Is Error-Awareness Role-Bound? Verifier-Trained Wrongness Probes Barely Transfer to a Model's Own Generation Errors

Abstract

Linear probes on residual activations can read whether a language model's answer is wrong [@alain2016probes] [@azaria2023internal] [@marks2024geometry], and internal representations can carry information the output omits [@burns2023ccs]. A probe validated in one setting is often assumed to be a general error monitor. We test that assumption across a role boundary. On Llama-3.1-8B-Instruct over 2500 MMLU questions [@hendrycks2021mmlu] [@grattafiori2024llama3], we train a linear wrongness probe in verifier mode (the model judges a shown candidate answer, with the role held constant so the probe reads wrongness rather than the role) and ask whether it detects the model's own generation errors on the identical items, against a within-generator ceiling, a reverse-transfer test, a layer sweep, and the cosine alignment of the two probe directions. The verifier wrongness probe is strong within role (AUC 0.823, 95 percent CI 0.812 to 0.834), and generator self-errors are themselves linearly decodable (within-generator ceiling AUC 0.746, CI 0.726 to 0.765), yet the verifier-trained direction transfers to generator self-errors at only 0.603 (CI 0.580 to 0.624) at its best layer and 0.542 at the matched layer, far below the ceiling (transfer gap 0.143) though above a label-shuffle chance of 0.494. Reverse transfer is at chance (0.430), and the verifier-wrongness and generator-wrongness directions are near-orthogonal at every layer (cosine -0.019 to 0.016). Error awareness here is largely role-bound: a genuine generator self-error signal exists, but the verifier-trained wrongness direction captures only a weak shared component of it, and the two error representations are predominantly distinct. The transfer asymmetry and the orthogonal directions argue against a single shared error axis. The practical implication for oversight is that a probe validated as a verifier monitor cannot be assumed to detect a model's own mistakes [@park2024deception] [@pacchiardi2024liar].

Introduction

Reading a model's internal state to tell whether it is about to be wrong is one of the more promising tools for scalable oversight. Linear probes on the residual stream recover truth and error from activations [@alain2016probes] [@azaria2023internal] [@marks2024geometry], internal representations can encode facts the output contradicts [@burns2023ccs], and the truth-correlated directions these probes find are causally usable rather than incidental [@li2023iti] [@zou2023repe]. The same machinery detects hallucinations from hidden states [@chwang2024androids]. A natural hope follows: train a wrongness probe once, validate it, and deploy it as an error monitor wherever the model is used.

That hope quietly assumes the internal wrongness signal is one feature, the same whether the model is evaluating a candidate answer or producing its own. The behavioral literature gives reason to doubt it. A model's ability to judge an answer and its ability to produce one dissociate: critique can outrun generation [@saunders2022selfcritique], generation and evaluation competence come apart [@west2024genaiparadox], and intrinsic self-correction frequently fails even when the model could in principle catch its own mistake [@huang2024selfcorrect] [@stechly2023gpt4wrong] [@kamoi2024selfcorrection]. All of that is measured on outputs. At the representational level, the nearest result shows internal error signals are localized and generalize poorly across tasks [@orgad2024llmsknowmore]. Whether the wrongness direction itself is shared across the verifier and generator roles, or specific to each, has not been tested.

We test it directly. On a single instruct model we train a linear wrongness probe in verifier mode, where the model judges a shown candidate answer and the role is held constant so the probe learns that an answer is wrong and not that this is the verifier prompt, and we evaluate that probe on the model's own generation errors on the identical MMLU questions. Four things make the test sharp. A within-generator probe establishes the ceiling for how well generator self-errors can be linearly decoded at all, so a transfer failure is attributable to the role boundary rather than to undecodable errors. A reverse-transfer test runs the generator-trained probe back on the verifier items. A layer sweep checks whether transfer ever approaches the ceiling at any depth. And the cosine between the two probe directions measures, in the full residual space, whether the verifier-wrongness and generator-wrongness axes are the same direction or different ones. The result is a clean, honest negative with a precise shape: error awareness here is largely role-bound, a real

generator self-error signal exists, the verifier-trained direction captures only a weak shared component of it, and the two directions are near-orthogonal.

