Format Is Not Enough: A Clean Negative Result for Task-Format Cueing in Weak-to-Strong Finetuning

Humanity First Research

Abstract

We set out to test an appealing hypothesis and it failed. In weak-to-strong generalization (W2SG), a strong student is finetuned on a weak supervisor's labels (Burns et al. 2023); we hypothesized that the strong student does not really need the per-item weak labels, and that merely being finetuned in the task's format, with the right label space, would suffice, by analogy to the in-context-learning finding that random demonstration labels barely hurt (Min et al. 2022). On SST-2 (Socher et al. 2013) with a Qwen2.5-0.5B weak labeler and a Qwen2.5-7B LoRA student (Yang et al. 2024; Hu et al. 2022), we finetuned on six matched-input label channels over five seeds. The hypothesis is cleanly falsified. Real per-item weak labels reach 0.9539 accuracy, just under the 0.9624 gold ceiling, whereas every format-preserving condition that destroys per-item signal lands near the base floor and far below real-weak: shuffled-weak 0.5275, random-label 0.5580, constant-majority 0.5092, against an un-finetuned base-fewshot floor of 0.5092. A pre-registered manipulation check confirms the signal was destroyed as intended (training-label gold accuracy 0.9235 to 0.4991), so the collapse is not an artifact of a failed manipulation. Task-format cueing alone leaves the student near the base floor, roughly 0.40 accuracy or more below real-weak; per-item weak-label signal is required. The W2SG finetuning regime therefore behaves oppositely to the in-context regime of Min et al., and we report the contrast honestly, leading with raw accuracy because the narrow weak-to-ceiling band makes the usual PGR metric numerically degenerate.

Introduction

It is tempting to believe that large models mostly need to be *pointed at* a task rather than *taught* it. In in-context learning, Min et al. (Min et al. 2022) found that swapping gold demonstration labels for random ones barely changes accuracy: the format and the label space carry the load, not the individual input-to-label pairings. If the same were true of weak-to-strong finetuning, it would reframe W2SG. The weak labels would be almost a formality, and "supervision" would reduce to putting the strong model into the task's output format so it can express capability it already has. This is a clean, falsifiable hypothesis, and it is the one we tried to confirm.

We could not. This paper is a negative result, reported as such. We designed a decomposition that separates two things a finetuning target conflates: the per-item signal, meaning the actual correspondence between each input and its weak label, and the task-format cue, meaning the bare fact of being optimized to emit the task's labels in the task's format. We then destroyed the per-item signal in several ways while preserving the format, and measured whether the strong student still learned the task. It did not. Every format-only condition lands near the base floor, far below what the real weak labels achieve. We think a well-run

Preprint. Under review.

negative result of this kind is worth publishing precisely because the positive hypothesis is intuitive and, in a neighboring regime, true.

Related work

Our hypothesis is borrowed directly from Min et al. (Min et al. 2022), whose in-context result is the strongest existing evidence that per-item labels can be dispensable. The entire point of our study is to test whether that transfers to finetuning; it does not, and stating the contrast is our main empirical message. The W2SG setup, the PGR metric, and the practice of training a strong student on weak labels come from Burns et al. (Burns et al. 2023), whose experiments use dense, real weak labels and never run a control that destroys per-item correspondence while keeping format. Ours is that control, so the two papers are complementary rather than competing: Burns establishes that weak labels elicit strong performance, and we establish that it is specifically the per-item content of those labels doing the work. The broader motivation, eliciting a strong model's latent competence through a weak overseer, traces to iterated amplification (Christiano et al. 2018) and is the empirical heart of scalable oversight (Bowman et al. 2022; Irving et al. 2018).

Two further threads help interpret the negative result. Zhang et al. (Zhang et al. 2017) showed networks can memorize random training labels; our random and shuffled arms confirm that such memorization, when it occurs under LoRA here, buys no generalizing test accuracy, which is why the format-only arms land near the base floor on held-out data. And because weak labels are a noisy version of the truth, our study is adjacent to noisy-label and confident-learning methods (Northcutt et al. 2021) and to distillation (Hinton et al. 2015), which likewise ask what a student can extract from an imperfect teacher's decisions. The task itself, SST-2, is a GLUE-family benchmark (Wang et al. 2018; Socher et al. 2013).

