The Cross-Format Advantage of Contrastive Error-Awareness Probes Does Not Replicate Across Models

Ephraiem Sarabamoun, Independent Research, ephraiemsam16@gmail.com

Abstract

A model's hidden states carry a linear trace of whether its own statement is true or false, and a cheap probe can read it out [@azaria2023internal; @burns2022discovering]. Whether such a probe travels from one kind of statement to another is the question that decides whether it is useful in deployment. We compare two readouts of error awareness at the commitment position, a standard L2 logistic regression on standardized hidden states and a contrastive mass-mean direction, training each on arithmetic true-or-false statements and testing cross-format on capital-city statements in Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct. The first thing to say is the positive one: both probes transfer cross-format well above chance on both models, with cross-format AUC between 0.8877 and 0.9873, so neither readout collapses. What does not replicate is the contrastive probe's advantage over the standard probe. On Llama the contrastive readout beats the standard one by a paired difference of +0.0362 with a 95 percent interval of [0.0160, 0.0609], and zero of 2000 bootstrap resamples favoured the standard probe (p=0.0000 to resampling resolution, that is p below 0.0005). On Qwen the sign reverses to -0.0401 with a 95 percent interval of [-0.0743, -0.0078]. Two models of comparable scale give contrastive advantages of opposite sign, and both intervals exclude zero. The reversal is not an artifact of comparing different layers: it holds when both probes are read at a common layer, and the Qwen comparison is already same-layer. We also report a sharper honesty check. At several mid layers the arithmetic-trained probe scores higher cross-format on capitals than in-format on arithmetic, which is impossible if it is reading a transferred truth direction and instead signals a surface cue in the capitals format that the arithmetic direction happens to align with. We therefore base our conclusions on the principled in-format-selected layer rather than mid-layer maxima, and we read the result as an honest cautionary negative for the probing and representation-reading literature: claims about which readout transfers best should be paired across models, and across layers and preprocessing, before they are believed.

Introduction

Several lines of work establish that a language model keeps an internal record of whether its own output is correct. Hidden activations linearly encode whether a produced statement is true or false [@azaria2023internal]. An unsupervised method recovers a truth direction from the geometry of contrasting activations [@burns2022discovering]. Models can estimate the probability that their own answers are correct with usable calibration [@kadavath2022language]. These results motivate an inexpensive error detector: run one forward pass, read a direction in activation space, and flag statements the model itself treats as wrong. Such detectors would matter precisely because capable models assert falsehoods confidently [@lin2021truthfulqa].

The catch is generalization. A detector is trained on whatever data is convenient and then trusted on data that is not, and recent work finds that activation-based error detectors transfer poorly across tasks and may track features other than truth [@orgad2024llms; @levinstein2023still; @farquhar2023challenges]. The sharpest form of this concern is cross-format transfer, where a probe trained to judge one kind of true-or-false statement is asked to judge another kind that shares the underlying truth task but almost none of its surface vocabulary. An earlier study in this setting trained a token-distribution probe on arithmetic statements and tested it on capital-city statements in Qwen2.5-7B-Instruct [@qwen2025qwen25]. The probe scored near ceiling within arithmetic and within capitals, but collapsed across the two, reaching only about 0.649 cross-format AUC, and the published verdict was that the model's token-level error awareness is largely format-specific. A subsequent replication on Llama-3.1-8B-Instruct [@grattafiori2024llama], with byte-identical data and an unchanged pipeline, found no such collapse, locating the difference in the model rather than the method.

Two threads suggest that the collapse might be tied to the readout rather than the representation. First, the token-distribution probe reads the next-token distribution at the commitment position, a thin and surface-

coupled feature, whereas the hidden state at the same position carries far more. Second, the standard probe is a fitted logistic regression, and Marks and Tegmark argue that the mean-difference, or mass-mean, direction isolates a more causal and more general truth direction than a fitted classifier, because it averages away nuisance variation rather than exploiting it [@marks2023geometry]. The same difference-of-means construction underlies a broad family of representation-reading and steering directions [@li2023inference; @zou2023representation; @tigges2023linear]. Put together, the threads predict that a contrastive mass-mean probe on hidden states should transfer across formats at least as well as a logistic probe, and perhaps better.

