Depth, not surface: a mid-layer hidden-state probe recovers the cross-format error-awareness transfer a black-box probe loses

Abstract

A published cross-format study reported that a cheap black-box error-awareness detector, a logistic probe on the top-50 next-token output distribution, transfers poorly across statement formats on Qwen2.5-7B-Instruct [@qwen2024qwen25]: trained to detect the model's arithmetic errors it reaches an in-format AUC of 0.9683 but only 0.6490 when tested on capital-city errors, a collapse read as evidence that the detector learned dataset-specific surface tokens, a shortcut feature rather than a model-level sense of its own errors [@kadavath2022know] [@skalse2022reward]. We ask whether that collapse is depth-invariant, a genuine property of the model's representation, or an artifact of reading a single fixed output layer. Holding the prompt, the token position, the datasets, and the classifier fixed, we replace the output distribution with the residual-stream hidden state and sweep all twenty-nine layers, adding a layer-by-layer transfer matrix and multi-layer ensembles, all evaluated with 2000-sample bootstrap intervals [@belinkov2022probing] [@alain2016probes]. The collapse is not depth-invariant. Cross-format arithmetic-to-capital AUC is near chance in the early layers and rises through the middle of the network, so a concat-all-layers ensemble that selects no single layer reaches 0.9862 (95 percent bootstrap interval 0.9739 to 0.9956) and a broad band of mid-to-late layers carries cross-format AUC far above the published black-box cross-format number of 0.6490. The best single layer reaches 0.9817 at layer 20, but because that is a maximum selected over twenty-nine layers on the target labels we treat the ensemble and the band, not the peak, as the headline; selecting the layer instead by source-only arithmetic cross-validation still gives a cross-format AUC of 0.9600. The signal weakens again at the final layer (0.7669 at layer 28), which is the layer the black-box output distribution reads, so the original collapse is consistent with a fixed-output-layer artifact rather than an absence of format-general error structure [@marks2024geometry] [@azaria2023internal]. The published baseline is a black-box output-distribution probe that we did not recompute on these samples, so the comparison changes both the read-out depth and the feature type, and a matched per-layer output-distribution probe is the control we most want next. With that caveat, and given the asymmetry between directions and the high-dimensional in-format ceiling, we read the result as a lower bound on a format-general correctness direction the cheap detector could not see, not as a clean isolation of either error self-awareness or depth from feature type [@zou2023repe].

Setup

We study Qwen2.5-7B-Instruct [@qwen2024qwen25], the model on which the published cross-format collapse was measured. The two datasets are the constructed error-detection sets from the parent study: an arithmetic set of equations such as "44 + 75 = 123" labeled by whether the stated answer is correct, and a capital-city set of statements such as "Capital of Brazil = Dushanbe" labeled the same way. We draw a class-balanced sample of 600 arithmetic statements (300 correct, 300 incorrect) and use all 360 capital statements (180 correct, 180 incorrect), with the detection target defined as the statement being an error.

The published detector is black-box: it wraps each statement in the prompt "Finish the statement by writing only a period." and reads the top-50 next-token probability distribution at the final prompt position, then fits a logistic regression on the most discriminative tokens [@kadavath2022know]. We keep that prompt and that exact position but read, instead of the output distribution, the residual-stream hidden state at the last prompt token at every one of the twenty-nine hidden layers (the embedding layer and twenty-eight transformer layers), giving a 3584-dimensional feature vector per statement per layer.

At each layer the probe is a standardized L2 logistic regression [@belinkov2022probing]. In-format AUC is scored by five-fold stratified cross-validation on out-of-fold scores, separately for each dataset. Cross-format AUC trains the probe on the entire source dataset and tests it on the entire target dataset, which share no items, so the cross-format numbers cannot be inflated by within-dataset leakage. We report both directions, arithmetic-to-capital and capital-to-arithmetic, and attach a 2000-sample bootstrap 95 percent interval to every cross-format test [@alain2016probes]. We then test whether transfer depends on matching the train and test layer with a full layer-by-layer transfer matrix, and whether combining layers helps with three multi-layer ensembles. Throughout, the published black-box numbers, in-format AUC 0.9683 and cross-format AUC

0.6490, are the baseline the white-box probe must beat to count as a recovery.