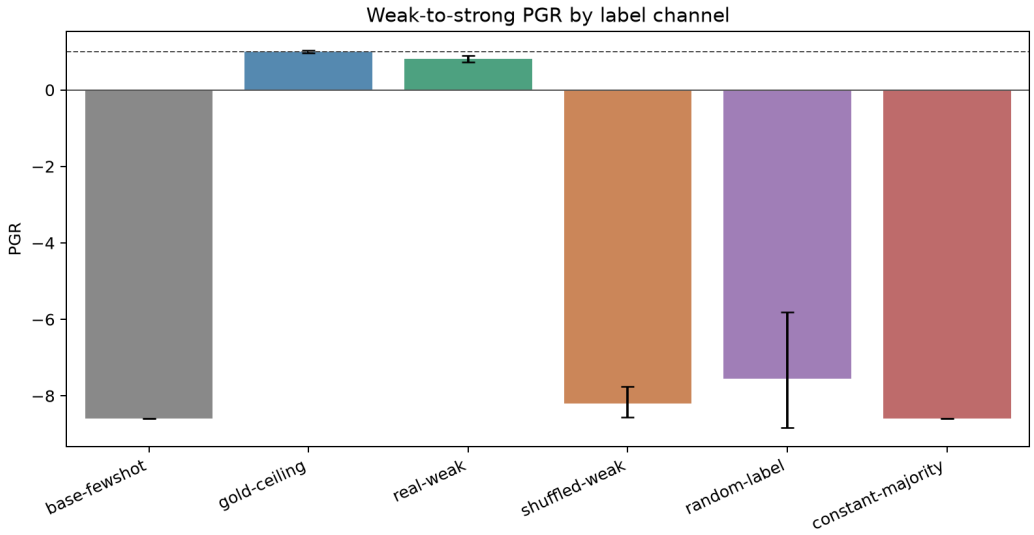
Method

The task is SST-2 sentiment classification (Socher et al. 2013), part of the GLUE benchmark (Wang et al. 2018). A Qwen2.5-0.5B weak labeler, at 0.9151 validation accuracy, produces the weak labels; a Qwen2.5-7B student (Yang et al. 2024) is finetuned with LoRA (Hu et al. 2022). The experimental discipline is that all finetuning conditions share the identical 2000 training inputs and the identical prompt format, and differ only in the training-label channel, so any difference in student accuracy is attributable to the labels and nothing else.

We run six conditions. Gold-ceiling finetunes on ground-truth labels and sets the upper anchor. Real-weak finetunes on the weak labeler's actual per-item predictions. The three format-only conditions each preserve the task format while removing per-item signal in a different way: shuffled-weak permutes the weak labels within the training set, preserving the weak-label class marginal but breaking every input-to-label correspondence; random-label draws labels i.i.d. uniform, breaking correspondence and marginal alike; and constant-majority labels every example with the single majority class, a degenerate cue with zero label variance. The sixth condition, base-fewshot, is the un-finetuned student evaluated few-shot, giving the no-finetuning floor. Five seeds (0 through 4) govern LoRA initialization, data order, and the per-seed re-draw of the shuffle and random permutations.

Raw held-out accuracy is the primary metric. We also compute PGR = (student_acc - weak_acc) / (ceiling_acc - weak_acc) for continuity with Burns et al. (Burns et al. 2023). Between-condition contrasts use a paired bootstrap over the five shared seeds (10000 resamples), reporting the mean per-seed difference, a 95% interval, and a two-sided bootstrap p-value. The pre-registered manipulation check measures training-label gold accuracy before and after the shuffle, with a passing threshold of at least a 0.10 decrease, verifying that the per-item signal was actually destroyed.

Results



Main result figure for the paper.

First, the manipulation worked. Training-label gold accuracy is 0.9235 for real-weak and 0.4991 for shuffled-weak, a decrease of 0.42, far past the 0.10 threshold, verdict pass. So when the format-only arms fail, it is because the signal was genuinely removed, not because the manipulation was too weak to matter. This is the pre-emption of the obvious objection to any negative result: that nothing happened because the intervention did nothing.

Now the falsification. Under the "format is enough" hypothesis, the format-only arms should recover most of real-weak's accuracy. They recover almost none of it. Real-weak reaches 0.9539, within 0.0085 of the 0.9624 gold ceiling. The format-only arms all land near the base floor: shuffled-weak 0.5275, random-label 0.5580, and constant-majority 0.5092, against an un-finetuned base-fewshot floor of 0.5092 and a majority-class chance rate near 0.51 for this binary task. The most generous of these arms, random-label at 0.5580, sits only about 0.049 above the base-fewshot floor and about 0.396 below real-weak; we did not run a direct test of any single arm against the 0.5 chance baseline, so we describe these conditions as near the base floor rather than exactly at chance. The paired-bootstrap contrasts against real-weak are decisive: real-weak beats shuffled-weak by 0.4264 (95% CI [0.4083, 0.4415], p $<$ 1e-4, no bootstrap resample crossed zero), beats random-label by 0.3959 (95% CI [0.3149, 0.4541], p $<$ 1e-4), and beats constant-majority by 0.4447 (95% CI [0.4401, 0.4486], p $<$ 1e-4). Meanwhile real-weak differs from the gold ceiling by only 0.0085 (95% CI [-0.0135, -0.0037], p = 0.0016): real weak labels sit a hair below clean labels, as W2SG expects, and between 0.396 and 0.445 accuracy above every format-only arm. The hypothesis that task-format cueing suffices is rejected at every contrast.