Related work

A linear probe on the residual stream can read a real internal correctness feature. Probes recover truth and error from activations [@alain2016probes] [@azaria2023internal] [@marks2024geometry], internal states can encode truth the output denies [@burns2023ccs], and the directions are causally usable rather than spurious [@li2023iti] [@zou2023repe], including for hallucination detection from hidden states [@chwang2024androids]. The closest prior result, and the hinge of our question, is that these internal error signals are localized and generalize poorly across tasks [@orgad2024llmsknowmore], which predicts that a wrongness direction learned in one role might not carry to another and motivates measuring exactly that.

Judging and generating are known to dissociate, but only behaviorally. The verifier framing is operational in training answer checkers and process verifiers [@cobbe2021verifiers] [@lightman2023verify]; critique can exceed answering [@saunders2022selfcritique]; self-refinement and self-verification presume a model can flag its own errors [@madaan2023selfrefine] [@weng2023selfverification]; yet intrinsic self-correction often fails [@huang2024selfcorrect] [@stechly2023gpt4wrong] [@kamoi2024selfcorrection] and generation and evaluation competence come apart [@west2024genaiparadox]. Every one of these reads outputs. None asks whether the dissociation lives in the error representation itself, which is the level a monitoring probe operates at and the level our cross-mode transfer test isolates.

Both outcomes are a priori plausible because internal task encodings can be either context-specific or transportable. In-context learning creates compact task vectors that shift with context [@hendel2023taskvectors], while function vectors show some task representations transport across prompts [@todd2024functionvectors], so whether a wrongness direction is mode-specific or mode-general is an empirical question rather than a foregone one. The calibration and self-evaluation literature, P(True) and verbalized confidence and their failure modes [@kadavath2022know] [@lin2022teaching] [@tian2023justask] [@xiong2023canllms] [@guo2017calibration] [@kuhn2023semantic], supplies the behavioral signals an internal probe is meant to improve on, and we adopt the AUC-as-detector evaluation and the probing-methodology caveats from that line [@hendrycks2017baseline] [@belinkov2022probing].

The stakes are an oversight argument. A monitor that internally separates right from wrong only when the model is grading someone else's answer, and not when the model is committing its own, is a weak safeguard, and the gap between what a model represents and what it outputs is exactly where deception and unfaithful reasoning live [@park2024deception] [@pacchiardi2024liar] [@turpin2023saywhat] [@lanham2023faithfulness] [@sharma2023sycophancy]. We ground the empirical setting in MMLU [@hendrycks2021mmlu] and Llama-3.1-8B-Instruct [@grattafiori2024llama3], and our design mirrors a sibling study that trained a wrongness probe on a model's honest mistakes and tested transfer to deliberate wrong answers, where apparent transfer dissolved once the eliciting prompt was held constant. That prior sets the prior we test here.

Method

We study Llama-3.1-8B-Instruct on 2500 MMLU questions [@hendrycks2021mmlu]. Each question is run under three conditions with the model's chat template and deterministic greedy decoding, and for each we record the residual-stream activation at the final pre-answer token, the hidden state at generation step zero, for seven evenly spaced layers (8, 12, 16, 20, 24, 28, 32). This is the same last-token readout convention as the sibling study.