We test that prediction with a model-paired comparison. On each of Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct we extract hidden states at the commitment position, fit a standard logistic probe and a contrastive mass-mean probe on arithmetic statements, select the layer by in-format five-fold cross-validation on arithmetic alone, and report cross-format AUC on capital-city statements at the selected layer with a paired bootstrap. The honest result is that both probes transfer, the contrastive advantage reverses sign across the two models, and a cross-format-exceeds-in-format anomaly warns that part of what the probes read in the capital format is a surface cue rather than a transferred error signal. The model-paired design is what prevents the tempting single-model conclusion.

Related Work

The two probes we compare descend from two traditions of reading information out of a network. The standard probe is a linear classifier probe in the sense of Alain and Bengio, fit on representations and read as a measurement of what a layer encodes [@alain2016understanding]. Gurnee et al. catalog both the reach of this approach and its central hazard, that a supervised probe can latch onto spurious, dataset-specific structure that separates the training classes without generalizing, which is exactly the failure a cross-format test exposes [@gurnee2023finding]. The contrastive probe takes the mass-mean direction of Marks and Tegmark, who argue this construction recovers a more causal and more transferable truth direction than logistic regression [@marks2023geometry]. The mean-difference construction is not specific to truth: Li et al. locate a steerable truthfulness direction from differences in attention-head activations [@li2023inference], Tigges et al. recover a causal sentiment direction the same way [@tigges2023linear], and Zou et al. generalize the pattern into a program of reading and controlling concepts as directions extracted from contrasting stimuli [@zou2023representation].

Where our work parts company with the optimism of the direction-finding literature is on generalization. Levinstein and Herrmann argue that activation probes for truthfulness can fail out of distribution and may track features other than truth [@levinstein2023still], and Farquhar et al. show that unsupervised methods can recover directions that are not the model's knowledge of truth at all [@farquhar2023challenges]. Orgad et al. sharpen the warning with evidence that error information concentrates in particular tokens and that activation-based error detectors generalize poorly across tasks and datasets [@orgad2024llms]. These results recommend judging a probe by its transfer behavior rather than its in-distribution score, and they frame our two-probe comparison as a direct question about which readout, if either, escapes the generalization failures they document. Our per-layer analysis follows the view, refined by the tuned lens, that information can be read from intermediate layers and that the layer at which a feature is most legible is itself an object of study [@belrose2023eliciting]. We compare two model families, Llama and Qwen, to avoid the single-model trap that this kind of question invites; a third family, Mistral 7B, anchored our earlier model-generalization run [@jiang2023mistral].

Method

We probe error awareness at the commitment position, the point where the model decides whether to end a statement. Following the cross-format study we use two balanced datasets of true and false statements. The arithmetic set holds 600 statements of the form "A op B = C" with addition, subtraction, and multiplication, half true and half false. The capital set holds 300 statements of the form "The capital of X is Y", half true and half false, over a fixed list of countries. The two formats share the truth task and share almost no surface vocabulary, which is what makes transfer between them informative. We report n_arith of 600 arithmetic

statements and n_capitals of 300 capital statements.

For each statement we build the same commitment prompt used in the prior work, run one forward pass through the model in bfloat16 on a single RTX 5090, and record the residual-stream hidden state at the final position at every layer. Llama-3.1-8B-Instruct exposes 33 layers and Qwen2.5-7B-Instruct exposes 29, so the per-layer sweep covers n_layers of 33 on Llama. The model never generates; the probe reads a single hidden vector per statement.

We compare two probes. The standard probe is an L2-regularized logistic regression fit on hidden states standardized by the training-split mean and variance. The contrastive probe is the mass-mean direction d equal to the mean wrong-statement hidden state minus the mean correct-statement hidden state, with each statement scored by the projection of its hidden state onto d. The two readouts differ in two ways at once, the fitting rule (a discriminative classifier versus a difference of class means) and the geometry in which they operate (the logistic probe on standardized features, the mass-mean direction in raw activation space), and we return to that confound in the limitations. Both probes are fit on the arithmetic set only. Layer selection is the methodological pivot of the study: for each probe we choose the single layer that maximizes five-fold cross-validated AUC within arithmetic, using the training format alone and never touching capitals, which is the only layer-selection rule a deployer who has labels for one format could actually follow. At the selected layer we report the in-format AUC on held-out arithmetic and the cross-format AUC on the full capital set.

