Format-Specific Error Awareness Is Not Model-General: An Arithmetic-Trained Wrongness Probe Transfers Cleanly in Llama-3.1-8B-Instruct

Ephraiem Sarabamoun
Independent Research
ephraiemsam16@gmail.com

Abstract

A recent transfer test on Qwen2.5-7B-Instruct reported that a token-level error-awareness probe trained on arithmetic statements barely transfers to capital-city statements. The probe read the next-token distribution at the commitment position, reached an AUC of 0.968 on held-out arithmetic, and fell to 0.649 on capitals, a drop of 0.319, even though a probe trained within capitals reached 0.990. The conclusion was that token-level error awareness in that model is real and largely format-specific. We ask whether the format-specificity is a property of the probing method or a property of the model, by replicating the pipeline exactly on meta-llama/Llama-3.1-8B-Instruct. We use byte-identical datasets, the same feature selection by Cohen's d on an arithmetic training split, the same logistic regression, and the same bootstrap procedure. The collapse does not replicate. On Llama the probe reaches an AUC of 0.992 (95% CI 0.981 to 0.999) on held-out arithmetic and 0.982 (95% CI 0.964 to 0.996) on capital-city statements, a transfer gap of 0.0105 against Qwen's 0.319. The mechanism is visible in the commitment channel. In Llama the mean probability of ending a statement with a period separates true from false in both formats, by 0.184 in arithmetic and 0.078 in capitals, while in Qwen the capitals gap was slightly negative. Llama writes wrongness into format-general hedging tokens, question marks, withheld periods, and continuation words, so a probe tuned on one format reads the other almost as well. Format-specificity of the token-level error signal is therefore model-dependent, not intrinsic to the readout. For deployment the practical lesson tightens rather than relaxes: whether a cheap error probe will travel across topics cannot be predicted from the method alone, so it must be validated per format on the model actually being deployed.

1 Introduction

Several lines of work show that a model carries an internal trace of whether its own output is correct. Hidden activations linearly encode whether a statement the model produced is true or false, and a probe on that state detects lies [Azaria and Mitchell, 2023]. An unsupervised method recovers a truth direction from contrastive activations alone [Burns et al., 2022]. Models can estimate the probability that their own answers are correct with usable calibration [Kadavath et al., 2022]. The geometry of those truth representations, however, generalizes unevenly across datasets and topics [Marks and Tegmark, 2023].

The cross-format error-awareness study turned that unevenness into a behavioral test with the cheapest possible readout. It built six hundred arithmetic statements and three hundred capital-city statements, took one forward pass of Qwen2.5-7B-Instruct per statement, recorded the top fifty next-token probabilities at the position where the model decides whether to commit with a period, trained

Preprint. Under review.

a logistic regression on arithmetic, and tested it both in-format and on capitals. The probe lost most of its discriminative power across formats while a within-capitals control showed the wrongness information was present all along. The published verdict was that token-level error awareness in that model is largely format-specific.

A single model cannot distinguish two readings of that verdict. The strong reading is that the token-level readout is inherently format-bound, so any deployer training a cheap probe on one topic should expect it to fail on others. The weak reading is that Qwen2.5-7B-Instruct happens to write arithmetic wrongness into arithmetic-specific tokens, and another model trained by another lab might write wrongness into tokens that travel. Replication on a second model family is the cheapest way to separate the readings, and the two answers have different consequences for how error detectors should be qualified before deployment.

We replicate the study on meta-llama/Llama-3.1-8B-Instruct [Grattafiori et al., 2024], a model of comparable scale from a different family than Qwen2.5 [Team, 2024]. Everything except the model is held fixed: the datasets are byte-identical copies of the original files, the feature selection, classifier, splits, seeds, and bootstrap settings are unchanged, and the hardware is the same class of GPU. The transfer collapse does not replicate.

2 Method

We reuse the published pipeline without modification to its statistical procedure, so the numbers are directly comparable to the Qwen run. The arithmetic set holds six hundred statements of the form "A op B = C" with operators drawn from addition, subtraction, and multiplication, operands between zero and one hundred, and false statements perturbed from the correct answer by a nonzero offset between one and twenty. The capital set holds three hundred statements of the form "The capital of X is Y" over one hundred fifty countries, each appearing once with its real capital and once with a real capital of a different country. Both sets are balanced and seeded; we copied the original study's data files directly and verified the copies by checksum.

