Life of Research · Published 4 November 2026 · DOI 10.5281/zenodo.23951230

A Lift Is Not a Selection Gradient: What the Amplification-Lift Gap in Thinking Models Measures, Why an Answer-Aware Label and a Pooled Contrast Account for Much of It, and the Within-Prompt Test That Would Show Training Is Misdirected

Pranay Mahendrakar
SONYTECH · ORCID 0009-0003-7224-029X
pranaymahendrakar@sonytech.in · mahendrakarpranay@gmail.com · +91 6361723454

Abstract. Reasoning-oriented post-training makes model traces look more deliberative, and a 2026 study asks whether it amplifies the behaviours that go with correct answers. Annotating 15,282 traces from 15 models, it finds that thinking models express self-correction, hypothesis testing and uncertainty several times more often than instruction-tuned models, while the behaviours with the largest accuracy difference between traces that show them and traces that do not, knowledge alignment and confidence calibration above all, are no more common (self-awareness only modestly so, 1.1 to 1.9 times across six benchmarks). It reads this as training that rewards surface form over the behaviours that predict success. In this analysis the author reviews and synthesises published results to show that the gap, as measured, is not yet evidence that training is misdirected. Three of its ingredients predict much of it without any misdirection. First, the top-ranked labels were assigned by a judge that saw the answer key: the author derives from the study's tables that at least 98 percent of incorrect thinking-model traces that voice doubt are still labelled uncalibrated, so the label behaves like "confident and correct". Second, the headline difference is pooled across questions; within a question, in the study's own control, self-correction is associated with 26.7 points higher accuracy against 30.5 for calibration, which ranks third of nine behind knowledge alignment (59.6) and goal tracking (40.5). Third, outcome-rewarded training acts on within-prompt covariance with reward, and multiplying that control by q(1 - q), where q is prevalence, gives near-equal covariances for self-correction, calibration and goal tracking (0.066, 0.068 and 0.064), against 0.0015 for uncertainty acknowledgment; the largest, 0.099, belongs to knowledge alignment, an outcome-referenced label like calibration. Goal tracking, which is not amplified, is not explained by these three ingredients and counts against this reading. What remains is narrower and real: uncertainty acknowledgment is amplified about five-fold with no within-question association, which imitation of teacher traces, length and clipping biases, hitchhiking on shared tokens, or a training-time role all predict. The paper consolidates 28 published results and two derived quantities in a table and a figure, states the predictions that separate six accounts, gives a reading procedure, and specifies the blinded, within-prompt and stage-separated tests that would settle the claim.

Keywords: reasoning models; chain of thought; reinforcement learning from verifiable rewards; self-correction; confidence calibration; process supervision; selection differential; evaluation methodology

1. Introduction

Large reasoning models are trained to write long traces before they answer. DeepSeek-R1 showed that reinforcement learning against a rule-based reward could produce traces with reflection and revision without supervised traces as a starting point (DeepSeek-AI et al., 2025), and families such as Qwen3 now ship a thinking mode beside a fast mode in the same release (Yang et al., 2025a). The behaviours these traces display have been named and counted: verification, backtracking, subgoal setting and backward chaining in one account (Gandhi et al., 2025), twenty-eight cognitive elements in another (Kargupta et al., 2025), clusters discovered bottom-up in a third (Lee et al., 2025). A recurring finding is that the most visible behaviours are not obviously the useful ones. Self-reflection becomes more frequent under R1-Zero-style training without being positively correlated with accuracy (Liu et al., 2025); most reflections after a candidate answer confirm it rather than change it (Kang et al., 2025); and suppressing the tokens that open a reflection shortens traces substantially without loss of accuracy (Wang et al., 2025a).

A 2026 study by Nyandwi, Mathur, Bisk, Hawkins and Neubig turns this pattern into a measured claim (Nyandwi et al., 2026). It annotates 15,282 traces from 15 open-weight language and vision-language models on six benchmarks with nine behaviours, and defines Behavioral Lift as the accuracy of traces that show a behaviour minus the accuracy of traces that do not. Thinking models show self-correction, hypothesis testing and uncertainty acknowledgment far more often than instruction-tuned models, while the behaviours with the highest Lift - confidence calibration, knowledge alignment and self-awareness - are about as common in both (self-awareness modestly more common in thinking models, 1.1 to 1.9 times across the six benchmarks). Calibration carries a Lift of +79.6 points in the language-model pool; uncertainty acknowledgment carries -13.9. The authors call this an Amplification-Lift

© 2026 Pranay Mahendrakar. Licensed under CC BY 4.0.

Gap, conclude that reasoning-oriented training does not preferentially amplify the highest-Lift behaviours, and recommend process-level objectives that reward calibrated and grounded reasoning.

The study is careful. It states that Lift is descriptive and not causal, validates its judge against three others and against manual labels, reports a within-question control, a length-controlled prevalence analysis, linear probes and a prompting experiment, and notes that its released checkpoints differ by whole training recipes rather than single stages. This paper does not dispute any of its measurements. Its question is what the comparison the headline rests on - prevalence ratio set beside Lift - can show about training.

The thesis is that Lift is not a selection gradient, and that a gap between Lift and amplification is therefore not, by itself, evidence that training is misdirected. Training that rewards correct outcomes does select for behaviours, but it selects on a specific quantity: to first order, whether by policy gradient or by fine-tuning on filtered correct samples, the covariance between the behaviour and the reward among samples from the same prompt (Williams, 1992; Shao et al., 2024), not the covariance pooled over prompts. Evolutionary biology made the same distinction half a century ago. Price's covariance identity gives the change in a trait's mean under selection (Price, 1970); Lande and Arnold separated the selection differential, a trait's raw covariance with fitness, from the selection gradient, which holds correlated traits fixed (Lande and Arnold, 1983); and the response to selection depends on how much heritable variation there is and on how traits are coupled, so that a trait under no direct selection can still change through its correlation with one that is (Lande, 1979; Maynard Smith and Haigh, 1974). Lift is closest to a selection differential, measured on released endpoints, pooled across prompts, and computed on labels some of which were assigned with knowledge of the outcome.

Six accounts predict a gap between Lift and amplification. The first is the study's: training amplifies surface behaviour without regard to its contribution. The second is answer-aware labelling: the high-Lift labels are in part redescriptions of correctness. The third is confounding by indication: the amplified behaviours are triggered by difficulty and by error, so they appear on traces that were already more likely to fail. The fourth is selection with constrained response: training selects on within-prompt covariance, and a label defined by the relation between confidence and correctness is not a behaviour a policy can move independently of correctness. The fifth is amplification without selection: imitation of teacher traces, optimisation biases in the algorithm, and hitchhiking on tokens shared with a selected behaviour. The sixth is training-time value: a behaviour can raise the learning signal, or another behaviour's quality, without raising the accuracy of the trace that contains it.

The paper makes five contributions. First, it derives from the study's own conditional tables that the calibration label is close to a detector of correctness, with at least 98 percent of doubting incorrect thinking traces labelled uncalibrated, and connects this to outcome bias in human and model judgement (Section 4). Second, it reads the study's within-question control as a test of confounding and shows that it moves self-correction close to calibration (Section 5, Figure 1). Third, it derives which form of Lift each family of training acts on, and computes from the study's numbers that within-question covariance with correctness ranks self-correction with calibration and goal tracking, behind knowledge alignment, and puts uncertainty acknowledgment near zero (Section 6, Figure 1). Fourth, it assembles the evidence on what did the amplifying - imitation, reinforcement, optimiser bias - and on training-time roles for low-Lift behaviours (Sections 7 and 8), and the interventional record on the amplified behaviours (Section 9). Fifth, it consolidates 28 published results and two derived quantities in Table 1, states the six accounts' predictions in Table 2, gives a reading procedure as Algorithm 1, and specifies five studies (Section 14).

Section 2 fixes terms. Section 3 reads the anchor's design. Sections 4 to 8 take one account each, Section 9 collects interventions, Sections 10 to 12 hold the table of evidence, the predictions and the procedure, and Sections 13 to 17 cover objections, proposed studies, open questions, limitations and the conclusion.

2. Scope and terms

The paper concerns studies that annotate model reasoning traces with behaviour labels and relate those labels to answer correctness and to training. Its central case is the anchor study; the argument applies to any comparison of a behaviour's frequency across training conditions with its association with success. It does not concern whether reasoning traces are faithful to a model's internal computation, which is a separate question with its own literature (Chen et al., 2025; Boppana et al., 2026), except where faithfulness changes what a label can mean.

A **behaviour** b is a binary property of a trace assigned by an annotator, human or model. A **form-referenced** label can be assigned from the trace alone, for instance whether it contains an explicit revision. An **outcome-referenced** label is defined, wholly or partly, by the relation between the trace and the correct answer, for instance whether the trace uses the domain facts appropriate to reaching that answer.

Prevalence is the share of traces showing b. **Amplification** is the ratio of prevalence in a thinking model to prevalence in a comparison model; the anchor reports prevalence gaps and multiplicative ratios.