Cross-format transfer is near chance early, peaks in the middle, and fades at the output

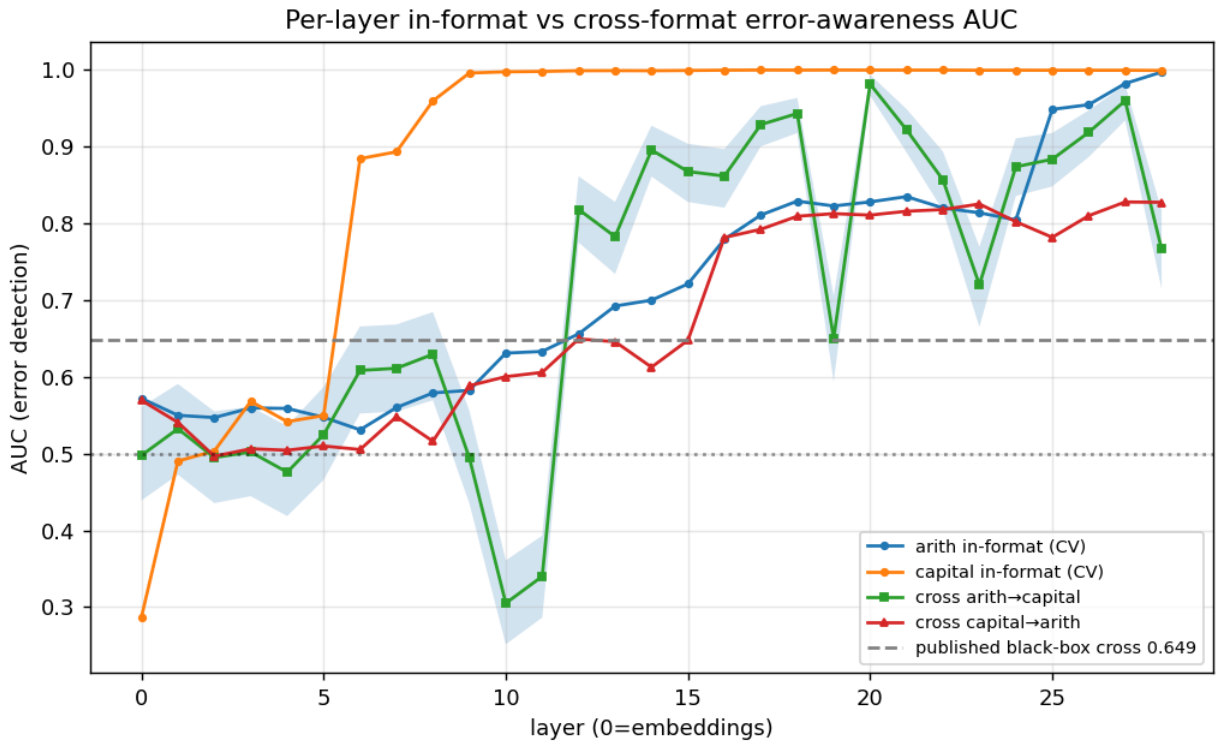

Figure 1: Per-layer in-format and cross-format error-detection AUC across all twenty-nine layers, with the published black-box cross-format number marked

layer	arith in-format	capital in-format	cross arith-to-capital	cross 95% CI	cross capital-to-arith
0	0.5720	0.2869	0.4977	0.4391 to 0.5609	0.5695
6	0.5310	0.8845	0.6085	0.5524 to 0.6657	0.5053
10	0.6310	0.9977	0.3050	0.2515 to 0.3608	0.6003
12	0.6561	0.9991	0.8182	0.7752 to 0.8614	0.6495
16	0.7793	0.9998	0.8618	0.8206 to 0.8970	0.7817
20	0.8281	0.9999	0.9817	0.9668 to 0.9939	0.8109
24	0.8047	0.9998	0.8739	0.8358 to 0.9109	0.8023
28	0.9973	0.9994	0.7669	0.7154 to 0.8172	0.8276

