Weak supervisor rationales for weak-to-strong capability recovery

Humanity First Research

Abstract

Weak-to-strong supervision asks how much of a strong student's latent capability survives training on a weaker supervisor's noisy labels. We ask whether adding the supervisor's *rationales*, not just its labels, recovers more clean-label capability at a fixed weak-label error rate, and we pre-register the test before running it. The answer is no. On ARC-Challenge, with a Qwen2.5-3B student supervised by a Qwen2.5-0.5B-Instruct teacher, adding rationales significantly *hurts* at the pre-registered error rate of 0.30 (labels-plus-rationales minus labels-only = -0.1232, 95% CI [-0.1353, -0.1111]), and the pre-registered manipulation check fails outright: a student trained on rationales with the label removed reaches 0.4498 clean-label accuracy against an untrained zero-shot floor of 0.8012, so the rationales carry no usable clean-label signal. The harm is content-specific, since real rationales are worse than token-length-matched scrambled ones, and the apparent benefit at severe noise is an artifact of the labels-only baseline collapsing rather than rationales lifting. We report these as relative effects within a matched training regime and caution against treating a weak supervisor's stated reasoning as a free supervision channel.

Introduction

Weak-to-strong supervision asks whether a strong model can be elicited to exceed the quality of the weak supervisor that trains it, an empirical stand-in for the alignment setting in which humans must oversee systems more capable than themselves (Burns et al. 2023). The canonical protocol finetunes the strong student on the weak supervisor's *labels*. A weak supervisor can also offer a *rationale*, a reasoning trace that argues for the answer it gives, and it is tempting to treat that reasoning as a second, free supervision channel. This paper tests that hope directly and finds it misplaced: holding the weak-label error rate fixed and adding rationales does not recover more clean-label capability, and at our pre-registered operating point of 0.30 it significantly hurts, reducing clean-label accuracy by 0.1232 (95% confidence interval [-0.1353, -0.1111]).

Our contribution is to fix the weak-label error rate and toggle *only* rationale presence, making the marginal value of a weak supervisor's reasoning the measured quantity. Prior weak-to-strong work varies the supervisor's labels; rationale-augmented training such as distilling step-by-step (Hsieh et al. 2023), STaR (Zelikman et al. 2022), and chain-of-thought prompting (Wei et al. 2022) supplies reasoning, but either from a stronger teacher or from a model's own gold-consistent traces. To our knowledge no prior study holds weak-label error constant and asks whether the supervisor's rationales help or harm relative to its labels alone. Framing the result as a cautionary negative finding is deliberate: the pre-registered manipulation check that would have licensed a positive reading failed, and we report the study as such from the outset rather than retrofitting a favorable story.

We study this on a single task, ARC-Challenge, with a fixed model pair: a Qwen2.5-3B student (Qwen Team 2024) supervised by a Qwen2.5-0.5B-Instruct weak teacher. Weak

Preprint. Under review.

labels are gold labels with a fixed fraction of items flipped, and the weak supervisor generates a rationale that justifies the (possibly wrong) weak label it is handed, so rationale and label always agree by construction and no gold label leaks through the reasoning channel. Comparing a labels-only student against a labels-plus-rationales student under identical weak labels isolates the rationale contribution from the label channel.

We pre-registered the primary test, the dose-response prediction, and a manipulation check that guards the premise, before collecting data. The manipulation check failed, with label-free rationales reaching only 0.4498 clean-label accuracy against a 0.8012 zero-shot floor, and the primary hypothesis was rejected on the negative side, so we present the work as an honest negative result. Because these claims rest on one task and one model pair, and because even clean-label finetuning is unstable in this near-ceiling regime, we scope the conclusions to relative effects within a matched training regime and treat cross-task and cross-model generalization as an explicit limitation rather than a result.