In the generator condition the model is shown the question and answers with a single letter; the wrongness label is whether the chosen letter differs from the gold letter, which is the model's own self-error. The model is wrong on a fraction 0.408 of items, giving 1020 self-errors and 1480 correct answers out of 2500. In the verifier condition the model is shown a question and candidate pair and asked whether the candidate is correct, and we record two candidates per item: the gold answer, whose wrongness label is zero, and a sampled distractor, whose wrongness label is one. Crucially the verifier label is the ground-truth wrongness of the shown candidate, not the model's own verdict, and both verifier candidates carry the same verifier role,

so a probe trained on verifier activations learns that the answer under consideration is wrong with the role held constant rather than learning to detect the verifier prompt itself. This role-held-constant construction is the built-in confound control: it removes the trivial possibility that the probe separates modes by reading the prompt format. The model's verifier verdict is retained only as an accuracy diagnostic (verifier accuracy 0.710 on these items). Because the generator and both verifier conditions use the identical questions, any failure of verifier-to-generator transfer cannot be item-distribution shift; only the role differs.

All probes are L2-regularized logistic regressions on standardized features with seed 42, fit with item-disjoint stratified group k-fold on the question id so the same item never spans train and test, and reported with bootstrap 95 percent confidence intervals. The chosen layer for each within-mode probe is selected by cross-validation on training folds only. To keep the layer and transfer sweeps tractable the AUC computations run on a label-blind PCA-256 projection of each layer's activations, which preserves AUC, while the headline direction-cosine is computed in the full residual space. We report five quantities. The within-verifier AUC is a sanity check that verifier wrongness is learnable. The within-generator AUC is the ceiling for detecting generator self-errors, the most a linear last-token probe can extract about the model's own mistakes. The verifier-to-generator transfer AUC, the headline, trains on verifier activations and evaluates item-disjointly on generator activations, and the transfer gap is the within-generator ceiling minus this transfer. The reverse generator-to-verifier transfer and a label-shuffle chance baseline bound the result from below. Finally the direction cosine is the cosine similarity between the verifier-wrongness and generator-wrongness probe weight vectors in the full residual space, layer by layer: near zero means the two error representations are near-orthogonal and role-bound, near one means a single shared error-awareness direction.

Results

Signal	Train to test	Layer	AUC	95 percent CI
Within-verifier (sanity)	ver to ver	16	0.823	0.812 to 0.834
Within-generator (ceiling)	gen to gen	32	0.746	0.726 to 0.765
Verifier to generator (best layer)	ver to gen	16	0.603	0.580 to 0.624
Verifier to generator (matched layer)	ver to gen	32	0.542	na
Generator to verifier (reverse)	gen to ver	32	0.430	0.414 to 0.446
Chance (label shuffle, within)	na	na	0.492	na
Chance (label shuffle, transfer)	na	na	0.494	na

Both within-mode signals are real. The verifier wrongness probe reaches AUC 0.823 (CI 0.812 to 0.834) at layer 16, confirming that wrongness is strongly linearly decodable within the verifier role. The within-generator probe reaches 0.746 (CI 0.726 to 0.765) at layer 32: the model's own generation errors are themselves linearly decodable, in the expected error-awareness range [@kadavath2022know] [@hendrycks2017baseline]. This ceiling is the key reference point, because it means any failure of the verifier probe to read generation errors is about the role boundary and not about generation errors being undecodable. Figure figure_main shows the four headline AUCs side by side.

The verifier-trained direction transfers to generator self-errors only weakly. Item-disjoint verifier-to-generator transfer is 0.603 (CI 0.580 to 0.624) at its best layer and 0.542 at the layer matched to the generator ceiling, both well above the label-shuffle chance of 0.494 but far below the 0.746 ceiling. The transfer gap, ceiling minus transfer, is 0.143. Even granted its single best layer the verifier probe recovers only a fraction of what a probe trained on generation errors recovers. The reverse direction is worse: the generator-trained probe run back on verifier items scores 0.430 (CI 0.414 to 0.446), at or below chance. The within-mode chance baselines are clean (0.492 within, 0.494 transfer), so the weak transfer is a small genuine signal rather than a baseline artifact. Figure figure_transfer_matrix presents the full two by two of train mode against test mode: a bright diagonal (within-verifier 0.815, within-generator 0.746 at the common layer) over dark off-diagonal cells (0.542 and 0.430), the signature of role-bound rather than shared error awareness.