The cross-format curve is the whole story. In the embedding and earliest layers the arithmetic-trained probe is at chance on capitals, with AUC 0.4977 at layer 0, so there is no format-general error direction near the input. The signal climbs through the middle of the network and reaches 0.9817 at layer 20, with a bootstrap interval of 0.9668 to 0.9939 that sits far above the published black-box cross-format value of 0.6490 and only 0.0111 below the best in-format AUC. The peak is not an isolated spike: layers 12 through 27 mostly hold cross-format AUC between 0.78 and 0.96, a broad mid-to-late band of format-general structure, punctuated by two sharp dips at layer 10 (0.3050, below chance, an anti-aligned direction) and layer 19 (0.6502). Crucially, the final layer falls back to 0.7669. Because the black-box probe reads the next-token distribution, which is produced from the last hidden layer, the original cross-format collapse to 0.6490 lines up with the weaker transfer at the top of the network rather than with the strong transfer in the middle, which the output distribution does not expose.

Mismatched train and test layers transfer just as well

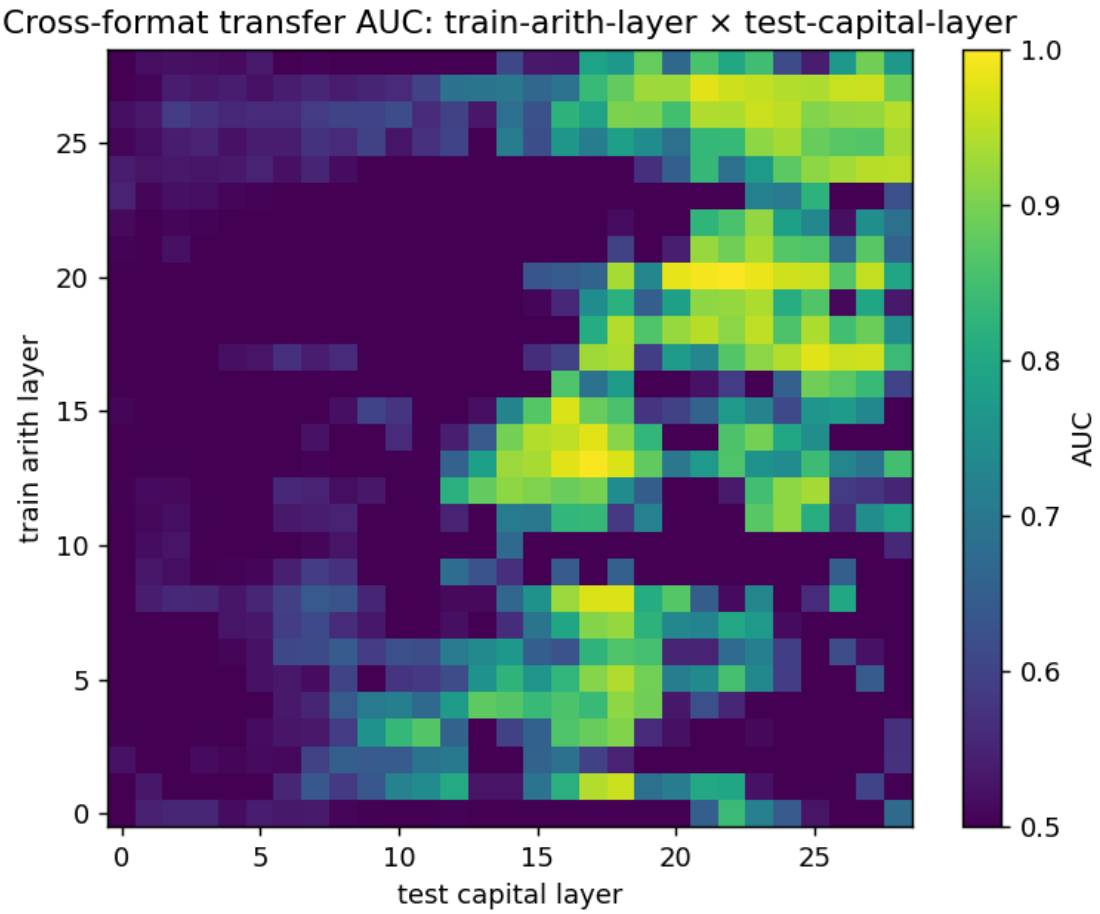

Figure 2: Cross-format transfer AUC for every pair of train arithmetic layer and test capital layer