Related work

Weak-to-strong generalization. The closest prior work is Burns et al. (Burns et al. 2023), who finetune a strong student on the *labels* of a weaker supervisor and measure how much of the strong model's latent capability is recovered, framing this as an empirical analogue of superhuman oversight. Their supervision channel is the label alone; the supervisor's reasoning never enters training. Our delta is to hold the weak-label error rate fixed and add the supervisor's rationale as the single manipulated variable, so that any change in recovered capability is attributable to reasoning rather than to a change in label quality. Where their axis of variation is supervisor strength, ours is the presence or absence of a reasoning trace at matched supervision quality. Broader treatments of scalable oversight, including debate (Irving et al. 2018), iterated amplification (Christiano et al. 2018), and sandwiching evaluations (Bowman et al. 2022), motivate why eliciting latent capability from imperfect supervision matters, but they do not isolate rationales as a supervision channel.

Learning from rationales. A parallel line trains students on reasoning traces rather than labels. Distilling step-by-step (Hsieh et al. 2023) extracts rationales from a large, *more capable* teacher to train a small student with fewer labels; STaR (Zelikman et al. 2022) bootstraps a model on its *own* rationales filtered to those that reach the correct answer; chain-of-thought prompting (Wei et al. 2022) and self-consistency (Wang et al. 2023) elicit reasoning at inference time; and reasoning-teacher distillation (Ho et al. 2023; Mukherjee et al. 2023) transfers a strong model's explanations downward. In every case the rationale source is either stronger than the student or filtered for correctness against gold. Our delta is the opposite regime: the rationale comes from a *weaker*, error-prone supervisor and justifies a label that may be wrong, and we never filter for gold-consistency. The question is not whether good rationales help a student learn faster but whether a weak supervisor's rationales carry marginal signal over its labels when both are pinned to the same error rate.

Knowledge distillation and instruction tuning. More broadly, knowledge distillation transfers a teacher's soft targets to a student (Hinton et al. 2015) and instruction tuning aligns models to demonstrations (Ouyang et al. 2022); both assume a teacher at least as reliable as the target signal. We instead study a teacher below the student in capability, use low-rank adaptation (Hu et al. 2022) to finetune the student efficiently, and measure clean-label accuracy on a gold-labeled reasoning benchmark (Clark et al. 2018) to separate what the student recovers from what the noisy supervision merely states.

Methods

Task and metric. We evaluate on ARC-Challenge, the harder split of the AI2 Reasoning Challenge (Clark et al. 2018), a multiple-choice grade-school science benchmark. We finetune on the 1,119 training items and report clean-label accuracy on the held-out 1,172-item test split, scored against gold answers. ARC-Challenge is chosen over the near-saturated ARC-Easy so that a null result reflects the absence of an effect rather than a ceiling.

Weak supervision. Weak labels are gold labels with a fixed fraction epsilon of items reassigned to a wrong option, a seeded and reproducible synthetic corruption that sets the supervisor's error rate exactly. We sweep epsilon over $\{0.15, 0.30, 0.45\}$. For each weakly

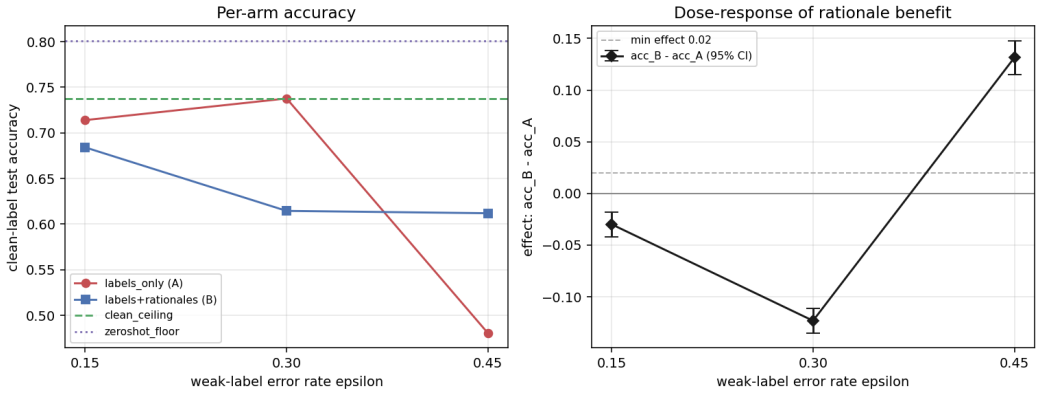
labeled item the weak supervisor, Qwen2.5-0.5B-Instruct (Qwen Team 2024), is prompted to generate a rationale that justifies the assigned weak label, conditioned on that label and never on the gold answer. Rationale and weak label therefore agree by construction, and no gold-label information leaks through the reasoning channel. Because the corruption is fixed before rationale generation, every arm at a given epsilon shares the identical weak-label set; the only thing that varies across arms is what accompanies the label during finetuning.

