THE RULER DECIDES: TWO STANDARD REASONING DIAGNOSTICS MEASURE OTHER THAN CLAIMED

Anonymous authors
Paper under double-blind review

ABSTRACT

Same-weight dual-mode models such as Qwen3 offer a controlled pairing: one weight set, one prompt, and only an `enable_thinking` switch. We audit two standard reasoning diagnostics on this pairing: the leading variance share $1 - \text{top1_var}$ of the cross-problem difference matrix $\delta = h_{\text{think}} - h_{\text{no-think}}$, and binary agreement $\mathbf{1}[a > 0.5] == \mathbf{1}[b > 0.5]$ read as metrics of divergence. In this paper, we ask what each actually measures. On MATH500 (0.6B–32B) the leading direction is mainly the **question-invariant (mean) direction**: $|\cos(u_1, \text{mean})| \geq 0.994$ **at L8** across the transition ($j \leq 6$; 0.6B–14B; ≥ 0.95 question-disjoint split-half). The familiar curve is therefore a **question-specific energy-share curve**, not a measure of rank structure; a **non-thinking control** shows the same drop, so its existence is not evidence of a phase transition. We build a content-graded ladder of five controlled continuations. Its crossing position orders consistently with the contrast's real content, but monotonicity does not survive a family-wise joint test. That position shows no monotone scaling across six sizes, and a pre-registered difficulty axis fails outright.

1 INTRODUCTION

Qwen3 ships a same-weight dual-mode design: one `enable_thinking` switch decides whether the model writes out a chain of thought, with weights, prompt, decoding and seed held fixed, so any representational difference between the branches comes only from whether the model reasons. This pairing is a standard protocol on Qwen3 1.7B–32B (Senthil Kumar, 2026). We use it to ask what the standard summary statistics of the resulting contrast actually report, along the chain.

Sahoo et al. (2026) showed that linear probes reading "reasoning mode" collapse to chance once the *format* direction is residualised out; Wang et al. (2026) argued that no-think is "softened thinking", so δ cannot collapse to a single clean low-rank *computational* direction. Both say what δ fails to be; neither settles what the standard summary statistic *of* δ reports. We claim the following.

1. **Two standard diagnostics measure other than claimed, by one mechanism.** In both, one side of a constructed pair is near-constant, so a quantity named for "the difference between the two sides" silently degenerates into a monotone function of the other side.
2. **Instance $\mathbb{A}$.** The leading direction of δ is the question-invariant (mean) direction: $|\cos(u_1, \text{mean})| \geq 0.994$ at L8 ($j \leq 6$; 0.6B–14B; question-disjoint split-half ≥ 0.95). Hence $1 - \text{top1_var}(\delta_j)$ is a *question-specific energy-share curve*, and a non-thinking control decays just as much. The crossing position x_0 is *directionally consistent* with the contrast's real content ($\rho = +0.89$ to $+0.96$) but *does not survive a family-wise joint test* (Holm 0/15), and its scale stability *depends on the layer*.
3. **Instance $\mathbb{B}$.** $\mathbf{1}[a > 0.5] == \mathbf{1}[b > 0.5]$, with a verbal confidence and b a probe, is claimed to measure *divergence*; it re-encodes *the probe's sign*, equal to $\mathbf{1}[\text{probe} > 0.5]$ on every record where the verbal side is above the threshold and so differing from it on only 5/128 and 6/126 records (4B/8B), inverting on the records where verbal confidence varies and falling to chance (0.469) there (0.6B).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

4. **Two supporting negatives**: the pre-registered difficulty axis fails outright, and δ norm does not distinguish answer-changing from answer-preserving questions.

2 Related Work

Sahoo et al. (2026) is our primary methodological reference: a Qwen3-14B probe reads "reasoning mode" at 100% but collapses to chance once the format direction is residualised. We confirm its warning — in our chain's opening segment the leading direction of δ is exactly such a format direction — and reply by decomposing that direction along the chain. Of our two probes the *mode* probe's conclusions are void by pre-registration, and the *answer* probe's are retained, evaluated out-of-fold on a question-disjoint split.

Wang et al. (2026) says no single clean low-rank *computational* direction separates the modes; we say the chain-opening leading direction of δ is the question-invariant mean direction, which is precisely *not* a computational structure — the direction that work predicts must exist. Yang et al. (2026) shows the thinking switch is a template-level input signal, the obvious deflationary explanation for x_0: is $x_0 \approx$ 5–6 simply the prefix length? §5.2 answers with the zero-content prefix. Senthil Kumar (2026) is our closest neighbour in substrate — Qwen3 1.7B–32B same-weight ON/OFF, plus cross-family models and activation patching — and finds that thinking *redistributes* errors rather than uniformly helping; we occupy its experiment, not its conclusion. Seddik & Fard (2026) finds representations separate task types reliably while failing to separate problems within a task, hence the three pure-CPU controls of §5.1, which cap the late segment at "reproducible structure".

The performative sub-question is crowded (Boppana et al., 2026; Li et al., 2026a;b; Lewis-Lim et al., 2025; Parekh & Goyal, 2026); we make no independent performative claim, and our position is the Qwen3 same-weight `enable_thinking` pairing with per-position hidden states and a format/question-specific decomposition.

3 Method

3.1 Paired collection

Same weights, same prompt, same seed, fixed decoding; only `enable_thinking` is toggled. Seven branches are collected: `c1` (think), `c2` (no-think), `c3` (chain emptied), `c4` (same-length unrelated filler), `c5` (no-think prefixed with the `<think>` marker), and `c6` (random content, §5.4). Hidden states are recorded at every layer for the last token, plus a per-position subset: the first $K = 24$ continuation tokens at layers $\{4, 8, 12\}$, which is the greedy collection's window — the sampling leg of §5.4 recomputes the common-prefix length per token and is not the same window. Fig. 1 draws the design and the contrast it feeds.

3.2 δ and low-rank statistics

For each problem q and continuation position j, $\delta_j(q) = h^{\mathtt{c1}}_{L,j}(q) - h^{\mathtt{c2}}_{L,j}(q)$. $\mathrm{top1_var}(\delta_j)$ is the largest eigenvalue divided by the eigenvalue sum of the $n \times n$ Gram matrix over problems (the fast path, since $n \ll \dim$). `effective_rank` is the effective rank in the eigenvalue-entropy sense.

3.3 Crossing-position fitting

The curve is $y_j = 1 - \mathrm{top1_var}(\delta_j)$ over $j = 0 \ldots 23$, fitted with a bounded sigmoid $\mathrm{sigmoid}(j) = lo + (hi - lo)/\big(1 + \exp(-k\,(j - x_0))\big)$, rising from lo to hi with $k > 0$. The reported R^2 is the squared Pearson correlation between the observed series and the fitted sigmoid, not $1 - \mathrm{RSS}/\mathrm{TSS}$; for Qwen3-0.6B at L8 the two definitions give 0.985 and 0.941. The fit is globally rising but no series is monotone: `top1_var`(j) is non-increasing in j in 18/18 model–layer combinations across the three layers (129 violations).