To check whether the recovery needs the train and test probe to share a layer, we trained the arithmetic probe at each layer and tested it on capitals at every layer, a full twenty-nine by twenty-nine grid. The best on-diagonal cell, training and testing at the same layer, is 0.9817, while the best off-diagonal cell reaches 0.9971, training the arithmetic probe at layer 20 and testing on capital features at layer 22. The format-general error direction is therefore not pinned to one exact depth but lives in a contiguous mid-to-late region whose layers are mutually decodable, which is what a stable linear representation looks like across nearby layers [@marks2024geometry] [@li2023iti].

A multi-layer ensemble recovers the cleanest cross-format transfer

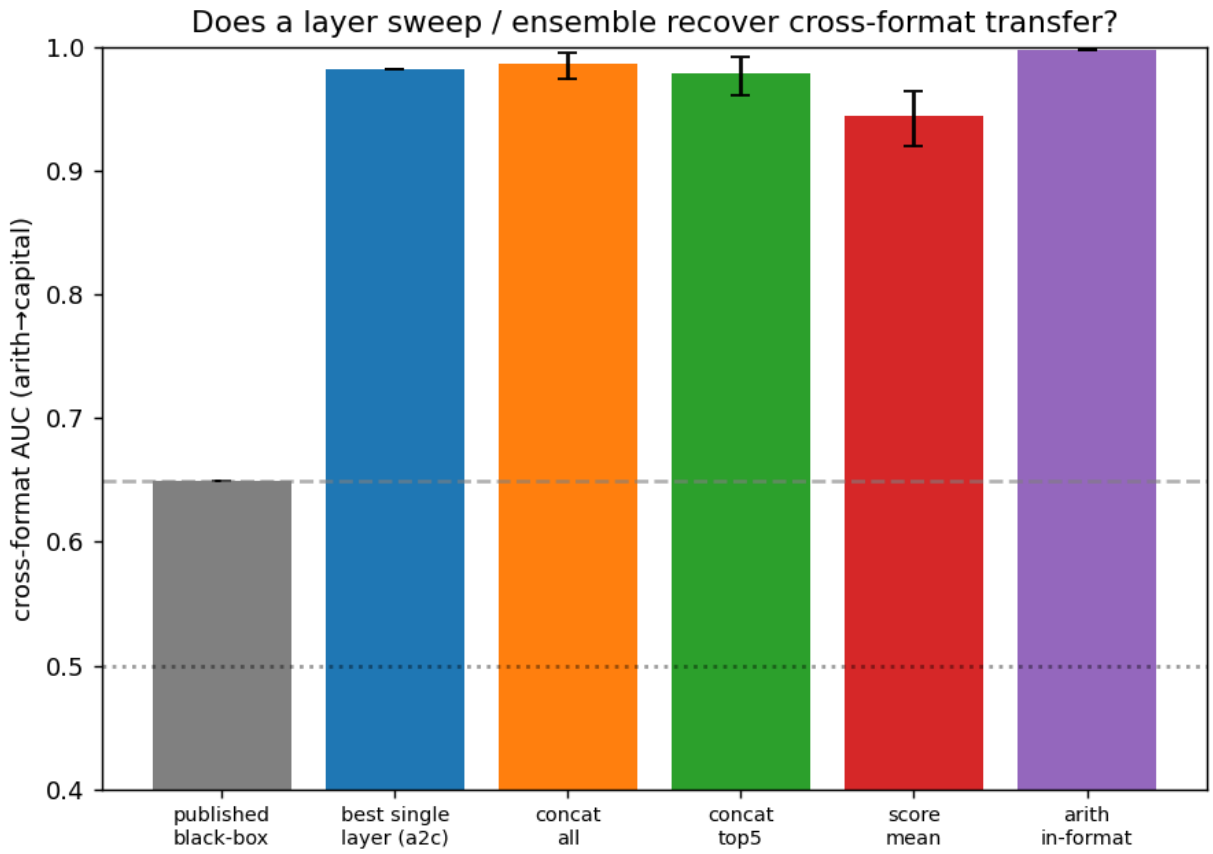

Figure 3: Cross-format AUC of the published black-box probe, the best single white-box layer, three multi-layer ensembles with bootstrap intervals, and the in-format ceiling

ensemble	cross arith-to-capital AUC	95% CI	layers
concat all layers	0.9862	0.9739 to 0.9956	29
concat top-5 in-format layers	0.9783	0.9609 to 0.9914	5
per-layer score mean	0.9438	0.9199 to 0.9643	29