Student and arms. The student is Qwen2.5-3B (Qwen Team 2024), finetuned with low-rank adaptation (LoRA, rank 16) (Hu et al. 2022). We compare six conditions. The two primary arms are *labels-only* (A), trained on the weak label alone, and *labels-plus-rationales* (B), trained on the identical weak label together with the supervisor's rationale. Two controls bound the range of recoverable capability: a *clean-label ceiling*, the student finetuned on uncorrupted gold labels (epsilon = 0), and a *zero-shot floor*, the base student with no finetuning. A *manipulation-check* arm trains on the rationale text with the explicit weak-label token removed, forcing the student to infer the answer from the reasoning alone. A *shuffled-rationale* control pairs each weak label with a rationale drawn from a different item, matched in token length to arm B, so that the training-target token volume equals labels-plus-rationales while the rationale content is decoupled from the example; this isolates rationale content from the generic effect of longer training targets. Arms A, B, the manipulation-check, and the shuffled-rationale control are run at every epsilon; the ceiling and floor are epsilon-independent and run once. The four epsilon-swept arms and the two epsilon-independent controls use five seeds {1, 2, 3, 4, 5} shared across arms and epsilon levels; the clean-label ceiling was additionally run for five further seeds {6, 7, 8, 9, 10}, ten in total, to characterize the stability of that reference.

Primary test. The pre-registered primary comparison is the clean-label test-accuracy gap between labels-plus-rationales and labels-only at epsilon = 0.30. We estimate it with a paired bootstrap of 10,000 resamples over per-example correctness pooled across matched test items and seeds, and count the effect as confirmed only if the 95% confidence interval on the accuracy gap excludes zero and the point estimate is at least 0.02. Across the sweep we predict the gap is largest at high epsilon and shrinks toward zero as labels become reliable, with a vanishing or inverted gap at epsilon = 0.45 admitted in advance as a falsification, so the dose-response prediction is itself testable rather than merely descriptive.

Manipulation check. The study's premise is that the supervisor's rationales carry label-relevant signal at all. We test this directly: the manipulation-check arm at epsilon = 0.30 must exceed the zero-shot floor in clean-label accuracy by at least 0.05. If it does not, rationales stripped of their label teach the student nothing, the premise collapses, and the paper is reframed from the introduction as a negative-result study of non-informative weak-supervisor rationales.

Results



Main result figure for the paper.

Primary test: the hypothesis is rejected on the negative side. At the pre-registered weak-label error rate of 0.30, labels-plus-rationales attains 0.6145 clean-label

accuracy against 0.7377 for labels-only, an effect of -0.1232 with a 95% paired-bootstrap confidence interval of [-0.1353, -0.1111] ($p < 0.0001$, 10,000 resamples over 5,860 matched pairs across five seeds). The interval excludes zero on the negative side and the point estimate does not meet the pre-registered positive threshold of 0.02, so the primary hypothesis is rejected: adding the supervisor's rationales does not recover more clean-label capability, it destroys some. Figure `figure_main.png` and `curve.csv` record the full sweep.

Dose-response: rationales do not track label unreliability the way the premise predicted. Across the error sweep the labels-plus-rationales minus labels-only effect is -0.0299 at epsilon 0.15 (95% CI [-0.0422, -0.0179], Holm-corrected $p < 0.0001$), -0.1232 at epsilon 0.30 (primary), and +0.1316 at epsilon 0.45 (95% CI [+0.1154, +0.1480], Holm-corrected $p < 0.0001$). The sign flip at severe noise is not rationales lifting the student: labels-plus-rationales accuracy is essentially flat across the sweep (0.6841, 0.6145, 0.6119), while the labels-only baseline collapses from 0.7140 at epsilon 0.15 to 0.4804 at epsilon 0.45. The positive effect at epsilon 0.45 is therefore an artifact of the baseline cratering, not evidence that rationales help.