For each statement we build the prompt "Finish the statement by writing only a period." followed by the statement, run one forward pass through meta-llama/Llama-3.1-8B-Instruct in bfloat16 on one RTX 5090, and record the fifty most probable next tokens with their probabilities at the final position. The model never generates. The arithmetic set is split eighty twenty into train and test, stratified by label, with the same seed as the original study. On the training split alone we compute Cohen's d of each token's probability between true and false statements and keep the fifty tokens with the largest absolute d. Each statement becomes a fifty-dimensional vector of those token probabilities, standardized on the training split, and a logistic regression is fit on arithmetic train. The in-format arm scores the held-out arithmetic test and the cross-format arm scores all three hundred capital statements. Every reported AUC carries a 95 percent confidence interval from two thousand bootstrap resamples with a fixed seed, the same count as the published run and double the floor this replication was specified to meet.

Two diagnostics from the original study are retained. The select-and-train procedure is repeated within the capital set under five-fold cross-validation, which bounds how much wrongness signal capitals carry at all. And the mean probability mass the model places on period tokens after true versus false statements is computed per format, which tracks the original scalar notion of error awareness, the willingness to commit.

Every per-example output of the run, including the probe score and period mass for each of the nine hundred statements, ships with this paper, and a standalone script recomputes every headline number from those raw outputs and reproduces the summary file byte for byte.

3 Results

In format, the probe behaves as expected and slightly better than in the Qwen run. On held-out arithmetic the AUC is 0.992 with a 95 percent confidence interval of 0.981 to 0.999 over one hundred twenty test statements, against 0.968 for Qwen.

Across format, the published result inverts. On the three hundred capital-city statements the arithmetic-trained probe reaches an AUC of 0.982 with a confidence interval of 0.964 to 0.996.

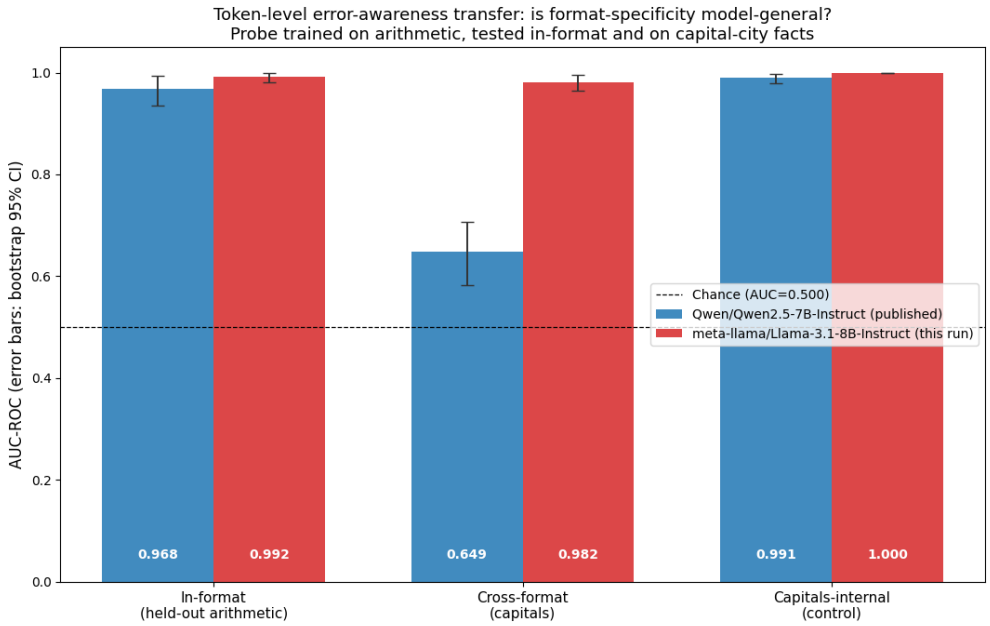

Figure 1: AUC of the arithmetic-trained error-awareness probe on Qwen2.5-7B-Instruct (published run) and meta-llama/Llama-3.1-8B-Instruct (this run), with 95% bootstrap confidence intervals. In format the models are nearly identical (0.968 vs 0.992). Across formats they diverge sharply: Qwen falls to 0.649 while Llama holds 0.982. The within-capitals control is near ceiling for both, so the divergence is in the portability of the readout, not the presence of signal.

The transfer gap is 0.0105 in AUC, where Qwen's was 0.319. The probe is also well separated in score, not merely in rank: the median probe score is 0.998 for true capitals and 0.000 for false ones, and no false capital scores above 0.068. The handful of true statements the probe misranks involve capitals the model is plausibly less certain of, such as Dodoma and Mbabane. Figure 1 shows both models side by side, with the in-format, cross-format, and within-capitals control arms and their bootstrap intervals.

