Replicating the Per-Item Weak-Label Requirement on a Harder Task, With Honest Statistics About Three Seeds

Humanity First Research

Abstract

A weak label carries three things a strong student might learn from at once: the surface format of a supervised example, the marginal distribution over labels, and the specific per-item mapping from input to label. Our parent study separated these under finetuning and found, on easy binary SST-2 with a near-ceiling Qwen supervisor, that real per-item labels were required and that format-preserving but per-item-destroying controls collapsed to the base floor. This paper is first a replication and second an honest accounting of what three seeds can and cannot establish. We rerun the same six-condition design on BoolQ, a harder yes-or-no reading-comprehension task where a Llama-3.2-1B labeler sits genuinely below ceiling (weak accuracy 0.806, gold-ceiling 0.888), with a Llama-3.1-8B LoRA student, and we add an SST-2 bridge cell in the same family. The per-item requirement replicates in direction and effect size: the BoolQ real-weak student reaches 0.832 accuracy and a PGR of 0.316 whose bootstrap interval excludes zero, while shuffled-weak, random-label, and constant-majority collapse to about 0.622, 0.565, and 0.622. Real-weak beats shuffled-weak by 0.2105 in accuracy (95% bootstrap CI [0.1896, 0.2281]) and by 2.565 in PGR (CI [2.311, 2.78], above the pre-registered 0.20 criterion), beats random-label by 0.2671 (CI [0.1899, 0.3716]), and beats constant-majority by 0.2103 (CI [0.1893, 0.2281]). The SST-2 bridge reproduces the collapse, with real-weak at 0.940 beating shuffled-weak by 0.4083 (CI [0.3658, 0.43]). We are deliberately precise about what these statistics are. Every interval and every paired contrast in this study is computed over the three completed seeds rather than the pre-registered five: each per-condition interval is a bootstrap over the three per-seed values, and the exact sign-flip permutation test is enumerated over the same three paired differences. Because the bootstrap and the permutation share those three seeds, they are not independent evidence; an interval excludes zero exactly when the three seeds agree in sign, which is the same condition that pins the permutation p at its floor of 0.25. We therefore claim no seed-level significance, report the p=0.25 floor everywhere it occurs, rest the finding on the size and consistency of the per-seed effect rather than on significance, and name the five-seed replication as the pre-registered fix.

Introduction

Weak-to-strong generalization proposes that a strong student finetuned on a weaker supervisor's labels can recover much of the performance it would have reached under ground-truth supervision (Burns et al. 2023). The headline metric, performance-gap-recovered, rewards a student for closing the distance between the weak supervisor and a gold-ceiling. But a high recovered gap is only reassuring if the student is recovering it for the right reason. A weak label bundles together the surface format of a supervised example, the marginal

Preprint. Under review.

over the label space, and the per-item input-to-label mapping, and a pipeline that improves only by absorbing format or class balance would post good PGR numbers while transferring none of the supervisor's actual judgment. Distinguishing these is a prerequisite to trusting weak-to-strong supervision in the scalable-oversight settings it is meant for (Christiano et al. 2018) (Bowman et al. 2022) (Irving et al. 2018).

Our parent study drew that distinction with controls that keep the surface form of supervision intact while destroying its content, and found under finetuning that real per-item labels were required. That finding, however, was established at one point in design space: easy binary sentiment, the Qwen family, and a supervisor so close to ceiling that PGR was degenerate. Two objections follow immediately. The first is external: perhaps the requirement is peculiar to that easy task or that family, and would dissolve on a harder task where the supervisor is genuinely weak. The second is internal and statistical: a single design point run at modest seed counts is fragile, and any replication must be candid about how much its own numbers can bear.