One fitting convention is load-bearing. `c1`, `c3`, `c4` and `c5` all begin with `<think>` while `c2` does not, so any pairing of two `<think>` arms is token-identical at position 0, where δ is float noise

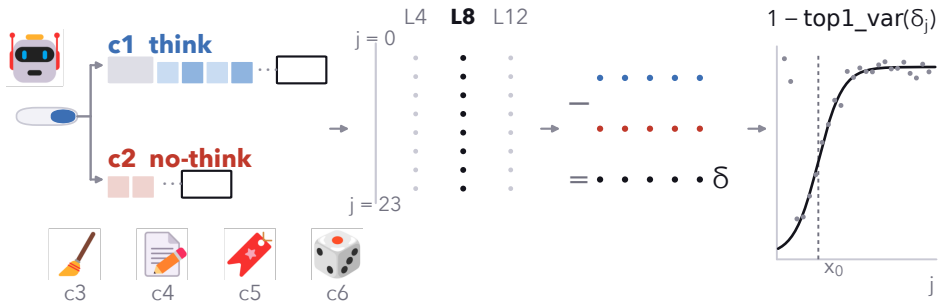

Figure 1: The pairing and the contrast. One weight set, one switch: `c1` writes a chain and then the boxed answer, `c2` writes the answer directly, so the same answer sits at a different position in each branch; `c3`–`c6` are the controlled continuations. Hidden states are recorded at every position (vertical axis j) for layers 4, 8 and 12, and the contrast $\delta_j = h^{c1} - h^{c2}$ is formed position by position. The curve on the right is $1 - \mathrm{top1_var}(\delta_j)$, with its crossing position x_0.

(norm ~ 0.2 against ~ 20 elsewhere) and y spikes to ≈ 0.93 against a curve otherwise in 0.02–0.43. The fix is structural rather than threshold-based: tokenise both arms, drop the leading positions whose tokens are identical, and add the number dropped, d_{drop}, back for cross-pairing comparison. The comparable coordinate is $x_0^{\text{fix}} + d_{\text{drop}}$, and it is the only one we use.

4 Experimental Setup

Models and data. Qwen3-0.6B / 1.7B / 4B / 8B / 14B / 32B, all same-weight dual-mode. MATH500; difficulty is the model-independent `math_level` field. The chain-axis x_0 fitting covers 0.6B–32B ($n = 100$, cap 8192, on the first 100 problems of the 14B set), and so does the leading-direction alignment $|\cos(u_1, \text{mean})|$, which we extended to 32B after Table 1 was frozen (§5.1, appendix). The remaining statistics — the `c1`–`c6` pairing fits, the difficulty stratification and the reliability statistics — cover 0.6B–14B only.

Problems are drawn without replacement with a fixed seed, and all six scales share the same problems nested (0.6B–8B take 200; 14B/32B the first 100 of those), with the five smaller models' prompts byte-identical to 8B's, so cross-scale comparison has no "different problems per scale" confound. There is no train/test split; held-out support is question-disjoint split-half and grouped cross-validation.

Decoding. Every table except the sampling leg is a single greedy run (`do_sample: false`, temperature 0.0, top_p 1.0, seed 42); the sampling leg uses temperature 0.7, top_p 0.8 and is never placed in the same table as a greedy number. The generation budget is 8192 for the main and large-model legs and 2048 for the external-validity rerun, measured from the data rather than read off the configuration (appendix).

Sample sizes. Nominal n is 200 for 0.6B–8B and 100 for 14B and 32B; effective n differs per statistic, and every table carries its own axis's n: per-position/δ 200/200/200/200/100/100 with no attrition, answer-comparison 140/165/172/167/87/89, random-content control 140/165/172/167/87 (no 32B artifact), probe-alignment 73/–/128/126/–. The last two axes coincide numerically but are different.

Conventions. Answers are re-parsed cleanly throughout (the earlier substring match of the gold answer is retired), and the per-position layers $\{4, 8, 12\}$ are a pre-fixed study-design choice.

5 Results

Two diagnostics, one failure: §5.1–§5.4 audit the first, §5.5 the second, and §5.6 collects the supporting negatives.

Table 1: Instance $\mathbb{A}$: the leading direction is the question-invariant direction. $\mathrm{shared_frac}_j$ is the fraction of δ_j energy along the between-problem mean; $|\cos(u_1, \mathrm{mean})|$ is the angle between the leading Gram eigenvector and that mean, taken as the minimum over $j \leq 6$. Held-out values use a question-disjoint random split: the mean is fitted on one half and u_1 on the other, five times at a fixed seed, reporting the minimum over both splits and positions (readout `p0_heldout_alignment_20260926`). The null for a random Gaussian direction is 0.006–0.035. All statistics are at L8 unless the column says otherwise, and cover 0.6B–14B. The δ axis carries $n = 200$ problems at 0.6B–8B and $n = 100$ at 14B, with no attrition; the alignment statistic also holds on 32B (§5.1 and appendix), measured after this table was frozen.

	$\mathrm{shared_frac}_j$ (L8)					$\lvert\cos(u_1, \mathrm{mean})\rvert$, $j \leq 6$		
Model	j=0	j=4	j=8	j=12	j=23	L8	L8 held-out	L12 held-out
Qwen3-0.6B	0.9865	0.5481	0.2688	0.0600	0.0205	0.9970	0.9872	0.9801
Qwen3-1.7B	0.9884	0.5268	0.2338	0.0566	0.0234	0.9942	0.9711	0.9623
Qwen3-4B	0.9865	0.5308	0.2865	0.0536	0.0250	0.9965	0.9817	0.9822
Qwen3-8B	0.9698	0.5152	0.2841	0.0519	0.0221	0.9941	0.9679	0.9691
Qwen3-14B	0.9464	0.5513	0.2754	0.0602	0.0310	0.9971	0.9751	0.9792

5.1 Instance $\mathbb{A}$: the curve is a question-specific energy share

The first diagnostic is the leading variance share $y_j = 1 - \mathrm{top1_var}(\delta_j)$, read as a measure of rank structure. To ask what it actually tracks we measure the fraction of δ_j energy along the single direction shared by all problems,

$$\mathrm{shared_frac}_j = \frac{\|\mathrm{mean}_q\, \delta_j(q)\|^2}{\mathrm{mean}_q \|\delta_j(q)\|^2},$$

and $|\cos(u_1, \mathrm{mean})|$, the angle between the leading eigenvector of the uncentred Gram matrix and the between-problem mean, mapped back to feature space. A direction common to all problems cannot by definition be question-specific computation; it can only be template or format. Both statistics use the same uncentred Gram as the main curve, so they are on one ruler.