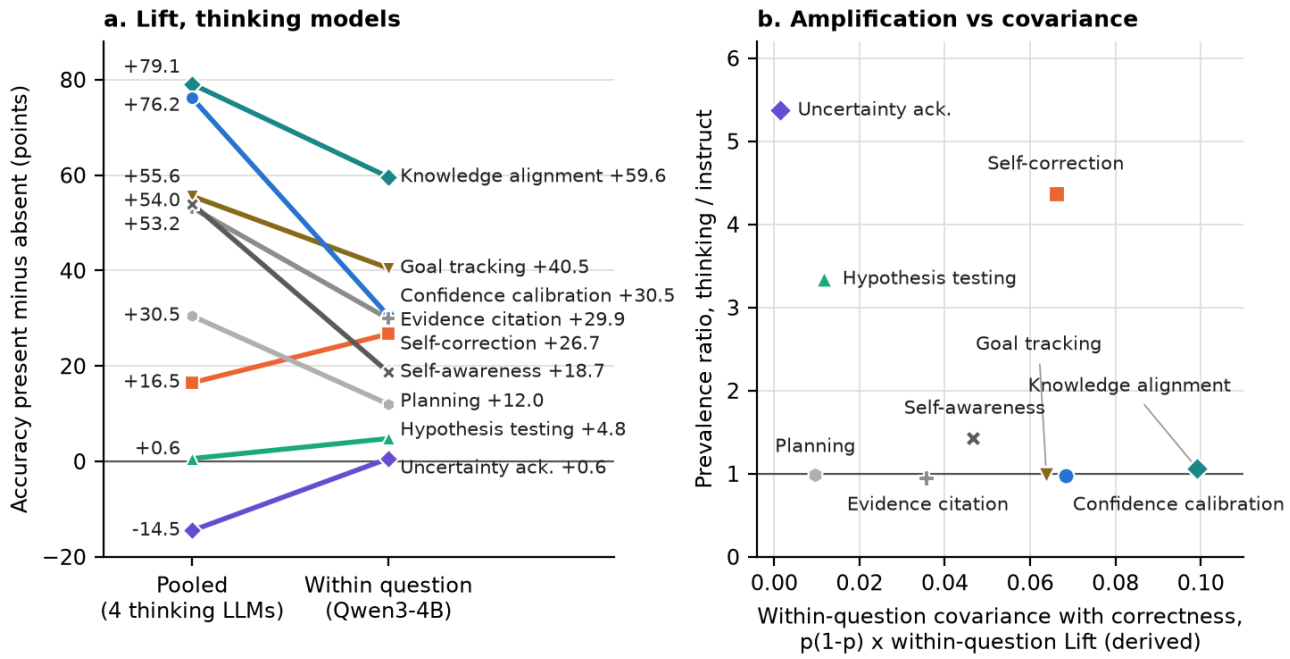


Figure 1. (a) Accuracy difference between thinking-model traces with and without each of nine behaviours, pooled over four language models and all questions (Nyandwi et al., 2026, Table 12, thinking-pool column) and within question for Qwen3-4B-Thinking on 250 MATH-500 problems (Nyandwi et al., 2026, Tables 2 and 17). (b) Prevalence ratio, thinking to instruct models on MATH-500 (Nyandwi et al., 2026, Table 7), against within-question covariance with correctness, p(1-p) times the within-question difference, derived by this paper using the pooled thinking prevalence as p. Self-correction is both amplified and among the most selected-for; knowledge alignment has the largest covariance and goal tracking about the same as self-correction, neither amplified; uncertainty acknowledgment is amplified with near-zero covariance.

Pooled Lift is P(correct | b) - P(correct | not b) over all traces in a pool, which mixes questions and, in the anchor's main analysis, models. **Within-prompt Lift** for prompt x, written D_x, is the same difference computed among samples from one model on one prompt. The anchor's within-question analysis reports the average of D_x over questions.

The **within-prompt covariance** of b with a binary reward R is C_x = q_x (1 - q_x) D_x, where q_x is the probability that a sample from prompt x shows b. The **pooled covariance** is the covariance of b and R over all prompts; by the law of total covariance it equals the average of C_x plus the covariance across prompts between q_x and the prompt's success rate p_x. Pooled Lift carries both terms; Section 6 shows that outcome-based training acts on the first.

Following quantitative genetics, a **selection differential** is a trait's raw covariance with fitness, a **selection gradient** is its partial regression on fitness holding correlated traits fixed, and a **correlated response** is a change in a trait caused by selection on another trait with which it covaries (Lande and Arnold, 1983; Lande, 1979). The analogy is used for its algebra, which carries over exactly in the cases derived in Section 6; nothing is claimed about biology.

3. The anchor, read at the resolution of its design

Models. The anchor evaluates 15 open-weight models from 3B to 9B parameters: four thinking and three non-thinking vision-language models, and four thinking and four non-thinking language models. Five pairs are same-family, same-scale thinking and non-thinking endpoints (Qwen3-VL, Kimi-VL, InternVL3.5, Qwen3-4B and OLMo-3-7B). The others are not matched pairs: DeepSeek-R1-Distill-Qwen-7B, a smaller model fine-tuned on samples from DeepSeek-R1, the route that model's authors describe for passing its reasoning patterns to smaller models (DeepSeek-AI et al., 2025), is an unmatched thinking model; GLM-4.1V is another; Qwen2.5-7B-Instruct is an extra baseline; and the Nemotron pair compares a reasoning model with a base model decoded greedily. The anchor's model table states that the role labels do not assert that any pair differs by a single training stage. Decoding settings also differ by role: thinking language models were sampled at temperature 0.6 and the Qwen2.5 instruct model at 0.7, with family-specific limits on generation length.

Benchmarks. Six benchmarks span a range from structure to knowledge: VisualPuzzles, MathVista and MMMU for vision-language models, and LogiQA2, MATH-500 and MMLU-Pro for language models, with about 350 items each.

Annotation. Every trace was labelled by GPT-4o, which received the question, the ground-truth answer and the model's full output, and returned binary labels for all behaviours. For confidence calibration the judge marks true only when the model's expressed certainty is consistent with the observed strength of its reasoning, and false when the trace is clearly overconfident or underconfident. Agreement with three other judges on 600 MATH-500 traces is reported as kappa 0.53 to 0.82 for self-correction, 0.71 to 0.79 for uncertainty acknowledgment, 0.51 to 0.59 for hypothesis testing and 0.56 to 0.81 for calibration, and a manual check of 120 traces gave 95.1 percent agreement on six focal behaviours.

Metrics. Lift is computed on pooled samples across all models in a modality. A within-question analysis generates eight traces per question from Qwen3-4B on 250 MATH-500 problems and averages the per-question difference in accuracy between traces with and without each behaviour.

What is compared. The amplification half of the gap compares released endpoints. The Lift half pools all models. The two halves are then set side by side, behaviour by behaviour, in the anchor's Figure 1. Three features of this design matter for what follows: the judge saw the answer; the headline Lift is pooled over questions and models; and the amplification is a difference between whole recipes, several of which include distillation from a stronger model's traces.

4. Account two: labels that read the answer key

The anchor's top three behaviours by Lift are knowledge alignment (uses the appropriate domain facts or framework), confidence calibration (certainty tracks the strength of reasoning) and self-awareness (recognises missing information or limits of support). In the language-model pool, the bottom three are self-correction, hypothesis testing and uncertainty acknowledgment, each defined by an observable move in the text. The anchor's full Lift table for language models also includes "context understanding" at +76.8 and "logical steps valid" at +69.9 - labels that are assessments of reasoning quality rather than behaviours a model chooses - and "shortcut" at -49.1. The ordering of Lift follows, to a large degree, how much each label is a judgement about whether the trace is on the way to the right answer.

That ordering is partly unavoidable. A label that asks whether a trace uses the right framework is hard to assign without knowing what the right answer is. But it means that the high-Lift end of the ranking is not a list of behaviours that a model could adopt and thereby become more accurate; it is in part a list of descriptions of accurate traces.

The calibration label behaves like "confident and correct". The anchor's conditional table for language-model traces gives, for thinking models, P(calibration | correct) = 82.0 percent and P(calibration | incorrect) = 1.0 percent, and P(uncertainty acknowledgment | incorrect) = 49.7 percent. Half of the incorrect thinking traces voice doubt, and one in a hundred is judged calibrated. The overlap between the two sets cannot be larger than the smaller of them, so at least 49.7 - 1.0 = 48.7 percent of incorrect thinking traces voice doubt without being labelled calibrated: at least 98 percent (48.7 / 49.7) of the doubting incorrect traces. Pooled over all language-model traces, the calibration label is present on 84.7 percent of correct traces and 0.6 percent of incorrect ones. Read as a classifier of correctness, it has a true-positive rate of 0.847 and a false-positive rate of 0.006, a positive likelihood ratio of about 141.

A calibrated forecaster is one whose stated probabilities match observed frequencies over a set of forecasts (DeGroot and Fienberg, 1983; Guo et al., 2017). Such a forecaster is wrong mostly when it says it is unsure. Under that definition, an incorrect trace that expresses low confidence is evidence of calibration, not of its absence. The anchor's label is a per-trace judgement rather than a property of a set of forecasts, and the judge may have read doubt voiced in the middle of a trace and then a confident final answer as overconfidence, which would be a fair reading. The point is narrower: the published tables cannot distinguish the label from "the trace was confident and the answer was right", and a label with that profile will have a Lift near the ceiling whatever the model's actual calibration.

The literature on language-model confidence measures calibration differently, and its findings do not map onto the label. Kadavath et al. (2022) found larger models well calibrated on multiple-choice and true/false questions presented in a suitable format, and scored the probability a model assigns to its own proposed answer being true against outcomes across many questions. Studies of verbalised confidence score stated confidence against accuracy over a dataset and find language models overconfident when they put their confidence into words (Xiong et al., 2024), and find that for models tuned with human feedback the stated numbers can be better calibrated than the token probabilities (Tian et al., 2023). For dialogue agents, whether expressed doubt matches the likelihood of being wrong was found to be poor, and was much improved by training a controllable generator on a predicted likelihood of correctness (Mielke et al., 2022). One analysis found deployed models reluctant to express uncertainty even when wrong, an error rate of 47 percent on average among answers given with confidence when models were asked to state it, and a bias against uncertain text in human preference data used for post-training (Zhou et al., 2024). Every one of these is a property of a set of answers. None licenses reading a single incorrect trace

that hedges as a failure of calibration.

