Wrongness lives in the reasoning, not just the answer: cross-position transfer of a token-level error probe in chain-of-thought

Abstract

A linear probe on a language model's hidden states can detect whether a stated arithmetic fact is wrong, and recent work on chain-of-thought faithfulness has made it urgent to know where in a reasoning trace such a wrongness signal lives [@chen2025reasoning] [@turpin2023language]. If an overseer trains an error probe only on the final answer of a multi-step solution, does the same direction detect errors made earlier, at intermediate reasoning steps, or is wrongness encoded position by position so that monitoring the answer alone is blind to mistakes in the reasoning? We study Qwen2.5-7B-Instruct [@qwen2025technical] on controllable multi-step arithmetic chains of thought in which each step's stated result is independently correct or corrupted, and we read the residual-stream hidden state at the result token of every step. Probes are fit with chain-disjoint cross-fitting so the intermediate and final representations of the same chain never straddle the train and test split, and every transfer is evaluated layer by layer with bootstrap intervals [@alain2016understanding] [@belinkov2022probing]. Three findings hold. First, step wrongness is near-perfectly linearly decodable from a single hidden layer, with in-position AUC 0.980 for the final computation step and 0.991 for intermediate steps at layer 27. Second, the wrongness direction transfers substantially across reasoning positions in both directions, a probe trained on intermediate steps reaching 0.979 on the final answer and one trained on the final answer reaching 0.879 on intermediate steps; the gap is not a training-set-size artifact, since it persists at a matched cap of 320 examples, but it is partly one of training diversity, because the pooled intermediate set spans three step positions and no single intermediate position transfers as well, leaving only a modest position-type asymmetry concentrated at the steps nearest the answer and absent at the step furthest from it. Third, the token where the model restates its committed answer is a blind spot: it carries only a weak in-position wrongness signal (AUC 0.796) and computation-trained probes barely reach it at all (0.528 from the final computation step, 0.578 from intermediate steps, both near chance). The safety reading is that a token-level error monitor should be trained on reasoning steps rather than answers, that intermediate-step supervision generalizes best, and that the committed-answer token, the natural place an overseer would look, is the worst place to read.

Setup

We study Qwen2.5-7B-Instruct [@qwen2025technical], the open-weight model used by the parent error-awareness line, which has twenty-eight transformer layers and so twenty-nine hidden states including the embedding layer, each of width 3584. The stimuli are controllable multi-step arithmetic chains of thought. Each chain presents a worked solution of four steps, each step a stated equation of the form "lhs op rhs = result" continuing from the previous line, followed by an "Answer:" line that restates the last step's result. Every step's stated result is, independently with probability one half, replaced by a wrong value offset by a nonzero delta, so a step is an error exactly when its stated result does not equal the arithmetic of its as-written operands. Because each step computes on the stated result of the previous line, the four steps are independently correct or incorrect; across 400 chains this yields 1200 intermediate step instances (steps one through three), 400 final computation instances (step four), and 400 answer-restatement instances, each near balanced at roughly 51 percent errors.

We read the all-layer residual-stream hidden state at the terminal token of each stated result, located by the fast tokenizer's character offset mapping. Steps one through three are the intermediate positions, step four is the final computation answer, and the "Answer:" line is a separate diagnostic position. The final computation step and the intermediate steps are both genuine arithmetic claims, so comparing them is a like-for-like test of position; the answer-restatement token is a copy rather than a computation, and we treat it apart.

At each layer we standardize the hidden state and project it to 256 principal components, a label-blind reduction that denoises the 3584-dimensional state and makes the layer-by-layer sweep tractable [@belrose2023eliciting]; the probe is then a balanced L2 logistic regression on the projection [@alain2016understanding] [@belinkov2022probing]. Every AUC, in-position and cross-position, is computed

with five-fold cross-fitting grouped by chain identity, so no chain contributes rows to both the train and the test side of any comparison and the cross-position transfer cannot be inflated by the shared context of a single forward pass. In-position AUC trains and tests at the same position under this grouped scheme; cross-position transfer trains a probe at one position and scores it on the held-out chains' rows at the other. We attach a 2000-sample bootstrap 95 percent interval over the pooled out-of-fold scores to every headline number. The headline layer is the layer of maximum mean in-position AUC excluding the final, output-adjacent hidden state, whose representation is essentially the next-token logit direction and would conflate a hidden-state finding with a logit readout.

