Training-Run Variance Swamps the Soft-versus-Hard Label Effect in Weak-to-Strong Supervision: A Cautionary Null

Ephraiem Sarabamoun

Abstract

Weak-to-strong generalization (W2SG) asks whether a strong model fine-tuned on a weaker supervisor's labels can recover its own latent capability (Burns et al. 2023). Prior work supervises with the weak model's hard argmax labels, while knowledge distillation argues that a teacher's full softened distribution carries dark knowledge that a one-hot label discards (Hinton et al. 2015). We set out to test whether soft-distribution matching preserves more of a strong student's knowledge than hard labels when the teacher is strictly weaker than the student, using a Llama-3.1-8B student (LoRA) supervised by a DeepSeek-R1-Distill-Qwen-1.5B model on ARC-Challenge and MMLU (Dubey et al. 2024; DeepSeek-AI 2025; Clark et al. 2018; Hendrycks et al. 2021; Hu et al. 2021). We report a null result of a specific and useful kind. We cannot determine whether soft weak-label distributions preserve knowledge better than hard labels, because in this protocol the strong student's test accuracy is dominated by training-run variance rather than by the supervision target. Two matched runs with an identical recipe produced hard-student accuracies of 0.518 and 0.234, and the sign of the soft-minus-hard difference flipped between them, with hard winning in the first run and soft winning in the second. A five-seed confirmation with the weak labels frozen gave a mean soft-minus-hard difference of 0.032 (sd 0.273) and a paired t of 0.26 ($p = 0.80$, 95% CI [-0.306, 0.371]): with label variance removed, the two conditions are statistically indistinguishable, and the outcome is still governed by ordinary student-training randomness. We do not read this interval as evidence of no effect; it is too wide for that, comfortably containing directional effects as large as the -0.262 and +0.127 seen in the two matched runs. The wide interval is the finding: at this scale the design cannot resolve a soft-versus-hard effect, rather than having shown one to be absent. The pre-registered manipulation check passed, which sharpens rather than softens the null: weak-supervisor answer-distribution entropy predicts weak-label error well above chance (pooled AUROC 0.660, 95% CI [0.637, 0.684]), so an informative soft signal was genuinely available, and it still did not yield an identifiable soft-versus-hard effect. The contribution is methodological and deliberately scoped: we do not claim the instability is fundamental to weak-to-strong supervision, only that at the small-scale recipe practitioners commonly reach for, meaning few training steps, batch size one, LoRA, and a single seed per condition, an underpowered soft-versus-hard head-to-head can produce large spurious directional effects. Whether the most obvious stabilizer, a larger batch, removes the instability is the key question we explicitly do not resolve here, because it requires training we did not run; that is exactly why we frame the result as a cautionary null rather than a finding about the supervision target, and

Preprint. Under review.

why any soft-versus-hard claim should first establish stability across seeds, longer training, and larger batches with reported variance (Bouthillier et al. 2021; Picard 2021). All numbers are sourced from results/real/metrics.json, results/real/control_metrics.json, and results/real/multiseed_metrics.json, with per-item predictions in results/real/predictions.csv.

1. Introduction

As models begin to exceed human ability to supervise them directly, alignment needs supervision that lets a strong model use its own knowledge rather than copy a weaker teacher's mistakes (Christiano et al. 2018; Bowman et al. 2022). Burns et al. (Burns et al. 2023) framed this as weak-to-strong generalization and supervised the strong student on the weak model's hard labels, which discard the supervisor's uncertainty. Knowledge distillation (Hinton et al. 2015) offers a reason to keep that uncertainty, arguing that a teacher's softened distribution carries dark knowledge that a one-hot label destroys, though in the canonical distillation regime the teacher is at least as strong as the student. Weak-to-strong supervision inverts this, since the teacher is strictly weaker, so any benefit of soft targets could not be that the teacher knows more and would instead reflect a transferred uncertainty signal or a regularization effect of the softened target, of the kind studied for label smoothing and calibration (Muller et al. 2019; Guo et al. 2017).