Every cross-format AUC carries a 95 percent confidence interval from 2000 bootstrap resamples. The contrastive-minus-standard difference is computed as a paired bootstrap, resampling statements jointly for both probes so the interval reflects their correlation, and we report the bootstrap tail count for the one-sided hypothesis that the contrastive probe transfers better than the standard probe. A 2000-resample bootstrap cannot deliver a p value of exactly zero; when no resample favours one probe we report this as p=0.0000 to resampling resolution, meaning p below 0.0005, and we treat the percentile intervals on the high-AUC arms as approximate because percentile bootstrap is skewed against the upper bound of 1.0. Every per-example score ships with the paper and a standalone script recomputes every headline number from the raw outputs and reproduces the summary files.

Results

Both probes transfer cross-format well above chance on both models. Table 1 gives the four-cell comparison at the in-format-selected layer. The lowest cross-format AUC anywhere in the table is the Qwen contrastive probe at 0.8877 [0.8472, 0.9247], and the highest is the Llama contrastive probe at 0.9873 [0.9744, 0.9964]; the standard probe transfers at 0.9511 [0.9246, 0.9740] on Llama and 0.9278 [0.8969, 0.9551] on Qwen. Notably, the Qwen standard hidden-state probe transfers at 0.9278, far above the roughly 0.649 reported for the token-distribution probe on the same model and the same formats. We read this as consistent with the motivating collapse being a property of the token-distribution feature rather than of the representation, but we do not assert that dissociation as established, because we did not re-run the token-distribution probe inside this pipeline on these splits; the 0.9278 and the 0.649 come from different runs, and isolating the feature as the cause requires the matched experiment we flag as the most important followup.

Table 1. Two models, two probes, in-format and cross-format AUC with 95 percent bootstrap intervals on the cross-format arm. Layer selected by in-format five-fold cross-validation on arithmetic.

Model	Probe	Layer	In-format AUC	Cross-format AUC [95% CI]
Llama-3.1-8B-Instruct	Standard (logistic)	31	0.9932	0.9511 [0.9246, 0.9740]
Llama-3.1-8B-Instruct	Contrastive (mass-mean)	30	0.9914	0.9873 [0.9744, 0.9964]
Qwen2.5-7B-Instruct	Standard (logistic)	28	0.9980	0.9278 [0.8969, 0.9551]

Model	Probe	Layer	In-format AUC	Cross-format AUC [95% CI]
Qwen2.5-7B-Instruct	Contrastive (mass-mean)	28	0.9743	0.8877 [0.8472, 0.9247]

The headline finding is that the contrastive probe's advantage over the standard probe reverses sign between the two models, illustrated at the selected layers in figure_compare.png. On Llama-3.1-8B-Instruct the contrastive probe transfers at 0.9873 [0.9744, 0.9964] against the standard probe at 0.9511 [0.9246, 0.9740], a paired difference of +0.0362 with a 95 percent interval of [0.0160, 0.0609], and zero of 2000 resamples favoured the standard probe (p=0.0000 to resampling resolution). On Qwen2.5-7B-Instruct the contrastive probe transfers at 0.8877 [0.8472, 0.9247] against the standard probe at 0.9278 [0.8969, 0.9551], a paired difference of -0.0401 with a 95 percent interval of [-0.0743, -0.0078]; the one-sided bootstrap for contrastive beating standard returns 0.9930, so the data point firmly the other way. Two models of comparable scale, probed identically on identical statements, give contrastive advantages of opposite sign, and both confidence intervals exclude zero. The claim that mass-mean probes transfer better cross-format is therefore not robust; it holds on one model and reverses on the other.