Table 1 gives the measurement. At the start of the continuation 95–99% of the energy of δ is the shared (question-invariant) direction; by $j = 23$ only 2–3% is. Over the whole sigmoid transition region, $j \leq 6$, the leading direction is the shared direction to within $|\cos| \geq 0.9941$ at L8 in all five scales, and the alignment is not an in-sample artefact: with a question-disjoint split it is ≥ 0.9679 at L8 and ≥ 0.9623 at L12. Consequently $y_j \approx 1 - \mathrm{shared_frac}_j$ to first order, and the familiar curve is a *question-specific energy share*, not a measure of rank structure or computational complexity; the approximation can also be made exact rather than numerical, since the shared direction is the all-ones vector in sample space, so $\mathrm{top1_var} \geq \mathrm{shared_frac}$ with no exception in 120 model-by-position combinations. The alignment holds on the 32B chain-axis collection too, which uses the same protocol at $n = 100$: the minimum over $j \leq 6$ is 0.9947 at L8 and 0.9970 at L12, so across six scales the L8 minimum remains the 8B value of 0.9941. That reading was taken after the table was frozen and is recorded rather than folded into it (appendix).

The decline of $|\cos|$ is not confined to the thinking comparison: non-thinking controls decay to the same magnitude, so the existence of the decay is a general property of generation along a chain; what survives is the crossing position x_0, where the format share is drowned by question specificity.

5.2 A content-graded ladder

We now ask whether x_0 — the crossing position of Fig. 2(a) — is informative, by building a content-graded ladder of seven contrasts between controlled continuations that differ in how much real problem content separates their two arms (Table 2; the rung-by-rung figure is Fig. 5 in the appendix). The design establishes one thing. Structural contrasts cross earliest (filler 1.7–2.9, empty chain 2.9–3.7); the naive think/no-think delta sits in the middle (5.0–6.1); the contrasts that isolate content cross latest (6.0–8.1); and the separation holds under the paired bootstrap over questions in all but 9 of the 90 steps, those 9 being written as $\approx$ or merged by the same rule. That is the reliable content of the ladder: x_0 separates a structural contrast from a content-isolated one.

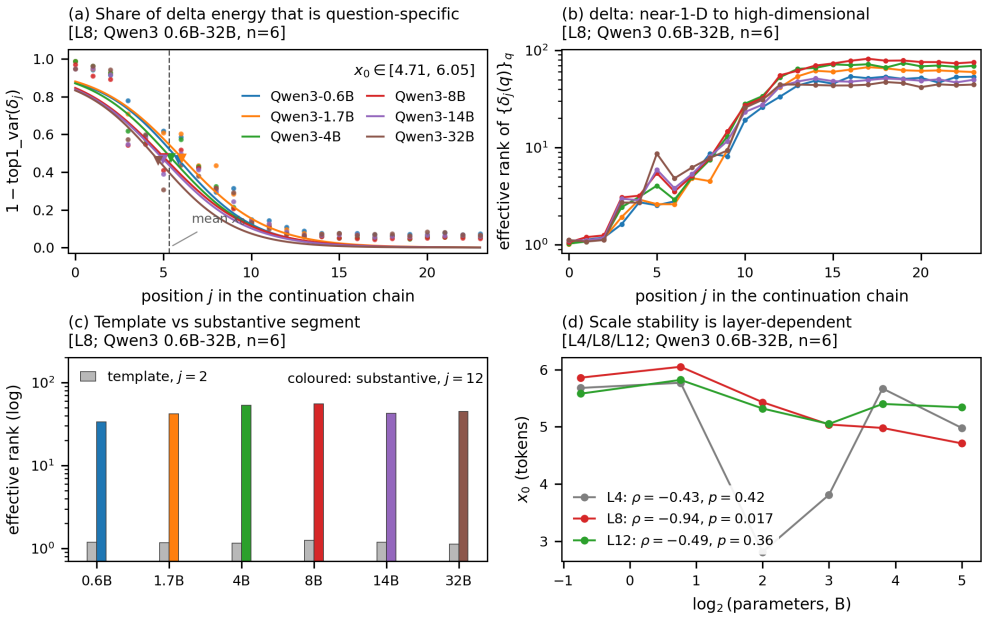

Figure 2: The chain-axis curve and its crossing position, six scales, L8 unless marked. (a) $1 - \text{top1_var}(\delta_j)$; points are the measured series, lines the fitted sigmoid, triangles the fitted x_0. (b) Effective rank of $\{\delta_j(q)\}_q$ against j, log axis. (c) Effective rank at the template position $j = 2$ against the substantive position $j = 12$. (d) Fitted x_0 against $\log_2$ parameters, one line per layer. The ρ and p in the legend are Spearman correlations over $n = 6$ scales with an *exact* two-sided permutation p over all 720 orderings; they are not asymptotic.

Table 2: The content-graded ladder. Each cell is $x_0^{\text{fix}} + d_{\text{drop}}$, the corrected fit mapped back to original token coordinates, at L8; the last row is the range over both layers. Contrasts are ordered by how much real content separates the two arms. R^2 is over the 70 fits in the source table; the permutation null is the same sigmoid fitted to position-shuffled data. The ordering is directionally consistent but does not survive a family-wise joint test (§5.4). Every cell is one fit on the same $n = 200$ / $n = 100$ problems as §5.1.

Pair	Construct	0.6B	1.7B	4B	8B	14B
`c4--c2`	same-length unrelated filler	1.71	2.14	2.82	2.64	2.85
`c3--c2`	empty think chain	2.94	3.70	3.50	3.31	3.26
`c5--c2`	empty marker prefix (zero content)	5.22	5.68	5.16	4.83	4.61
`c1--c2`	naive delta (think vs no-think)	5.86	6.05	5.43	5.04	4.98
`c1--c3`	content-isolated, vs empty chain	6.42	6.56	6.32	6.23	6.00
`c5--c1`	content-isolated, both templated	7.12	7.72	7.27	7.12	7.27
`c1--c4`	content-isolated, vs filler	7.65	7.43	8.00	8.10	7.96
range over L8 and L12		1.7–8.1	2.0–7.7	2.7–8.0	2.4–8.1	2.9–8.1
R^2 (corrected fit) / permutation null			0.931–0.992 / 0.107–0.147			

The ordering of the rungs is weaker and exploratory. The per-shard Spearman correlation between the fitted and the construct ordering is $+0.89$ to $+0.96$, but over 6 adjacent steps $\times$ 5 models $\times$ 3 layers $= 90$ steps it does not survive the pre-registered family-wise test: uncorrected comparison reaches $\alpha = 0.05$ on 14/15 shards, while Holm gives *0/15* (0/10 over the L4+L8 family) and Bonferroni fails with the smallest $p = 0.02401$ against $\alpha/15 = 0.00333$; Benjamini–Hochberg would give 14/15, but the pre-registration says family-wise, so Holm decides. With six comparisons per shard the test has little power, which bounds what the ladder may be used to claim rather than telling against the ordering: we claim *direction, not monotonicity*, not that x_0 is a linear reading of content.