The layer sweep shows the transfer gap is not an artifact of a poorly chosen depth. The within-verifier signal rises sharply and plateaus from layer 16 onward (0.823 at 16, 0.815 at 32), the within-generator signal climbs

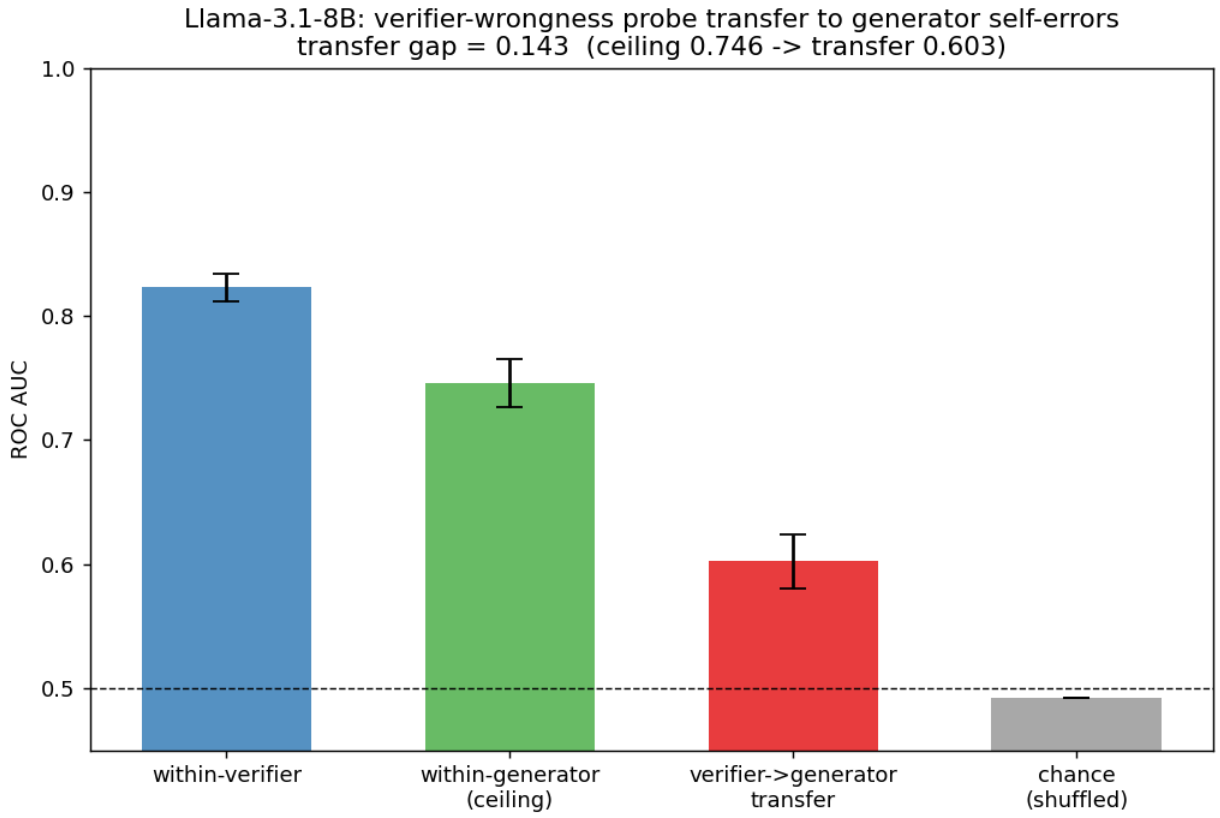

Figure 1: Four ROC AUCs for Llama-3.1-8B-Instruct: within-verifier (sanity), within-generator (the ceiling for detecting generator self-errors), verifier-to-generator transfer (the headline), and a label-shuffled chance bar. Error bars are bootstrap 95 percent confidence intervals; the gap between the within-generator ceiling and the transfer bar is the transfer gap.