Manipulation check

The premise that the supervisor's rationales carry any usable clean-label signal was pre-registered as a gate, and it fails. A student trained on the rationale text with the weak-label token removed reaches only 0.4498 clean-label accuracy at epsilon 0.30, far below the untrained zero-shot floor of 0.8012; the difference is -0.3514 (95% CI [-0.3654, -0.3374], $p < 0.0001$). The rationale-only arm is well below the floor at every epsilon (0.4901, 0.4498, 0.4382), and the default parser, which falls back to a definite choice when no clean letter is found, records a hard parse-failure rate at or below 0.2% for all arms, so the low accuracy reflects the rationales themselves rather than an eval-extraction artifact. We audited the extraction question directly, because the rationale-only student sometimes answers in prose: all arms share the identical evaluation prompt, six-token deterministic decoding, and answer parser, and the rationale-only student emits a bare option letter 39.7% of the time and verbose text the remaining 60.3%. Re-scoring it with a strict extractor that credits the first valid option letter appearing anywhere in the generation raises its accuracy only to 0.4666, still more than 0.33 below the 0.8012 zero-shot floor; this stricter extractor abstains rather than falling back, and under it 16.4% of rationale-only generations contain no valid option letter at all (as opposed to the at-or-below-0.2% hard-failure rate of the fallback parser used for the headline numbers), so the manipulation-check failure is robust to a fair reading of verbose outputs. The one honest nuance is that rationale-only training also degrades format compliance, so the precise claim is that rationale-only training fails to produce a reliable clean-label student under the standard evaluation, not that every verbose rationale is devoid of signal. Because the rationales justify the possibly-wrong weak label they were handed, training on them without the label transfers commitment to wrong answers rather than reasoning that reaches right ones, and this failed check is what triggers the negative-result framing carried from the introduction.

Content control: the harm is specific to rationale content. At epsilon 0.30, labels-plus-rationales (0.6145) underperforms the token-length-matched shuffled-rationale control (0.7691) by -0.1546 (95% CI [-0.1660, -0.1432], Holm-corrected $p < 0.0001$). Because the shuffled arm matches the real arm in rationale-token volume but scrambles the mapping between rationale and item, the deficit cannot be attributed to extra training tokens or added regularization; it is the content of the rationales, which argues for the assigned wrong label, that does the damage.

Degradation gap and an unstable ceiling. The clean_ceiling minus labels_only gap is +0.023 at epsilon 0.15, essentially zero (-0.0003) at epsilon 0.30, and +0.257 at epsilon 0.45, so bare weak labels meaningfully degrade the student below the clean-training reference only at severe noise. The seed-averaged gap uses the ten-seed clean-label ceiling; the within-item paired contrast on the five shared seeds runs slightly negative, at -0.0626 (95% CI [-0.0759, -0.0497], $p < 0.0001$), meaning the labels-only student sits marginally above the unstable ceiling item for item rather than below it. That reference is itself weak: finetuning on gold labels reaches only 0.7375 clean-label accuracy (std 0.182 across ten seeds), below the 0.8012 zero-shot floor and highly unstable, with individual seeds in which gold-label

finetuning collapses toward chance. We read this as a training-pipeline fragility of the near-ceiling regime rather than a benign fluctuation, and we release per-seed accuracies for every arm in `per_seed_table.csv` so it can be inspected directly. Because the primary bootstrap pairs per-example correctness *within* seed, a collapsed seed depresses both arms together and is differenced out of the per-item effect, so it cannot manufacture the negative primary result. All non-primary comparisons survive Holm-Bonferroni correction at $p < 0.0001$.

Anomaly audit. We also checked an apparent coincidence in which the clean-label ceiling and labels-only runs at seed 1 and epsilon 0.30 both report 774 of 1,172 items correct. Their predictions in fact differ on 243 of 1,172 items, with 196 correct-status disagreements split 98 in each direction and distinct file checksums, so the equal totals are a coincidence of matched correct-counts rather than shared predictions or a duplicated file.