Table 3: Fitted crossing position x_0 (tokens) for the naive think/no-think contrast, by model and layer, with the per-layer scale correlation. ρ is Spearman over the six scales against $\log_2$ parameters with an *exact* two-sided permutation p over all 720 orderings; the L4/L8/L12 row gives that per-layer value, while the first row of statistics is the three-layer mean that the submitted abstract reports and uses the asymptotic t approximation ($p = 0.40$, exact $p = 0.42$). Each cell is one greedy fit, on $n = 200$ problems at 0.6B–8B and $n = 100$ at 14B and 32B.

Model	0.6B	1.7B	4B	8B	14B	32B
L4	5.68	5.77	2.81	3.81	5.67	4.98
L8	5.86	6.05	5.43	5.04	4.98	4.71
L12	5.58	5.82	5.32	5.05	5.40	5.34
three-layer mean						
mean	5.71	5.88	4.52	4.63	5.35	5.01
ρ / p	-0.43 / 0.40 (exact 0.42), $n = 6$					
per layer						
ρ	L4 -0.43		L8 -0.94		L12 -0.49	
p	L4 0.42		L8 **0.017**		L12 0.36	

The zero-content prefix is the sharpest control: `c5` is `c2` with a text-only `<think>\n\n</think>` marker prepended, so `c5`–`c2` differs by a marker carrying no reasoning content, and its crossing position, 4.61–5.68, is indistinguishable from the naive delta's 5.0–6.1. A zero-content prefix therefore reproduces most of the crossing position, and x_0 is not evidence that reasoning has begun. Our pre-registered expectation was mis-specified in form: a length-L prefix shifts the two sequences against each other by L tokens, so the template explanation predicts $x_0 \approx$ prefix length, which is what is measured, not the collapse to ≈ 0 we predicted.

5.3 Scale stability depends on the layer

Table 3 gives the fits. The naive contrast's x_0 is stable across scales at L8/L12 over the five smaller models (4.98–6.05), with $k \in [0.401, 0.462]$ and $R^2 \geq 0.945$. Whether that counts as scale stability depends on the layer, so the two calibers must be reported together. The caliber we can state as a result is the coarse one, and it is a negative: on the three-layer mean, the caliber the submitted abstract quotes, the correlation with $\log_2$ size is -0.43 at $p = 0.40$, far from significant at $n = 6$. We therefore claim neither a scaling law nor its absence. What the per-layer split offers is exploratory and is labelled as such: L12 is flat ($\rho = -0.49$, $p = 0.36$) while L8 moves downward across scales ($\rho = -0.94$, exact permutation $p = 0.017$), with the 0.6B and 14B bootstrap intervals disjoint ($[5.60, 6.05]$ against $[4.53, 5.12]$). L8's ρ is one of three layers chosen after seeing this table, and the multiplicity correction it has not received would take p to approximately 0.050; the layer-dependent reading is therefore a hypothesis this table raises, not a result it establishes.

5.4 Out-of-sample reproduction and joint testing

Two pre-registered stress tests bound what the ladder can carry, and both are pure-CPU reanalyses: one resamples the *generation*, the other resamples *questions* out of the existing hidden states with the same indices for both contrasts of a step. The instrument gate passes exactly (same seed on the first 24 problems, hidden states bit-for-bit, $|\Delta x_0| = 0$ on all five models). Under sampling the reading is a split, not a failure: the structural segment is generation-caliber-independent, while adjacent steps on the content side are not separable (1/5 models meet the strict per-seed criterion and 30 of the 90 steps fail a cross-seed band) and are merged or written as $\approx$. One contrast is systematically sensitive — `c1`–`c4`'s greedy values lie above the sampling band on 8/8 cells — so every citation of it carries the interval. Protocol and remaining bootstrap detail are in the appendix.

5.5 Instance $\mathbb{B}$: the agreement diagnostic degenerates

The second diagnostic is binary agreement, $\mathbf{1}[a > 0.5] == \mathbf{1}[b > 0.5]$, with a a verbal confidence and b a linear probe, read as a measure of divergence between two cheap confidence estimators. It is not one: the quantity re-encodes the probe's hard label, with the verbal side acting only as a sign

Table 4: Instance $\mathbb{B}$. The third and fourth columns are equal by the identity above. Non-overlap is #disagreeing$/n$, and each cell also prints its complement, the overlap rate; the frozen source table labels the same number the other way round, so we print both. All four counts are deterministic functions of the records and carry no interval or p-value. Guard rows are retained rather than deleted: 1.7B and 14B fail guards (2)/(3), and 14B has $n = 0$.

Model	n	$a \leq 0.5$	disagreeing	non-overlap	AUC(hard)	AUC(agree)	guard
Qwen3-0.6B	73	40	40	54.8% (45.2%)	0.9011	**0.4688**	(2)(3) ok
Qwen3-1.7B	44	15	15	34.1% (65.9%)	0.8221	0.6421	**FAIL**
Qwen3-4B	128	5	5	3.9% (96.1%)	0.8070	0.7768	(2)(3) ok
Qwen3-8B	126	6	6	4.8% (95.2%)	0.8654	0.8358	(2)(3) ok
Qwen3-14B	0	0	0	—	—	—	**FAIL**

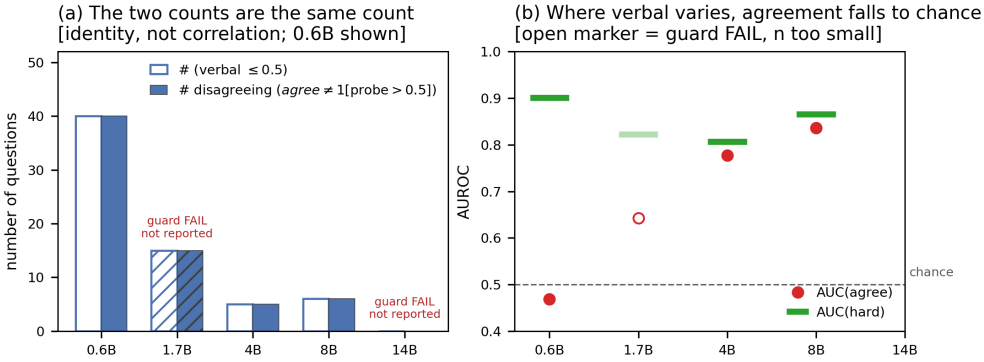

Figure 3: Instance $\mathbb{B}$. (a) The number of disagreeing questions and the number of questions with $a \leq 0.5$ are the same count on every model — an identity, not a correlation. Hatched bars are guard failures, retained and labelled rather than deleted; 14B has $n = 0$. (b) AUROC of the probe's hard label (bars) against AUROC of agreement (points). Where the verbal side sits above the threshold the two track each other, because there agreement *is* the probe's label; on 0.6B, where verbal confidence varies, the sign flips cancel and agreement drops to chance. Open markers are guard failures.

flip, equal to $\mathbf{1}[b > 0.5]$ on every record with $a > 0.5$ and to its negation below. Hence the number of *disagreeing* questions equals the number of questions with $a \leq 0.5$ exactly, with no assumption about either side — an identity, not a correlation, and what it leaves is one estimator's sign.