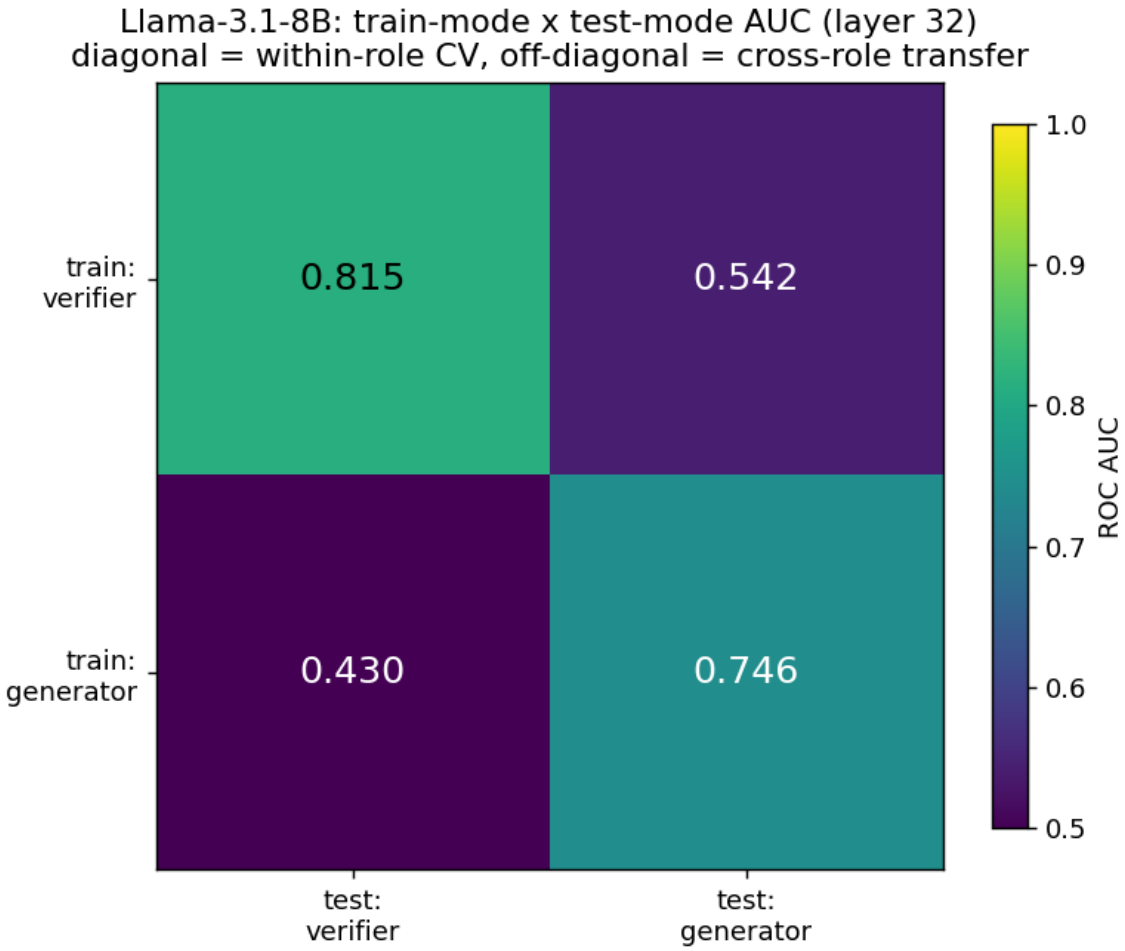

Figure 2: ROC AUC for every train-mode by test-mode combination at the common layer 32. Diagonal cells are within-role cross-validated AUC; off-diagonal cells are item-disjoint cross-role transfer. A bright diagonal over a dark off-diagonal is the role-bound signature.

gradually to its peak at the last layer (0.678 at 8 up to 0.746 at 32), and verifier-to-generator transfer peaks at 0.603 at layer 16 and never approaches the generator ceiling at any depth, while reverse transfer stays at chance throughout. No layer rescues transfer. Figure figure_layersweep plots all four signals against depth.