The within-capitals control reaches an AUC of 1.000, with a degenerate bootstrap interval because every resample preserves the perfect separation of the out-of-fold scores. As in the original study this number selects features on the full capital set before cross-validating, so it is a mild upper bound and we treat it as a ceiling estimate rather than a precise quantity. Its role is qualitative: capitals carry an essentially saturating wrongness signal in Llama, as they did in Qwen, so neither model's cross-format number can be excused by the target format lacking signal.

The commitment channel explains the difference between the models. In Llama the mean period mass after a true arithmetic statement is 0.684 against 0.500 after a false one, a gap of 0.184, and after a true capital statement it is 0.746 against 0.668, a gap of 0.078. Both gaps point the same way: wrong statements make Llama withhold the period. In the published Qwen run the arithmetic gap was 0.354 and the capitals gap was -0.014, so the single most transferable feature carried no usable cross-format signal there. Consistent with this, the tokens selected on Llama's arithmetic training split are dominated by format-general commitment and hedging tokens, period variants, question marks, conjunctions like "but", and continuation operators, rather than by anything arithmetic-bound, which is why the same feature set reads capital-city wrongness almost as well.

4 Discussion

The replication gives a clean answer to the question it was designed to ask. The format-specificity reported for Qwen2.5-7B-Instruct is real but it is a property of that model, not of token-level error probing as a method. Two instruction-tuned models of similar scale, probed identically on identical

statements, place their wrongness signal in differently shaped output channels. Qwen expresses arithmetic wrongness through arithmetic-specific continuations and leaves capitals wrongness out of the period channel entirely. Llama expresses wrongness in both formats through a shared hedging vocabulary, a withheld period, raised question-mark mass, and discourse tokens that would let it walk the statement back. A probe is exactly as portable as the channel it happens to read.

This sharpens rather than softens the deployment lesson of the original study. If format-specificity were intrinsic to the method, a deployer could at least plan around a known failure mode. Model-dependence is worse to plan around: the same cheap probe recipe yields a detector that travels on one model and silently fails to travel on another, and in-format validation looks equally excellent in both cases. The Qwen and Llama in-format AUCs differ by 0.024 while their cross-format AUCs differ by 0.333. Nothing visible at training time predicts which behavior you get. Per-format validation on the deployed model remains the only honest qualification, and our result adds that the validation cannot be inherited from a sibling study on a different model family.

The contrast also offers a small interpretability lead. Whatever instruction-tuning or pretraining difference causes Llama to hedge with a format-general vocabulary where Qwen hedges format-specifically is a concrete, behaviorally measurable target, and the activation-level analogue of this comparison, whether the truth directions of Marks and Tegmark [2023] cohere across formats more strongly in Llama than in Qwen, is a natural next experiment.

5 Limitations

This is a two-model comparison, which can refute generality but cannot establish it; a third family could behave like either, and the honest claim is only that format-specificity is not universal. The comparison against Qwen relies on the published numbers of the original run rather than a re-execution inside this run's environment, although the pipeline, data, seeds, and library versions match. Both arms share a single 80/20 arithmetic split and a fixed seed, as in the original, so split-to-split variance is not measured. We tested transfer in one direction, arithmetic to capitals, on two formats; a different format pair or the reverse direction could show a gap that this pair does not. The capitals-internal control selects features on the full capital set before cross-validating and its perfect AUC should be read as a ceiling, not a calibrated estimate. Finally, both models are instruction-tuned chat models probed through one prompt convention, and base models or other prompt framings could relocate the wrongness signal.

6 Conclusion

An arithmetic-trained token-level error-awareness probe that collapsed from 0.968 to 0.649 AUC across formats on Qwen2.5-7B-Instruct reaches 0.992 in format and 0.982 across formats on meta-llama/Llama-3.1-8B-Instruct under a byte-identical replication. The transfer collapse is a fact about the model, not the method. Error-detection probes still need per-format validation, and now also per-model validation, because the portability of a model's sense of its own errors is itself one of the things that varies between models.

References

Amos Azaria and Tom Mitchell. The internal state of an llm knows when it's lying, 2023.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision, 2022.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The llama 3 herd of models, 2024.

Saurav Kadavath, Tom Conerly, Amanda Askell, et al. Language models (mostly) know what they know, 2022.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets, 2023.

Qwen Team. Qwen2.5 technical report, 2024.