Table 4 is that identity in data, and Fig. 3(a) is the same identity as bars: the two columns are equal cell by cell (40/40, 15/15, 5/5, 6/6, 0/0). Two consequences follow. First, where the verbal side is nearly always above the threshold — 4B and 8B, at 5/128 and 6/126 records below it — agreement and $\mathbf{1}[\text{probe} > 0.5]$ are bit-identical on 96.1% and 95.2% of records, so the quantity's variance is the probe side's. Second, where verbal confidence genuinely varies the sign flips cancel and agreement falls to chance: 0.6B is the only model whose verbal side moves (40/73 records, a non-overlap rate of 54.8%), and there agreement equals the probe's sign on 45% of records and its negation on the rest, so AUC(agree) is 0.469 while the probe's hard label scores 0.9011; the signed gap between the two AUROCs tracks verbal activity (-0.43 against -0.030), with no interval or p-value attached.

5.6 Supporting results

The pre-registered difficulty axis fails, and x_0 does not encode difficulty. Fitting the same sigmoid against `math_level` gives x_0 values of 0.405, 1.013, 0.550 and 0.762 tokens for 0.6B, 4B, 8B and 14B with no consistent position, and 1.7B is fully degenerate ($x_0 = 1.37 \times 10^4$): the abstract's "difficulty axis fails outright". The converse threat — that x_0 is merely a difficulty reading — does not survive refitting within bins, whose ranges are at most 0.62 tokens where four bins are usable (0.6B–8B) against construct separations of 2.9–3.6 tokens.

The late segment is reproducible structure. 97–98% of δ energy at $j = 23$ is question-specific, as is 0.61–0.77 of the question-by-question similarity structure under split-half (permutation null at

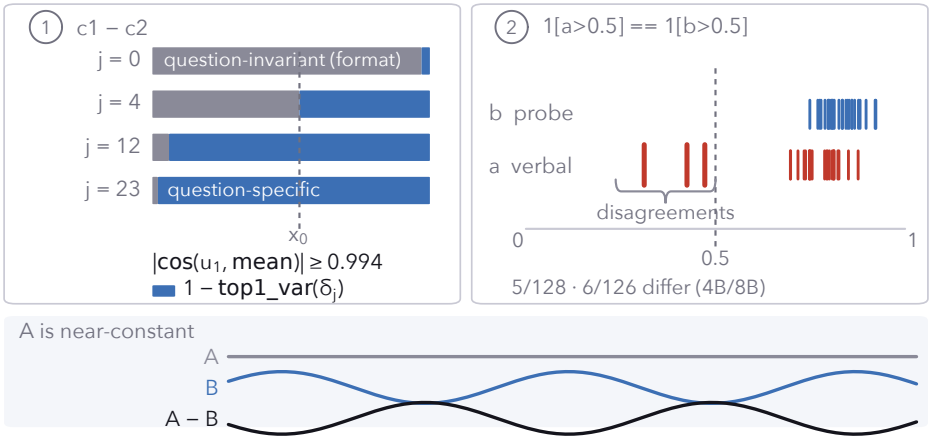

Figure 4: Two diagnostics, one failure. (1) The energy of δ splits into a question-invariant part (grey) and a question-specific part (blue); the grey share falls from 97% at $j = 0$ to 2% at $j = 23$, the two cross at x_0, and $1 - \mathrm{top1_var}(\delta_j)$ is the blue part. (2) Agreement between a verbal confidence a and a probe label b equals the probe's label wherever $a > 0.5$, so where the verbal side is saturated — 4B and 8B, at 5/128 and 6/126 records below the threshold — it is the probe's label, and only those few records disagree. Below, the mechanism the two share: one side near-constant, so the difference reads the other side alone.

most 0.008), with cross-position RSA against $j = 0$ at 0.01–0.06. So the late segment is neither noise nor a continuation of the input — though the same reproducibility appears in non-thinking controls, which caps it at within-task question distinguishability. The nuisance analysis and two further negatives are in the appendix.

6 Discussion

Both instances are the same failure. Instance $\mathbb{A}$ reads $1 - \mathrm{top1_var}(\delta_j)$ over the pair think/no-think, where the format share is near-constant at the crossing region; instance $\mathbb{B}$ reads $\mathbf{1}[a > 0.5] == \mathbf{1}[b > 0.5]$ over the pair verbal/probe, where the verbal side's hard label is near-constant above the threshold (> 0.5 on 96% of records at 4B and 8B). Both quantities are named for "the difference between the two sides" of a constructed pair, and in both one side is near-constant, so the quantity degenerates into a function of the other — the question-specific energy share in one case, the probe's sign in the other. Neither needs a model fitted to the representation: the first is an eigenvalue share and an angle, the second a count. That is what "the ruler decides" means here, and Fig. 4 states both instances and the mechanism they share in one panel.

Two checks cost almost nothing and are worth running before either quantity is interpreted. For $1 - \mathrm{top1_var}$, report the alignment of the leading direction with the between-problem mean: one angle separates a rank-structure reading from a question-specific share. For an agreement between two confidence estimates, report each side's rate above the threshold, because a side that is almost always above it reduces the quantity to the other side's sign. Neither check fits anything to the representation.

The zero-content prefix fixes what x_0 can mean: a five-token marker with no reasoning content reproduces a crossing position indistinguishable from the naive delta, so x_0 marks where a given contrast's two continuations stop sharing fixed template text and start depending on the problem, not where reasoning begins. Its boundary is that scale stability depends on the layer, so neither "we found a scaling law" nor "scale effects are excluded" follows.

7 Limitations

Format deconfounding failed, and we turned the failure into the measurement. The pre-registered criterion returns `format_confound: true` on 5/5 models, every non-thinking con-

trol decays as the thinking contrast does, and the leading direction over the transition region is the question-invariant one. The response is §5.2: the confound is decomposed, not avoided.

The abstract's scale caliber is superseded by the body. The submitted abstract reports the three-layer-mean caliber ($\rho = -0.43$, $p = 0.40$); §5.3 reads the same table per layer and finds L8 at $\rho = -0.94$ with exact $p = 0.017$ and disjoint intervals, while L12 is flat. The abstract is the earlier, coarser reading of one table; §5.3 replaces it and states it as exploratory.