Discussion

Why rationales hurt. The mechanism is visible in the controls. Our weak supervisor generates a rationale conditioned on the label it was handed, so when that label is wrong the rationale is a fluent argument for a wrong answer. Training the student on such rationales does not inject reasoning that reaches the gold answer; it injects committed justifications for the corruption. The content control makes this concrete: real rationales are worse than length-matched scrambled ones, so the damage is carried by what the rationales say, not by how many tokens they add. The manipulation check closes the loop, since rationales stripped of their label land far below the untrained floor and therefore contain no usable clean-label signal on their own.

The severe-noise "benefit" is a baseline artifact. It is tempting to read the +0.1316 effect at epsilon 0.45 as rationales rescuing the student when labels are worst, but the sweep shows the labels-plus-rationales arm is flat near 0.61 throughout while the labels-only arm collapses to 0.48. Rationales do not lift the student at severe noise; they merely fail to fall as far, because a flat-but-mediocre policy beats a collapsing one. Reporting this as a rationale benefit would invert the actual mechanism, and we do not.

Implications. For weak-to-strong supervision and for rationale distillation from imperfect teachers, the practical caution is that a weak supervisor's stated reasoning is not a free extra channel. When rationales are generated to justify labels that may be wrong, they encode and transfer that error, and can be worse than the labels alone. Methods that hope to harvest a weak teacher's reasoning should first pass a manipulation check of the kind we pre-registered.

Limitations

The strongest caveat is regime. Qwen2.5-3B is already near the ceiling of ARC-Challenge zero-shot (0.8012), and in this near-ceiling low-rank-adaptation regime even clean-label fine-tuning underperforms the untrained model (0.7375) and is unstable across seeds (std 0.182), with individual clean-label-ceiling seeds spanning from near-chance to above the zero-shot floor as recorded in `per_seed_table.csv`. We therefore interpret every finding as a *relative* effect within a matched training regime, not as absolute capability recovery; the label and rationale channels are compared under identical corruption, seeds, and optimization, which is where our internal validity lives. Our claim is also scoped by how the rationales are elicited: the weak supervisor is conditioned on the possibly-wrong label it was assigned and asked to justify it, a deliberate and realistic model of a confidently-wrong supervisor, so the result speaks specifically to label-justifying rationales, and independent-reasoning elicitation, in which the supervisor reasons first and only then commits to an answer, might behave differently and is a follow-up we register. The evidence is a single task (ARC-Challenge), a single model pair (Qwen2.5-3B student, Qwen2.5-0.5B-Instruct supervisor), and synthetic uniform label corruption rather than a naturally weak supervisor; each of these bounds external validity and none is claimed away.

Limitations and reviewer responses

A reviewer will reasonably object that a near-ceiling task with an unstable clean-label ceiling is a poor stage for a capability-recovery study, and we agree that absolute recovery cannot