Outcome knowledge shifts judgement. People who know the outcome of a decision rate the decision as better when the outcome was good, holding the information available to the decision-maker fixed (Baron and Hershey, 1988), and people who know an outcome judge it to have been more predictable than they would have without that knowledge, while being unaware of the shift (Fischhoff, 1975). Model judges of reasoning share the hazard. Process reward models evaluated by best-of-N selection tolerate responses that reach the right answer through a flawed process, and many place their minimum scores on the final-answer step, a drift from judging the process toward judging the outcome (Zhang et al., 2025b). Automated evaluation of step quality finds that higher final-answer accuracy need not come with better intermediate steps (Xia et al., 2024), and on a benchmark of human-annotated step errors in competition mathematics, existing process reward models generalise poorly beyond easier problem sets (Zheng et al., 2024). Judges of open-ended responses have documented position, verbosity and self-enhancement biases (Zheng et al., 2023).

What the anchor did about it. The anchor offers two defences. Its lucky-guess analysis shows that calibration is rare on traces judged to have reached a correct answer through flawed reasoning (3.4 percent of language-model lucky guesses) and common on traces judged sound (95.9 percent). But the judge that decides whether reasoning is sound is the same judge, with the same answer key. Calibration that tracks the judge's verdict of soundness more closely than final correctness is what a label shaped by the judge's overall impression of the trace would also show. Its probe analysis finds that calibration is linearly decodable from Qwen3-4B hidden states on correct-only traces (area under the curve 0.771 for the thinking model), which shows that the label tracks something in the representation, but on incorrect traces calibration is present in 1.7 percent of thinking traces, too few to evaluate. Neither defence separates a label about calibration from a label about confident correctness.

The second account therefore predicts most of the high end of the gap: an outcome-referenced label has a large Lift because it encodes the outcome, and it is not amplified because it is not a behaviour a model can adopt separately from being right. Whether calibration in the forecasting sense is amplified by thinking training is a different question, and the evidence on it, taken up in Section 8, runs the other way.

5. Account three: confounding by indication

In clinical epidemiology, a treatment given to sicker patients can look harmful in observational data because the indication for the treatment is itself a cause of the outcome; the term for this is confounding by indication (Salas et al., 1999). A related bias arises when an analysis conditions on a variable that is itself an effect of both exposure and outcome (Hernan et al., 2004; Elwert and Winship, 2014). And associations that hold within every subgroup can reverse when the subgroups are pooled, the phenomenon named for Simpson (Blyth, 1972).

The amplified behaviours have obvious indications. Self-correction presupposes something to correct. Uncertainty acknowledgment is a response to a hard or confusing step. The anchor itself makes this argument: it notes that a model may hedge more because it is confused and self-correct more because it erred earlier, that self-correction is associated with recovery (+39 points) among traces with at least one detected failure, and that uncertainty acknowledgment falls from 66.7 to 43.7 percent as Qwen3-VL thinking models grow from 2B to 32B, which it reads as uncertainty tracking difficulty. Independent evidence agrees that these behaviours are state-dependent. Backtracking and similar inflection points occur almost only in traces where probes show a large shift in the model's belief about the answer (Boppana et al., 2026); "aha" moments come with uncertainty that rises and falls with problem difficulty (Yang et al., 2025b); and models that reason longer than needed on easy problems and switch between approaches too often on hard ones show the same behaviours for different reasons (Chen et al., 2024; Wang et al., 2025c).

What the within-question control shows. The anchor's Table 2 is the right control for difficulty, and its results are more informative than the summary that "the pattern persists". For Qwen3-4B-Thinking on MATH-500, the mean within-question difference in accuracy is +0.305 for calibration, +0.267 for self-correction, +0.048 for hypothesis testing and +0.006 for uncertainty acknowledgment, the first three with bootstrap intervals excluding zero. For the pooled thinking-model language pool, the corresponding Lift values are +76.2, +16.5, +0.6 and -14.5 points. The appendix version of the table (Table 17) reports all nine behaviours; for the other five the values are +0.596 for knowledge alignment, +0.405 for goal tracking, +0.299 for evidence citation, +0.187 for self-awareness and +0.120 for planning, against pooled thinking Lift of +79.1, +55.6, +53.2, +54.0 and +30.5. Within question, calibration therefore ranks third of nine, behind knowledge alignment and goal tracking. Setting the two side by side (Figure 1a), controlling for the question cuts calibration's association by more than half, raises self-correction's by ten points, raises hypothesis testing's by four, and removes uncertainty acknowledgment's negative association entirely. The two columns are not from the same models - the pooled values cover four thinking language models and the within-question values one - nor from the same benchmarks: the pooled values cover LogiQA2, MATH-500 and MMLU-Pro, the within-question values MATH-500 alone, so part of each shift

may reflect benchmark composition rather than the control. Each within-question mean is also taken only over the questions on which both present and absent traces occurred, and these differ by behaviour (189 questions for calibration, 132 for self-correction, 214 for hypothesis testing, 222 for uncertainty acknowledgment). The comparison is therefore indicative rather than exact. But within a question, among thinking traces, self-correction is associated with nearly as large an accuracy difference as calibration, and the negative Lift of uncertainty acknowledgment, which the anchor's abstract singles out, is not present.

The anchor's appendix states the limit of this control plainly: comparing traces on the same question removes difficulty but not the latent quality of the path a trace has taken, since a trace already on a productive path may both express calibrated confidence and reach the right answer. That caveat cuts both ways. It is a reason not to read calibration's within-question advantage as causal, and an equally good reason not to read self-correction's smaller pooled Lift as evidence that self-correction does not help. For self-correction, the latent state works against the behaviour: a trace self-corrects when it has gone wrong, so its within-question association with success is, if anything, an underestimate of whatever benefit the correction brings.

The third account therefore predicts the low end of the gap for the pooled measure. Once difficulty is held fixed, the only one of the three amplified behaviours with a clearly null association is uncertainty acknowledgment.

6. Account four: what outcome-rewarded training selects

The anchor's recommendation - reward the behaviours associated with correctness - presupposes that current training does not already do so. Outcome-rewarded training does reward behaviours associated with correctness; the question is which association.

Policy gradients act on within-prompt covariance. In REINFORCE-style training the expected update is the reward-weighted gradient of the log-probability of each sampled trace, and subtracting a baseline that does not depend on the sampled trace leaves the expectation unchanged (Williams, 1992). GRPO, the algorithm behind DeepSeek-R1, computes each sample's advantage relative to the other samples from the same prompt and divides by their spread (Shao et al., 2024; DeepSeek-AI et al., 2025). Consider a policy that differs from a reference only by a tilt toward traces showing b, applied on every prompt: the probability of trace y on prompt x is proportional to the reference probability times exp(t b(y)). Differentiating the expected reward on prompt x with respect to t at t = 0 gives exactly the within-prompt covariance of b and R, which for binary reward is q_x (1 - q_x) D_x, the quantity C_x of Section 2. Summed over prompts, the first-order pressure on the tilt is a weighted sum of within-prompt covariances; the covariance across prompts, between how often b appears and how often a prompt is solved, does not enter, because the policy conditions on the prompt. Algorithms differ in the weights. A prompt whose samples are all right or all wrong contributes nothing, and a recent analysis shows that GRPO, Dr. GRPO and DAPO amount to dividing by, ignoring, and filtering on the spread of rewards within each group (Bay and Yearick, 2026; Liu et al., 2025; Yu et al., 2025).

Filtering on correct samples does too, though the corpus does not. When training data are built by keeping only correct samples - rejection sampling, self-taught reasoning, or distillation datasets filtered for correct answers - the share of retained traces showing b is E[bR] / E[R]. Writing p for the overall success rate, this equals q + Cov(b, R) / p, which is Price's identity with correctness as fitness and no transmission term (Price, 1970). Over a pool of prompts, that covariance is the pooled one, so the composition of the filtered corpus does shift with pooled Lift: a behaviour that appears mostly on hard prompts becomes rarer in the corpus because hard prompts contribute fewer correct traces. But a model fitted to the corpus conditions on the prompt. For the same tilt, the expected maximum-likelihood gradient on the retained samples is the number of samples per prompt times the sum over prompts of q_x (1 - q_x) D_x: within-prompt covariance again, with prompts weighted equally rather than by their spread. The between-prompt part of pooled Lift reaches the model only to the extent that it learns a behaviour's frequency without reference to the prompt, for instance when it generalises a trace style from easy training prompts to hard test prompts.

So within-prompt Lift, not pooled Lift, is the quantity outcome-based training acts on to first order, whether it uses policy gradients or filtered fine-tuning. The anchor's headline reports pooled Lift; its control reports the within-question version.

The derived covariances. Combining the within-question differences for Qwen3-4B-Thinking with the thinking models' MATH-500 prevalences gives within-prompt covariances of 0.068 for calibration (prevalence 66.1 percent, difference 0.305), 0.066 for self-correction (54.6 percent, 0.267), 0.012 for hypothesis testing (42.8 percent, 0.048) and 0.0015 for uncertainty acknowledgment (43.5 percent, 0.006). Because these four prevalences lie between 40 and 70 percent, the factor q(1 - q) varies only from 0.224 to 0.248 among them, and their ranking is set by the within-question differences, so the approximation of using pooled prevalence does not drive it. The same calculation for the other five gives 0.099 for knowledge alignment (78.9 percent, 0.596), 0.064 for goal tracking (80.4 percent, 0.405), 0.047 for self-awareness (52.6 percent, 0.187), 0.036 for evidence citation (86.1 percent, 0.299) and 0.0096 for planning (91.2 percent, 0.120). Across all

nine, q(1 - q) ranges from 0.080 to 0.249, so for the high-prevalence labels the approximation matters more. The prevalence ratios on MATH-500 are 0.97, 4.37, 3.34 and 5.37 for the first four, and 1.06, 0.98, 1.43, 0.94 and 0.99 for the other five (Figure 1b).