External validity is still open, and one published reading of it was wrong. Reliability here is internal consistency, which every deterministic function of the problem satisfies, not semantic content. Our wording is therefore capped at "reproducible structure": no second independent collection of the same problems exists here, and the one earlier reading that appeared to settle this was an invalid test (appendix). The window is the first 24 continuation tokens, token-identical between the two collections for 93–100% of problems, so the statistic and its null were both correct even though the window was shorter than the difference it was to capture.

Four further caveats are recorded in the appendix. Steering does not pass its pre-registered criterion and is not a claim of this paper. Difficulty stratification has uneven coverage and an impure proxy, and its confidence intervals are unusable and are cited nowhere. Three measurement conventions bound which statistics may be compared: `effective_rank` is bounded by the number of problems; the sampling leg's window is not the greedy window; and the abstract's $|\cos| \geq 0.994$ holds at L8 only, with L12's held-out minimum at 0.9623, so a two-layer statement must write ≥ 0.95.

Reproducibility statement

Each result above names the statistic's caliber, its sample axis and, where a fitting convention matters, the convention. Three are stated in the body because a recomputation would otherwise differ from our numbers without either being wrong: the reported R^2 is a squared Pearson correlation rather than $1 - \mathrm{RSS}/\mathrm{TSS}$ (§3.3); the comparable crossing coordinate is the corrected fit plus the number of dropped degenerate positions; and every family-wise statement uses Holm, as the pre-registration specifies. Statistics that are deterministic functions of the records are reported as counts with no interval, and the permutation and bootstrap checks use fixed seeds. Code, frozen readouts and the mapping from each table to its producing script are released as an anonymized supplement.

AI use statement

Generative AI tools assisted with code development, analysis scripting, generating graphs and the editing of drafts. All of the main-body part of the paper is written by the authors.

References

Siddharth Boppana, Annabel Ma, Max Loeffler, Raphael Sarfati, Eric Bigelow, Atticus Geiger, Owen Lewis, and Jack Merullo. Reasoning theater: Disentangling model beliefs from chain-of-thought. *arXiv preprint arXiv:2603.05488*, 2026. URL https://arxiv.org/abs/2603.05488.

Yiğit Utku Bulut. It's the problem, not the path: Budget and difficulty confounds in llm reasoning trajectories. Zenodo preprint, September 2026. URL https://doi.org/10.5281/zenodo.22261107.

Esakkivel Esakkiraja, Sai Rajeswar, Denis Akhiyarov, and Rajagopal Venkatesaramani. Therefore i am. i think. *arXiv preprint arXiv:2604.01202*, 2026. URL https://arxiv.org/abs/2604.01202.

Samuel Lewis-Lim, Xingwei Tan, Zhixue Zhao, and Nikolaos Aletras. Analysing chain of thought dynamics: Active guidance or unfaithful post-hoc rationalisation? In *Proceedings of EMNLP 2025 Main Conference*, 2025. URL https://arxiv.org/abs/2508.19827. arXiv:2508.19827.

Wenkai Li, Fan Yang, Ananya Hazarika, Shaunak A. Mehta, and Koichi Onoue. When reasoning traces become performative: Step-level evidence that chain-of-thought is an imperfect oversight channel. *arXiv preprint arXiv:2605.11746*, 2026a. URL https://arxiv.org/abs/2605.11746.

Yubo Li, Ramayya Krishnan, and Rema Padman. The chain holds, the answer folds: Trace-answer dissociation in reasoning models under adversarial pressure. *arXiv preprint arXiv:2605.29087*, 2026b. URL https://arxiv.org/abs/2605.29087.

Subhadeep Pal, Shashwat Sourav, Tirthankar Ghosal, and Markus J. Buehler. Graph-native reinforcement learning enables traceable scientific hypothesis generation through conceptual recombination. *arXiv preprint arXiv:2607.00924*, 2026. URL https://arxiv.org/abs/2607.00924.

Swapnil Parekh and Naman Goyal. Drop the act: Probe-filtered rl for faithful chain-of-thought reasoning. *arXiv preprint arXiv:2605.11467*, 2026. URL https://arxiv.org/abs/2605.11467.

Subramanyam Sahoo, Vinija Jain, Aman Chadha, and Divya Chaudhary. Linear probes detect task format, not reasoning mode in language model hidden states. In *6th Workshop on Trustworthy NLP, ACL 2026*, 2026. URL https://arxiv.org/abs/2606.02907. arXiv:2606.02907.

Fahd Seddik and Fatemeh Fard. Formalizing latent thoughts: Four axioms of thought representation in llms. *arXiv preprint arXiv:2606.27378*, 2026. URL https://arxiv.org/abs/2606.27378.

Sai Adith Senthil Kumar. When built-in thinking helps and hurts: Constraint-level error shifts in instruction following. *arXiv preprint arXiv:2606.09662*, 2026. URL https://arxiv.org/abs/2606.09662.

Shouren Wang, Wang Yang, Chuang Ma, Debargha Ganguly, Vikash Singh, Chaoda Song, Xinpeng Li, Xianxuan Long, Vipin Chaudhary, and Xiaotian Han. Path-lock expert: Separating reasoning mode in hybrid thinking via architecture-level separation. *arXiv preprint arXiv:2604.27201*, 2026. URL https://arxiv.org/abs/2604.27201.

Wang Yang, Debargha Ganguly, Xinpeng Li, Chaoda Song, Shouren Wang, Vikash Singh, Vipin Chaudhary, and Xiaotian Han. Mid-think: Training-free intermediate-budget reasoning via token-level triggers. *arXiv preprint arXiv:2601.07036*, 2026. URL https://arxiv.org/abs/2601.07036.

Jiachen Zhao, Yiyou Sun, Weiyan Shi, and Dawn Song. Can aha moments be fake? towards quantifying decorative and true thinking in chain-of-thought. *arXiv preprint arXiv:2510.24941*, 2025. URL https://arxiv.org/abs/2510.24941.

A Audit record: corrections, retired wording, and items resolved by decision

This appendix preserves the corrections made while preparing this paper, and the detail that the 9-page limit pushed out of the main text. It is recorded rather than deleted, because a correction that is dropped is indistinguishable from an error that was never found, and it is placed after the references so that it does not consume main-text space.

Detail relocated from the main text

Setup detail relocated from §4. Hardware: four A800 80GB and two A40 46GB, and all analysis is pure CPU — the CUDA visible-device index is not the `nvidia-smi` index, which is how an earlier cross-GPU comparison was mistaken for a stronger test than it was. Prompt-level comparison shows the five smaller models are byte-identical to Qwen3-8B at 200/200 and 100/100 with zero differences, and 14B's problem identifiers are exactly `MATH500_00000...00099`; this is the check behind the main text's claim that cross-scale comparison has no "different problems per scale" confound. The generation-budget configuration file gives 4096, which is stale — it stopped at an abandoned first round and is overridden by the collection scripts.