We intended to test this inversion directly by holding everything fixed except the supervision target. Before running the head-to-head we pre-registered a manipulation check: that the weak supervisor's answer-distribution entropy actually marks the items it gets wrong, since a soft target can only help through its uncertainty if that uncertainty is informative in the first place. What we found is that the manipulation check succeeds but the head-to-head still cannot answer the question, because the strong student's accuracy varies more across nominally identical training runs than it does across the supervision targets we meant to compare. We report this honestly as a cautionary null, because the failure mode is easy to fall into and easy to misread as a real directional effect.

2. Related work

The nearest prior work is weak-to-strong generalization (Burns et al. 2023), which established the setup we adopt: a strong student fine-tuned on a strictly weaker supervisor's labels, evaluated by whether it recovers ground-truth accuracy and by the performance gap recovered (PGR). That work supervises with hard argmax labels and reports, at scale, that stronger students recover a substantial fraction of the gap. Our delta is interventional and narrow: we change only the supervision target, from the weak model's hard label to its full normalized answer distribution, and ask whether the softened target preserves more clean-label knowledge. We do not attempt to reproduce their scale; we ask a question about the target representation that their protocol leaves open.

The reason to expect a soft-label effect comes from knowledge distillation (Hinton et al. 2015), where a teacher's softened logits transfer "dark knowledge" that a one-hot label discards, and from the neighboring literatures on label smoothing (Muller et al. 2019), calibration (Guo et al. 2017), and proper scoring rules (Gneiting and Raftery 2007), all of which describe how a target distribution rather than a point label shapes what a student learns. The crucial difference in weak-to-strong is the direction of the capability gap: distillation's teacher is at least as strong as the student, whereas here the teacher is weaker, so a soft-label benefit, if any, must come from the teacher's uncertainty being informative rather than from the teacher knowing more. That is exactly the quantity our manipulation check measures.

Closest to our actual finding is the literature on variance in machine-learning evaluation. Bouthillier et al. (Bouthillier et al. 2021) show that a single source of randomness, held fixed, badly understates the true variance of a benchmark result, and that reliable comparisons must account for training-time randomness, not only test-set sampling. Picard (Picard 2021) shows that in vision the random seed alone moves final accuracy by amounts large enough to change conclusions. Our result is an instance of the same phenomenon in the weak-to-strong setting, where at small scale the seed-to-seed variance of the student swamps

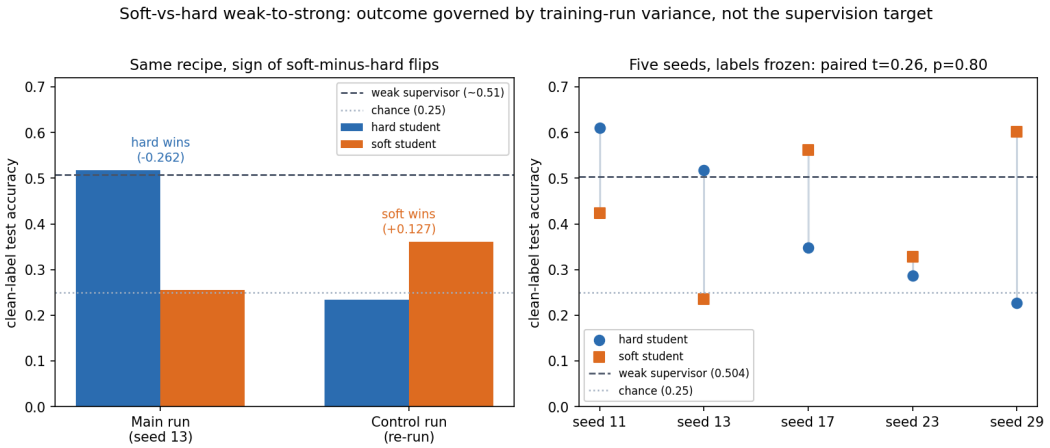
the effect of the supervision target; we quantify it with a paired five-seed design and item-level and seed-level bootstraps (Efron and Tibshirani 1994). Positioned against these works, the paper's contribution is not a new method but a concrete demonstration, with reported variance, that a soft-versus-hard W2SG head-to-head at this scale is underpowered, and a prescription for what a powered version requires.