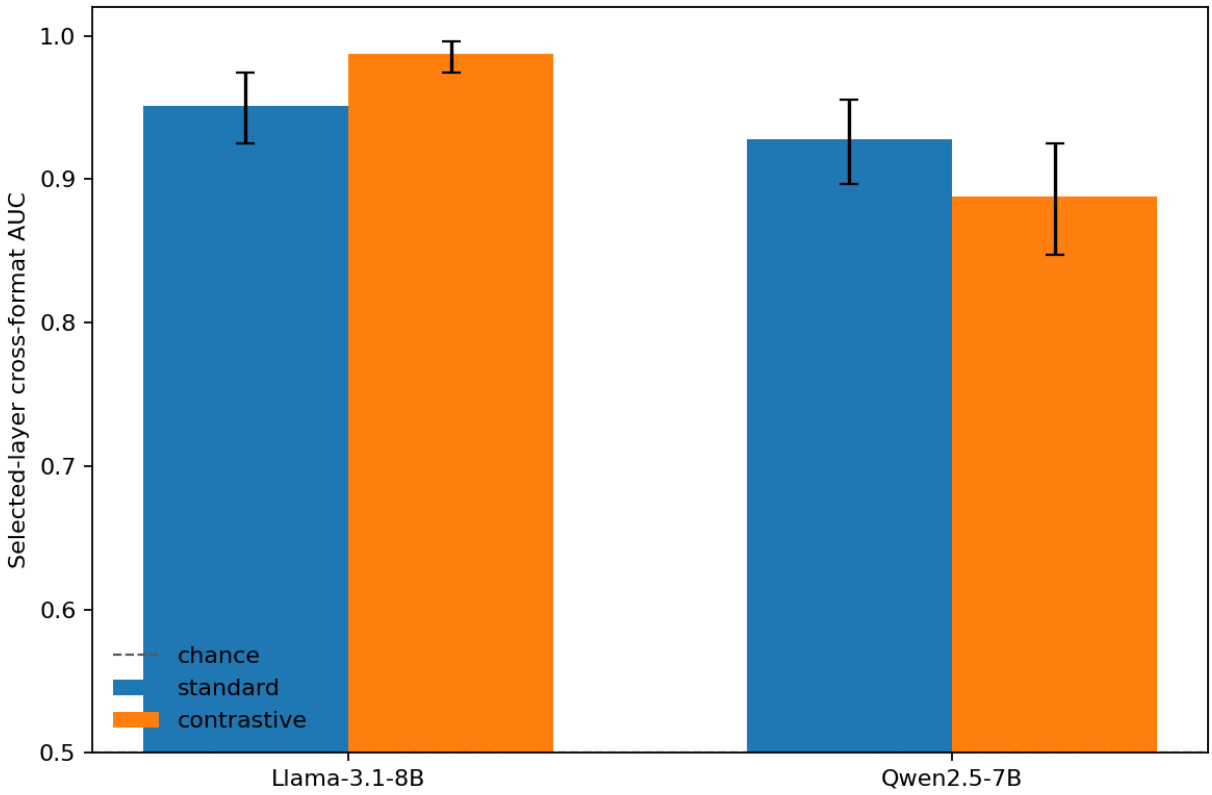

Figure 1: Cross-format AUC at the in-format-selected layer for the standard and contrastive probes on both models, with 95 percent bootstrap intervals. The contrastive advantage reverses sign between Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct.

The reversal is not an artifact of comparing the two probes at different layers. Table 1 reads the Llama probes at their separately selected layers, 31 for standard and 30 for contrastive, but the Llama advantage survives when the layer is held fixed. At the standard-selected layer 31, the contrastive probe scores 0.9911 against the standard probe's 0.9511, a same-layer gap of about +0.040, and at the contrastive-selected layer 30 the contrastive probe scores 0.9873 against the standard probe's 0.8049, an even larger same-layer gap.

The Qwen comparison is already same-layer, since in-format cross-validation selects layer 28 for both probes, and there the standard probe wins at 0.9278 against 0.8877. Holding the layer fixed, then, the contrastive probe wins on Llama and loses on Qwen, so the opposite-sign result is a property of the model and the probe pair rather than of a depth mismatch.

A more important caveat concerns what the probes are reading at all. The per-layer curves for Llama in figure_main.png and for Qwen in figure_qwen.png show that at several mid layers the arithmetic-trained probe scores higher cross-format on capitals than in-format on arithmetic. On Llama the contrastive probe at layer 11 separates arithmetic truth at an in-format AUC of only 0.7111 yet separates capital truth at a cross-format AUC of essentially 1.0, and the same inversion appears at adjacent mid layers and for the standard probe near layer 10. Figure figure_inout.png places the in-format and cross-format curves on a common axis so these crossings are visible. A direction fit as the mean wrong-arithmetic state minus the mean correct-arithmetic state that barely separates arithmetic truth cannot be reading a shared error-awareness representation when it perfectly separates capital truth; the most parsimonious explanation is that the capital format is separable by a surface cue, such as the identity or scale of the answer token, that the arithmetic-trained direction happens to align with. This reframes the mid-layer near-ceiling band as a confound signature rather than a virtue. For that reason we do not treat the contrastive probe's mid-layer robustness as evidence in its favour, and we base every conclusion on the in-format-selected layer. We note in addition that the mid-layer maxima are selected on the evaluation data, choosing the layer that maximizes cross-format AUC and then reporting it, which biases those maxima upward, and that several of them sit at AUC exactly 1.0, where the metric saturates and cannot rank one perfectly separated layer above another.

