Format-Specificity of Error Awareness Is Model-Dependent

Ephraiem Sarabamoun
Independent Research
ephraiemsam16@gmail.com

Abstract

Token-level error-awareness probes read a model's next-token distribution at the moment it would commit to a statement and ask whether the model knows the statement is wrong. A recent study on Qwen2.5-7B-Instruct found this signal largely format-specific, with an arithmetic-trained probe collapsing from 0.968 AUC in format to 0.649 on capital-city statements, while an exact replication on Llama-3.1-8B-Instruct found the same probe transferring almost perfectly, with a gap of 0.011 against Qwen's 0.319. This paper asks what that disagreement means and how a deployer should act on it. We extend the comparison along three axes held to the identical pipeline. We add a third family, Mistral-7B-Instruct-v0.3, which transfers as cleanly as Llama (0.980 across formats against 0.966 in format), making Qwen the outlier among three families whose arms all carry in-pipeline provenance, the published Qwen pipeline re-executed inside this codebase and reproduced to four decimals. We test the reverse training direction, capitals to arithmetic, and find the disagreement is not even a stable property of the models. Qwen's collapse is bidirectional (0.678 reverse against 0.649 forward), while Llama's near-perfect forward transfer inverts to reliably below chance in reverse (0.432, with the wrongness signal still present at 0.981 for an in-format control), because the features capitals training selects flip their class direction on arithmetic. And we rebuild the target task from same-country confusable cities so its internal probe lands below ceiling, giving the transfer gap a meaningful denominator; there Llama transfers at its own internal ceiling (0.989 against 0.988) while Qwen falls reliably below chance (0.337 against an internal 0.976). A training-time-visible statistic over the selected tokens separates the forward collapse from the clean transfers but misses both below-chance inversions. The practical conclusion is that probe portability belongs to the (model, training format, target task) triple, predicting it at training time is open, and a deployed error probe must be validated on the model, format pair, and error difficulty it will actually face.

1 Introduction

A growing body of evidence says language models carry an internal trace of whether their own statements are correct. Hidden activations linearly encode truth [Azaria and Mitchell, 2023], unsupervised methods recover truth directions from contrastive activations alone [Burns et al., 2022], models estimate the reliability of their own answers with usable calibration [Kadavath et al., 2022], and yet the geometry of these representations generalizes unevenly across topics [Marks and Tegmark, 2023]. The cheapest behavioral readout of this trace sits at the output layer. Prompt the model to finish a statement with a period, take one forward pass, and read the next-token distribution at the commitment position. A model that believes the statement puts its mass on the period; a model that knows better hedges.

Preprint. Under review.

The cross-format error-awareness study built that readout into a transfer test on Qwen2.5-7B-Instruct [Team, 2024] and reported a collapse. A logistic probe over the top fifty next-token probabilities, trained on six hundred arithmetic statements, reached 0.968 AUC on held-out arithmetic and fell to 0.649 on three hundred capital-city statements, even though a probe trained within capitals reached 0.990, so the wrongness signal was present and the probe simply could not read it across formats. A byte-identical replication on Llama-3.1-8B-Instruct [Grattafiori et al., 2024] then inverted the result. Llama's arithmetic-trained probe reached 0.992 in format and 0.982 on capitals. One model writes wrongness into format-specific continuations, the other into format-general hedging tokens, and nothing visible at training time distinguishes them, since the two in-format AUCs differ by only 0.024.

A one-against-one disagreement is an anecdote, and an anecdote with three open confounds. First, two models cannot establish a pattern; the field needs to know whether clean transfer or collapse is the typical case at this scale. Second, both studies trained in only one direction, so the Qwen collapse could reflect what arithmetic training selects rather than what the model represents. Third, the capitals task was trivially easy for the within-format control, which hit a degenerate ceiling of 1.000 AUC on Llama, so the transfer gap had no internal reference to be measured against. This paper closes all three gaps with the pipeline held fixed, then asks whether anything observable at training time, the commitment channel or the selected features themselves, predicts when the probe travels.

2 Method

All experiments share one probing pipeline, inherited unchanged from the published study. Each statement is wrapped in the prompt "Finish the statement by writing only a period." followed by the statement without its trailing period. One forward pass in bfloat16 on an RTX 5090 records the fifty most probable next tokens and their probabilities at the final position; the model never generates. The training format is split eighty twenty with label stratification under a fixed seed. On the training split alone, each vocabulary token is scored by the absolute Cohen's d of its probability between true and false statements, the top fifty tokens become the feature set, and a logistic regression over standardized token probabilities is fit. Every reported AUC carries a 95 percent confidence interval from two thousand bootstrap resamples under the same fixed seed, and every per-example output ships with the paper together with a standalone script that recomputes each headline number from raw outputs byte for byte.