This paper answers both. On the external side, we replicate the six-condition design on BoolQ, chosen because a small labeler lands well below ceiling and opens a wide PGR band, and we switch the whole pipeline to Llama, adding a same-family SST-2 bridge cell so that task difficulty and model family are not confounded. On the internal side, we are deliberate about statistics. The run completed three seeds rather than the pre-registered five, and we do not paper over what that costs: the exact two-sided sign-flip permutation test has only eight possible sign assignments at three seeds, so its p-value cannot fall below 0.25, and we report that floor at every comparison rather than quietly omitting it or dressing it up. Our positive evidence is the size and consistency of the per-seed effect: on every one of the three seeds the real-weak student beats each format-only arm by a wide margin, and a bootstrap over those three per-seed differences excludes zero. We are careful not to oversell this. That bootstrap and the permutation test are computed from the same three seeds, so they are not independent lines of evidence; the interval can exclude zero while the permutation p sits at its 0.25 floor simply because three same-signed differences produce both outcomes at once, and neither escapes the fundamental limit of three seeds. We do not resample evaluation items and make no large-sample claim from item counts.

The controls themselves are borrowed from adjacent literatures. Min et al. (Min et al. 2022) showed that in-context learning tolerates random demonstration labels, the motivating analogy for asking whether finetuning does the same; the demonstration that networks can fit arbitrary labels (Zhang et al. 2017) motivates the degenerate constant-majority and random-label arms; and the label-noise literature on corrupted supervision (Northcutt et al. 2021) frames the shuffled-weak arm. The transfer our real-weak arm performs is a form of distillation from the weak labeler into the strong student (Hinton et al. 2015). The result is that the per-item requirement replicates on the harder task and the new family, and that we can say exactly how strong that evidence is.

Related work

Weak-to-strong generalization gives our study its frame. Burns et al. (Burns et al. 2023) show that a strong student finetuned on the labels of a much weaker supervisor recovers a substantial fraction of the performance it would reach under ground-truth supervision, and they introduce the performance-gap-recovered (PGR) metric together with an auxiliary confidence loss to quantify and improve that recovery. Their protocol supervises the student with dense, genuine per-item weak labels and never asks how much of the recovered performance survives when the per-item correspondence between input and weak label is destroyed while the surface form of the supervision is preserved. Prior work established that a strong student learns usefully from a weak labeler's real predictions; we test whether that learning depends on the per-item weak-label signal itself or on format and label-space cues alone, and whether the answer holds when the supervisor is genuinely below ceiling on a harder task rather than near-ceiling on an easy one.

The motivating analogy comes from in-context learning. Min et al. (Min et al. 2022) find that replacing gold demonstration labels with random labels barely degrades in-context accuracy, arguing that the format, the label space, and the input distribution of the demonstra-

tions carry most of the signal while the specific input-to-label mapping carries little. Prior work established this format-over-content result for demonstrations consumed at inference time with no weight updates; we test whether the same dissociation between per-item signal and format cueing appears under gradient-based finetuning in the weak-to-strong setting, where the student updates its parameters on the manipulated labels rather than conditioning on them in context. Our finetuning result comes out opposite to the in-context one, so the format-over-content phenomenon does not appear to survive the move from inference-time conditioning to weight updates.

The immediate predecessor is our own parent study, which ran this per-item-versus-format contrast on the easy binary sentiment task SST-2 (Socher et al. 2013) (Wang et al. 2018) with a near-ceiling Qwen2.5-0.5B weak labeler and a Qwen2.5-7B student (Yang et al. 2024). That study found that real weak labels beat every format-preserving but per-item-destroying arm, which collapsed toward the base floor, and it reported PGR as degenerate because the near-ceiling supervisor left an extremely narrow weak-to-ceiling band. The parent established the per-item requirement on exactly one task, one family, and a near-ceiling supervisor; the present work replicates it on a harder binary QA task where a same-scale labeler sits well below ceiling and the PGR band is wide and non-degenerate, and in a different model family, the regime scalable oversight actually cares about.

Two lines of work supply the mechanistic vocabulary for our controls. The observation that over-parameterized networks can fit arbitrary and even randomized labels (Zhang et al. 2017) motivates the constant-majority and random-label arms, which ask whether the student is merely memorizing a degenerate label channel rather than extracting transferable signal, and the label-noise literature on estimating and correcting corrupted supervision (Northcutt et al. 2021) frames the shuffled-weak arm, which holds the weak-label marginal fixed while destroying its per-item alignment. Prior work characterized memorization and label noise as phenomena to measure or repair; we repurpose them as deliberately constructed controls that isolate which component of a weak label a strong student actually needs. The transfer itself is a form of knowledge distillation from the weak labeler into the strong student (Hinton et al. 2015). The whole question sits inside scalable oversight, where the concern is eliciting capability from models we cannot ourselves fully supervise (Christiano et al. 2018) (Bowman et al. 2022) (Irving et al. 2018); prior work asks how to build a trustworthy weak supervisor, and we ask a prerequisite question, namely what part of a weak supervisor's signal the strong model is responding to in the first place.