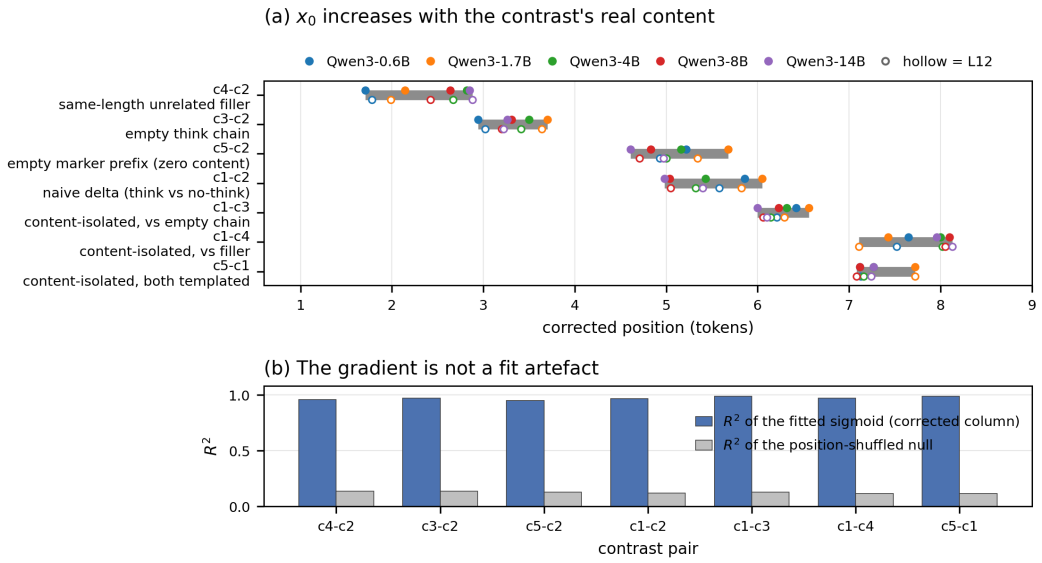

Figure 5: The content-graded ladder of Table 2, relocated here by the page limit. (a) Each grey bar is the range over L8 and L12 for one contrast; filled markers are L8, hollow L12, coloured by scale. (b) R^2 of the fitted sigmoid against the R^2 of the same fit on position-shuffled data. The null reaches at most 0.147 on every rung.

Limitations detail relocated from §7. *An external-validity reading that does not hold.* An earlier result that appeared to answer whether reliability here has anything to do with semantic content was a rerun on a different GPU with a retest correlation of 0.9997–1.0000. That measures the numerical robustness of one greedy trajectory read twice, not external validity, and it is cited as such nowhere. *Steering.* The steering experiment covers two models and does not pass its pre-registered criterion: across six (direction, step-size) cells the differences run -0.050 to $+0.000$, every interval contains zero, and p ranges from 0.126 to 1.000. Three checks establish that this is not an instrument failure — the random control direction is near-orthogonal to the real ones (maximum $|\cos| = 0.022$), the degeneracy rate is 0.000 across all nine cells, and generation length does not collapse. *Difficulty stratification.* Four bins are usable on 0.6B–8B, but 14B at $n = 100$ has only L5 above the minimum bin size, so the claim that within-bin refitting leaves the ordering intact rests on one bin there. The correct/incorrect split is not a pure difficulty proxy: it moves `c5–c1` and `c1–c2` by 0.86 and 0.77 tokens while barely moving the two ablation contrasts, and since 0.86 is of the same order as the 0.89-token separation between adjacent constructs, we claim robustness only for the separation between structural and content controls. The confidence intervals for this stratification are not usable and are cited nowhere: point estimates fall outside their own intervals and 97.5% of bootstrap replicates lie above the point estimate. *Three conventions.* `effective_rank` is bounded by the number of problems and so is not comparable across bins; $K = 24$ is the greedy collection's window, while the sampling leg recomputes the common-prefix length per token (4B, 8B and 14B diverge from greedy inside the window on 3–7% of problems, the earliest at token 4, so the two collections are not the same object past that point); and within L8 the alignment is a transition-region statement ($j \leq 6$), not an all-position statement.

Two supporting negatives relocated from §5.6. *δ norm does not distinguish answer-changing from answer-preserving questions.* If thinking really changed the computation, the substantive segment's δ norm should be larger when the answer changes. It is not: at $j = 12$ the per-model p-values run 0.086–0.90 with the direction inconsistent across models, and the template segment is significant on two models in the *opposite* direction to the hypothesis. This restricts the reading "δ is answer-directed computation"; it does not establish that δ is decoration, since a norm is not a content. *A random-content control.* To separate "is it content" from "is it text of that length and diversity", `c6` pairs the think continuation against a think-length continuation assembled from other problems' chains. The pre-registered criterion asks whether `c1–c6` is indistinguishable from the naive delta (voiding the content component) or from the content-isolated contrast (establishing it

independently). The component is not voided: the two are separable on 14 of 15 cells, and at the level at which the criterion was written — independent standing — on 12 of 15. 1.7B's L8 and L12 fail it and 4B's L4 is not judgeable (a degenerate fit, $R^2 = 0.153$, with x_0 pinned at the parameter boundary), which we record as degenerate rather than as a negative. The trivial explanation — that x_0 is just the boilerplate length — is not met at any scale, and 1.7B is the one outlier, as it is in §5.1 and §5.4.

Detail relocated from the main text. Five further items were in an earlier draft's main text and are here because the limit is hard. (1) The three-item account of how instance $\mathbb{B}$'s degeneration was detected, including the per-model guard table's full failure reasons. (2) The two-component decomposition of δ into an answer-aligned and a mode-aligned part: an out-of-fold answer-component AUROC of 0.873–0.930 (the in-sample version is identically 1.0 and is not used; permutation AUROCs fall in 0.484–0.517, so the pipeline does not leak), about 4–6% of δ energy answer-aligned with 99.8% orthogonal to it, subject to the format deconfounding of §5.2 being applied first. (3) The 32B truncation counts and the tokenisation evidence for the 8192 generation budget: 0.6B 66/200 truncated, 1.7B 41, 4B 36, 8B 39, 14B 15, 32B 16, with continuations reaching a maximum of exactly 8192 and a pile-up there. (4) The per-model detail of the `c1–c4` sampling sensitivity (4B/8B reach 8.0–8.13 greedily against 7.37–7.60 under sampling) and of the four models that fail the strict out-of-sample criterion. (5) The uncorrected fit's R^2 range, which is 0.28–0.47 for `c1–c4` and 0.73–0.78 for `c1–c3`, and which ranks two pairings in the wrong order — the reason the main text quotes only the corrected column. (6) The positive control that falsified the early-window independence claim: at 64 against 2048 tokens the maximum absolute hidden-state difference is 5.46×10^1 with only 4/5 problems token-identical.