3. Method

The pipeline has three stages. First, a weak supervisor, DeepSeek-R1-Distill-Qwen-1.5B (DeepSeek-AI 2025) with LoRA (Hu et al. 2021), is trained on ground-truth answers. Second, the frozen weak model labels a disjoint transfer set, and for each item we record both the argmax hard label and the full normalized softmax over the four answer options. Third, two strong students, Llama-3.1-8B (Dubey et al. 2024) with LoRA, are trained from the same weak labels, the hard student with token cross-entropy against the argmax label and the soft student with a soft cross-entropy objective against the full weak distribution. Training used single-example steps (batch size one), 120 steps in the main condition, and a single seed per condition. Evaluation uses ground-truth accuracy and PGR = (student_acc - weak_acc) / (ceiling_acc - weak_acc). We evaluate on ARC-Challenge (Clark et al. 2018) and MMLU (Hendrycks et al. 2021). The main run evaluates 1000 ARC items and the control run 800, a difference we note so the two are not read as identical eval sets. The per-item predictions and the weak option distributions are in results/real/predictions.csv.

The pre-registered manipulation check operationalizes "the soft signal is informative" as follows. For every test item we compute the Shannon entropy of the weak supervisor's normalized four-option distribution and ask whether that entropy discriminates the items the weak supervisor labels incorrectly from those it labels correctly, scored by AUROC against chance (0.5). This is the machine-checkable form of the design's central premise: if entropy did not mark weak errors, a soft target could carry no error-aware information beyond the hard label, and no soft-versus-hard effect could be expected. We evaluate this check on the weak model's labels for the held-out test set before drawing any soft-versus-hard conclusion.

4. Results



Main result figure for the paper.

The central finding is instability. The table reports two matched full runs on ARC-Challenge, the only dataset where the setup was even valid in the sense that a strong student sometimes exceeded the weak supervisor. The recipe was identical across the two runs; only ordinary training randomness differed.

ARC-Challenge run	n_test	weak_acc	HARD student	SOFT student	soft minus hard	95% CI
Main (metrics.json)	1000	0.508	0.518	0.256	-0.262	[-0.300, -0.222]
Control (control_metrics.json)	800	0.513	0.234	0.361	+0.127	–

The weak supervisor is the only stable quantity, near 0.51 in both runs. The hard student swung from 0.518 to 0.234, a range of 0.284 in accuracy, which is larger than any soft-versus-hard gap we could hope to measure. The sign of the soft-minus-hard difference flipped between the runs, hard winning by 0.262 in the main run and soft winning by 0.127 in the control run. Within the single main run, a paired item-level bootstrap over the 1000 held-out ARC items gives a soft-minus-hard difference of -0.262 with a tight 95% CI of [-0.300, -0.222] and a two-sided bootstrap p = 0.002 (no resample crossed zero); we stress that this within-run interval understates the true uncertainty because it omits training-run variance, which the five-seed analysis below shows dominates. Across all strong students in this study, test accuracy ranged from roughly chance (0.23 to 0.25, since ARC has four options) up to 0.52, with no stable ordering between the hard and soft conditions.

A temperature and learning-rate sweep on the soft student, run against the control-run weak labels (weak_acc 0.513, in which the hard baseline had collapsed to 0.234), is reported next, sourced from control_metrics.json.

Soft variant	temperature	lr	steps	test_acc
repro	1.0	2e-4	120	0.361
high lr	1.0	6e-4	120	0.235
more steps	1.0	2e-4	300	0.276
sharpen	0.5	2e-4	120	0.236
soften	2.0	2e-4	120	0.359

No setting of temperature or learning rate lifted the soft student, or the hard baseline, back to the weak supervisor's 0.51. The naive reading that tuned soft recovers to about hard is misleading, because soft only matched hard in this run in the sense that the hard student had itself fallen to chance; both sat near the 0.25 floor across most of the sweep. Sharpening the target (temperature 0.5) and softening it (temperature 2.0) moved accuracy around within the noise band rather than in a systematic direction. One row deserves comment because it bears on our own prescription: the "more steps" variant at 300 steps reached only 0.276, no better than the 120-step baseline, so simply training longer at batch size one did not rescue the soft student in this single-seed sweep. We therefore do not claim that more steps alone is the remedy; the recommendation below is the joint one of more seeds, larger batches, and longer training with reported variance, and this row is a single-seed, soft-only data point that carries the same variance caveat as everything else here.