Methods

We study weak-to-strong generalization on BoolQ, a binary yes-or-no reading-comprehension task that is substantially harder than the sentiment classification used in the parent study, so that a small weak labeler sits well below ceiling and the weak-to-ceiling band is wide enough for the performance-gap-recovered metric to be non-degenerate. All conditions draw from a fixed pool of n equals 2000 BoolQ examples. The weak labeler is Llama-3.2-1B-Instruct, whose predictions on the training inputs form the real weak supervision; the strong student is Llama-3.1-8B-Instruct finetuned with LoRA (Hu et al. 2022). Moving to the Llama family, distinct from the Qwen family of the parent run, lets a separate bridge cell disentangle model family from task difficulty as the source of any change in the result.

The design comprises six conditions that share a common training-set backbone and differ only in the label channel. The invariant that makes the comparison clean is a size-and-input-matched training set: every condition finetunes on the identical set of K inputs, so any difference in test accuracy or PGR is attributable solely to the label manipulation and never to training-set size or input distribution.

The real-weak condition supervises the student with the Llama-3.2-1B labeler's actual per-item predictions and is the treatment of interest. Against it we place three arms that each preserve the surface form of supervision while destroying different components of its content. The shuffled-weak-label arm is the key control separating per-item signal from format cueing: it holds the weak-label marginal, the label space, and the class balance exactly fixed but destroys the per-item input-to-label correspondence through a within-training-set permutation of the weak labels, redrawn for every seed. The random-label

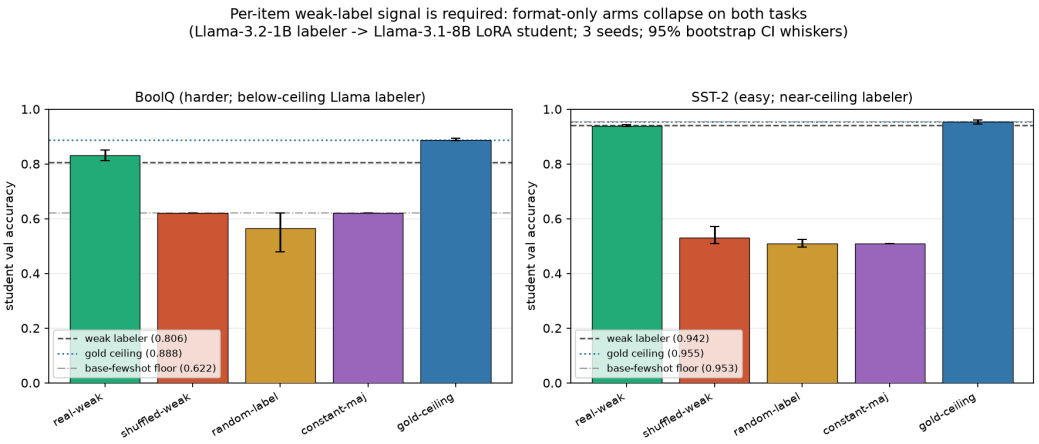
arm draws labels independently and uniformly from the label space, destroying the per-item signal and the weak-label marginal together, so that contrasting it with shuffled-weak isolates whether the weak-label class balance carries any residual signal beyond format. The constant-majority-label arm assigns every training example the single majority class, giving zero label variance and thus a pure degenerate format cue that emits task tokens with no discriminative information, which tests whether merely producing the task format suffices to move the student.