The datasets are those of the published study, verified by checksum. The arithmetic set holds six hundred balanced statements of the form "A op B = C" with single offsets corrupting the false answers. The easy capital set holds three hundred balanced statements of the form "The capital of X is Y" over one hundred fifty countries, with false statements drawn from other countries' capitals. The hard capital set, new in this work, holds two hundred thirty-two balanced statements over one hundred sixteen countries in the same wording, with each false statement naming a prominent non-capital city of the same country, a former capital, the largest city, or the famous commercial hub, for example Istanbul for Turkey, Toronto for Canada, and Bujumbura for Burundi. Countries with contested or split capitals were dropped or assigned an unambiguous non-seat city, so every label is clean. Random wrong capitals can be rejected on a country-city mismatch alone; confusable cities require knowing which prominent city is actually the capital, which is what pulls the internal probe off its ceiling.

Three model families are compared under this pipeline. Qwen2.5-7B-Instruct numbers were re-executed inside this codebase, reproducing the published run to four decimals with a byte-identical extraction. Llama-3.1-8B-Instruct numbers come from our replication. Mistral-7B-Instruct-v0.3 [Jiang et al., 2023] is the third family, chosen for maximal independence from the first two, a third lab, a disjoint pretraining recipe, and a 32k SentencePiece vocabulary against Qwen's 152k BPE and Llama's 128k vocabulary, which directly stresses the token-level feature selection the probe rides on. The reverse-direction experiment trains the same probe on the easy capital set and evaluates on all arithmetic statements, with an arithmetic-internal five-fold control, for Qwen and Llama. The hard-task experiment keeps the arithmetic-trained probe of the original direction and evaluates on the hard capital set, with a hard-capitals-internal five-fold control, for Qwen and Llama. Finally, for every probe we classify each of its fifty selected tokens by the sign and presence of its class signal on the target format and compute a weight-weighted sign-agreement score, asking whether transfer

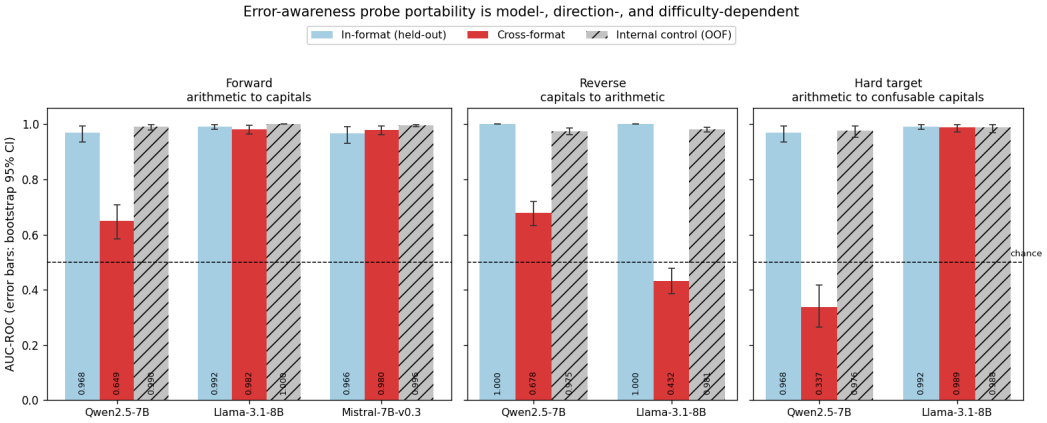

Figure 1: Error-awareness probe transfer across three conditions. Each panel shows, per model, the in-format AUC on the held-out training format, the cross-format AUC on the evaluation format, and a within-evaluation-format control (five-fold out-of-fold), with bootstrap 95% confidence intervals and the chance line at 0.5. Left: arithmetic-trained probes evaluated on capital-city statements; Qwen collapses while Llama and Mistral transfer. Middle: capitals-trained probes evaluated on arithmetic; Qwen's collapse is bidirectional and Llama inverts below chance. Right: arithmetic-trained probes on same-country confusable-city capitals; Llama transfers at its internal ceiling while Qwen falls below chance. Every internal control is high, so failed transfer is never explained by a missing signal.

is predictable from quantities observable at training time. Every headline AUC in this paper was additionally reproduced with an independent rank-based implementation, and every run's summary regenerates byte-identically from its per-example outputs.