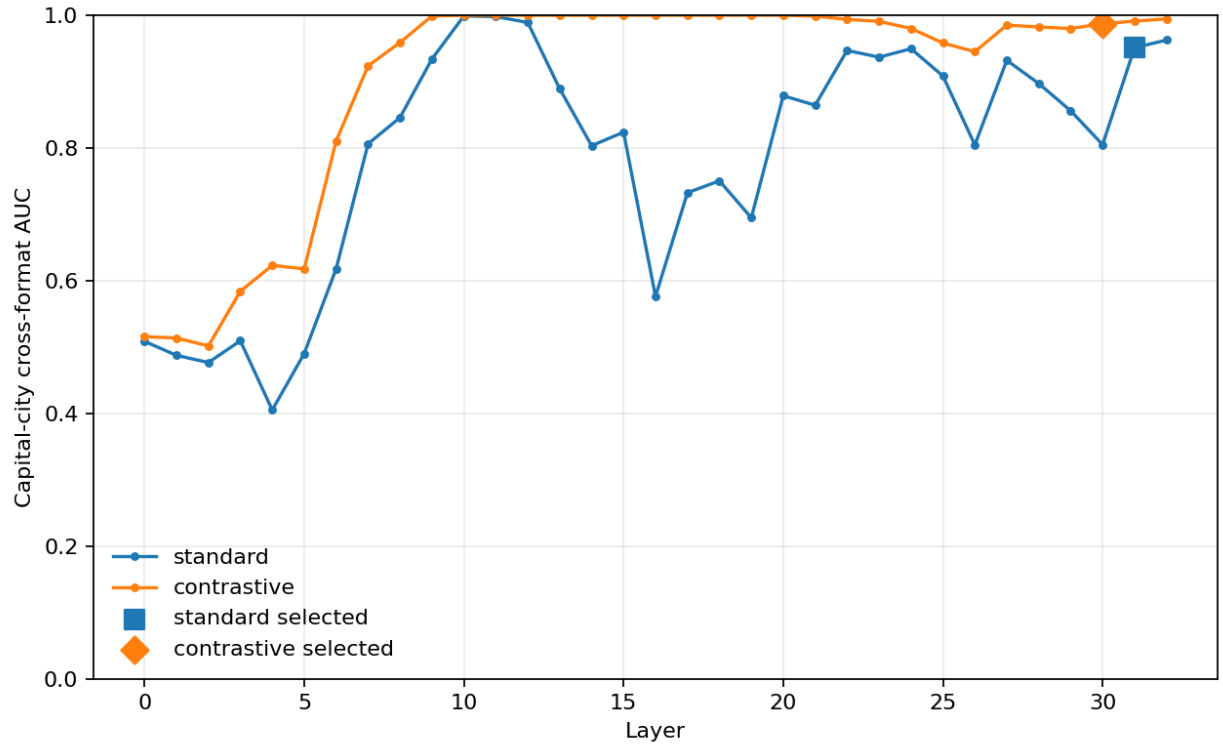

Figure 2: Per-layer cross-format AUC on Llama-3.1-8B-Instruct for the standard and contrastive probes.

Discussion and Limitations

The study answers the question it was designed to ask, and the answer is the unglamorous one. Both readouts transfer across formats, so neither collapses, and the Qwen standard hidden-state probe at 0.9278 is consistent with the motivating token-distribution collapse being a feature-level effect rather than a representational one.