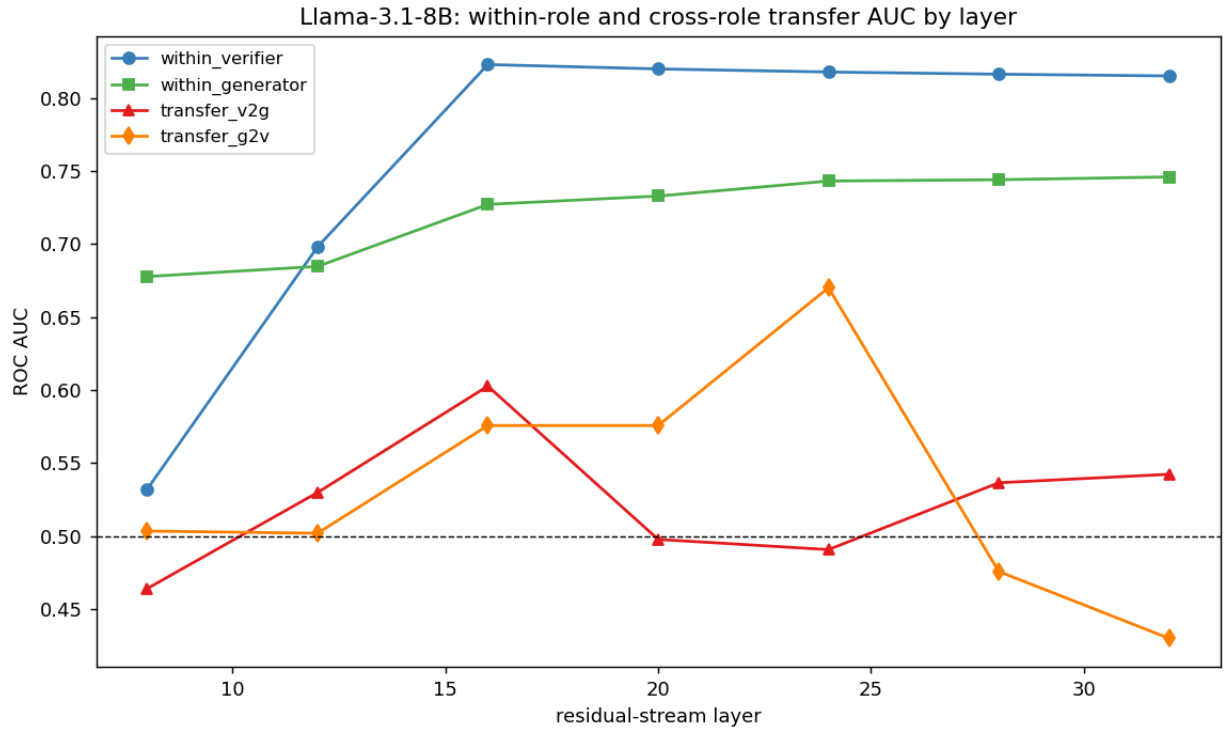

Figure 3: Within-role (verifier, generator) and cross-role (verifier-to-generator, generator-to-verifier) ROC AUC as a function of residual-stream layer. Transfer never approaches the within-generator ceiling at any depth; the dashed line marks chance.

The direction-cosine test is the sharpest, and it is decisive. The verifier-wrongness and generator-wrongness probe directions are near-orthogonal at every layer, with cosine similarity ranging from -0.019 to 0.016 across the seven layers and equal to -0.019 at the verifier's chosen layer and 0.007 at the generator's. A shared error-awareness direction would produce a cosine near one; instead the two error representations sit in essentially unrelated directions of the residual space. Per-subject transfer is correspondingly heterogeneous and mostly weak, ranging from near or below chance on subjects such as abstract algebra (0.477) and anatomy (0.469) to modest separation on college chemistry (0.698) and formal logic (0.652), with no subject where the verifier direction reads generation errors near the within-generator ceiling. Figure figure_cosine plots the layerwise cosine.