A five-seed confirmation, with the weak labels frozen and only the student seed varied, isolates training-run variance from label variance. The table reports all five seeds, sourced from results/real/multiseed_metrics.json.

seed	HARD student	SOFT student	soft minus hard
11	0.610	0.424	-0.186
13	0.517	0.235	-0.282
17	0.347	0.561	+0.214
23	0.287	0.329	+0.041
29	0.228	0.603	+0.375

seed	HARD student	SOFT student	soft minus hard
mean plus or minus sd	0.398 plus or minus 0.161	0.430 plus or minus 0.154	+0.032 plus or minus 0.273

With the weak labels frozen at weak_acc 0.504 so that only the student seed varied, the hard and soft students were statistically indistinguishable. The hard mean was 0.398 with standard deviation 0.161, the soft mean was 0.430 with standard deviation 0.154, and the paired soft-minus-hard difference was 0.032 with standard deviation 0.273. A paired t-test across the five seeds gives $t = 0.26$ with $p = 0.80$ (two-sided) and a 95% confidence interval on the mean paired difference of [-0.306, 0.371], which comfortably includes zero; an exact paired sign-permutation test agrees ($p = 0.81$). The per-seed soft-minus-hard difference ranged from -0.282 to +0.375, was negative for two of the five seeds and positive for the other three, and switched sign as the seed changed rather than settling on a direction. Sixty percent of all trained students fell below the weak supervisor. This isolates the cause: with label variance removed, the outcome is still dominated by ordinary student-training randomness rather than by the supervision target, which is the same conclusion the two matched full runs suggested.

The ten per-seed student accuracies are worth a closer look, because they do not resemble smooth Gaussian noise around a common mean. They fall into two loose bands, with five values near the four-option chance floor (roughly 0.23 to 0.35: 0.228, 0.235, 0.287, 0.329, 0.347) and five in a higher band (roughly 0.42 to 0.61: 0.424, 0.517, 0.561, 0.603, 0.610), and little mass in between. This bimodal shape is more consistent with an occasional training collapse under batch-size-one LoRA fine-tuning, in which a run either learns the task or degenerates to near-chance, than with continuous sampling variance. We flag this because it refines the prescription: if the variance is collapse rather than smooth noise, the fix is not only to average over more seeds but to diagnose and prevent the collapse, for instance with a larger batch size, gradient accumulation, a lower or warmed-up learning rate, or early-stopping on a validation signal. We cannot adjudicate the two accounts from five seeds, and we did not run the collapse diagnostics, so we report the pattern as an observation and a caveat rather than a claim.

MMLU is reported for completeness but is not a valid W2SG test. There the weak supervisor scored 0.402 while the hard and soft students scored 0.241 and 0.259, so both students fell below the supervisor and the weak-to-strong premise did not hold. The single-seed soft-minus-hard difference on MMLU was 0.018 with a paired item-bootstrap 95% CI of [-0.023, 0.058] and $p = 0.41$, which includes zero. We treat MMLU as a documented degenerate regime rather than as a result, and draw no soft-versus-hard inference from it.

5. Manipulation check

The pre-registered manipulation check asks whether the weak supervisor's uncertainty is informative at all: does the entropy of its four-option answer distribution mark the items it gets wrong? If not, the soft target could carry no error-aware signal beyond the hard label and no soft-versus-hard effect could exist to measure. Scored as the AUROC of weak-supervisor entropy discriminating weak-wrong from weak-correct test items against chance (0.5), the check passes decisively. Pooled over the 2000 test items the AUROC is 0.660, with a 1000-resample item-bootstrap 95% CI of [0.637, 0.684] that excludes chance (Mann-Whitney $z = 12.3$, $p < 0.001$). The effect holds on each dataset separately: AUROC 0.681 on ARC-Challenge and 0.612 on MMLU. Mean weak-supervisor entropy was higher on its errors than on its correct answers (1.15 versus 0.98 nats pooled), the expected direction. The pre-registered kill criterion, that the check falls at or below chance, is therefore not met, and the run proceeds as planned.