Read through this lens, the gap splits. Knowledge alignment has the largest covariance with correctness within a question, and calibration, self-correction and goal tracking follow close together. Of these four, only self-correction is strongly amplified; in Qwen3-4B-Thinking its profile is what selection on correctness would produce. Hypothesis testing has a small positive covariance and a large amplification. Uncertainty acknowledgment has none and the largest amplification. Calibration and knowledge alignment have large covariances and no amplification; both are outcome-referenced labels, so account two applies to them, and the next paragraph explains why such labels do not respond. Goal tracking is the exception this account does not explain. Its label is closer to form-referenced - the anchor's markers are phrases such as "the goal is…" and "so far we know…" - its covariance about equals self-correction's, and its prevalence is the same in both model types (80.4 against 81.7 percent). One possibility is that at about 80 percent in both it was near a ceiling before thinking training began; the anchor's tables cannot test this, and until it is tested goal tracking counts against account four.

One published within-question measurement exists for a model trained by reinforcement learning alone, and it is mixed. For DeepSeek-R1-Zero on MATH questions, Liu et al. (2025) sampled 100 responses per question, split them by whether they showed self-reflection, and compared accuracy within each question; they report that "nearly half responses with self-reflection do not achieve higher accuracy than those without self-reflection". Their self-reflection label is broader than the anchor's self-correction, but the result does not show, for a pure reinforcement-learning model, the clear within-question advantage that Qwen3-4B-Thinking shows.

Why calibration does not respond. In quantitative genetics the response of a trait to selection depends both on the selection on it and on how much heritable variation it has, and a trait under strong selection with little heritable variation barely moves (Lande, 1979). The analogue here is whether the policy can change b without changing correctness. For a form-referenced behaviour it can: a model can add or remove a revision independently of whether its answer is right. For a label that is close to "confident and correct", it cannot. Tilting toward calibrated traces is, under the label's empirical profile, tilting toward correct traces, and the only part of the label the policy controls separately is how confidence is expressed, which binary outcome reward does not pay for. Reinforcement learning with a binary correctness reward has been reported to degrade calibration, while adding a Brier-score term to the reward improves it without loss of accuracy (Damani et al., 2025). The anchor's own data fit this: thinking models are more accurate on four of six benchmarks, calibration prevalence is unchanged, and P(calibration | correct) is lower in thinking models (82.0 percent) than in instruct models (87.5 percent).

Why uncertainty acknowledgment might respond without selection. In multivariate selection, a trait with no direct selection on it changes when it is correlated with a trait that is selected (Lande and Arnold, 1983), and in population genetics a neutral variant linked to a favoured one rises with it, the hitchhiking effect (Maynard Smith and Haigh, 1974). Reasoning traces have an obvious linkage: the same tokens open both doubt and revision. The anchor's own list of surface markers gives "wait" and "actually" for self-correction and "I'm not sure" for uncertainty; a sentence-level taxonomy used in counterfactual resampling places expressions of confusion and backtracking in one category (Bogdan et al., 2025); and tokens such as "Hmm" and "Wait" mark points of high mutual information with the answer (Qian et al., 2025). Behaviour-specific steering directions exist for backtracking and for expressing uncertainty separately (Venhoff et al., 2025a), which shows the two are separable in principle; whether training moves them together is not known. Speculatively, selection on revision may carry the expression of doubt along with it. This is a hypothesis about mechanism; it is testable by estimating a selection gradient rather than a differential (Section 14).

7. Account five: amplification without selection

The anchor attributes the amplification to "reasoning-oriented training". Several mechanisms raise a behaviour's frequency that are not selection on correctness at all.

Imitation. Two of the four thinking language models were built largely by imitation. DeepSeek-R1-Distill-Qwen-7B was fine-tuned on samples from DeepSeek-R1 (DeepSeek-AI et al., 2025), and the Qwen3 report describes building its smaller models by transferring knowledge from its flagship models rather than by training each from scratch (Yang et al., 2025a). The anchor's own stage analysis is the clearest evidence: OLMo-3-7B-Think-SFT already shows self-correction in 34.5 percent of traces against 1.0 percent for the instruct SFT checkpoint, with the same pattern for hypothesis testing (22.7 against 1.0) and uncertainty (33.1 against 1.0), before any preference or reinforcement stage in the Olmo 3 pipeline (Olmo Team et al., 2025). Supervised fine-tuning on traces copies the style of the teacher. A study of bottom-up reasoning strategies found that the format of training data, free-form or multiple-choice, shapes reasoning behaviour far more than its domain (Lee et al., 2025). Model diffing finds that distillation

installs new reasoning mechanisms while reinforcement learning mainly teaches when to use mechanisms the base model already has, with hybrid models recovering about 76 percent of the gap to reinforcement-learned models but about 11 percent of the gap to distilled ones (Venhoff et al., 2025b). If the gap is visible after imitation alone, the selection the anchor's recommendation aims to redirect did not create it.

Base-model behaviour. Self-correction and reflection are present before post-training. A pretrained 7B model recovers from deliberately inserted errors on a suite of reflection tasks, and the ability grows over pretraining (Essential AI et al., 2025); error correction is present in base models but costly to elicit with reinforcement learning (Yeo et al., 2025); reflection keywords appear in DeepSeek-V3-Base before R1-Zero training (Liu et al., 2025); and reinforcement learning from different base models produces different behavioural trajectories, with response length rising in some runs without the emergence of verification (Zeng et al., 2025b). Analyses of pass@k across sample budgets indicate that the reasoning of models trained with verifiable rewards is drawn from, and bounded by, what the base model could already reach (Yue et al., 2025). Amplification in this regime is the reweighting of an existing repertoire, which is what selection does, but the repertoire, its coupling and its starting frequencies are set by pretraining.

Optimiser biases. Two biases in group-relative policy optimisation change trace composition independently of the reward's association with a behaviour. An optimisation bias in GRPO lengthens responses during training, incorrect ones especially (Liu et al., 2025), and a separate analysis traces verbosity in reinforcement-learned reasoners to the loss on incorrect answers rather than to deeper reasoning (Fatemi et al., 2025). The clipping term can amplify behaviours with a high prior even when rewards carry no information. With randomly assigned rewards, GRPO raised Qwen2.5-Math-7B's MATH-500 score by 21.4 points, against 29.1 points with true rewards; under such spurious rewards the share of Qwen2.5-Math responses that reason in code without executing it rose from 65 percent to over 90 percent; and the same rewards often did not help other model families such as Llama3 and OLMo2 (Shao et al., 2025). Behaviours that come with length - hedging, re-checking, enumerating alternatives - will be amplified by any bias toward length. The anchor's length-controlled analysis shows that length does not remove the prevalence gap within quintiles of response length; it does not show that length bias in training played no part in producing it.

8. Account six: a behaviour can pay at training time without paying per trace

Lift measures one thing: whether a trace that shows a behaviour is more likely to end in a correct answer. Training can value a behaviour for other reasons.

As a precondition for learning. In a controlled Countdown study, priming a model that lacked verification, backtracking, subgoal setting and backward chaining with examples of them enabled it to improve under reinforcement learning to the level of a model that had them naturally; priming with examples that showed the behaviours but reached wrong answers worked as well as priming with correct ones, while length-matched placeholder traces did not (Gandhi et al., 2025). The behaviours mattered for whether the model could learn, not for whether each primed example was right. In the same study, reinforcement learning on the primed model amplified backtracking and verification and suppressed backward chaining and subgoal setting, which is selection doing what the anchor wants it to do.

As exploration. A small minority of tokens in a reasoning trace have high entropy and act as forks between reasoning paths; restricting policy-gradient updates to roughly the highest-entropy fifth matched full updates on one base model and exceeded them on two larger ones (Wang et al., 2025b). Policy entropy falls during reinforcement learning at a rate set by the covariance between action probability and advantage, and performance gains are traded against that falling entropy (Cui et al., 2025). A behaviour that opens alternatives - hypothesis testing, or doubt that prompts a second look - may raise the share of prompts on which a group of samples disagrees, which Section 6 showed is where group-baselined learning happens, even if its per-trace association with success is zero.

As a mediator of calibration. The most direct evidence comes from a study of verbalised confidence: reasoning models were strictly better calibrated than their non-reasoning counterparts in 33 of 36 model and dataset settings, became better calibrated as their traces unfolded, and, in an ablation on one distilled 32B model on two knowledge datasets, lost calibration when the parts of the trace that explore alternatives and revise were removed; non-reasoning models guided to think slowly also gained calibration (Yoon et al., 2025). Another study finds that distillation and reinforcement learning improve verbalised calibration on reasoning-heavy tasks while reasoning models say "I don't know" less often on factual questions, and that confidence is higher after shorter reasoning (Zeng et al., 2025a). A third finds that reasoning behaviours do not by themselves make expressed confidence faithful to the model's internal uncertainty (Gani et al., 2026). The evidence is mixed, but on the first study's reading the amplified behaviours are part of how calibration in the forecasting sense is produced. If so, the anchor's two clusters are not competitors for training's attention: exploring alternatives and backtracking would be inputs to calibration, and rewarding calibration directly might amplify them further.