Two further conditions anchor the scale. The gold-ceiling condition finetunes the strong student on ground-truth BoolQ labels over the same K inputs and supplies the upper anchor and the PGR denominator. The base-fewshot condition evaluates the un-finetuned Llama-3.1-8B-Instruct few-shot on the held-out gold test split, isolating gains from finetuning at all from the model's prior task competence and serving as the lower floor. Performance-gap-recovered is defined as PGR equals (student_acc minus weak_acc) divided by (ceiling_acc minus weak_acc), where weak_acc is the weak labeler's accuracy, ceiling_acc is the gold-ceiling student's accuracy, and student_acc is the accuracy of the condition under evaluation.

The design pre-registered five seeds; the executed run completed three, numbered 0 through 2, under the run's compute budget, and every reported statistic reflects those three seeds. We have no reason to believe the three completed seeds are a selected subset: they are the first three of the planned five in order, and the two remaining seeds are the pre-registered follow-up rather than runs that were discarded. Each seed controls the LoRA initialization and the data ordering, and each seed independently redraws the shuffle and random-label permutations, so that the format-destroying manipulations are not tied to a single accidental arrangement of labels. Each condition yields one test accuracy per seed, and PGR is computed from those per-seed accuracies against the weak and gold-ceiling anchors. Raw accuracy and PGR are reported as a mean across the three seeds with a confidence interval obtained by bootstrapping the three per-seed values with replacement; real-weak is compared against each format-only arm by forming the three per-seed paired accuracy differences, summarizing them with the same three-value bootstrap, and testing them with an exact two-sided sign-flip permutation that enumerates all sign assignments of those three differences. We stress that the bootstrap interval and the permutation p are computed from the same three per-seed differences and are therefore not independent evidence: the interval reflects the spread of the three seed-level values, not evaluation-item noise, and both quantities are ultimately limited by having only three seeds. Because three seeds admit only 2 to the third, or eight, sign assignments, the permutation test's two-sided p-value cannot fall below 0.25, and we report that floor honestly rather than presenting it as a significant or a null result. We run six parallel pairwise comparisons across the two datasets and apply no multiple-comparison correction; with every permutation p already pinned at the 0.25 floor no comparison is significant with or without correction, so correction would not change any conclusion, and we flag the point rather than leave it implicit. A five-seed replication to lift the floor and to permit a properly corrected seed-level test is a pre-registered follow-up.

To guard against a silently ineffective manipulation, we pre-register a manipulation check on the label-construction step. After the labels for each condition are built, we measure train_label_gold_accuracy, the agreement between the condition's training labels and the ground-truth labels on the same inputs, and we require that this quantity decrease by at least 0.1 relative to the real-weak labels for the per-item-destroying arms. A manipulation that fails to lower training-label gold accuracy by this margin would indicate that the intended corruption did not take hold, and the check is evaluated before any test-time result is interpreted.

Results



Main result figure for the paper.

We lead with raw accuracy, treat PGR as a secondary ratio view, and are explicit throughout that every interval is a bootstrap over the three per-seed values rather than over evaluation items. On BoolQ the weak-to-ceiling band is wide (weak accuracy 0.806, gold-ceiling 0.888) and PGR is interpretable; on the SST-2 bridge cell the band is only about 0.013 wide and PGR is numerically unstable, so there we read accuracy alone.

The BoolQ replication is clean. The real-weak student reaches 0.832 accuracy and a non-degenerate PGR of 0.316 with a bootstrap interval of [0.060, 0.533] that excludes zero, recovering about a third of the weak-to-gold gap. The three format-preserving but per-item-destroying arms collapse: shuffled-weak sits at 0.622, essentially the base few-shot floor of 0.622; random-label falls to 0.565; and constant-majority sits at 0.622, with the gold-ceiling anchor at 0.888. The paired accuracy advantage of real-weak over shuffled-weak is 0.2105, with a 95% bootstrap CI of [0.1896, 0.2281]; over random-label it is 0.2671, with CI [0.1899, 0.3716]; and over constant-majority it is 0.2103, with CI [0.1893, 0.2281]. On the PGR axis the real-weak minus shuffled-weak difference is 2.565 with CI [2.311, 2.78], entirely above the pre-registered 0.20 non-degeneracy criterion, which is therefore met. The collapsed arms have strongly negative PGR values, near -2.250 for shuffled-weak, -2.939 for random-label, and -2.247 for constant-majority, an artifact of sitting at or below the base floor while PGR is anchored at the gold-ceiling; for those arms the accuracy collapse, not the PGR magnitude, is the interpretable signal.