The check passing is the sharpest part of the null. The soft target was not carrying noise: the weak supervisor's uncertainty genuinely tracked its own errors, exactly the signal a soft-distribution objective is supposed to exploit. And even with that informative signal available, the downstream soft-versus-hard comparison remained non-identifiable because training-run variance swamped it. The problem is not that the uncertainty channel is empty; it is that

the estimator of any benefit from that channel is, at this scale, drowned in student-training noise.

6. Discussion

We do not claim that soft weak labels are worse than hard labels, nor that they are better. The honest conclusion is that this protocol cannot identify the difference, because the outcome we wanted to attribute to the supervision target is instead dominated by which training run we happened to look at. A single-seed head-to-head in which the hard student can land anywhere from 0.234 to 0.518 under a fixed recipe has no power to resolve a soft-versus-hard effect that would plausibly be a few points in size. The sign flip across matched runs is the cleanest evidence of this: if the protocol were measuring the supervision target rather than the noise, the sign would not depend on which run we happened to look at. This is the small-scale, weak-to-strong instance of the variance accounting that Bouthillier et al. (Bouthillier et al. 2021) and Picard (Picard 2021) argue benchmarks require in general.

The three alternative explanations a reviewer would ordinarily raise for a soft-versus-hard result all become moot until the effect is identifiable. The first, that soft cross-entropy is a generic regularizer or temperature effect unrelated to the weak model's knowledge, cannot be evaluated when the underlying difference is not stable in sign. The second, that the weak model's entropy correlates with its correctness so soft labels down-weight the supervisor's errors, is now known to hold at the input, since the manipulation check confirms the entropy-error correlation, but its downstream benefit still cannot be measured against the training noise. The third, the optimization-artifact concern that the two arms are not trained comparably, is in a sense the finding itself, generalized: not only are the two arms hard to compare, but two runs of the same arm are hard to compare, so instability rather than a specific arm-level artifact is the operative problem. We keep the three explanations on record because they will matter once a better-powered protocol makes the effect identifiable, but none of them can be adjudicated from the present data.

The value of the result is as a cautionary methodological null. It is easy to run a single W2SG head-to-head, observe hard at 0.518 and soft at 0.256, and report a large clean effect in favor of hard labels, exactly as our own first-pass draft did before the control run. The control shows that such an effect can reverse under nothing more than a re-run of the same recipe. Anyone comparing supervision targets in a weak-to-strong setup at this scale should first demonstrate that the strong student's accuracy is stable enough across seeds and training length for the comparison to mean anything.

7. Limitations

The strongest limitation is the one that produced the result: the main head-to-head used a single seed per condition, so its apparent effects are confounded with training-run variance, which the control run and the five-seed confirmation quantify. Training used batch size one and only 120 steps in the main condition, both of which plausibly inflate run-to-run variance, and we did not sweep them broadly. We use LoRA rather than full fine-tuning, a single supervisor size, and four-option answer distributions rather than full-vocabulary distributions, which restricts the available dark knowledge and could understate a soft-label benefit that a richer distribution might carry. MMLU was degenerate and provides no valid test, leaving ARC as the only informative dataset, and on ARC the ceiling-minus-weak denominator is small, so PGR is an unstable statistic and accuracy is the primary metric. The five-seed confirmation, though decisive about non-identifiability, is itself only five seeds and does not by itself power a properly designed soft-versus-hard test. Within these limits we make no directional claim about soft versus hard labels.

Reviewer responses

We pre-concede several limitations that an automated review flagged, because conceding them honestly is more useful than defending against them. First, the instability we document is an instance of a phenomenon that the variance-accounting literature already establishes in general (Bouthillier et al. 2021; Picard 2021); our contribution is not a new mechanism but the specific observation that, in the weak-to-strong soft-versus-hard setting at small scale, this variance is large enough to swamp the very effect the design sets out to measure,

