Error Awareness Is Format-Specific: An Arithmetic-Trained Wrongness Probe Does Not Transfer to Capital-City Facts

Ephraiem Sarabamoun
Independent Research
ephraiemsam16@gmail.com

Abstract

A language model often assigns a different next-token distribution to a statement it has just completed depending on whether that statement is true or false, and a small classifier reading that distribution can recover whether the statement was correct. We ask a narrower question than prior work on whether such a signal exists. We ask whether it is the same signal across topics. We build two balanced datasets, six hundred arithmetic statements of the form "A op B = C" and three hundred capital-city statements of the form "The capital of X is Y", and for each statement we take one forward pass through Qwen2.5-7B-Instruct and record the top fifty next-token probabilities at the commitment position. We select the fifty most discriminative tokens by Cohen's d on an arithmetic training split, train a logistic regression on arithmetic, and then test it two ways. On held-out arithmetic the classifier reaches an AUC of 0.968 (95% CI 0.935 to 0.993). On the capital-city statements it falls to 0.649 (95% CI 0.583 to 0.708), a drop of 0.319 in AUC. The collapse is not because capitals carry no wrongness signal. A classifier trained and tested within capitals reaches an AUC of 0.990, and the standard period-commitment probability separates true from false arithmetic by 0.35 while separating true from false capitals by essentially nothing. The model knows when a capital is wrong, but it writes that knowledge into different tokens than it uses for arithmetic, so a probe tuned on one format does not read the other. Token-level error awareness in this model is real and strong, and it is largely format-specific.

1 Introduction

Several lines of work show that a model carries an internal trace of whether its own output is correct. The hidden state of a model linearly encodes whether the statement it is producing 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 be asked to estimate the probability that their own answer is correct and are reasonably calibrated when they do [Kadavath et al., 2022]. Each of these results establishes that a wrongness signal is present somewhere in the network.

For a safety tool the more useful question is not whether the signal exists but whether it is one signal or many. If a deployer trains a lightweight error detector on whatever labeled data is cheap to produce and then runs it on the open-ended traffic a model actually sees, the detector is only useful insofar as the signal it learned on the training topic is the same signal that appears on every other topic. The geometry-of-truth analysis already hints that this assumption is shaky, since truth directions recovered on one dataset generalize unevenly to others [Marks and Tegmark, 2023]. We test the behavioral version of that claim with the cheapest possible readout, the next-token distribution, and the sharpest possible contrast, two formats that share almost no surface vocabulary.

Preprint. Under review.

We study arithmetic and capital-city facts because they are both easy to label with certainty and maximally different in form. An arithmetic error is a number that does not follow from two operands and an operator. A capital-city error is a real city attached to the wrong country. If error awareness were a single format-general property of the model, a probe trained to catch the first kind of error should catch the second. We find that it largely does not.

2 Method

We generate two datasets with a fixed seed so the experiment is reproducible. The arithmetic set holds six hundred statements written as "A op B = C" with the operator drawn from addition, subtraction, and multiplication and operands between zero and one hundred. Half are true. The false half take the correct answer and perturb it by a nonzero offset between one and twenty in either direction, so a false statement is always close to right rather than absurd. The capital set holds three hundred statements written as "The capital of X is Y" over one hundred fifty countries, each appearing once with its real capital and once with a real capital belonging to a different country. Both sets are balanced exactly fifty-fifty and every statement is unique. No statement carries a trailing period, because the period is exactly the token we probe.

For each statement we build the prompt "Finish the statement by writing only a period." followed by the statement, and we run a single forward pass through Qwen2.5-7B-Instruct in bfloat16 on one RTX 5090. We never let the model generate. We take the next-token distribution at the final position, the position where the model decides whether to commit by writing a period or to hedge, and we record the fifty most probable tokens with their probabilities. This probing convention follows an existing error-awareness pipeline whose method we reuse.

The classifier is deliberately simple. We split the arithmetic set into eighty percent train and twenty percent test, stratified by label. On the training split alone we compute, for every token that appears in any statement's top fifty, the Cohen's d of its probability between true and false statements, and we keep the fifty tokens with the largest absolute d as features. Each statement becomes a fifty-dimensional vector of those token probabilities, with zero where a selected token is absent from that statement's top fifty. We standardize the features on the training split and fit a logistic regression. We then score the held-out arithmetic test, which is the in-format arm, and all three hundred capital statements, which is the cross-format arm. For every arm we report AUC-ROC with a ninety-five percent confidence interval from two thousand bootstrap resamples.

To interpret a cross-format collapse we add two diagnostics. First we repeat the whole select-and-train procedure within the capital set using five-fold cross-validation, which tells us whether capitals carry any learnable wrongness signal at all. Second we compute the mean probability the model places on a period for true and for false statements in each format, which is the original scalar error-awareness measure and a check on whether the commitment signal itself transports.

3 Results

In format the classifier works very well. Trained on arithmetic and tested on held-out arithmetic, it reaches an AUC of 0.968 with a ninety-five percent confidence interval of 0.935 to 0.993 over one hundred twenty test statements. The model's next-token distribution after an arithmetic statement is highly informative about whether the arithmetic is correct, and fifty token-probability features are enough to read it.

Across format the same classifier degrades sharply. On the three hundred capital-city statements its AUC is 0.649 with a confidence interval of 0.583 to 0.708. That is above chance, and the interval excludes 0.5, so the arithmetic-trained probe is not completely blind to capital errors. But it has lost most of its discriminative power, and the in-format minus cross-format gap is 0.319 in AUC. Figure 1 shows the two ROC curves on one axis.