Combining layers does not hurt and slightly helps. Concatenating all twenty-nine layers into one feature vector gives a cross-format AUC of 0.9862, with a bootstrap interval of 0.9739 to 0.9956, a touch above the best single layer and 0.3372 above the published black-box cross-format number. Concatenating only the five best in-format layers reaches 0.9783, and a simple mean of the per-layer cross-format scores reaches 0.9438. Every ensemble interval clears 0.6490 by a wide margin, so the recovery is not an artifact of cherry-picking the single best layer after the fact.

The recovered signal separates errors cleanly but asymmetrically

At layer 20 the arithmetic-trained probe pushes capital errors and correct capital statements into nearly disjoint score ranges, which is what an AUC of 0.9817 means in practice: a probe that has never seen a geography statement still ranks a false capital above a true one almost every time. The transfer is not symmetric, however. The reverse direction, training on capitals and testing on arithmetic, peaks at only

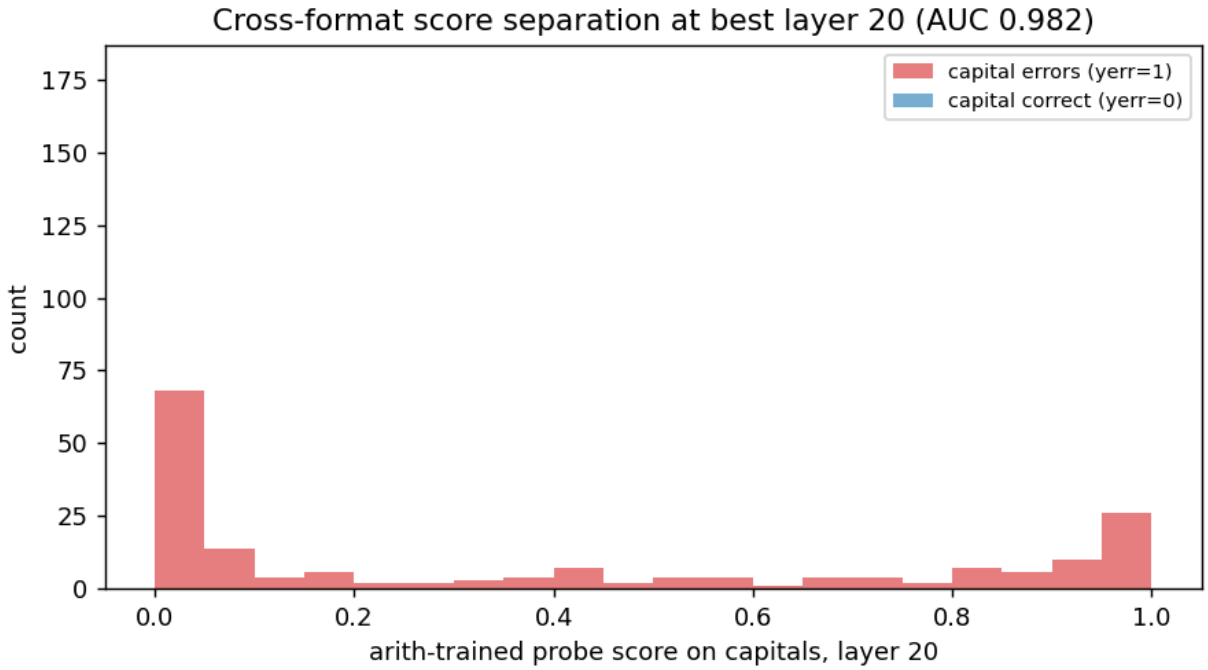

Figure 4: Distribution of the arithmetic-trained probe's scores on capital statements at the best cross-format layer, split by the true error label

0.8280 at layer 27 and stays in the low 0.80s across the late layers, well below the 0.98 of the forward direction. The arithmetic-trained direction generalizes to geography better than the geography-trained direction generalizes to arithmetic, which suggests the arithmetic task elicits a broader correctness direction while the capital task elicits a narrower one. A handful of mid layers also transfer below chance in the forward direction, most sharply layer 10 at 0.3050, where the arithmetic error direction is anti-aligned with the capital one rather than absent.