Step wrongness is near-perfectly decodable, and transfers across positions but asymmetrically

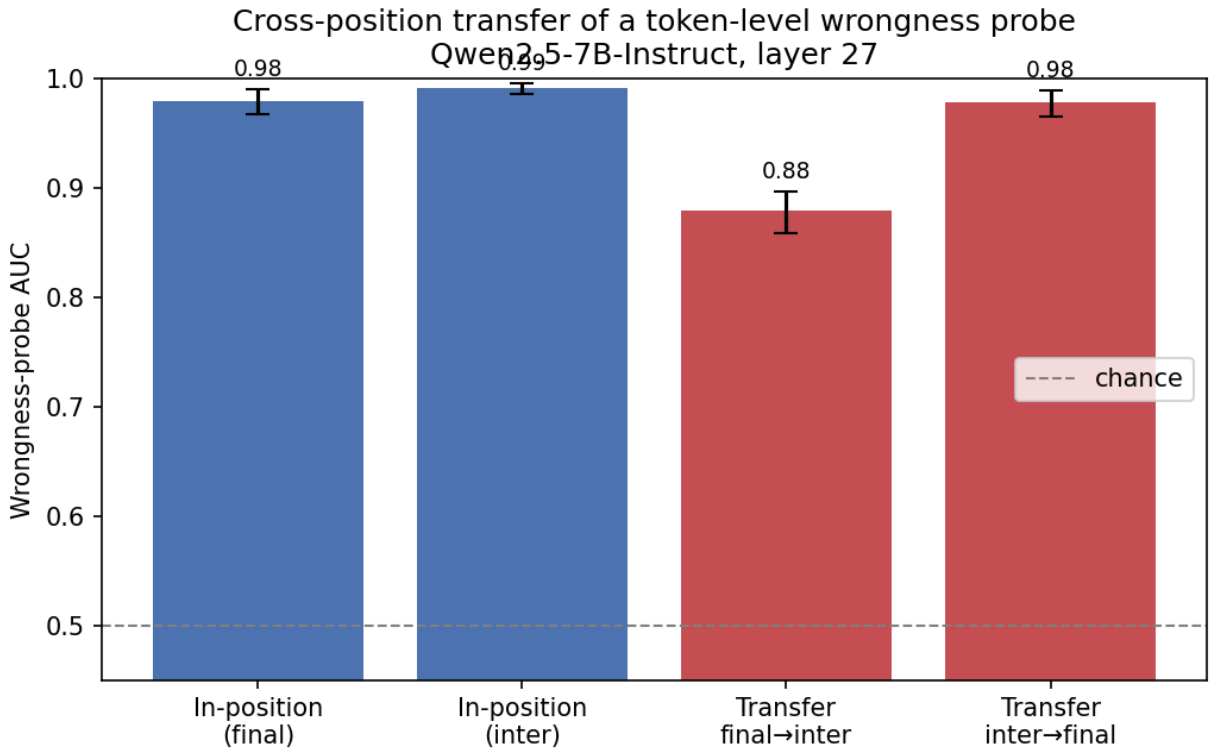

Figure 1: Best-layer AUC for in-position probes and the two cross-position transfer directions, with bootstrap 95 percent intervals and the chance line

analysis	AUC	95% CI	n eval
in-position, final computation	0.9799	0.9674 to 0.9908	400
in-position, intermediate steps	0.9912	0.9858 to 0.9956	1200
transfer, final to intermediate	0.8789	0.8589 to 0.8971	1200
transfer, intermediate to final	0.9789	0.9654 to 0.9898	400

At the headline layer 27 the model represents the wrongness of a single arithmetic step almost perfectly: a probe scores final-step errors at AUC 0.980 and intermediate-step errors at 0.991 in position. This is not surprising for two-digit arithmetic read near the top of the network, and we treat these in-position numbers as a ceiling rather than a result. The load-bearing comparison is the cross-position transfer, where the probe is trained at one reasoning position and tested on chains it has never seen at the other. The wrongness

direction does transfer: a probe trained on intermediate steps ranks final-answer errors at 0.979, almost its in-position ceiling, and a probe trained on the final answer ranks intermediate-step errors at 0.879, well above chance but a clear nine points below the intermediate ceiling. The two directions are not equal. Wrongness learned at intermediate steps generalizes to the final answer better than wrongness learned at the final answer generalizes to intermediate steps, which says the intermediate-step representation of an error is the more general one and the final-step representation is a partial special case of it.