A statistic measured after the table was frozen. The leading-direction alignment on 32B was measured on 2026-09-25, after Table 1 was frozen: the minimum of $|\cos(u_1, \text{mean})|$ over $j \leq 6$ is 0.9947 at L8 and 0.9970 at L12 on the 32B chain-axis collection ($n = 100$, cap 8192), whose per-position hidden states were already on disk. It is a pure-CPU reanalysis of an existing collection; the script reproduces the five-scale values of Table 1 bit-for-bit before reporting 32B; and the readout is frozen as `p0_32b_alignment_20260925/readout.txt`. The submitted abstract's parenthetical range therefore describes the caliber at submission time, not the scope of the result.

The held-out column, re-frozen. The question-disjoint split-half column of Table 1 had no producing script in the working tree: it was computed in the writing environment and could not be re-derived from a frozen artifact. On 2026-09-26 it was re-implemented to the caption's own description — five random question-disjoint halves at a fixed seed, mean fitted on one half and u_1 on the other, minimum over the five splits and over $j \leq 6$ on the same position grid as the in-sample column — and re-frozen as `p0_heldout_alignment_20260926/readout.txt`, covering 0.6B–32B. The table's ten held-out cells are that readout's values; on the same protocol and readout the 32B minimum is 0.9649 at L8 and 0.9758 at L12. The re-implementation gives slightly higher minima than the values it replaces (L8: 0.9679–0.9872 against 0.9588–0.9799), which is the expected direction when the held-out column keeps the in-sample position grid; the claim it supports — that the alignment survives a question-disjoint split — is unchanged, and it is now checkable from a frozen readout.

AUDIT RECORD: CORRECTIONS, RETIRED WORDING, AND DECISIONS

A two-cluster reading of the ladder is refuted. An earlier version of this analysis endorsed a two-component picture with a format component near 3 and a content component near 7.3. It rested on `c1–c3` and `c5–c1` agreeing, which was an artefact of the degenerate first position. On corrected coordinates the two are 6.0–6.6 and 7.1–7.7 with disjoint bootstrap intervals (14B L8: $[5.58, 6.26]$ against $[6.99, 7.46]$), so the picture is a single monotone gradient, not two clusters.

Retired wording. Two formulations used earlier in this project are not used anywhere in the main text, including not as assertions in quotation marks. First, no "phase transition", "register-to-computation" or "thinking becomes computation" reading is asserted; the term appears only where this paper rejects it or where the protocol's original prediction is quoted in order to be falsified. Second, monotonicity is claimed nowhere: the ladder is directionally consistent, and the family-wise

joint test does not reject, so the wording is *direction, not monotonicity*. A third retired phrase, that the crossing position shows no monotone scaling across scales, is superseded by "scale stability depends on the layer" (§5.3).

A column label that was printed backwards. In the frozen source table for §5.5, the column header reads "overlap rate" while the printed value is the non-overlap rate n_{disc}/n, because the printing expression follows the count of disagreeing records rather than the count of agreeing ones. The same number appears under both labels elsewhere in the project. We print both rates in Table 4, state which one each sentence of prose uses, and do not modify the frozen file, which is a readout record.

Earlier readings that this paper reverses. (1) An in-sample answer AUROC of 1.0 is an artefact of fitting and scoring on the same records; the out-of-fold value is 0.873–0.930. The answer fraction, by contrast, is robust to the fix, moving by only 0.003–0.012. (2) The claim that the correction leaves the content-isolated contrast's coordinates exactly unchanged does not hold literally: on 5 of its 10 rows the corrected and uncorrected values differ, by at most 0.10 tokens. The correct statement, and the one the verifiability argument actually needs, is that the change is very small. (3) A relaxed out-of-sample criterion giving 4/5 monotone scales originates in the greedy run and could not be reproduced under sampling; restating it in the sampling caliber would create a third caliber, so it is recorded as unreviewed (§5.4). (4) The literal permutation statistic — how often a random relabelling of the seven rungs produces a strictly increasing sequence — is constant at $1/5040$, so it carries no information about the data at all. Vacuity of this kind is not visible in the value; it is visible only in the observation that all 5040 relabellings give the same answer. The family-wise test in §5.2 is what decides.

Items decided for this paper. Where the project's own records left a choice open, the following decisions were taken and are applied consistently above. The per-position layers $\{4, 8, 12\}$ are described as a pre-fixed study-design choice, because no written selection rationale exists. The 32B scope words are stated explicitly: chain-axis fitting covers 0.6B–32B while all other statistics cover 0.6B–14B. The 4B L4 cell of the random-content control is reported as degenerate rather than as a negative result. The related-work entry (Pal et al., 2026) is cited neutrally, for the fact that a different line of work segments traces by structural boundaries; its abstract states verbatim that it performs layer-wise hidden-state analyses, so no claim in either direction is made about whether it overlaps ours. Esakkiraja et al. (2026) is likewise cited in the main text for its decoding result only; its steering protocol, which we aligned to and which did not pass, is confined to §7. Zhao et al. (2025)'s report that steering can flip answers grounds a prediction the failed difficulty axis leaves untested, and it is cited in the reference list by its paper title rather than its method alias. And the question-only baseline requested by Bulut (2026) is *not* claimed to be equivalent to the within-difficulty refitting in §5.6; whether the two coincide was not determined in this project, and we say so rather than asserting an equivalence.

The closest paradigm paper: the clauses of the comparison. The main text keeps only the points that bear on the argument; the rest is here, because the naive versions of these clauses are false and a reader repeating the comparison needs them. Senthil Kumar (2026) finds that thinking redistributes errors rather than uniformly helping, and its own text records that (i) smaller models flip more often, the 32B result breaks a simple monotonic scale pattern, and the patching gap grows with scale, so it does have a scale-trend analysis and we do not claim it lacks one; (ii) it has two position axes at once — a descriptive peak location at first/middle/last-third granularity along the trace, and five equal-width relative-depth bins along depth, where the gap is not concentrated in early or late layers — so the difference from our continuous crossing parameter is one of granularity and statistic, not of whether an axis exists; (iii) its dependent variable is IFEval pass rate and cross-encoder trace relevance, where ours is representation geometry, and it uses patching for causal verification where we use the `c1`–`c6` controls to decompose the confound; and (iv) its appendix performs a matched-length analysis on the same axis as our core deconfounding. Two further points from its own text are stated in our favour in §6 only if they can be verified there: it self-declares that its activation patching is not a symmetric head-to-head test, which our symmetric `c1`–`c5` controls are, and its appendix documents token-by-token that OFF mode injects an empty `<think></think>` block, exactly the prefix `c5` simulates.

Unresolved. The generation budget cannot be read off the configuration file, which stopped at an abandoned first round and is overridden by the collection scripts; we report the budget measured from the data itself (§4). The early-window independence claim is likewise not made, having been falsified by our own positive control. The reliability statistics cover 0.6B–14B; the same-weight pairing covers 1.7B–32B in the source protocol, but the paired hidden states here are our own collection and their coverage is as stated in §4.