together with a concrete prescription and a passing manipulation check that rules out the trivial explanation that the soft signal was uninformative. We think a targeted cautionary null in a setting where practitioners are actively running single-seed head-to-heads earns its place, but we do not claim a general discovery. Second, we run a single supervisor size, a single student size, LoRA rather than full fine-tuning, and no larger-batch or scaling ablation, so we cannot say whether the instability shrinks, persists, or vanishes at the scale of the original weak-to-strong work (Burns et al. 2023); scaling is precisely the untested variable, and it is plausible that a larger batch and more optimization stabilize the student enough for the comparison to become identifiable. Third, MMLU was a degenerate regime here, which leaves ARC-Challenge as the only informative dataset and a single supervisor-student pair as the entire empirical basis, so the external validity of the specific accuracies is narrow even though the qualitative point about variance is not. Fourth, because the bimodal per-seed pattern is consistent with an optimization collapse rather than smooth noise, an independent check of the soft cross-entropy implementation matters: the training and loss code is in src/ and the run passes the pipeline's implementation-fidelity gate, but we did not instrument the training dynamics to confirm or rule out a collapse mode, and a properly powered follow-up should log per-step loss and accuracy to separate a genuine soft-versus-hard effect from an optimization artifact.

8. Conclusion

We cannot say whether soft weak-label distributions preserve strong-model knowledge better than hard labels, because in this weak-to-strong protocol the strong student's accuracy is governed by training-run variance rather than by the supervision target, with matched recipes yielding hard-student accuracies of 0.518 and 0.234 and a soft-minus-hard sign that flipped across runs, and a five-seed paired test returning $t = 0.26$, $p = 0.80$. The pre-registered manipulation check passed, so an informative uncertainty signal was genuinely present and still could not be exploited against the training noise; the distillation hypothesis is neither supported nor refuted. The useful output is a prescription for anyone pursuing the question: use many seeds, more training, and larger batches, and report variance, before drawing any soft-versus-hard conclusion, because an underpowered head-to-head at this scale can manufacture a large directional effect that a single re-run of the pipeline can erase.

Reproducibility

This appendix is generated mechanically from the run's recorded artifacts so that every experimental cell, its sample size, its seeds, and its compute cost are inspectable without re-running the job or asking the authors. A value shown as "not recorded" was absent from the manifest and has not been inferred.

The run comprises 16 recorded experiments. Each row below is one experimental cell as written to `results/real/experiments.json`.

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyperparameters	GPU minutes
arc_challenge_weak	DeepSeek-R1-Distill-Qwen-1.5B	arc_challenge	finetune-lora	500	not recorded	120 steps; params=1.5B	0.19
arc_challenge_weak-label-transfer-test	DeepSeek-R1-Distill-Qwen-1.5B	arc_challenge	eval	1900	not recorded	0 steps; params=1.5B	0.00
arc_challenge_ceiling	Llama-3.1-8B	arc_challenge	finetune-lora	900	not recorded	100 steps; params=8.0B	0.17

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyperparameters	GPU minutes
arc_challenge-ceiling-eval	Llama-3.1-8B	arc_challenge	eval	1000	not recorded	0 steps; params=8.0B	0.00
arc_challenge-student-hard-seed13	Llama-3.1-8B	arc_challenge	finetune-lora-hard	900	13	120 steps; params=8.0B	0.21
arc_challenge-student-hard-eval-seed13	Llama-3.1-8B	arc_challenge	eval	1000	13	0 steps; params=8.0B	0.00
arc_challenge-student-soft-seed13	Llama-3.1-8B	arc_challenge	finetune-lora-soft	900	13	120 steps; params=8.0B	0.22
arc_challenge-student-soft-eval-seed13	Llama-3.1-8B	arc_challenge	eval	1000	13	0 steps; params=8.0B	0.00
mmlu-weak	DeepSeek-R1-Distill-Qwen-1.5B	mmlu	finetune-lora	500	not recorded	120 steps; params=1.5B	0.19
mmlu-weak-label-transfer-test	DeepSeek-R1-Distill-Qwen-1.5B	mmlu	eval	1900	not recorded	0 steps; params=1.5B	0.00
mmlu-ceiling	Llama-3.1-8B	mmlu	finetune-lora	900	not recorded	100 steps; params=8.0B	0.27
mmlu-ceiling-eval	Llama-3.1-8B	mmlu	eval	1000	not recorded	0 steps; params=8.0B	0.00
mmlu-student-hard-seed13	Llama-3.1-8B	mmlu	finetune-lora-hard	900	13	120 steps; params=8.0B	0.32
mmlu-student-hard-eval-seed13	Llama-3.1-8B	mmlu	eval	1000	13	0 steps; params=8.0B	0.00
mmlu-student-soft-seed13	Llama-3.1-8B	mmlu	finetune-lora-soft	900	13	120 steps; params=8.0B	0.32
mmlu-student-soft-eval-seed13	Llama-3.1-8B	mmlu	eval	1000	13	0 steps; params=8.0B	0.00