As a route to first-try accuracy. Training on traces with more reflection steps improved the accuracy of the first

answer more than the ability to correct a wrong first answer, and training on traces with corrective reflections was no better than training on traces with only confirmatory ones (Kang et al., 2025). A behaviour whose value appears in the first answer would show little Lift within the trace that contains it, because the trace with the better first answer would often not need it.

9. What interventions on the amplified behaviours show

Associations cannot settle what a behaviour does; interventions can. The record on the amplified behaviours is mixed, and it supports the anchor in part.

Removing them often costs little. Suppressing tokens such as "Wait" and "Hmm" during decoding shortened traces by 27 to 51 percent in five model series across ten benchmarks without loss of utility (Wang et al., 2025a). Over 90 percent of reflections after a candidate answer left it unchanged, corrective changes from wrong to right were mostly under 2 percent, and truncating reflection after a candidate answer saved 24.5 percent of tokens for under 2.9 points of accuracy (Kang et al., 2025). Repeated re-verification was mostly confirmatory, and suppressing it when past experience suggested it was unnecessary cut tokens by up to 20.3 percent without loss of accuracy (Long et al., 2026). Probes on hidden states can tell when an intermediate answer is right well before the model stops, and exiting on the probe's signal saved 24 percent of tokens without loss (Zhang et al., 2025a). In vision-language models, behaviours of the "aha moment" type gave no reliable gains from inference-time scaling (Wu et al., 2025a).

Adding or extending them sometimes helps, up to a point. Appending "Wait" when a model tries to stop thinking led it to double-check and raised one model's AIME 2024 accuracy from 50 to 57 percent (Muennighoff et al., 2025). Across models and benchmarks, extending thinking in this way improves accuracy at first and then lowers it (Ghosal et al., 2025); accuracy follows an inverted U in chain length, with the optimum longer for harder tasks and shorter for more capable models (Wu et al., 2025b); and DeepSeek-R1 has a range of reasoning length beyond which more thinking hurts, with a tendency to ruminate on formulations it has already explored (Marjanovic et al., 2025). Within a question, the shortest sampled chain was up to 34.5 percent more accurate than the longest (Hassid et al., 2025).

Some reflective moves are pivotal. Resampling the continuation of a trace from each sentence finds that sentences that plan or manage uncertainty, including backtracking, have more influence on the final answer distribution than computation or fact retrieval (Bogdan et al., 2025); this is a measure of influence, not of benefit. Tokens of reflection and transition coincide with peaks in mutual information between the model's representation and the correct answer (Qian et al., 2025). Penalising switches between lines of thought at decoding time improved accuracy on hard mathematics problems, after frequent switching had been found to go with incorrect answers (Wang et al., 2025c).

The anchor's prompting experiment. The anchor prompted Qwen3-4B-Thinking to calibrate its confidence, use the right framework and notice missing information (high-Lift prompt), or to voice doubt at every step (low-Lift prompt). The high-Lift prompt raised accuracy by 5.8 points on MATH-500 and 3.6 on MMLU-Pro Math, with a 0.7-point difference on GPQA that was not significant; the low-Lift prompt lowered it by 4.6, 12.1 and 3.5 points, the last not significant. The anchor describes this experiment as preliminary because each prompt changes several properties at once. Two further points bear on interpretation. The low-Lift prompt drove uncertainty acknowledgment to 94 to 99 percent of traces, well beyond the 25 to 85 percent seen in thinking models without it, so it tests a dose of hedging the models do not produce on their own. And instructing a model to doubt every step changes what the model does, not only what it says.

Self-correction in general. Earlier work found that models could not reliably improve their reasoning by reviewing it without external feedback, and sometimes made it worse (Huang et al., 2024); a critical survey concluded that self-correction works with reliable external feedback or with large-scale fine-tuning for the purpose (Kamoi et al., 2024); and multi-turn reinforcement learning on a model's own correction traces improved self-correction where fine-tuning on offline correction traces collapsed into ineffective modes (Kumar et al., 2024). Rewarding the reflection that precedes a successful second attempt improved performance by as much as 34.7 percent on writing math equations and 18.1 percent on function calling (Bensal et al., 2025). These results show that self-correction trained for its own sake can work. They do not show that the self-correction amplified in current thinking models is the trained kind.

Read together, the interventions support part of the anchor's reading. Much amplified reflection is confirmatory and can be cut without loss; doubt that is voiced at every step hurts. They also show that some reflective moves are among the most influential in a trace, that extending them can help when a model would otherwise stop early, and that their removal costs calibration. What they cannot show, because none of them is a training intervention matched to the anchor's comparison, is whether reasoning-oriented training would have produced better models by amplifying different behaviours.

10. Consolidated evidence

Table 1 collects the results the argument uses, one per row, with its source. Values are those the source reports, except the rows marked "This paper, derived", which are computed from the anchor's published tables as described in Sections 4 and 6. Rows use different metrics and should not be compared across rows. The last column names the account each row bears on: A1 misdirected training, A2 answer-aware labels, A3 confounding by indication, A4 selection with constrained response, A5 amplification without selection, A6 training-time value.

11. Predictions that separate the accounts

The six accounts are not exclusive; more than one can hold for the same behaviour, and the evidence above suggests that they do. What makes them useful is that they disagree about observable quantities. Table 2 lists five.

The most informative single comparison is the third column. If the amplification of uncertainty acknowledgment appears in a model trained only by group-baselined reinforcement learning from a base model, with no distillation, account five's imitation branch is ruled out for it and the question becomes whether optimiser bias, hitchhiking or training-time value explains it. If it appears only after supervised fine-tuning on another model's traces, the gap says something about the teacher, and nothing about what reinforcement learning rewards.

The second most informative is the first column. If the Lift of calibration falls toward that of self-correction when the judge does not see the answer, account two carries the high end of the gap.

12. A procedure for reading a prevalence-Lift gap

Before a gap between how often training makes a behaviour appear and how strongly it is associated with success is read as evidence that training is rewarding the wrong thing, it should pass the steps in Algorithm 1. The steps run from cheapest to most expensive; the first three need only data such studies already collect.

A behaviour that is amplified, has a positive within-prompt covariance with reward, and rises under true rewards but not random ones is being selected on correctness, and its gap with pooled Lift is not evidence of misdirection. A behaviour that rises under random rewards too, or rises only after imitation, is being amplified for another reason, and the anchor's concern applies to it with full force.

13. The strongest case against this reading

"**The anchor already said Lift is not causal; this paper attacks a claim it did not make.**" The anchor does say that Lift is descriptive. But its central conclusion - that reasoning-oriented training does not preferentially amplify the highest-Lift behaviours, and that process objectives should therefore reward them - is a claim about training, and it is drawn from the descriptive measure. The point here is not that Lift fails to be causal; it is that Lift is the wrong descriptive measure for a claim about what training selects, because training selects on a different covariance and the high-Lift labels are not behaviours in the sense the claim needs.

"**The within-question control confirms the ranking; calibration is still first.**" It is not first: within question it ranks third of nine, behind knowledge alignment ($+0.596$) and goal tracking ($+0.405$). The margin between calibration and self-correction shrinks from about 60 points pooled to under 4 points within question, though the two averages are taken over different sets of questions (189 and 132), and the anchor does not report a test of the difference between them. In Qwen3-4B-Thinking, the within-question control removes the negative association of uncertainty acknowledgment, which was the abstract's sharpest example; in the matched instruct model it remains (-0.094, with an interval excluding zero), as does a large calibration/self-correction gap ($+0.518$ against $+0.078$). For account three, that instruct row means difficulty does not explain the whole negative association: there, traces that voice doubt do worse than other traces on the same question, which the latent-path caveat of Section 5 allows and confounding by question difficulty alone does not. And calibration's within-question advantage, like knowledge alignment's, still carries the answer-aware labelling of Section 4.

"**Imitation is reasoning-oriented training too, so the gap stands.**" It does stand as a description of the released models. The anchor's recommendation, however, concerns objectives: reward calibrated and grounded reasoning. If the amplification of uncertainty came from copying a teacher's traces, a change in the reward used in a later stage is not the lever; the lever is the choice of teacher, or the filtering of its traces.

"**The amplified behaviours really are mostly wasteful; the efficiency literature agrees.**" Much of it is. Confirmatory reflection can be truncated cheaply (Kang et al., 2025; Long et al., 2026), and suppressing reflection tokens loses little (Wang et al., 2025a). This paper agrees that thinking models produce more deliberative surface than they need, and that efficiency work is right to cut it. It disputes only the inference that because surface is abundant and outcome-referenced labels are not amplified, the training signal is pointed at the wrong behaviours. The confirmatory reflection that can be cut at inference may still have had value during

Table 1. Published results on reasoning behaviours, their associations with correctness, and the training that amplifies them, with the accounts each informs.

