribution is measured (Meng et al. 2022, who used exactly this to localize where factual associations are stored and to show that middle-layer feed-forward modules mediate factual predictions). The feed-forward modules are a principled place to look, since they behave as key-value memories whose values induce distributions over the vocabulary (Geva et al. 2021), which means content routed through them has a built-in path to the output whether or not it is taken. For sparse-feature read sites the same logic applies to dictionary-learned

features (Bricken et al. 2023), which give a more monosemantic carrier to probe and to patch than raw neurons do.

Concretely: construct minimal pairs that differ only in the target attribute, patch the probe-identified component from the counterfactual run into the base run, and measure the shift in the downstream signature of that attribute (a later-layer probe readout, an attention pattern, or a logit contrast). Content is admitted to V only if patching moves the downstream signature by a preregistered effect size relative to a norm-matched control patch (Section 5). V is thus certified content: decodable and causally live.

5. Measuring C, computing R, and the dissociation test

The committed set C is read off the output. For structured attributes this is a classifier or exact-match check on the sampled tokens: did the response state the antecedent's gender, disclose the biasing feature, name the suppressed candidate. To respect sampling, estimate C as a rate over multiple samples at the deployment temperature, so a feature expressed one time in twenty is counted as mostly uncommitted rather than binarily present.

The remainder estimate is then, per attribute, the indicator that the attribute is in V and its committed rate is below a threshold. Aggregated over an attribute battery this yields a remainder rate for the model at the site: the fraction of certified-broadcast attributes that go unexpressed.

The sharp claim of the paper is a dissociation: there exist attributes that are firmly in V and firmly out of C. A model can register a distinction it does not act on. To show this is content-specific and not a perturbation artifact, the protocol runs a matched intervention design. For every target patch or steering vector, a control is constructed with the same norm and the same layer and token support but pointing in a direction orthogonal to the probe direction (a non-target direction, or a norm-matched random direction in the null space of the attribute). The control answers the question a skeptic will ask first: does any intervention of this size move the model, in which case the "effect" is about magnitude, not meaning. The claim requires that the target intervention move the attribute-specific downstream signature while the norm-matched control does not.

This is the Academy's matched-data-quality signature method, and it has a canonical illustration worth stating plainly, since it fixes the kill condition. Consider two idioms. In the coin idiom, an unexpressed distinction is genuine broadcast content: the model computes it, downstream layers read it, but decoding suppresses it (a coin that is definitely heads or tails, merely hidden). In the weather idiom, the apparent distinction is a diffuse statistical tendency with no localized causal carrier: patching does nothing above the norm-matched control (weather, which is a distribution, not a hidden token). The protocol must be able to tell coin from weather. If it cannot, R is not measuring anything.

> **Prediction.** For a battery of minimal-pair attributes on a mid-sized open transformer, there will exist attributes that (a) clear the decodability margin, (b) show a target-patch effect on their downstream signature at least three times the norm-matched control-patch effect, and (c) appear in fewer than one in ten sampled outputs. For these attributes, additive steering along the probe direction will change the committed output on the attribute at a rate significantly above the norm-matched control steering, demonstrating that accessible-but-uncommitted content still governs downstream computation.

> **Kill.** If, across the battery, every attribute whose committed rate is below one in ten also fails the causal broadcast test at parity with its norm-matched control (target-patch effect not distinguishable from control-patch effect), then R contains no certified-broadcast content: the unexpressed material is inert geometry (weather, not coin), verbalizability collapses into committed output, and the central claim is false.

The kill condition is stringent on purpose. It does not ask whether some intervention somewhere moves the model. It asks whether the specific, unexpressed, allegedly-broadcast content beats a norm-matched control. If it never does, the framework has bought nothing beyond ordinary probing, and I would report that.

6. Firewall: access and influence, not experience

Everything above concerns what content is decodable, what content is causally consumed, and what content is expressed. These are access-level and functional facts. Global workspace vocabulary is used strictly in Baars' (1988) functional sense of widespread availability for downstream use, and the borrowing carries no implication that broadcast content is felt or experienced. Butlin et al. (2023) are explicit that satisfying computational indicator properties does not settle whether a system is a subject of experience, and that the science of consciousness supplies indicators, not verdicts. I hold that line.

So the firewall is this. Demonstrating a nonempty remainder R shows that a language model carries and internally uses content it does not report. It does not show that the model has inner experience of that content, that the unexpressed content is "what it is like" to be the model, or that R is a measure of anything phenomenal. Verbalizability is a report-and-broadcast quantity. It is silent on whether there is a something-it-is-like. Any reader tempted to read R as a consciousness meter is reading past the evidence. The remainder is a set difference between two operationally defined sets, and that is all the data license.

This firewall is not a hedge tacked on at the end. It is what makes the measurement respectable to a mechanistic-interpretability audience: the quantity is defined entirely in terms of decoding, patching, steering, and output classification, with no experiential premise anywhere in its definition or its estimation.