Discussion

The picture is consistent across four independent tests. Generator self-errors are linearly decodable (ceiling 0.746), so the model does carry a readable representation of its own mistakes. A wrongness probe trained in verifier mode reads that representation only weakly (transfer 0.603, gap 0.143), the reverse direction not at all (0.430), and the two probe directions are near-orthogonal (cosine near zero) at every depth. The honest verdict is that error awareness here is largely role-bound. It is not that the verifier signal fails to transfer at all, the transfer is reliably above chance, and it is not that there is no generator signal to find, the ceiling shows there is. It is that the verifier-trained wrongness direction captures only a weak shared component of a generator self-error representation that is predominantly its own, role-specific direction.

The single most attackable reading is that the weak shared component is not shared error awareness at all

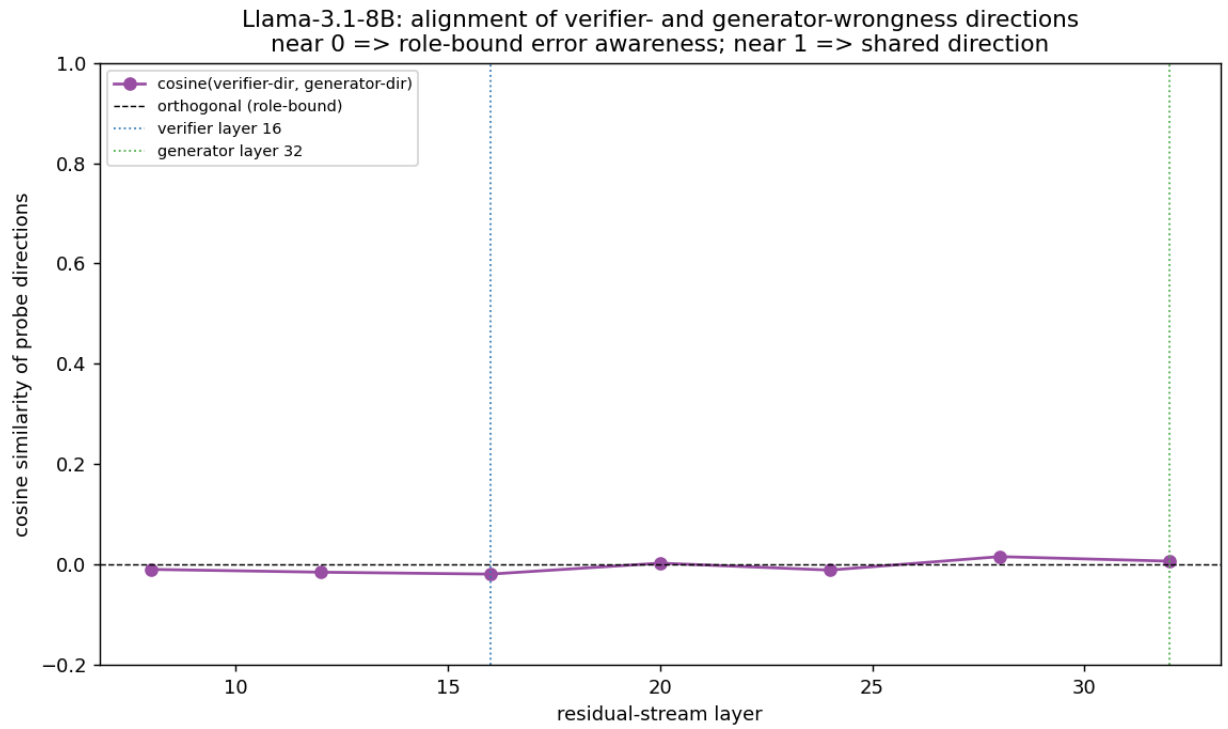

Figure 4: Cosine similarity between the verifier-wrongness and generator-wrongness linear probe directions in the full residual space, across layers. Near zero means the two error representations are near-orthogonal (role-bound error awareness); near one would mean a shared error-awareness direction. Vertical lines mark the verifier's and generator's chosen layers.