How much of the asymmetry is real, and how much is training diversity

The intermediate pool is three times larger than the final pool and spans three step positions rather than one, so an intermediate-trained probe sees both more data and more positional variety, either of which could explain its edge. We separate these. Capping every probe's training rows to 320, the number a final-trained probe has per fold, lowers the intermediate-to-final direction from 0.979 to 0.955 while the final-to-intermediate direction is unchanged at 0.879, so raw training-set size is not the explanation. Diversity is part of it, however. When the intermediate probe is restricted to a single step position, matched to the final pool in both count and positional variety, no single position reaches the pooled 0.979: step three transfers to the final answer at 0.960, step two at 0.926, and step one at only 0.828. The pooled intermediate advantage is therefore partly the benefit of spanning several positions, not a uniformly more general single-position representation. What remains is a modest, position-dependent asymmetry: for the steps nearer the answer a single-step probe still transfers to the final answer better than the reverse, 0.926 against 0.847 at step two and 0.960 against 0.936 at step three, but at the step furthest from the answer the asymmetry vanishes and slightly reverses, 0.828 against 0.843 at step one. We therefore read the asymmetry as real but modest and concentrated near the answer, not as a clean statement that intermediate representations are globally more general. The transfer is, separately, not an artifact of position-specific mean offsets in the features: subtracting each position's own feature mean before transfer leaves both pooled directions essentially unchanged, at 0.879 and 0.979, so a probe is reading a shared wrongness direction rather than a constant shift that happens to separate the classes.

The committed-answer token is a blind spot

analysis	AUC	layer
in-position, answer restatement	0.7963	27
transfer, final computation to answer	0.5276	27
transfer, intermediate to answer	0.5779	27

The "Answer:" line restates the value the model already computed at step four, a copy rather than a fresh computation, and it behaves unlike any computation token. In position it carries a real but much weaker wrongness signal, AUC 0.796 against the 0.98 to 0.99 of the computation tokens, so the commit site partly knows the answer is wrong but encodes it far less linearly. More striking is that computation-trained probes barely reach it at all: a probe trained on the final computation step scores the restatement token at 0.528 and an intermediate-trained probe at 0.578, both close to chance. The wrongness representation at the commit token lives in a different subspace from the one at the computation tokens, so the single position an overseer would most naturally monitor, the place where the model announces its answer, is the position a probe trained on the reasoning is most blind to.

Transfer into the reasoning does not collapse with step depth

We also asked whether the final-answer probe's reach into the reasoning depends on how deep the intermediate step sits in the chain, since a step further from the answer might be represented more differently. The final-trained probe reaches the step immediately before the answer best, at AUC 0.936, and the two earlier steps less well, at 0.843 and 0.847, so its reach is strongest just behind the answer and weaker deeper in the chain, but it stays well above chance at every depth. The final-trained probe is thus partially but consistently blind