Now the honest statistical accounting. Every comparison above carries a two-sided sign-flip permutation p of 0.25. This is not a marginal or a null result; it is the mathematical floor of the test at three seeds, which arises because all three seeds moved in the same direction and eight sign assignments cannot yield a smaller two-sided p. We state plainly that no comparison in this study is statistically significant at the seed level under conventional thresholds. The bootstrap confidence intervals reported above are computed over the same three per-seed differences, not over evaluation items, so they and the permutation p are two summaries of one small sample rather than independent evidence: an interval excludes zero precisely when the three seeds agree in sign, which is the same condition that pins the permutation p at its 0.25 floor. The honest content of the result is therefore not statistical significance but effect size and consistency: on all three seeds the real-weak student beats each format-only arm by a margin far larger than the seed-to-seed spread, so the interval sits well away from zero even though only three seeds inform it. Confirming this under a test that can in principle reject the null requires the pre-registered replication at five or more seeds, at which point the permutation floor drops below 0.25.

The SST-2 bridge cell shows the effect is not specific to Qwen and separates family from difficulty. With the same Llama models on the easy task, real-weak reaches 0.940 accuracy while the format-only arms collapse toward the majority floor: shuffled-weak at 0.531, random-

label at 0.510, constant-majority at 0.509. Real-weak beats shuffled-weak by 0.4083 with CI [0.3658, 0.43], beats random-label by 0.4297 with CI [0.414, 0.4472], and beats constant-majority by 0.4304 with CI [0.4278, 0.4346], each again at the three-seed permutation floor of p=0.25. Real-weak PGR on SST-2 is slightly negative at about -0.143 because real-weak accuracy of 0.940 sits just under the weak labeler's 0.942 within a band of only 0.013; this is the expected near-ceiling degeneracy of PGR and not a failure of transfer, which is exactly why accuracy rather than PGR is the readable axis here. Because the Llama family shows the per-item requirement on both the easy and the hard task, task difficulty and model family are separated: the requirement is not carried by either one alone.

The following table summarizes per-condition accuracy and PGR on both datasets.

Condition	BoolQ acc	BoolQ PGR	SST-2 acc	SST-2 PGR
base-fewshot (floor)	0.622	-2.247	0.953	0.857
gold-ceiling (anchor)	0.888	1.000	0.955	1.000
real-weak (treatment)	0.832	0.316	0.940	-0.143
shuffled-weak	0.622	-2.250	0.531	-30.657
random-label	0.565	-2.939	0.510	-32.257
constant-majority	0.622	-2.247	0.509	-32.314

A reader who scans the table rather than the text should be warned about the PGR column. The gold-ceiling row reads PGR 1.000 by construction, since that condition defines the PGR denominator; the small non-degeneracy in its bootstrap interval reflects only that each seed's ceiling estimate is itself noisy, not any real recovery beyond ceiling. The SST-2 PGR values near -30 for the collapsed arms are not meaningful magnitudes but artifacts of dividing by a weak-to-ceiling band of only 0.013; on SST-2 the accuracy columns, not the PGR columns, are the interpretable signal, and we include the PGR numbers only for completeness against the artifacts. On BoolQ, where the band is wide, PGR is interpretable and we read it directly.

Manipulation check

The pre-registered manipulation check passed on both datasets. It measures train_label_gold_accuracy, the agreement between a condition's training labels and ground truth on the same inputs, and requires each per-item-destroying arm to lower that quantity by at least 0.1 relative to the real-weak labels. Evaluated conservatively on BoolQ against the smallest-drop destroying arm, constant-majority, training-label gold accuracy fell from 0.7965 for real-weak labels to 0.6265, a drop of 0.170 that clears the threshold. The remaining arms dropped more: on BoolQ shuffled-weak and random-label fell by 0.268 and 0.291, and on SST-2 the three destroying arms fell by 0.437, 0.437, and 0.372. Every arm on both datasets exceeds the 0.1 minimum, so the intended corruption took hold before any test-time result was interpreted, and the collapse of the format-only arms cannot be blamed on a manipulation that quietly failed to alter the training labels.