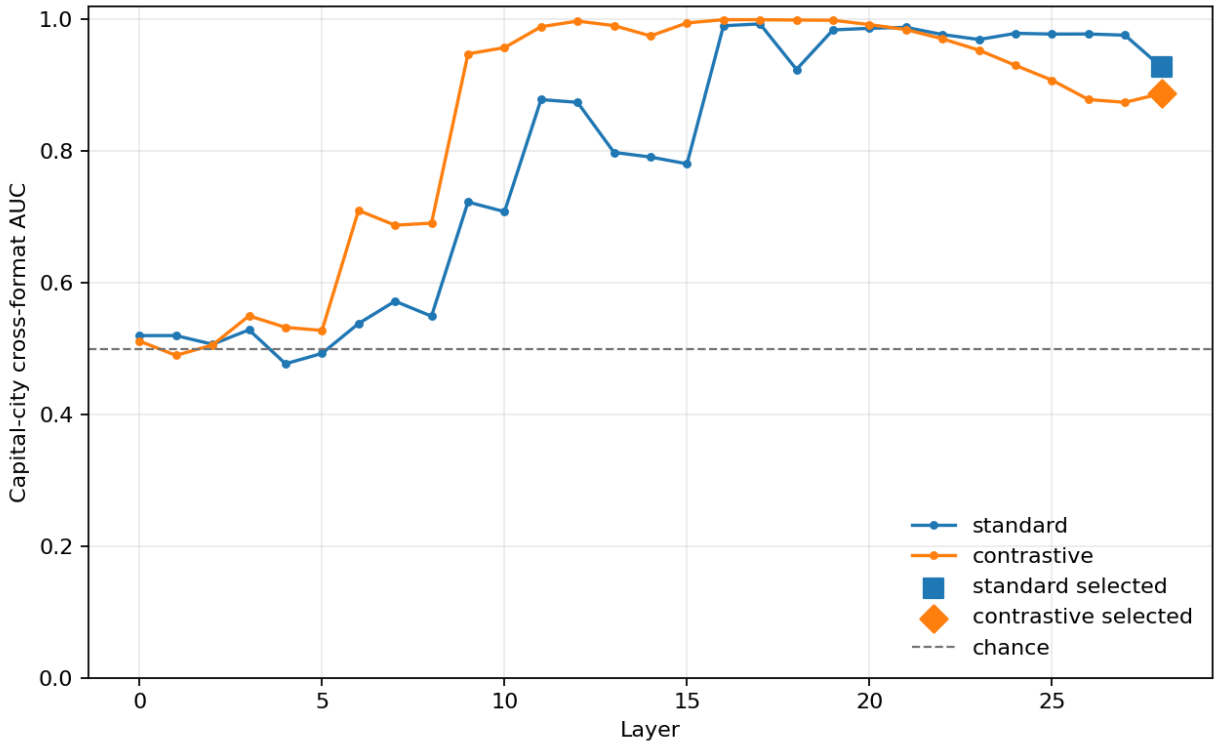

Figure 3: Per-layer cross-format AUC on Qwen2.5-7B-Instruct for the standard and contrastive probes.

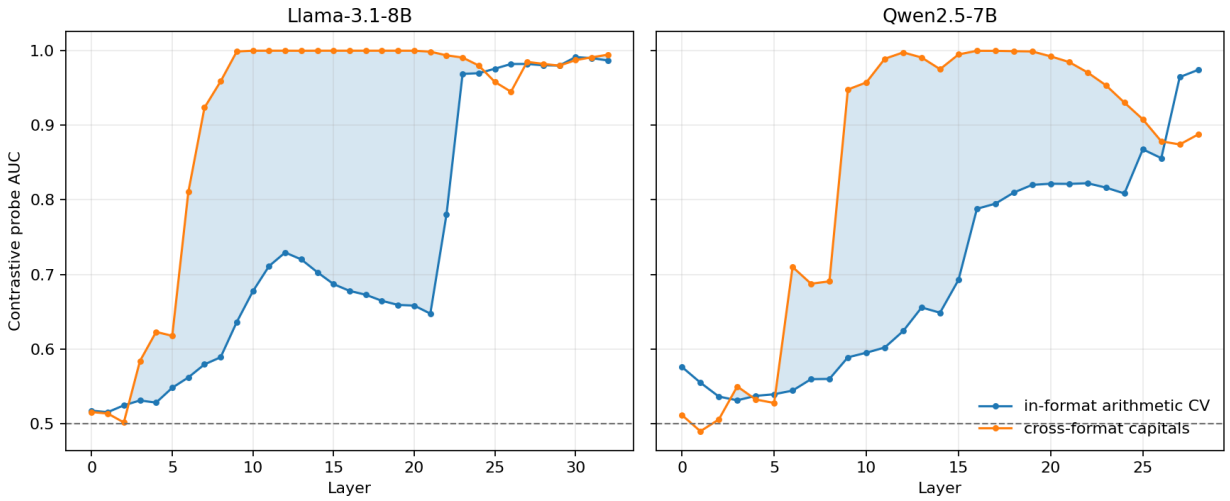

Figure 4: In-format versus cross-format AUC per layer for the contrastive probe on both models; mid-layer crossings where cross-format exceeds in-format signal a surface cue.