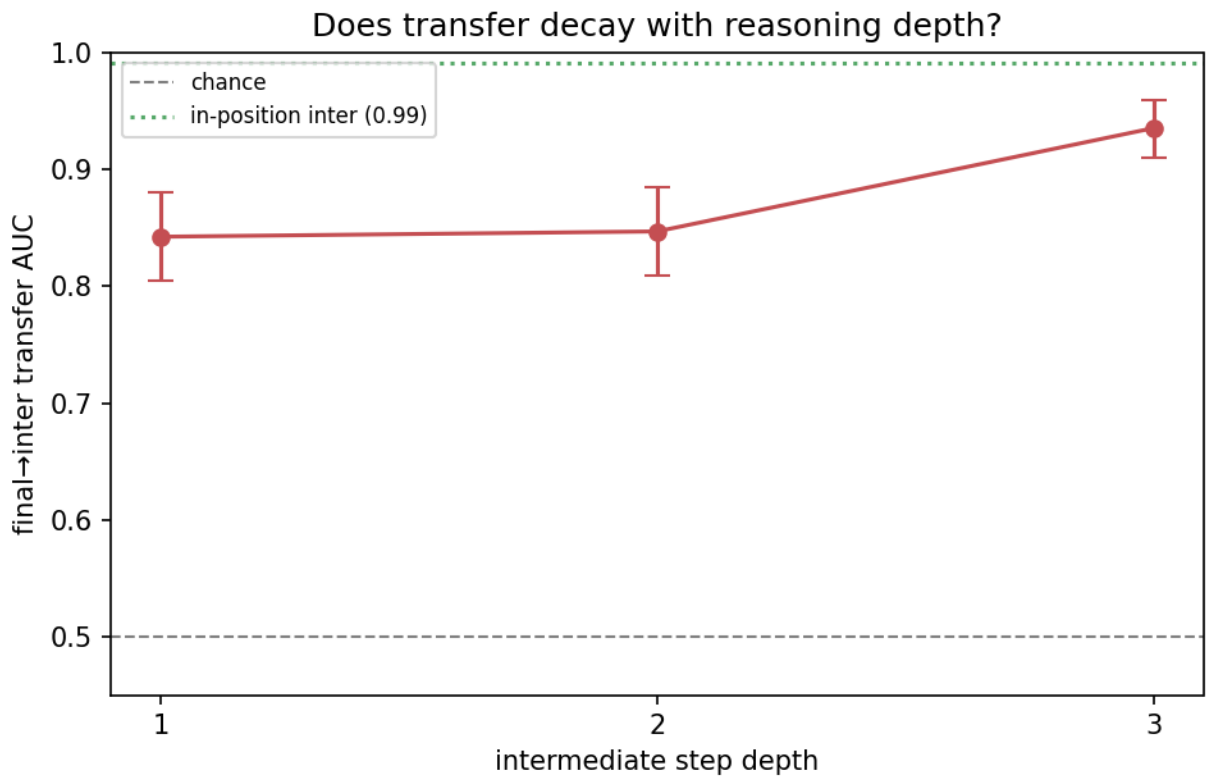

Figure 2: Final-trained probe AUC evaluated at each intermediate step depth, with bootstrap intervals and the in-position intermediate ceiling

across the reasoning, more so at the steps furthest from the answer, rather than reaching only the step it sits next to.