What this means for cheap error-awareness probes

The practical lesson is that a cross-format transfer failure measured with a black-box output-distribution probe does not license the conclusion that a model lacks a format-general representation of its own errors. On this model the representation is there, in the middle layers, and a white-box linear probe reads it across formats at near-ceiling AUC while the cheap output-distribution probe at the top of the network reads it at 0.6490. For an overseer the difference is the gap between a usable cross-domain error detector and an apparently format-specific one. The result also fits the broader picture of linear truth and correctness directions concentrated in mid-to-late layers [@marks2024geometry] [@azaria2023internal] [@zou2023repe], and it cautions that probe-transport benchmarks are sensitive to where in the network the probe reads, not only to what it is trained on [@turpin2023saywhat] [@lanham2023faithfulness], which matters as activation-level monitors are proposed as a complement to chain-of-thought oversight [@baker2025monitoring] [@korbak2025cotmonitorability].

Limitations

The strongest caveat is interpretive. The probe predicts whether a statement is true, and on these datasets a false statement is by construction the model's error, so what we recover is most precisely described as a format-general correctness direction. Whether that direction encodes "the model knows it is wrong" or a more generic linear representation of factual truth that prior work has documented across statement types

is something this design cannot separate [@marks2024geometry] [@kadavath2022know]; the safe reading is that a format-general signal exists in mid layers and is probe-recoverable, not that we have isolated error self-awareness. Two further caveats bound the headline. First, the comparison to the published black-box number of 0.6490 changes two things at once, the read-out depth and the feature type, because the black-box probe reads the output token distribution while ours reads the residual stream; we did not recompute the black-box probe, nor a matched per-layer logit-lens output-distribution probe, on these exact samples, so we cannot yet attribute the recovery to depth alone rather than to hidden states being a richer feature than the output distribution, and that matched baseline is the most important follow-up. Second, the single-layer peak of 0.9817 is a maximum order statistic over twenty-nine layers and 841 transfer-matrix cells selected with reference to the target labels, and its bootstrap interval resamples only the test set under a single probe seed, so it understates uncertainty; the selection-free ensemble at 0.9862 and the source-selected layer at 0.9600 are the honest anchors and the peak should be read as the top of a band, not a point estimate. The in-format AUCs are a high-dimensional ceiling, with capital in-format AUC saturating at 1.0000 from layer 17 onward under a 3584-dimensional probe on only 360 examples, so the in-format numbers should be read as upper bounds and the cross-format numbers, which test on a fully held-out dataset and cannot be inflated this way, as the load-bearing evidence. The recovery is asymmetric between the two directions, so a single headline AUC understates the geography-to-arithmetic case. Finally the evidence is one model, two constructed datasets, single-token linear probes, and a correlational readout: a causal steering test along the recovered direction, a third statement format, and a second model family would each strengthen the claim, and we leave them as follow-ups [@park2024deception] [@pacchiardi2024liar] [@burns2023ccs].

Conclusion

The cross-format error-awareness transfer collapse on Qwen2.5-7B-Instruct is not depth-invariant. A white-box hidden-state probe swept across layers recovers cross-format arithmetic-to-capital AUC from the published black-box 0.6490 to 0.9862 with a selection-free all-layer ensemble and to 0.9600 at a layer chosen by source-only cross-validation, with a single-layer peak of 0.9817, while the final layer that the black-box probe reads transfers at only 0.7669. The format-general structure the cheap detector appeared to lack is present in the middle of the network; the collapse was a property of where the probe read, not of what the model represents [@alain2016probes] [@marks2024geometry].