But the contrastive mass-mean probe, which the representation-reading literature would predict to transfer best, helps significantly on Llama and hurts significantly on Qwen. The single most important lesson is methodological: a single-model experiment here would have produced a clean, citable, and wrong general claim, and only the model-paired design exposes that the effect changes sign. The point about layer selection is also model-specific rather than universal: on Qwen the standard probe beats the contrastive probe at the in-format-selected layer, while on Llama the contrastive advantage survives layer selection and even widens at a common layer. This echoes the broader warnings that activation probes generalize unevenly and can track features other than the one intended [@levinstein2023still; @orgad2024llms; @farquhar2023challenges], and it adds a specific instance where the failure is not a quiet loss of accuracy but a reversal of which method looks better.

Several limitations bound the claims, and we state them plainly because they are the weakest joints of the paper. First, this is a two-model comparison on a single cross-format pair, arithmetic to capitals, in one direction. Two models are enough to refute a universal advantage but not enough to characterize the population of models, and a different format pair or the reverse direction could reorder the probes. Second, and most consequential, the cross-format-exceeds-in-format inversion at mid layers shows that at least part of the cross-format separability is a surface cue in the capitals format rather than a transferred error-awareness direction; until that cue is controlled, by matching or regressing out answer-token identity and activation norm across true and false capitals, or by using a capital set whose false statements name real capitals of other countries so the answer token is always a valid city, the transfer numbers should be read as an upper bound on genuine error-awareness transfer. Third, the feature-versus-representation reading of the Qwen result is not established here, because we compared our hidden-state 0.9278 against a token-distribution 0.649 computed in a different pipeline; re-running the token-distribution probe on these exact statements, commitment positions, and splits is the single most important followup, and absent it we claim only that a hidden-state logistic probe does not reproduce the prior collapse on Qwen. Fourth, probe type is confounded with preprocessing: the standard probe operates on standardized features while the mass-mean direction is computed and scored in raw activation space, so the contrast mixes a fitting-rule difference with a geometry difference, and the clean test is a factorial of logistic versus mean-difference crossed with raw versus standardized versus whitened features at the selected layer. Fifth, the mid-layer maxima we report for context are test-selected and therefore upward biased, and several sit at AUC exactly 1.0 where the metric saturates; they should not be read as a fair robustness comparison, and a held-out layer-selection scheme would be needed to make robustness claims at all. Finally, both arms share a single arithmetic train-test split per model and a fixed seed, so split-to-split variance is not measured, the percentile bootstrap intervals are approximate near the AUC ceiling, and both models are instruction-tuned chat models probed through one commitment prompt, so base models or other prompt framings could relocate the signal.

Conclusion

We set out to test whether a contrastive mass-mean error-awareness probe transfers across formats better than a standard logistic probe, motivated by a published cross-format collapse and by the claim that mass-mean directions generalize better. Both probes transfer well above chance on both models, between 0.8877 and 0.9873 cross-format AUC, so the headline is not about transfer failing but about ranking. The contrastive advantage does not replicate across models, running to +0.0362 [0.0160, 0.0609] on Llama-3.1-8B-Instruct with zero of 2000 resamples favouring the standard probe (p=0.0000 to resampling resolution) and to -0.0401 [-0.0743, -0.0078] on Qwen2.5-7B-Instruct, opposite signs with both intervals excluding zero, and the reversal holds when the layer is held fixed. A cross-format-exceeds-in-format anomaly at mid layers further warns that some of the apparent transfer is a surface cue rather than a shared error direction, which is why our conclusions rest on the in-format-selected layer and why a surface-cue control and a matched token-distribution re-run are the necessary next steps. The tempting single-model conclusion is wrong, and the broader lesson for the probing and representation-reading literature is that claims about which readout transfers best should be paired across models, layers, and preprocessing before they are believed.

References

Citations use the keys in references.bib, which holds seventeen entries each verified against the arxiv API. The internal cross-format collapse study and our prior Llama model-generalization replication are described in the text as prior work in this line and are not arxiv-indexed.