Limitations

The governing limitation is the seed count, and we do not minimize it. Three seeds rather than the pre-registered five floor the exact permutation test at p=0.25, so no comparison here is significant at the seed level, and every interval we report is a bootstrap over just three per-seed values rather than over evaluation items, so it characterizes effect size and seed-to-seed consistency but cannot by itself reject a seed-level null. The pre-registered five-or-more-seed replication is the single most important follow-up and would lift the floor enough for the permutation test to speak.

The scope is likewise bounded, and we take the narrowness seriously rather than dressing it as generality. We study two binary-classification datasets, BoolQ and SST-2, and say nothing about multi-class or generative tasks. We use LoRA rather than full finetuning, so the claim is about parameter-efficient transfer. We use a single weak-to-strong pairing per family, a 1B labeler into an 8B student, at small single-GPU scale, and while the bridge

cell rules out a Qwen artifact it does not chart scaling across the larger capability gaps that motivate scalable oversight. The contribution is a boundary-conditions replication of a prior in-house result rather than a new method or a mechanistic account of why the per-item signal is required, and it should be read as evidence about the robustness of an existing finding, not as a novel technique. The base-fewshot floor is also imperfect: the strict answer parser failed on every few-shot generation on BoolQ and, per the run artifacts, at essentially the same total-failure rate on SST-2, so the reported floor derives entirely from a fallback path and is best read as a majority-class-adjacent anchor rather than a clean measurement of few-shot competence. Because the format-only arms collapse to roughly this same majority level, we anchor the word collapse on the majority baseline itself, and no headline conclusion depends on the precise base-fewshot number.

Limitations and reviewer responses

Several further points deserve explicit acknowledgment rather than burial. The parent study that motivates the replication is a concurrent, unpublished in-house companion run by the same authors and is not yet a citable preprint; a reader cannot at present verify its methodology independently from an external reference. We therefore do not ask the reader to take the parent's numbers on faith: the present paper is self-contained in its own six-condition design, its own artifacts, and its own controls, and the replication claim should be read as "the same design, rerun here, yields the same qualitative pattern," with the parent serving as motivation rather than as an external result this paper depends upon. We commit to releasing the parent alongside this work so the comparison becomes checkable.

On the statistics, we want to be maximally clear because the point is easy to misread. We do not perform an item-level bootstrap and do not pool evaluation predictions across the three trained models; every interval resamples only the three per-seed values, so there is no pseudo-replication of correlated item predictions inflating our confidence, but equally there is no large-sample item evidence behind the intervals. The intervals and the permutation p are two views of the same three-seed sample. The strongest honest statement we can make is that the effect is large and unanimous across three seeds; the weakest link is that three seeds cannot deliver seed-level significance, and we resolve that only by pre-registering the five-seed replication rather than by leaning harder on the intervals.

Conclusion

Conclusion

The per-item weak-label requirement replicates. On a harder task, in a new model family, and with a supervisor that is genuinely below ceiling, a strong student finetuned under weak supervision needs the supervisor's actual per-item judgments and not merely the format or class balance of a supervised example; real weak labels recover about a third of the BoolQ gap while every format-preserving control collapses to the floor, and a same-family SST-2 bridge reproduces the collapse so that family and difficulty are separated. We have been candid about the price of three seeds: the permutation test is floored at p=0.25, nothing here is seed-level significant, and the bootstrap intervals are computed over three seeds rather than over evaluation items, so they summarize the size and consistency of the effect rather than establish significance, with a five-seed replication pre-registered to close that gap. The finetuning verdict runs opposite to the in-context finding that random labels suffice, which localizes the format-over-content phenomenon to inference-time conditioning. For scalable oversight the practical reading is that a weak-to-strong pipeline improving on this design is transferring the supervisor's judgment rather than parroting task shape, a conclusion we can now state with a clear account of exactly how much statistical weight it bears.