3 Results

Figure 1 places all seven conditions side by side, in-format, cross-format, and internal control per model, and carries its data and provenance in a sibling CSV.

3.1 A third model family

Mistral-7B-Instruct-v0.3 patterns with Llama. Its arithmetic-trained probe reaches 0.966 (95% CI 0.930 to 0.992) on held-out arithmetic and 0.980 (95% CI 0.961 to 0.994) on capital-city statements, a gap of negative 0.014, with a within-capitals control of 0.996. In the forward direction the three-family score is two transfers and one collapse, and Qwen is the outlier. The Qwen arm of this comparison is not an imported number; we re-executed the published Qwen pipeline inside this codebase, fresh forward passes on the same class of GPU, and reproduced the published values to four decimals in every arm, with the fresh extraction byte-identical to the original run's cache, so all three families carry the same provenance.

Mistral also breaks the simplest mechanistic story. Llama's clean transfer came with a commitment-channel signal in both formats, while Qwen's collapse came with a capitals commitment gap near zero. Mistral transfers cleanly with a capitals commitment gap that is also near zero, 0.717 against 0.587 on arithmetic but 0.824 against 0.831 on capitals, so its probe must read capitals wrongness through selected tokens other than the period mass. Whatever makes a probe travel, it is not reducible to the single most interpretable feature.

3.2 The reverse direction

Training the same probe on the easy capital set and evaluating it on all six hundred arithmetic statements overturns the simplest reading of the forward results. Qwen's collapse is bidirectional. Its capitals-trained probe reaches 1.000 on held-out capitals and falls to 0.678 (95% CI 0.632 to 0.720)

on arithmetic, almost exactly mirroring its forward number of 0.649, while an arithmetic-internal control reaches 0.975 (95% CI 0.962 to 0.986), so the signal it fails to read is abundantly present. Llama, whose forward transfer was nearly perfect, does not transfer in reverse at all. Its capitals-trained probe reaches 1.000 in format and 0.432 (95% CI 0.387 to 0.479) on arithmetic, reliably below chance, against an arithmetic-internal control of 0.981 (95% CI 0.973 to 0.989).

The below-chance number has a concrete cause visible in the selected features. The capitals training set hands the probe a different vocabulary than arithmetic training did, dominated by punctuation variants and copular continuations rather than by the bare commitment tokens the forward probe rode, and the heaviest of those features reverse their class direction between formats, newline-terminated periods and non-breaking spaces that accompany false capitals but true arithmetic. A probe is exactly as portable as the channel it reads, and the channel is chosen by the training format, not by the model alone. Llama is not a model that transfers; it is a model whose arithmetic-selected channel happens to travel and whose capitals-selected channel happens to invert.

3.3 A hard internal task

The confusable-city dataset does what it was designed to do. The within-task control falls off its degenerate ceiling for both models, to 0.988 (95% CI 0.969 to 0.999) on Llama and 0.976 (95% CI 0.953 to 0.994) on Qwen, so the transfer gap finally has an internal reference, and the probe side of the pipeline reproduces its own anchors exactly, with Llama's in-format arm matching the replication run and Qwen's matching the published run to four decimals under identical seeds.

Against that reference the two models pull even further apart. Llama's arithmetic-trained probe reaches 0.989 (95% CI 0.972 to 1.000) on hard capitals, statistically indistinguishable from the 0.988 its own within-task ceiling allows, so its format-general signal survives plausible errors undiminished. Qwen's probe falls to 0.337 (95% CI 0.265 to 0.416), reliably below chance, against a within-task ceiling of 0.976. The commitment channel again accounts for the difference. After a hard capital statement Llama still withholds the period when the statement is false, a mean probability gap of 0.145, while Qwen commits almost fully to true and false alike, a gap of 0.002, leaving its arithmetic-tuned features nothing to read and what remains anti-correlated. Hard, plausible errors are exactly the regime an error monitor exists for, and they are the regime where the model difference is widest.