The collapse is not explained by capitals lacking a wrongness signal. A logistic regression selected and evaluated within the capital set by five-fold cross-validation reaches an AUC of 0.990 with a confidence interval of 0.980 to 0.998. The information needed to separate true from false capitals is present in the model's next-token distribution at nearly the same strength as for arithmetic. It simply lives in different tokens, which is why a feature set chosen on arithmetic cannot read it.

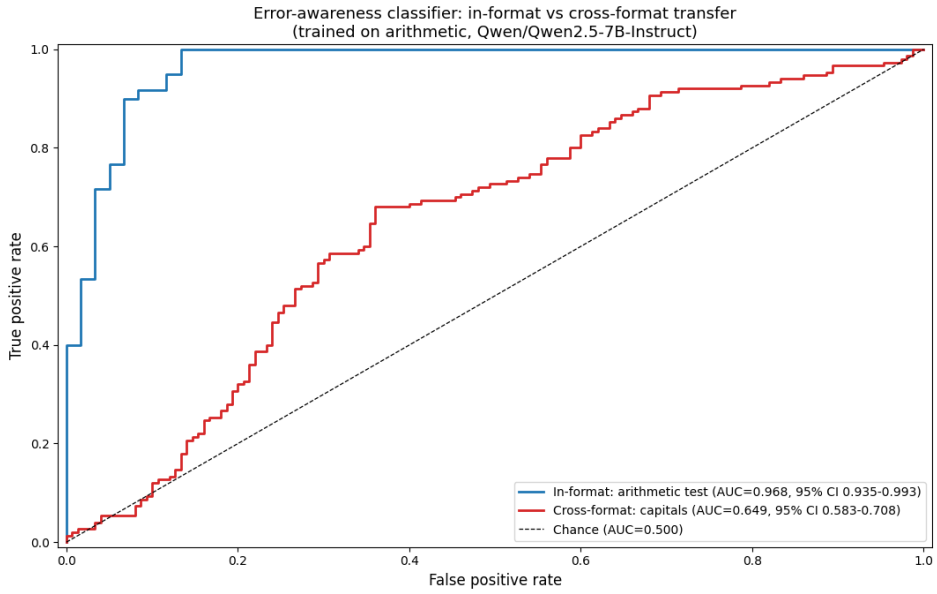

Figure 1: ROC curves for the arithmetic-trained error-awareness classifier. The in-format arm is held-out arithmetic (AUC 0.968); the cross-format arm is all capital-city statements (AUC 0.649). Confidence intervals are from two thousand bootstrap resamples. The classifier loses most of its discriminative power across formats even though capitals carry a near-ceiling internal signal.

The period-commitment measure makes the dissociation concrete. In arithmetic the mean probability of a period is 0.921 for true statements and 0.567 for false ones, a gap of 0.35 that by itself separates the classes. In capitals the mean probability of a period is 0.971 for true statements and 0.986 for false ones, a gap of about -0.01. The model commits to a wrong capital at least as readily as to a right one, so the single most transferable feature in arithmetic, the willingness to end the sentence, carries no usable signal for capitals and if anything points the wrong way.

4 Discussion

The headline is a dissociation. Qwen2.5-7B-Instruct has a strong internal record of whether a freshly completed statement is correct in both formats we tested, since an in-format probe reaches an AUC near 0.97 for arithmetic and near 0.99 for capitals. What does not survive is the mapping from that record to specific output tokens. Arithmetic wrongness shows up as reluctance to commit, a withheld period and raised mass on continuation tokens like a minus sign that would let the model keep computing. Capital wrongness does not show up in the period at all, and whatever it does show up in is orthogonal enough to the arithmetic features that a transferred probe operates near chance.

This matters for the way error detectors are likely to be deployed. A common and attractive plan is to train a cheap probe on a labeled slice of behavior and apply it broadly, on the theory that the model has a general sense of its own correctness. Our result is a small but clean counterexample. The general sense may exist internally, consistent with the activation-probing literature, and still fail to produce a format-general readout at the output distribution. A probe that looks excellent on its training distribution can quietly lose two thirds of its headroom on a distribution that differs only in topic, with no warning from in-format validation. The uneven cross-dataset generalization that Marks and Tegmark [2023] observed in the representation geometry has a direct behavioral consequence here.

The framing also sharpens what error awareness should be taken to mean. It is tempting to read a high in-format AUC as evidence that a model knows when it is wrong in a portable way. The capitals result shows the knowledge can be there, by the internal AUC of 0.990, while the portable readout is not. Awareness measured through one output channel is not the same as awareness the

model would surface through another, and a safety case that depends on reading wrongness off the output distribution has to be made per format rather than assumed once.

5 Limitations

This is a single model, a single probing convention, and two formats, so we show that format-specificity can happen rather than how common it is. We tuned features on arithmetic and tested on capitals, and the reverse direction may not be symmetric. The capital statements pair a country with a real capital of some other country, which is a particular kind of false; a wrong capital that is not a real capital anywhere might be easier to detect and could change the cross-format number. The next-token distribution at the commitment position is a deliberately shallow readout, and a probe on hidden activations rather than output probabilities might transfer better, which would itself be an informative contrast. The capitals-internal AUC selects features on the full capital set before cross-validating, so it is a mild upper bound on within-format performance rather than a fully held-out estimate. Finally, the model is instruction-tuned, and a base model probed the same way could place its wrongness signal differently.

6 Conclusion

A token-level error-awareness classifier trained on arithmetic reaches an AUC of 0.968 on held-out arithmetic and only 0.649 on capital-city statements, a drop of 0.319, even though a classifier trained within capitals reaches 0.990 and the model clearly retains the relevant information. The wrongness signal in Qwen2.5-7B-Instruct is real and strong in both formats, and it is largely format-specific at the output distribution. Error-detection probes should be validated on the formats they will actually see, because strong in-format performance does not certify that a model's sense of its own errors will travel.

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.

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.