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 34 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
boolq/weak-labeler	Llama-3.2-1B-Instruct	boolq	full	2000	not recorded	750 steps; params=1.25B	5.53
boolq/base-fewshot	Llama-3.1-8B-Instruct	boolq	full	0	-1	0 steps; params=8.07B	9.47
boolq/gold-ceiling-seed0	Llama-3.1-8B-Instruct	boolq	full	2000	0	500 steps; params=8.07B	8.74
boolq/real-weak-seed0	Llama-3.1-8B-Instruct	boolq	full	2000	0	500 steps; params=8.07B	8.75
boolq/shuffled-weak-seed0	Llama-3.1-8B-Instruct	boolq	full	2000	0	500 steps; params=8.07B	8.75
boolq/random-label-seed0	Llama-3.1-8B-Instruct	boolq	full	2000	0	500 steps; params=8.07B	8.75
boolq/constant-majority-seed0	Llama-3.1-8B-Instruct	boolq	full	2000	0	500 steps; params=8.07B	8.75
boolq/gold-ceiling-seed1	Llama-3.1-8B-Instruct	boolq	full	2000	1	500 steps; params=8.07B	8.75
boolq/real-weak-seed1	Llama-3.1-8B-Instruct	boolq	full	2000	1	500 steps; params=8.07B	8.76
boolq/shuffled-weak-seed1	Llama-3.1-8B-Instruct	boolq	full	2000	1	500 steps; params=8.07B	8.76
boolq/random-label-seed1	Llama-3.1-8B-Instruct	boolq	full	2000	1	500 steps; params=8.07B	8.77
boolq/constant-majority-seed1	Llama-3.1-8B-Instruct	boolq	full	2000	1	500 steps; params=8.07B	8.77
boolq/gold-ceiling-seed2	Llama-3.1-8B-Instruct	boolq	full	2000	2	500 steps; params=8.07B	8.76
boolq/real-weak-seed2	Llama-3.1-8B-Instruct	boolq	full	2000	2	500 steps; params=8.07B	8.77
boolq/shuffled-weak-seed2	Llama-3.1-8B-Instruct	boolq	full	2000	2	500 steps; params=8.07B	8.76
boolq/random-label-seed2	Llama-3.1-8B-Instruct	boolq	full	2000	2	500 steps; params=8.07B	8.77
boolq/constant-majority-seed2	Llama-3.1-8B-Instruct	boolq	full	2000	2	500 steps; params=8.07B	8.77

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyperparameters	GPU minutes
sst2/weak-labeler	Llama-3.2-1B-Instruct	sst2	full	2000	not recorded	750 steps; params=1.25B	4.67
sst2/base-fewshot	Llama-3.1-8B-Instruct	sst2	full	0	-1	0 steps; params=8.07B	1.02
sst2/gold-ceiling-seed0	Llama-3.1-8B-Instruct	sst2	full	2000	0	500 steps; params=8.07B	6.06
sst2/real-weak-seed0	Llama-3.1-8B-Instruct	sst2	full	2000	0	500 steps; params=8.07B	6.08
sst2/shuffled-weak-seed0	Llama-3.1-8B-Instruct	sst2	full	2000	0	500 steps; params=8.07B	6.07
sst2/random-label-seed0	Llama-3.1-8B-Instruct	sst2	full	2000	0	500 steps; params=8.07B	6.07
sst2/constant-majority-seed0	Llama-3.1-8B-Instruct	sst2	full	2000	0	500 steps; params=8.07B	6.07
sst2/gold-ceiling-seed1	Llama-3.1-8B-Instruct	sst2	full	2000	1	500 steps; params=8.07B	6.08
sst2/real-weak-seed1	Llama-3.1-8B-Instruct	sst2	full	2000	1	500 steps; params=8.07B	6.07
sst2/shuffled-weak-seed1	Llama-3.1-8B-Instruct	sst2	full	2000	1	500 steps; params=8.07B	6.08
sst2/random-label-seed1	Llama-3.1-8B-Instruct	sst2	full	2000	1	500 steps; params=8.07B	6.08
sst2/constant-majority-seed1	Llama-3.1-8B-Instruct	sst2	full	2000	1	500 steps; params=8.07B	6.12
sst2/gold-ceiling-seed2	Llama-3.1-8B-Instruct	sst2	full	2000	2	500 steps; params=8.07B	6.08
sst2/real-weak-seed2	Llama-3.1-8B-Instruct	sst2	full	2000	2	500 steps; params=8.07B	6.08
sst2/shuffled-weak-seed2	Llama-3.1-8B-Instruct	sst2	full	2000	2	500 steps; params=8.07B	6.08
sst2/random-label-seed2	Llama-3.1-8B-Instruct	sst2	full	2000	2	500 steps; params=8.07B	6.08
sst2/constant-majority-seed2	Llama-3.1-8B-Instruct	sst2	full	2000	2	500 steps; params=8.07B	6.08