Source	Setting	Reported value	Bears on
Nyandwi et al. 2026	8,282 LLM traces, pooled	Calibration Lift +79.6; uncertainty -13.9	A1
Nyandwi et al. 2026	MATH-500, Th vs In	Self-correction 54.6 vs 12.5 percent	A1, A5
Nyandwi et al. 2026	LLM traces	Calibration on 84.7% correct, 0.6% incorrect	A2
Nyandwi et al. 2026	Incorrect thinking traces	Uncertainty 49.7%, calibration 1.0%	A2
This paper, derived	From the row above	At least 98% of doubting errors uncalibrated	A2
Nyandwi et al. 2026	Qwen3-4B, within question	Know. align. +0.596, goal track. +0.405, calib. +0.305, self-corr. +0.267, unc. +0.006	A3, A4
This paper, derived	Within-prompt covariance	Know. align. 0.099, calib. 0.068, self-corr. 0.066, goal track. 0.064, hyp. 0.012, unc. 0.0015	A4
Nyandwi et al. 2026	OLMo-3 SFT checkpoints	Self-correction 34.5 vs 1.0 percent	A5
Nyandwi et al. 2026	Prompting, MATH-500	High-Lift +5.8, low-Lift -4.6 points	A1
Baron and Hershey 1988	Human raters	Outcome changes rated decision quality	A2
Zhang et al. 2025b	PRM evaluation	Minimum scores shift to final step	A2
Boppana et al. 2026	R1, GPT-OSS probes	Backtracking where belief shifts	A3
Yang et al. 2025b	Aha moments	Uncertainty adjusts to difficulty	A3
Bay and Yearick 2026	GRPO family	Unanimous groups give no update	A4
Damani et al. 2025	RL with Brier term	Calibration up, no accuracy loss	A4
Shao et al. 2025	Random rewards, Qwen2.5-Math	Code reasoning 65% to over 90%	A5
Liu et al. 2025	GRPO length bias	Longer incorrect responses	A5
Venhoff et al. 2025b	9 base/thinking pairs	Recovery 76% RL vs 11% SFT gap	A5
Essential AI 2025	OLMo2-7B pretraining	Self-correction before post-training	A5
Gandhi et al. 2025	Countdown priming	Wrong-answer priming matches correct	A6
Wang et al. 2025b	Forking tokens	20% of tokens match full update	A6
Yoon et al. 2025	6 models, 6 datasets	Better calibrated in 33 of 36	A6
Yoon et al. 2025	Ablation, R1-Distill-32B	Removing alternatives hurts calibration	A6
Kang et al. 2025	8 models, 5 datasets	Over 90% of reflections confirmatory	A1, A6
Wang et al. 2025a	NoWait, 5 series	27 to 51% shorter, utility kept	A1
Muennighoff et al. 2025	Budget forcing	AIME24 50% to 57%	A6
Hassid et al. 2025	Within question	Shortest chain up to 34.5% more accurate	A3
Bogdan et al. 2025	Resampling	Planning and uncertainty-management sentences most influential	A6
Liu et al. 2025	DeepSeek-R1-Zero, 100 samples per MATH question, within question	Reflection not positively tied to accuracy	A1, A4
Kargupta et al. 2025	192K traces, 18 models	Success-linked elements under-used	A1

Table 2. Predictions of each account for five measurements; "same" means the gap or ranking is unchanged from the anchor's.

Account	Answer-blind labels	Within-prompt Lift	RL-only vs SFT-only	Behaviour ablated	Calibration reward
A1 Misdirected	Same ranking	Same ranking	Gap in both	No accuracy cost	Different behaviours up
A2 Answer-aware	Top labels fall	n/a	n/a	n/a	n/a
A3 Indication	Same	Amplified rise	n/a	Cost when needed	n/a
A4 Selection	Same	Tracks amplification	Gap smaller in RL	n/a	Calibration rises
A5 Imitation, bias	Same	Unrelated	Gap mostly in SFT	No accuracy cost	n/a
A6 Train-time value	Same	Small	Gap in RL too	Learning slows	Amplified rise too

training (Kang et al., 2025; Gandhi et al., 2025), and the same reflective moves are implicated in producing calibration (Yoon et al., 2025).

"Process supervision is the obvious fix, whatever the cause of the gap." Step-level supervision from human labels outperformed outcome supervision for training reward models on MATH (Lightman et al., 2023), so rewarding the process is a reasonable direction. But a process reward built on the anchor's high-Lift labels, as annotated, would pay for traces a judge with the answer key recognises as correct, which is outcome supervision with extra steps, and optimising a proxy that is only correlated with the target is the textbook setting for the regressional and extremal failures catalogued under Goodhart's law (Manheim and Garrabrant, 2018). A related warning comes from rewarding confidence

Algorithm 1. Reading a gap between amplification and Lift for a trace behaviour

```
Input: traces with labels and correctness; model pairs; the training recipe of each model
step 1 – Classify each label as form-referenced or outcome-referenced.
   Re-annotate a stratified sample with the answer key withheld; report Lift both ways.
   Report P(label | incorrect) beside P(label | correct) for every label.
step 2 – Replace pooled Lift by within-prompt Lift. Sample k >= 8 traces per prompt;
   average D_x over prompts with mixed outcomes; report intervals.
step 3 – Compute the quantity the training acts on.
   for RL or filtered SFT: use the sum over prompts of q_x (1 – q_x) D_x,
   weighted as the algorithm weights prompts (e.g. 1 / reward spread in GRPO)
   report pooled Cov(b, R) only as a description of a filtered corpus
   Estimate partial (gradient) coefficients with all behaviours in one model.
step 4 – Locate the amplification by stage. Compare base, SFT and RL checkpoints of one
   family; mark distillation from another model's traces.
step 5 – Screen for amplification without selection. Rerun RL with random rewards and
   with length-unbiased loss; any behaviour that still rises is not selected for.
step 6 – Test training-time value. Ablate the behaviour in SFT data or penalise it in RL;
   report effects on learning rate and on final accuracy, not on per-trace Lift.
Output: for each behaviour, whether its amplification is selected, imitated, biased,
   coupled to a selected behaviour, or of training-time value, with the residual stated.
```

directly: using a model's own certainty as its only reward matched outcome-rewarded training on mathematics benchmarks in one study (Zhao et al., 2025), which shows that confidence-shaped rewards move behaviour, not that they move it toward calibration. The objective the anchor wants needs a target defined on the forecasting sense of calibration, as in a proper scoring rule (Damani et al., 2025), not on a per-trace label.

"**A judge with the answer key is standard practice and necessary for labels like knowledge alignment.**" It is necessary for some labels. That is the reason to report them separately: an outcome-referenced label is legitimate for describing why a trace succeeded, and unsuitable as a target a training objective should raise, since raising it is the same as raising accuracy.

14. Studies that would settle the open part

Study 1: blind relabelling. Re-annotate a stratified sample of the anchor's released traces with the answer key withheld from the judge, and with the model's final answer masked where feasible. Report Lift for every label under both conditions. If calibration, knowledge alignment and self-awareness lose much of their Lift while form-referenced labels keep theirs, account two carries the top of the ranking. This needs no new model runs.

Study 2: within-prompt selection analysis. For one model family, sample 16 traces per prompt on a few thousand prompts of mixed difficulty, label all nine behaviours, and fit a within-prompt model of correctness on all behaviours jointly. Report each behaviour's selection differential (marginal within-prompt covariance) and selection gradient (partial coefficient). If uncertainty acknowledgment has a zero or negative gradient but co-occurs with self-correction, its amplification is a candidate correlated response; if its gradient is positive once self-correction is held fixed, account one is weakened for it.

Study 3: stage-separated amplification. Using a fully open pipeline such as Olmo 3, which releases checkpoints after each stage (Olmo Team et al., 2025), measure behaviour prevalence at base, after supervised fine-tuning, after preference optimisation and after reinforcement learning. Repeat reinforcement learning from the base model with true rewards, with random rewards (Shao et al., 2025) and with a length-unbiased loss (Liu et al., 2025). A behaviour that rises under true rewards and not under random ones is selected; one that rises under both is amplified by the optimiser; one that rises only after imitation was copied.

Study 4: reward calibration directly. Train with a reward that adds a proper scoring rule for an explicit confidence to correctness, as in RLCR (Damani et al., 2025), and measure all nine behaviours before and after. If calibration-in-the-forecasting-sense improves and exploring alternatives and backtracking rise with it, as Yoon et al. (2025) predicts, the anchor's two clusters are complements. If the amplified behaviours fall while calibration rises, they were competitors, and account one gains support.

Study 5: training-time ablation. Remove one amplified behaviour from supervised traces, or penalise it during reinforcement learning, and measure learning speed and final accuracy at matched compute. A behaviour with near-zero per-trace Lift whose removal slows learning has training-time value that Lift cannot see (Gandhi et al., 2025; Wang et al., 2025b).

15. What is not known

It is not known how much of the Lift of the calibration, knowledge alignment and self-awareness labels survives when the judge does not see the answer. It is not known how calibrated, in the forecasting sense, the anchor's models are, since the anchor measures a per-trace label rather than the match between stated confidence and accuracy over a set of answers. The within-prompt covariances in Section 6 combine one model's within-question differences with four models' pooled

prevalences; the exact within-prompt quantities for each model have not been reported. No published study has estimated selection gradients, rather than differentials, for reasoning behaviours. It is not known which stage of which pipeline produced the amplification of uncertainty acknowledgment in any of the anchor's models, nor whether it persists when reinforcement learning starts from a base model without distillation. It is not known whether doubt and revision are coupled in the parameters that training moves. And it is not known whether a process reward for the anchor's high-Lift labels, as annotated, would do anything other than reward correct answers a second time.