3.4 Can transfer be predicted at training time?

The run that replicated Llama attributed its clean transfer to the commitment channel, the probability mass kept on the period token, which separates true from false in both of Llama's formats (gaps of 0.184 and 0.078) but only in Qwen's arithmetic (0.354 against negative 0.014). The new runs show this story is incomplete in both directions. Mistral transfers cleanly with no capitals commitment gap, and on hard capitals Llama keeps a 0.145 commitment gap where Qwen saturates to a 0.002 gap, consistent with their outcomes there but silent on Mistral.

We therefore tested the more general version of the idea directly. For each of the seven (model, training format, target task) conditions in this paper we classified the probe's fifty selected tokens by their class signal on the target format, no signal, same sign as training, or inverted sign, and computed a weight-weighted sign-agreement score that is observable at training time given one labeled sample of the target format. The score is honest but weak as a predictor, with a Spearman rank correlation of 0.357 against the observed transfer AUCs. Its failures are systematic. Strongly negative agreement does coincide with Qwen's forward collapse, and the three clean transfers score between 0.36 and 0.59, but both below-chance conditions carry positive agreement, because anti-transfer is driven by sign inversions concentrated in a handful of heavily weighted features rather than by the census of all fifty. Restricting the score to the heaviest features does not stabilize it. Predicting portability from training-time artifacts remains open, and until it closes, validation on the deployed format pair is the only qualification that means anything.

4 Discussion

The headline question, whether the format-specificity reported for Qwen is a property of the method or the model, now has a three-family answer with one twist. In the forward direction the method is

fine and Qwen is the outlier, since two of three families transfer essentially perfectly and the collapse is a fact about one model's output channels. The twist is that "a model that transfers" turns out to be the wrong abstraction. The same Llama that rides format-general hedging tokens when trained on arithmetic anti-transfers below chance when trained on capitals, because each training format hands the probe a different channel and the channels differ in portability. Portability belongs to the (model, training format, target task) triple, not to the model.

For deployment this compounds rather than relaxes the original lesson. In-format validation looks excellent in every condition of this paper, between 0.966 and 1.000, while cross-format outcomes span 0.337 to 0.989, and the spread is invisible at training time, where our best predictor, the selected-token sign-agreement score, separates one failure mode but misses the below-chance inversions entirely. The hard-task result sharpens the stakes, since plausible, confusable errors are exactly what an error monitor exists to catch, and they are where Qwen's probe goes from degraded to actively misleading, scoring false statements above true ones. An error probe qualified on easy targets can be worse than no probe at all on hard ones. Validation must therefore cover the deployed model, the deployed training format, the deployed target tasks, and target difficulty representative of real errors.

The interpretability lead from the replication also sharpens. The object to explain is no longer why Llama transfers and Qwen does not, but why arithmetic training selects a portable channel in two of three families while capitals training selects an inverting one in the same family that transferred, and whether the activation-level truth directions of Marks and Tegmark [2023] show the same triple-dependence that the output channel does.

5 Limitations

Three model families can refute universality but cannot establish a population claim about open-weight models in general; the honest claim is bounded by the sample, and only the forward direction has three families, with the reverse and hard-task arms run on two. All experiments share one prompt convention and instruction-tuned chat models, so base models or other framings could relocate the wrongness signal. The internal controls select features on the full target set before cross-validating, as in the published study, and should be read as mild upper bounds; Llama's hard-capitals control in particular sits at 0.988, only just below the preregistered 0.99 ceiling criterion, so its at-ceiling transfer reading leans on a control that barely cleared its own bar. The hard capital set, while label-clean by curation, narrows to countries with a famous non-capital city, which skews it toward better-known geography. The sign-agreement predictor was evaluated descriptively over seven conditions, far too few for a calibrated claim about training-time predictability; we report its rank correlation and its systematic failure mode, not a significance statement. Each probe is a single fitted model per (family, training format) under one fixed seed, as in the published study, so seed-to-seed variation in the selected features is not measured.

6 Conclusion

A token-level error-awareness probe that collapsed across formats on Qwen2.5-7B transfers essentially perfectly on Llama-3.1-8B and Mistral-7B-v0.3, so the collapse is a property of the model, not the method. But transfer is not a property of the model either. Llama's probe anti-transfers below chance when the training format is swapped, Qwen's falls below chance when the target task gets hard, and the cleanest training-time statistic we could construct does not predict which of these happens. Format-specificity of error awareness is model-dependent, direction-dependent, and difficulty-dependent, and a deployed error probe should be trusted exactly as far as it has been validated on the model, format pair, and error difficulty it will actually face.

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.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023.

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.