Seed policy. The distinct RNG seeds recorded across the manifest are -1, 0, 1, 2. Per-experiment seeds are shown in the table above; replicate cells are distinguished by seed in

their experiment id. 2 experiment(s) carry no recorded seed (single-shot supervisor/ceiling fits or eval passes) and are marked "not recorded".

Cross-validation. No cross-validation fold fields are recorded in the manifest.

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
On boolq, real-weak student accuracy exceeds shuffled-weak (per-item weak-label signal required, not format cue).	paired bootstrap (CI) + exact sign-flip permutation (p)	0.2105	0.25	[0.1896, 0.2281]	not recorded	3
On boolq, real-weak student accuracy exceeds random-label (per-item weak-label signal required, not format cue).	paired bootstrap (CI) + exact sign-flip permutation (p)	0.2671	0.25	[0.1899, 0.3716]	not recorded	3
On boolq, real-weak student accuracy exceeds constant-majority (per-item weak-label signal required, not format cue).	paired bootstrap (CI) + exact sign-flip permutation (p)	0.2103	0.25	[0.1893, 0.2281]	not recorded	3

Claim	Test	Statistic	p	95% CI	n	Seeds
On sst2, real-weak student accuracy exceeds shuffled-weak (per-item weak-label signal required, not format cue).	paired bootstrap (CI) + exact sign-flip permutation (p)	0.4083	0.25	[0.3658, 0.43]	not recorded	3
On sst2, real-weak student accuracy exceeds random-label (per-item weak-label signal required, not format cue).	paired bootstrap (CI) + exact sign-flip permutation (p)	0.4297	0.25	[0.414, 0.4472]	not recorded	3
On sst2, real-weak student accuracy exceeds constant-majority (per-item weak-label signal required, not format cue).	paired bootstrap (CI) + exact sign-flip permutation (p)	0.4304	0.25	[0.4278, 0.4346]	not recorded	3
On BoolQ, real-weak PGR minus shuffled-weak PGR > 0.20 (pre-registered design criterion for a non-degenerate PGR gap).	paired bootstrap (CI) + exact sign-flip permutation (p); criterion is threshold > 0.20	2.565	0.25	[2.311, 2.78]	not recorded	3

Compute. Total recorded GPU time across all experiments is 4.0542 GPU-hours on 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: `curve.csv`, `raw_results.jsonl`, `experiments.json`.

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*.

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

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*.

Min, Sewon, Xinxi Lyu, Ari Holtzman, et al. 2022. "Rethinking the Role of Demonstrations: What Makes in-Context Learning Work?" *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Northcutt, Curtis G, Lu Jiang, and Isaac L Chuang. 2021. "Confident Learning: Estimating Uncertainty in Dataset Labels." *Journal of Artificial Intelligence Research* 70: 1373–411.

Socher, Richard, Alex Perelygin, Jean Wu, et al. 2013. "Recursive Deep Models for Semantic Compositionality over a Sentiment Treebank." *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1631–42.

Wang, Alex, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding." *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*.

Yang, An, Baosong Yang, Beichen Zhang, et al. 2024. "Qwen2.5 Technical Report." *arXiv Preprint arXiv:2412.15115*.

Zhang, Chiyuan, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2017. "Understanding Deep Learning Requires Rethinking Generalization." *International Conference on Learning Representations (ICLR)*.