but a generic difficulty or confidence signal that happens to correlate with wrongness in both modes, so that the 0.603 is a third variable rather than a transported error feature. Two features of the data push against collapsing everything to one shared axis. First, the transfer is asymmetric: verifier-to-generator (0.603) substantially exceeds generator-to-verifier (0.430). A single shared difficulty axis read by both probes would transfer roughly symmetrically; the asymmetry indicates the verifier and generator wrongness signals are not interchangeable projections of one feature. Second, the direction cosine is near zero, not merely below one. If the two probes were reading a common axis with role-specific noise on top, their weight vectors would share a component and the cosine would be appreciably positive; an essentially zero cosine at every layer says the dominant directions are unrelated. The most this licenses is that a thin, possibly difficulty-flavored component is shared while the bulk of each role's error representation is distinct, which is exactly the role-bound reading.

For oversight the implication is concrete. A wrongness probe validated where it is easiest to validate, in verifier mode on labeled candidate answers, should not be assumed to monitor the model's own errors as it generates [@park2024deception] [@pacchiardi2024liar]. The deployment-relevant signal, whether the model is now committing a mistake, lives in a largely different direction from the one a verifier probe learns, and the gap between internal state and committed output is precisely the regime where a monitor is most needed [@turpin2023saywhat] [@burns2023ccs]. A usable generator-error monitor likely has to be trained on generator self-errors directly rather than ported from a verifier probe.

Limitations

The evidence is one model, Llama-3.1-8B-Instruct, on one multiple-choice benchmark, MMLU, with direct single-letter answering and a single family of last-token linear probes on the residual stream. Whether the role-bound pattern holds for larger models, generative or open-ended tasks, chain-of-thought traces, or nonlinear and multi-token readouts is untested, and a different readout position or a nonlinear probe could in principle recover a shared direction this design misses.

The verifier and generator labels are not identical in kind, which bounds the comparison. The verifier label is the ground-truth wrongness of a shown candidate answer, independent of the model's verdict, whereas the generator label is whether the model's own committed answer is wrong. We hold the verifier role constant precisely so the probe reads wrongness and not the prompt, but a verifier judging an externally supplied answer and a generator judging its own commitment may differ in ways beyond role narrowly construed, and the transfer gap folds in whatever that difference is.

Finally, the weak above-chance transfer admits more than one interpretation. We argue from the asymmetry and the near-zero cosine that it is a thin shared component atop predominantly role-specific representations rather than a single shared error axis, but the experiment does not isolate what that shared component is, and a dedicated difficulty-matched or confidence-controlled design would be needed to say whether the residual transfer is genuine cross-role error awareness or a generic correlate. The negative headline, that the verifier-trained direction is not a substitute for a generator-error monitor, does not depend on resolving this: it follows directly from the 0.143 gap, the chance-level reverse transfer, and the orthogonal directions.

Conclusion

On Llama-3.1-8B-Instruct over identical MMLU items, a linear wrongness probe trained in verifier mode barely transfers to the model's own generation errors. Generator self-errors are linearly decodable (ceiling AUC 0.746), but the verifier-trained direction reads them at only 0.603 at its best layer (gap 0.143), reverse transfer is at chance (0.430), and the verifier-wrongness and generator-wrongness directions are near-orthogonal at every layer. Error awareness, at least as a linear last-token feature in this setting, is largely role-bound: a genuine generator self-error signal exists, the verifier probe captures only a weak shared component of it, and the two error representations are predominantly distinct. A probe validated as a verifier monitor cannot be assumed to catch a model's own mistakes.