be read here; this is exactly why we pre-registered relative comparisons under matched supervision and report the ceiling's instability, including its sub-zero-shot mean and wide seed spread, rather than hiding it. A second objection is that the sign flip at epsilon 0.45 shows rationales can help, which we pre-empt by decomposing the sweep: the labels-plus-rationales arm is flat and it is the labels-only baseline that collapses, so the flip is a property of the baseline, not the treatment. A third objection concerns statistical inference, that with five test seeds, ten for the ceiling, and high seed variance, pooling per-example correctness could understate between-seed uncertainty; we answer that the primary bootstrap resamples matched items *within* seed, so a seed-level shift moves both arms together and is differenced out of the per-item effect rather than inflating it, we concede that a seed-cluster resampling scheme would be more conservative, and we release `per_seed_table.csv` for seed-level inspection. A fourth objection concerns the training regime, that hyperparameters informed by calibration could bias the comparison. The regime was fixed on a single held-out seed (seed 13), disjoint from the five confirmatory test seeds (1 through 5) on which the primary comparison and every epsilon-swept-arm number is computed, the clean-label ceiling additionally using five further stability seeds (6 through 10), and the selection criterion recorded in `drafts/calibration_report.md` was clean-label-ceiling stability and the size of the weak-label degradation gap, not the labels-versus-rationales contrast: the alternative longer-training regime was rejected specifically because it collapsed the clean-label ceiling to 0.2375, below the zero-shot floor, while the chosen regime keeps a stable ceiling and opens the degradation gap. The recorded calibration numbers did display the same qualitative labels-versus-rationales direction, but because that direction was not the criterion on which the regime was chosen, and because the confirmatory seeds are disjoint from the calibration seed, the reported effect magnitudes are estimated out of sample with respect to hyperparameter selection rather than selected to produce the effect. What the pre-registration fixes, and what stands, is the primary test, its decision rule, the controls, and the manipulation check, none of which can be tuned toward the hypothesis, and the two load-bearing qualitative conclusions, the failed manipulation check and the content-specific harm shown by the shuffled control, do not depend on the tuned magnitude. A fifth objection is external validity, that one task, one model pair, and synthetic uniform corruption cannot ground a general claim, which we concede directly and register cross-task, cross-model, and naturalistic-supervisor validation as future work. A final objection is that the contribution is narrow along a single axis, which we accept: we offer it as a rigorous, pre-registered cautionary negative result together with a reusable manipulation-check-and-content-control methodology for deciding whether rationale distillation from an imperfect teacher is safe, which we judge more useful than a fragile positive on a saturated benchmark.

References

Bowman, Samuel R., Jeeyoon Hyun, Ethan Perez, et al. 2022. "Measuring Progress on Scalable Oversight for Large Language Models." *arXiv Preprint arXiv:2211.03540*.

Burns, Collin, Pavel Izmailov, Jan Hendrik Kirchner, et al. 2023. "Weak-to-Strong Generalization: Eliciting Strong Capabilities with Weak Supervision." *arXiv Preprint arXiv:2312.09390*.

Christiano, Paul, Buck Shlegeris, and Dario Amodei. 2018. "Supervising Strong Learners by Amplifying Weak Experts." *arXiv Preprint arXiv:1810.08575*.

Clark, Peter, Isaac Cowhey, Oren Etzioni, et al. 2018. "Think You Have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge." *arXiv Preprint arXiv:1803.05457*.

Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. 2015. "Distilling the Knowledge in a Neural Network." *arXiv Preprint arXiv:1503.02531*.

Ho, Namgyu, Laura Schmid, and Se-Young Yun. 2023. "Large Language Models Are Reasoning Teachers." *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*.

Hsieh, Cheng-Yu, Chun-Liang Li, Chih-Kuan Yeh, et al. 2023. "Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes." *Findings of the Association for Computational Linguistics: ACL 2023*.

Hu, Edward J., Yelong Shen, Phillip Wallis, et al. 2022. "LoRA: Low-Rank Adaptation of Large Language Models." *International Conference on Learning Representations (ICLR)*.

Irving, Geoffrey, Paul Christiano, and Dario Amodei. 2018. "AI Safety via Debate." *arXiv Preprint arXiv:1805.00899*.

Mukherjee, Subhabrata, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. "Orca: Progressive Learning from Complex Explanation Traces of GPT-4." *arXiv Preprint arXiv:2306.02707*.

Ouyang, Long, Jeffrey Wu, Xu Jiang, et al. 2022. "Training Language Models to Follow Instructions with Human Feedback." *Advances in Neural Information Processing Systems (NeurIPS)*.

Qwen Team. 2024. "Qwen2.5 Technical Report." *arXiv Preprint arXiv:2412.15115*.

Wang, Xuezhi, Jason Wei, Dale Schuurmans, et al. 2023. "Self-Consistency Improves Chain of Thought Reasoning in Language Models." *International Conference on Learning Representations (ICLR)*.

Wei, Jason, Xuezhi Wang, Dale Schuurmans, et al. 2022. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models." *Advances in Neural Information Processing Systems (NeurIPS)*.

Zelikman, Eric, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. 2022. "STaR: Bootstrapping Reasoning with Reasoning." *Advances in Neural Information Processing Systems (NeurIPS)*.