16. Limitations

The derived quantities rest on the anchor's published tables. The bound on doubting incorrect traces uses only the marginal rates and is therefore a lower bound; the true share could be higher, and the judge's reading of mid-trace doubt followed by a confident final answer as miscalibration may be defensible for some traces. The covariance calculation pairs a single model's within-question differences with pooled prevalences; the approximation affects the scale of the covariances of the four behaviours compared most closely only slightly because q(1 - q) varies little over their prevalences, though more for high-prevalence labels such as planning and evidence citation, and a per-model calculation could change the ranking among calibration, self-correction and goal tracking, which is close. The tilting model in Section 6 holds the rest of the trace distribution fixed given the behaviour, which real policy updates do not; it identifies which covariance a baseline-corrected gradient responds to, not how fast a behaviour will rise. The quantitative-genetics analogy is algebraic: Price's identity and the differential-gradient distinction carry over exactly, but nothing like heritability has been measured for model behaviours, and "constrained response" is a description of the label's structure rather than a measured property of any model. The evidence on training-time value comes from small or narrow settings, Countdown in one case and particular model families in others, and may not transfer to the anchor's models. Of the paper's results, the reader should weigh the bound on doubting incorrect traces and the pooled-versus-within comparison most heavily, since both follow directly from the anchor's reported values; the covariance ranking less, because it combines two populations; and the mechanisms proposed in Sections 6 to 8 least, since none has been tested on the anchor's models. Finally, this paper relies on one anchor study for its central numbers; the anchor is a recent conference paper whose released annotations would allow every derived quantity here to be checked and refined.

17. Conclusion

Thinking models write more revisions, more alternatives and more doubt than their instruction-tuned counterparts, and the labels most strongly associated with correct answers are not among those that increase. The published evidence does not yet show that this is because training rewards the wrong behaviours. The labels at the top of the ranking were assigned by a judge that knew the answer, and one of them is present on almost no incorrect trace, including the half that voice doubt. The ranking is pooled over questions whose difficulty triggers the amplified behaviours; within a question, in Qwen3-4B-Thinking, self-correction is associated with almost as large an accuracy difference as calibration, and uncertainty's negative association is gone, though it remains in the matched instruct model. Outcome-rewarded training selects on within-prompt covariance, by which measure self-correction ranks with calibration (knowledge alignment, another outcome-referenced label, ranks first), and a label defined by confidence matching correctness is not something a policy can raise apart from being right. What remains is a real and narrower finding: the expression of doubt is amplified about five-fold with no within-question association, a pattern that imitation, optimiser bias, coupling to revision and training-time value each predict. Goal tracking, which is about as strongly tied to correctness within a question as self-correction and is not amplified, is a second residual that this paper does not explain. The tests that would decide among them - blind relabelling, within-prompt selection gradients, stage-separated checkpoints with random-reward controls, and a calibration reward - can be run with released models and the anchor's released annotations.

References

Baron, J., & Hershey, J. C. (1988). Outcome bias in decision evaluation. Journal of Personality and Social Psychology, 54(4), 569-579. doi:10.1037/0022-3514.54.4.569.

Bay, Y. Y., & Yearick, K. A. (2026). GRPO, Dr. GRPO, and DAPO Are Three Operations on One Number: The Group-Standard-Deviation Identity. arXiv:2607.00152. doi:10.48550/arXiv.2607.00152.

Bensal, S., Jamil, U., Bryant, C., Russak, M., Kamble, K., Mozolevskyi, D., et al. (2025). Reflect, Retry, Reward: Self-Improving LLMs via Reinforcement Learning. arXiv:2505.24726. doi:10.48550/arXiv.2505.24726.

Blyth, C. R. (1972). On Simpson's Paradox and the Sure-Thing Principle. Journal of the American Statistical Association, 67(338), 364-366. doi:10.1080/01621459.1972.10482387.

Bogdan, P. C., Macar, U., Nanda, N., & Conmy, A. (2025). Thought Anchors: Which LLM Reasoning Steps Matter? arXiv:2506.19143. doi:10.48550/arXiv.2506.19143.

Boppana, S., Ma, A., Loeffler, M., Sarfati, R., Bigelow, E., Geiger, A., et al. (2026). Reasoning Theater: Disentangling Model Beliefs from Chain-of-Thought. arXiv:2603.05488. doi:10.48550/arXiv.2603.05488.

Chen, X., Xu, J., Liang, T., He, Z., Pang, J., Yu, D., et al. (2024). Do NOT Think That Much for 2+3=? On the Overthinking of o1-Like LLMs. arXiv:2412.21187. doi:10.48550/arXiv.2412.21187.

Chen, Y., Benton, J., Radhakrishnan, A., Uesato, J., Denison, C., Schulman, J., et al. (2025). Reasoning Models Don't Always Say

What They Think. arXiv:2505.05410. doi:10.48550/arXiv.2505.05410.

Cui, G., Zhang, Y., Chen, J., Yuan, L., Wang, Z., Zuo, Y., et al. (2025). The Entropy Mechanism of Reinforcement Learning for Reasoning Language Models. arXiv:2505.22617. doi:10.48550/arXiv.2505.22617.

Damani, M., Puri, I., Slocum, S., Shenfeld, I., Choshen, L., Kim, Y., et al. (2025). Beyond Binary Rewards: Training LMs to Reason About Their Uncertainty. arXiv:2507.16806. doi:10.48550/arXiv.2507.16806.

DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., et al. (2025). DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. Nature, 645, 633-638. doi:10.1038/s41586-025-09422-z. arXiv:2501.12948. doi:10.48550/arXiv.2501.12948.

DeGroot, M. H., & Fienberg, S. E. (1983). The Comparison and Evaluation of Forecasters. The Statistician, 32, 12-22. doi:10.2307/2987588.

Elwert, F., & Winship, C. (2014). Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable. Annual Review of Sociology, 40, 31-53. doi:10.1146/annurev-soc-071913-043455.

Essential AI, Shah, D. J., Rushton, P., Singla, S., Parmar, M., et al. (2025). Rethinking Reflection in Pre-Training. arXiv:2504.04022. doi:10.48550/arXiv.2504.04022.

Fatemi, M., Rafiee, B., Tang, M., & Talamadupula, K. (2025). Concise Reasoning via Reinforcement Learning. arXiv:2504.05185. doi:10.48550/arXiv.2504.05185.

Fischhoff, B. (1975). Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under uncertainty. Journal of Experimental Psychology: Human Perception and Performance, 1(3), 288-299. doi:10.1037/0096-1523.1.3.288.

Gandhi, K., Chakravarthy, A., Singh, A., Lile, N., & Goodman, N. D. (2025). Cognitive Behaviors that Enable Self-Improving Reasoners, or, Four Habits of Highly Effective STaRs. arXiv:2503.01307. doi:10.48550/arXiv.2503.01307.

Gani, A., Meskin, A., Liu, G. K.-M., & Cohan, A. (2026). Quantifying Faithful Confidence Expression in Large Reasoning Models. arXiv:2606.03969. doi:10.48550/arXiv.2606.03969.

Ghosal, S. S., Chakraborty, S., Reddy, A., Lu, Y., Wang, M., Manocha, D., et al. (2025). Does Thinking More always Help? Mirage of Test-Time Scaling in Reasoning Models. NeurIPS 2025. arXiv:2506.04210. doi:10.48550/arXiv.2506.04210.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On Calibration of Modern Neural Networks. ICML 2017. arXiv:1706.04599. doi:10.48550/arXiv.1706.04599.

Hassid, M., Synnaeve, G., Adi, Y., & Schwartz, R. (2025). Don't Overthink it: Preferring Shorter Thinking Chains for Improved LLM Reasoning. arXiv:2505.17813. doi:10.48550/arXiv.2505.17813.

Hernan, M. A., Hernandez-Diaz, S., & Robins, J. M. (2004). A Structural Approach to Selection Bias. Epidemiology, 15(5), 615-625. doi:10.1097/01.ede.0000135174.63482.43.

Huang, J., Chen, X., Mishra, S., Zheng, H. S., Yu, A. W., Song, X., et al. (2024). Large Language Models Cannot Self-Correct Reasoning Yet. ICLR 2024. arXiv:2310.01798. doi:10.48550/arXiv.2310.01798.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., et al. (2022). Language Models (Mostly) Know What They Know. arXiv:2207.05221. doi:10.48550/arXiv.2207.05221.

Kamoi, R., Zhang, Y., Zhang, N., Han, J., & Zhang, R. (2024). When Can LLMs Actually Correct Their Own Mistakes? A Critical Survey of Self-Correction of LLMs. Transactions of the Association for Computational Linguistics, 12, 1417-1440. arXiv:2406.01297. doi:10.48550/arXiv.2406.01297.

Kang, L., Deng, Y., Xiao, Y., Mo, Z., Lee, W. S., & Bing, L. (2025). First Try Matters: Revisiting the Role of Reflection in Reasoning Models. arXiv:2510.08308. doi:10.48550/arXiv.2510.08308.

Kargupta, P., Li, S. S., Wang, H., Lee, J., Chen, S., Ahia, O., et al. (2025). Cognitive Foundations for Reasoning and Their Manifestation in LLMs. arXiv:2511.16660. doi:10.48550/arXiv.2511.16660.

Kumar, A., Zhuang, V., Agarwal, R., Su, Y., Co-Reyes, J. D., Singh, A., et al. (2024). Training Language Models to Self-Correct via Reinforcement Learning. arXiv:2409.12917. doi:10.48550/arXiv.2409.12917.

Lande, R. (1979). Quantitative Genetic Analysis of Multivariate Evolution, Applied to Brain: Body Size Allometry. Evolution, 33(1), 402-416. doi:10.2307/2407630.

Lande, R., & Arnold, S. J. (1983). The Measurement of Selection on Correlated Characters. Evolution, 37(6), 1210-1226. doi:10.2307/2408842.

Lee, S., Kim, S., Seo, M., Jo, Y., Go, D., Hwang, H., et al. (2025). The CoT Encyclopedia: Analyzing, Predicting, and Controlling how a Reasoning Model will Think. arXiv:2505.10185. doi:10.48550/arXiv.2505.10185.

Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., et al. (2023). Let's Verify Step by Step. arXiv:2305.20050. doi:10.48550/arXiv.2305.20050.