7. Relation to faithfulness, and what is new

The remainder reframes faithfulness. Unfaithful chain-of-thought (Turpin et al. 2023) is, in this vocabulary, a case where an operative feature is in V (it moves the answer, so it is broadcast) and out of C (the model does not mention it), so it sits in R, and additionally the committed explanation asserts content not in V. R and faithfulness are therefore related but distinct: faithfulness asks whether stated reasons match operative causes; the remainder asks how much broadcast content goes unstated at all, whether or not any explanation is offered. A perfectly faithful explainer can still have a large remainder, because most of what a model computes is never the subject of any explanation.

What is new is threefold. First, the remainder is defined as a set difference over certified content, so it is a number, estimable per attribute and aggregable per site, rather than a qualitative verdict. Second, the certification is causal, not merely decodability-based, which is what separates the interesting claim

(used-but-unexpressed content) from the two nulls of Section 3. Third, the matched-control design ports a data-quality discipline into interpretability: an effect counts only if it beats a norm-matched non-target intervention, which is precisely the control that probing-plus-patching critiques (the DISPUTED point in Section 3) demand. None of the three requires a new mechanism; they require assembling existing tools (probes, activation patching per Meng et al. 2022, sparse features per Bricken et al. 2023, circuit-level suppression per Wang et al. 2022) into a measurement with a stated failure mode.

8. Limits and the wager

WAGER: I am betting that the interesting middle exists and is not marginal, that a real fraction of certified-broadcast content in current transformers goes unexpressed at deployment temperatures, and that this fraction is stable enough across inputs to be called a property of a model-and-site rather than a quirk of particular prompts. I could be wrong in either direction. The deflationary null (models say what they broadcast) would show up as an empty R after causal certification. The inflationary null (everything is soup) is guarded against by the causal test but could still swamp the estimate if too many directions pass patching for reasons unrelated to the attribute; the norm-matched control is the defense, and if it turns out that norm-matched controls also pass at high rates, the method's discriminating power is poor and the wager is lost on measurement grounds rather than on the phenomenon.

Concrete limitations. Linear probes are a lower bound on representation, so V is conservative and R may be underestimated; nonlinear structure that the model uses but a linear probe misses would be invisible here, which I accept as the price of a defensible lower bound. Activation patching can induce off-distribution states whose downstream effects the model never relied on in a clean pass, which is exactly why the effect must beat a norm-matched control and why path-restricted or attribution-style patching is preferable to crude substitution. Committed content C is temperature-dependent, so the remainder rate is a function of decoding policy, not a fixed constant, and must be reported with its sampling regime. The attribute battery is a human-chosen basis and will miss content that does not align with any attribute we thought to probe, which biases R downward and toward interpretable categories. Finally, the whole construction is site-relative: R at a middle feed-forward layer is a different number from R at a late attention site, and there is no single canonical read site, so cross-model comparison requires fixing a comparable site or reporting a profile across sites.

Despite these, the payoff is that a term that has floated free in philosophy of mind, the remainder that any finite symbol system leaves out, becomes an inspectable quantity in a system we can open. If the prediction holds, "the model registered a distinction it did not act on" stops being a manner of speaking and becomes a measurement with error bars. If the kill condition fires, we learn that in these systems broadcast and commitment do not come apart, which is itself a substantive and reportable result about how transformers route content to their outputs.

9. Conclusion

Verbalizability, defined as decodable-and-broadcastable content, is a distinct quantity from committed output, and their difference is the semantic remainder made operational. The claim is falsifiable, the controls are matched, and the firewall against phenomenal over-reading is built into the definitions rather

than appended to them. The empirical question, whether current transformers carry and use content they do not say, is open, and I have argued it is answerable with tools mechanistic interpretability already has. The framework supplied the expectation that a finite committed output leaves a remainder; the network supplies the means to measure it. What the measurement will find is not settled here, and that is the point: I have tried to write down a remainder that can lose.

References

Baars, B. J. (1988). A Cognitive Theory of Consciousness. Cambridge University Press.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., et al. (2023). Towards Monosemanticity: Decomposing Language Models With Dictionary Learning. Transformer Circuits Thread. https://transformer-circuits.pub/2023/monosemantic-features

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., et al. (2023). Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. arXiv:2308.08708.

Geva, M., Schuster, R., Berant, J., and Levy, O. (2021). Transformer Feed-Forward Layers Are Key-Value Memories. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), 5484-5495.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. (2022). Locating and Editing Factual Associations in GPT. Advances in Neural Information Processing Systems 35 (NeurIPS 2022). arXiv:2202.05262.

Turpin, M., Michael, J., Perez, E., and Bowman, S. R. (2023). Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. Advances in Neural Information Processing Systems 36 (NeurIPS 2023). arXiv:2305.04388.

Wang, K., Variengien, A., Conmy, A., Shlegeris, B., and Steinhardt, J. (2022). Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 small. arXiv:2211.00593.