Wrongness emerges in the upper-middle layers

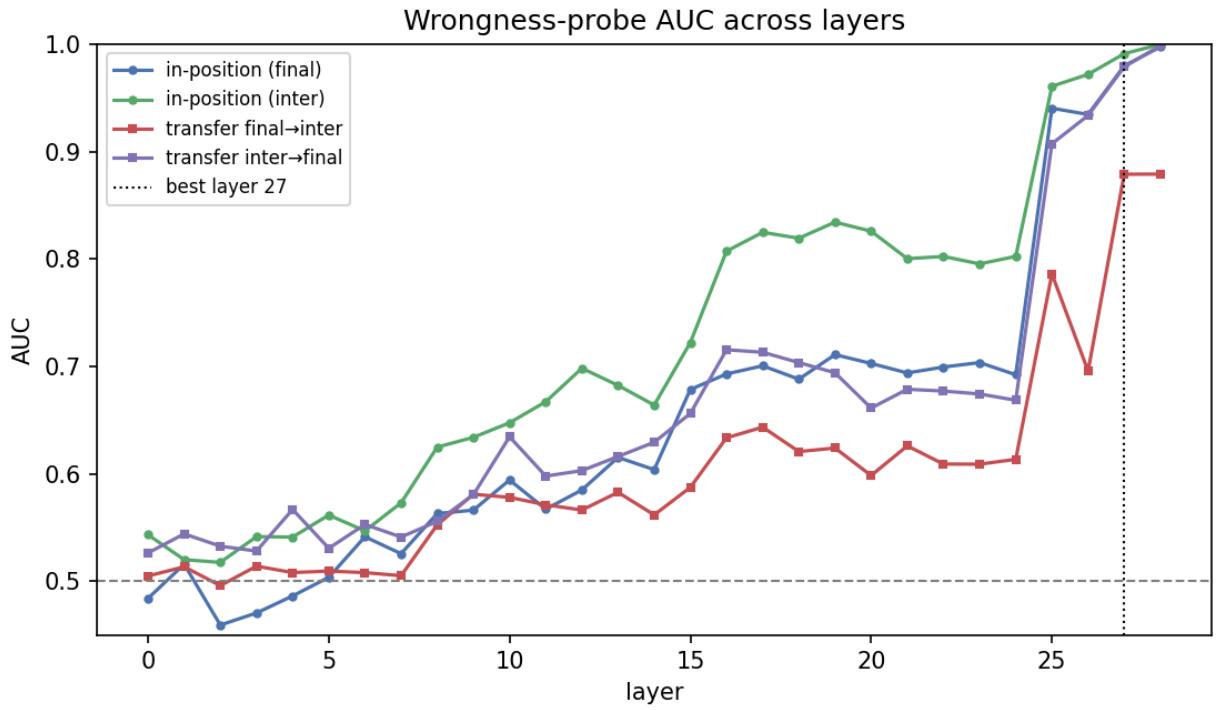

Figure 3: In-position and cross-position AUC as a function of layer depth

The depth profile shows where this structure lives. In the embedding and earliest layers all four curves sit near chance: there is no linearly decodable wrongness near the input. The in-position and transfer AUCs rise together through the middle of the network, crossing into clearly-above-chance territory around layer 15 and climbing to their near-ceiling values by the high twenties. The final-to-intermediate transfer curve tracks below the other three across the whole upper stack, the depth-resolved signature of the same asymmetry seen at the headline layer, and the gap is stable rather than confined to one layer. The very last hidden state, which we exclude from headline selection because it is essentially the next-token logit direction, reaches in-position AUC near 1.0 but its final-to-intermediate transfer stays at 0.879, so reading deeper does not close the asymmetry.

A layer-by-layer view confirms a contiguous representation

Training the final-answer probe at each layer and testing it on intermediate steps at every layer, on a strided grid of layers, shows the high-transfer cells concentrated on and near the diagonal in the upper part of the network. Matched-depth transfer is best, reaching 0.879, while mismatching the train and evaluation layer degrades it sharply, with the best off-diagonal cell only 0.637. The shared wrongness direction is therefore depth-aligned: the final-answer representation of an error at a given layer matches the intermediate representation at that same layer, and reading across depths loses the alignment, which is the signature of a linear feature that sits at a consistent depth across positions rather than one that is freely transportable across layers [@marks2023geometry] [@belrose2023eliciting].