Liu, Z., Chen, C., Li, W., Qi, P., Pang, T., Du, C., Lee, W. S., & Lin, M. (2025). Understanding R1-Zero-Like Training: A Critical Perspective. arXiv:2503.20783. doi:10.48550/arXiv.2503.20783.

Long, Q., Jiang, K. J., Chen, J., Guo, X., Gan, L., & Wang, W. (2026). Self-Verification Dilemma: Experience-Driven Suppression of Overused Checking in LLM Reasoning. arXiv:2602.03485. doi:10.48550/arXiv.2602.03485.

Manheim, D., & Garrabrant, S. (2018). Categorizing Variants of Goodhart's Law. arXiv:1803.04585. doi:10.48550/arXiv.1803.04585.

Marjanovic, S. V., Patel, A., Adlakha, V., Aghajohari, M., BehnamGhader, P., Bhatia, M., et al. (2025). DeepSeek-R1 Thoughtology: Let's think about LLM Reasoning. Transactions on Machine Learning Research. arXiv:2504.07128. doi:10.48550/arXiv.2504.07128.

Maynard Smith, J., & Haigh, J. (1974). The hitch-hiking effect of a favourable gene. Genetical Research, 23(1), 23-35. doi:10.1017/S0016672300014634.

Mielke, S. J., Szlam, A., Dinan, E., & Boureau, Y.-L. (2022). Reducing Conversational Agents' Overconfidence Through Linguistic Calibration. Transactions of the Association for Computational Linguistics, 10, 857-872. doi:10.1162/tacl_a_00494. arXiv:2012.14983. doi:10.48550/arXiv.2012.14983.

Muennighoff, N., Yang, Z., Shi, W., Li, X. L., Fei-Fei, L., Hajishirzi, H., et al. (2025). s1: Simple test-time scaling. arXiv:2501.19393. doi:10.48550/arXiv.2501.19393.

Nyandwi, J. de D., Mathur, L., Bisk, Y., Hawkins, R., & Neubig, G. (2026). Amplified Does Not Mean Predictive: Reasoning Behaviors in Thinking Models. COLM 2026. arXiv:2608.13760. doi:10.48550/arXiv.2608.13760.

Olmo Team, Ettinger, A., Bertsch, A., Kuehl, B., Graham, D., et al. (2025). Olmo 3. arXiv:2512.13961. doi:10.48550/arXiv.2512.13961.

Price, G. R. (1970). Selection and Covariance. Nature, 227, 520-521. doi:10.1038/227520a0.

Qian, C., Liu, D., Wen, H., Bai, Z., Liu, Y., & Shao, J. (2025). Demystifying Reasoning Dynamics with Mutual Information: Thinking Tokens are Information Peaks in LLM Reasoning. arXiv:2506.02867. doi:10.48550/arXiv.2506.02867.

Salas, M., Hofman, A., & Stricker, B. H. (1999). Confounding by Indication: An Example of Variation in the Use of Epidemiologic Terminology. American Journal of Epidemiology, 149(11), 981-983. doi:10.1093/oxfordjournals.aje.a009758.

Shao, R., Li, S. S., Xin, R., Geng, S., Wang, Y., Oh, S., et al. (2025). Spurious Rewards: Rethinking Training Signals in RLVR. arXiv:2506.10947. doi:10.48550/arXiv.2506.10947.

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., et al. (2024). DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. arXiv:2402.03300. doi:10.48550/arXiv.2402.03300.

Tian, K., Mitchell, E., Zhou, A., Sharma, A., Rafailov, R., Yao, H., et al. (2023). Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback. EMNLP 2023. arXiv:2305.14975. doi:10.48550/arXiv.2305.14975.

Venhoff, C., Arcuschin, I., Torr, P., Conmy, A., & Nanda, N. (2025a). Understanding Reasoning in Thinking Language Models via Steering Vectors. arXiv:2506.18167. doi:10.48550/arXiv.2506.18167.

Venhoff, C., Arcuschin, I., Torr, P., Conmy, A., & Nanda, N. (2025b). Base Models Know How to Reason, Thinking Models Learn When. arXiv:2510.07364. doi:10.48550/arXiv.2510.07364.

Wang, C., Feng, Y., Chen, D., Chu, Z., Krishna, R., & Zhou, T. (2025a). Wait, We Don't Need to "Wait"! Removing Thinking Tokens Improves Reasoning Efficiency. arXiv:2506.08343. doi:10.48550/arXiv.2506.08343.

Wang, S., Yu, L., Gao, C., Zheng, C., Liu, S., Lu, R., et al. (2025b). Beyond the 80/20 Rule: High-Entropy Minority Tokens Drive Effective Reinforcement Learning for LLM Reasoning. NeurIPS 2025. arXiv:2506.01939. doi:10.48550/arXiv.2506.01939.

Wang, Y., Liu, Q., Xu, J., Liang, T., Chen, X., He, Z., et al. (2025c). Thoughts Are All Over the Place: On the Underthinking of o1-Like LLMs. arXiv:2501.18585. doi:10.48550/arXiv.2501.18585.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine Learning, 8, 229-256. doi:10.1007/BF00992696.

Wu, M., Li, M., Yang, J., Jiang, J., Yan, K., Li, Z., et al. (2025a). Aha Moment Revisited: Are VLMs Truly Capable of Self Verification in Inference-time Scaling? arXiv:2506.17417. doi:10.48550/arXiv.2506.17417.

Wu, Y., Wang, Y., Ye, Z., Du, T., Jegelka, S., & Wang, Y. (2025b). When More is Less: Understanding Chain-of-Thought Length in LLMs. arXiv:2502.07266. doi:10.48550/arXiv.2502.07266.

Xia, S., Li, X., Liu, Y., Wu, T., & Liu, P. (2024). Evaluating Mathematical Reasoning Beyond Accuracy. AAAI 2025. arXiv:2404.05692. doi:10.48550/arXiv.2404.05692.

Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J., et al. (2024). Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. ICLR 2024. arXiv:2306.13063. doi:10.48550/arXiv.2306.13063.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., et al. (2025a). Qwen3 Technical Report. arXiv:2505.09388. doi:10.48550/arXiv.2505.09388.

Yang, S., Wu, J., Chen, X., Xiao, Y., Yang, X., Wong, D. F., et al. (2025b). Understanding Aha Moments: from External Observations to Internal Mechanisms. arXiv:2504.02956. doi:10.48550/arXiv.2504.02956.

Yeo, E., Tong, Y., Niu, M., Neubig, G., & Yue, X. (2025). Demystifying Long Chain-of-Thought Reasoning in LLMs. arXiv:2502.03373. doi:10.48550/arXiv.2502.03373.

Yoon, D., Kim, S., Yang, S., Kim, S., Kim, S., Kim, Y., et al. (2025). Reasoning Models Better Express Their Confidence. NeurIPS 2025. arXiv:2505.14489. doi:10.48550/arXiv.2505.14489.

Yu, Q., Zhang, Z., Zhu, R., Yuan, Y., Zuo, X., Yue, Y., et al. (2025). DAPO: An Open-Source LLM Reinforcement Learning System at Scale. arXiv:2503.14476. doi:10.48550/arXiv.2503.14476.

Yue, Y., Chen, Z., Lu, R., Zhao, A., Wang, Z., Yue, Y., Song, S., & Huang, G. (2025). Does Reinforcement Learning Really Incentivize Reasoning Capacity in LLMs Beyond the Base Model? NeurIPS 2025. arXiv:2504.13837. doi:10.48550/arXiv.2504.13837.

Zeng, Q., Xuan, W., Cui, L., & Voigt, R. (2025a). Thinking Out Loud: Do Reasoning Models Know When They're Right? EMNLP 2025. arXiv:2504.06564. doi:10.48550/arXiv.2504.06564.

Zeng, W., Huang, Y., Liu, Q., Liu, W., He, K., Ma, Z., et al. (2025b). SimpleRL-Zoo: Investigating and Taming Zero Reinforcement Learning for Open Base Models in the Wild. COLM 2025. arXiv:2503.18892. doi:10.48550/arXiv.2503.18892.

Zhang, A., Chen, Y., Pan, J., Zhao, C., Panda, A., Li, J., et al. (2025a). Reasoning Models Know When They're Right: Probing Hidden States for Self-Verification. arXiv:2504.05419. doi:10.48550/arXiv.2504.05419.

Zhang, Z., Zheng, C., Wu, Y., Zhang, B., Lin, R., Yu, B., et al. (2025b). The Lessons of Developing Process Reward Models in Mathematical Reasoning. arXiv:2501.07301. doi:10.48550/arXiv.2501.07301.

Zhao, X., Kang, Z., Feng, A., Levine, S., & Song, D. (2025). Learning to Reason without External Rewards. ICLR 2026. arXiv:2505.19590. doi:10.48550/arXiv.2505.19590.

Zheng, C., Zhang, Z., Zhang, B., Lin, R., Lu, K., Yu, B., et al. (2024). ProcessBench: Identifying Process Errors in Mathematical Reasoning. ACL 2025. arXiv:2412.06559. doi:10.48550/arXiv.2412.06559.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., et al. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. NeurIPS 2023 Datasets and Benchmarks. arXiv:2306.05685. doi:10.48550/arXiv.2306.05685.

Zhou, K., Hwang, J. D., Ren, X., & Sap, M. (2024). Relying on the Unreliable: The Impact of Language Models' Reluctance to Express Uncertainty. ACL 2024. arXiv:2401.06730. doi:10.48550/arXiv.2401.06730.

© 2026 Pranay Mahendrakar. Licensed under CC BY 4.0.

DOI: 10.5281/zenodo.23951230