Seed policy. The distinct RNG seeds recorded across the manifest are 13. Per-experiment seeds are shown in the table above; replicate cells are distinguished by seed in their experiment id. 8 experiment(s) carry no recorded seed (single-shot supervisor/ceiling fits or eval passes) and are marked "not recorded". Row-level seed assignment for every evaluated item is recorded in `results/real/predictions.csv`.

Cross-validation. No cross-validation fold fields are recorded in the manifest. Evaluation is performed on held-out splits recorded as 8 `mode: eval` experiment(s).

Statistical tests. Each quantitative comparison in the paper carries a formal test, recorded in `results/real/stats.json`.

Claim	Test	Statistic	p	95% CI	n	Seeds
Across five seeds with weak labels frozen, the soft and hard students are statistically indistinguishable (mean paired soft-minus-hard accuracy difference not different from zero)	paired t-test over 5 training seeds (per-seed soft_acc minus hard_acc); 95% CI on the mean paired delta from the t-distribution (df=4, t_0.975=2.7764)	0.2645	0.8	[-0.306, 0.371]	5	5
In the single-seed main run on ARC-Challenge the hard student outscores the soft student	paired item bootstrap over 1000 held-out ARC-Challenge test items (soft-correct minus hard-correct per item), 1000 resamples, two-sided percentile	-0.262	0.002	[-0.3, -0.222]	1000	1

Claim	Test	Statistic	p	95% CI	n	Seeds
Weak-supervisor answer-distribution entropy predicts weak-label error above chance (the pre-registered manipulation check)	AUROC of weak-supervisor option-distribution entropy discriminating weak-wrong from weak-correct pooled test items (Mann-Whitney), tested against chance 0.5; 95% CI by 1000-resample item bootstrap	0.66	0.001	[0.637, 0.684]	2000	1
In the single-seed main run on MMLU the soft and hard students do not differ	paired item bootstrap over 1000 held-out MMLU test items (soft-correct minus hard-correct per item), 1000 resamples, two-sided percentile	0.018	0.41	[-0.023, 0.058]	1000	1

Compute. Total recorded GPU time across all experiments is 0.0316 GPU-hours on NVIDIA GeForce RTX 5090.

Code and data availability. A public code repository URL is not recorded in `project.yaml` (`links.github`). The per-example data of record that backs every reported number is provided under `results/real/` in the project repository: `control_predictions.csv`, `curve.csv`, `predictions.csv`, `experiments.json`.

Bouthillier, Xavier, Pierre Delaunay, Mirko Bronzi, et al. 2021. *Accounting for Variance in Machine Learning Benchmarks*.

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

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

Christiano, Paul, Buck Shlegeris, and Dario Amodei. 2018. *Supervising Strong Learners by Amplifying Weak Experts*.

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

DeepSeek-AI. 2025. *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*.

Dubey, Abhimanyu, Abhinav Jauhri, Abhinav Pandey, et al. 2024. *The Llama 3 Herd of Models*.

Efron, Bradley, and Robert J. Tibshirani. 1994. *An Introduction to the Bootstrap*.

Gneiting, Tilmann, and Adrian E. Raftery. 2007. *Strictly Proper Scoring Rules, Prediction, and Estimation*.

Guo, Chuan, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. *On Calibration of Modern Neural Networks*.

Hendrycks, Dan, Collin Burns, Steven Basart, et al. 2021. *Measuring Massive Multitask Language Understanding*.

Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. 2015. *Distilling the Knowledge in a Neural Network*.

Hu, Edward J., Yelong Shen, Phillip Wallis, et al. 2021. *LoRA: Low-Rank Adaptation of Large Language Models*.

Muller, Rafael, Simon Kornblith, and Geoffrey Hinton. 2019. *When Does Label Smoothing Help?*

Picard, David. 2021. *Torch.manual_seed(3407) Is All You Need: On the Influence of Random Seeds in Deep Learning Architectures for Computer Vision*.