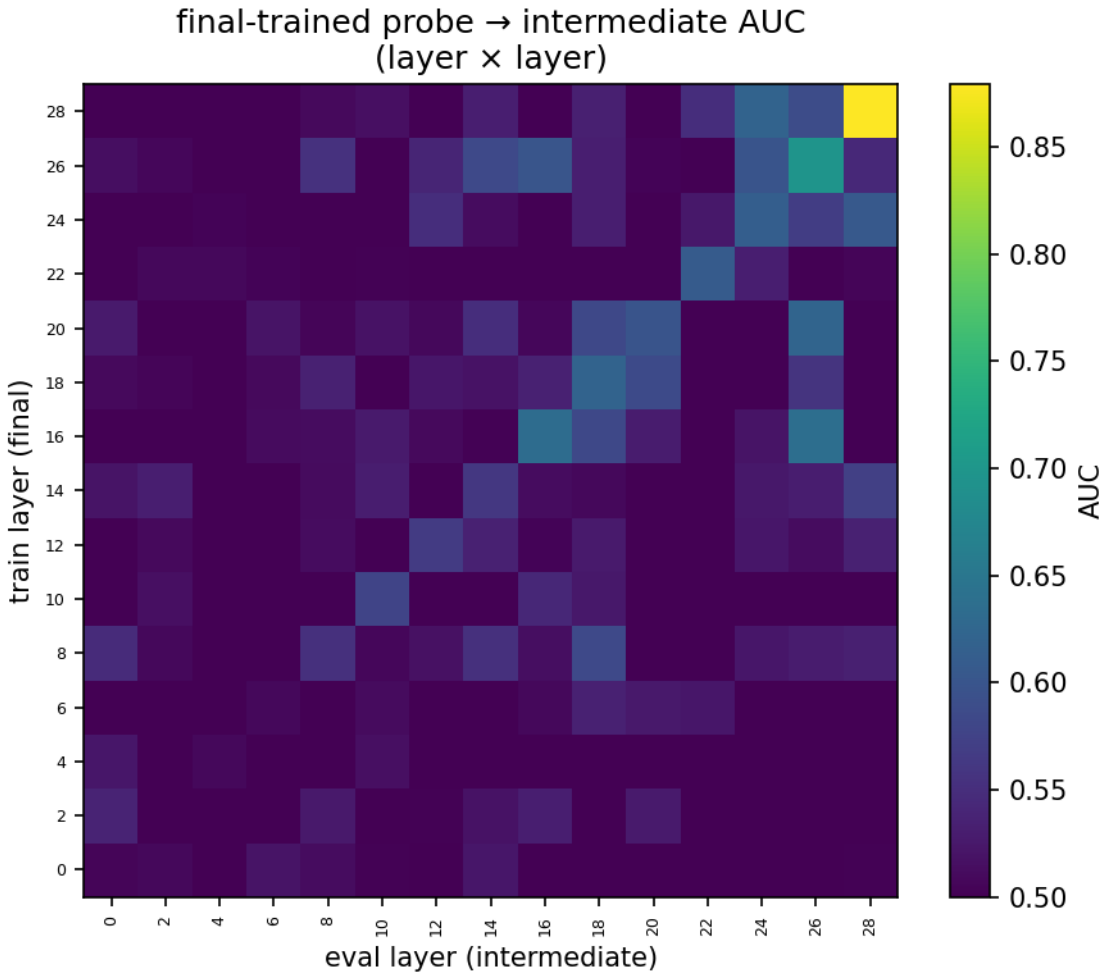

Figure 4: Layer-by-layer transfer: a probe trained on final-answer activations at one layer, evaluated on intermediate-step activations at another

What this means for error monitoring

The result speaks directly to where an activation-level error monitor should read. Wrongness at a reasoning step is strongly and linearly present in this model's hidden states, in line with work showing models carry internal correctness and truth signals concentrated in a few directions and positions [@azaria2023internal] [@burns2022discovering] [@kadavath2022language] [@zou2023representation] [@li2023inference], and that signal is largely shared across reasoning positions, so a single probe can in principle watch a whole chain rather than one token [@wei2022chain]. But the sharing is asymmetric and incomplete in a way that matters for oversight. A monitor trained only on final answers misses roughly a tenth of the rank ordering on intermediate errors, while a monitor trained on a spread of intermediate steps catches final errors almost perfectly, and the spread is what matters, since training across several step positions transfers better than any single position does. Process-style supervision that labels many intermediate steps is therefore the better source for a position-general detector, echoing the value of diverse step-level signals in training [@cobbe2021training] [@lightman2023lets]. Most cautionary is the answer-restatement blind spot: the committed-answer token, the obvious target for a cheap monitor and the token a faithfulness auditor reads, is where computation-trained probes are nearest to chance, so error monitoring that watches only the model's stated answer can be systematically blind to errors its own reasoning steps represent clearly [@lanham2023measuring] [@lyu2023faithful].

Limitations

The strongest caveat is interpretive. The probe predicts whether a stated step is arithmetically wrong, and on these constructed chains a wrong step is a presented error rather than one the model spontaneously made, so what we measure is the model's read-out of the correctness of content in its context, not its awareness of its own generative mistakes; the design cannot separate a genuine error-monitoring representation from a general linear representation of arithmetic truth [@marks2023geometry] [@kadavath2022language]. The in-position AUCs near 1.0 are a high-dimensional ceiling on easy two-digit arithmetic read in the upper layers, so they should be read as upper bounds, and the cross-position transfer numbers, which test on fully held-out chains under chain-grouped cross-fitting, as the load-bearing evidence. The headline layer is a maximum over twenty-nine layers selected on in-position AUC, and although we exclude the output-adjacent layer and report the full depth profile so the band rather than the peak carries the claim, a single layer chosen this way understates uncertainty. The evidence is one model, one task family of single-step arithmetic claims embedded in a chain, single-token linear probes, and a correlational readout; a causal steering test along the recovered direction, a non-arithmetic reasoning format, naturally generated rather than corrupted errors, and a second model family would each strengthen the claim and are the follow-ups we would run first [@chen2025reasoning] [@turpin2023language].

Conclusion

A token-level wrongness probe trained on one position of a chain of thought transfers to another, but not uniformly. On Qwen2.5-7B-Instruct, intermediate-step and final-answer computation tokens share a near-ceiling wrongness direction that transfers asymmetrically, better from intermediate steps to the final answer than the reverse, and the asymmetry survives matched training budgets. The token where the model restates its committed answer is a representational blind spot that computation-trained probes read at little above chance. Wrongness lives in the reasoning, and an error monitor that watches only the answer is looking in the one place the signal is hardest to find [@alain2016understanding] [@belrose2023eliciting].