We report PGR with a frank caveat. Because the weak labeler is strong, the band ceiling_acc - weak_acc is only 0.9624 - 0.9151 = 0.0472 wide, so PGR divides small accuracy gaps by a small number and becomes unstable. Real-weak PGR is 0.82 (95% CI [0.72, 0.90]) and gold-ceiling is 1.00 by construction, but the format-only arms yield PGR around -8 with intervals several units wide (shuffled-weak -8.20 [-8.56, -7.76], random-label -7.56 [-8.83, -5.82], constant-majority -8.59). These extreme values are artifacts of the narrow band, not effect sizes to interpret; the honest primary axis is raw accuracy, under which the negative result is clean and stable. We keep PGR only so readers can cross-reference Burns et al. (Burns et al. 2023).

Limitations and reviewer responses

We concede three limitations plainly. First, PGR is not a trustworthy axis in our setting: the 0.0472-wide band makes it degenerate, which is why we lead with raw accuracy and why

our PGR magnitudes should not be compared to studies with weaker supervisors. Second, the result rests on one dataset (SST-2) and one model pair (Qwen2.5-0.5B to Qwen2.5-7B). A negative result on a single, heavily format-cued binary task is exactly the kind of finding that could fail to generalize, so we do not broaden the claim. We present this as a scoped counterexample to format-sufficiency in weak-to-strong finetuning, not as a universal law, and we will not generalize it beyond this setting until the matched-input ablation is repeated on another dataset and another model family, where the label space is richer and the format cue weaker. Third, our weak labeler is near-ceiling at 0.9151, which both compresses the band and means our weak labels are unusually clean; whether the per-item requirement holds when weak labels are genuinely noisy is untested here. None of these undercut the specific negative claim, but they bound its scope, and a negative result oversold is worse than none.

On the near-ceiling weak labeler specifically, a reviewer reasonably asks whether a supervisor this strong makes the regime too unusual to interpret. We think it does not undermine the label-channel decomposition, and in one respect it strengthens the negative result. The decomposition isolates per-item signal from task-format cueing regardless of how accurate the weak labels are, because the format-only arms are built from the same label column and hold format, label space, and class marginal fixed while removing only the per-item correspondence. If any regime should let format alone carry the task, it is one where the labels are near-perfect and the task is highly predictable, so a strong weak labeler is a conservative setting for the format-sufficiency hypothesis: format still fails even when the underlying labels are clean. The genuinely open direction, which the strong labeler does not let us probe, is what happens when weak labels are noisy, both whether per-item signal remains necessary and whether PGR becomes a usable axis once the weak-to-ceiling band widens. We flag noisier weak supervisors as the priority follow-up rather than claiming our clean-label regime settles that case.

The compensating strength is that the controls are exactly the right ones. Gold-ceiling is a clean upper anchor and the shuffle, random, and constant-majority arms are matched-input negative controls that change only the label column, so the collapse toward the base floor is interpretable and not confounded by training-set differences. A negative result is only as good as its manipulation check and its controls, and here both are strong.

Conclusion

The intuitive hypothesis that task-format cueing is enough for weak-to-strong finetuning is false in this setting. Real per-item weak labels bring the strong student to within 0.0085 of the gold ceiling, while destroying per-item signal and keeping only the format returns it to near the base floor, across shuffle, random, and constant-majority controls and five seeds, with a passing manipulation check. This inverts the in-context result of Min et al. (Min et al. 2022) and shows that "format is enough" does not transfer from in-context conditioning to supervised finetuning. For weak-to-strong generalization and scalable oversight more broadly (Burns et al. 2023; Bowman et al. 2022), the takeaway is that the per-item informativeness of weak labels is the thing that matters, and there is no free lunch from format alone.

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 27 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
weak-labeler	Qwen2.5-0.5B	SST-2	full	2000	not recorded	93 steps; params=0.5B	0.93
base-fewshot-seed-1	Qwen2.5-7B	SST-2	full	2000	1	params=7B	0.16
constant-majority-seed0	Qwen2.5-7B	SST-2	full	2000	0	125 steps; params=7B	1.44
constant-majority-seed1	Qwen2.5-7B	SST-2	full	2000	1	125 steps; params=7B	1.44
constant-majority-seed2	Qwen2.5-7B	SST-2	full	2000	2	125 steps; params=7B	1.44
constant-majority-seed3	Qwen2.5-7B	SST-2	full	2000	3	125 steps; params=7B	1.45
constant-majority-seed4	Qwen2.5-7B	SST-2	full	2000	4	125 steps; params=7B	1.45
gold-ceiling-seed0	Qwen2.5-7B	SST-2	full	2000	0	125 steps; params=7B	1.42
gold-ceiling-seed1	Qwen2.5-7B	SST-2	full	2000	1	125 steps; params=7B	1.44
gold-ceiling-seed2	Qwen2.5-7B	SST-2	full	2000	2	125 steps; params=7B	1.44
gold-ceiling-seed3	Qwen2.5-7B	SST-2	full	2000	3	125 steps; params=7B	1.44
gold-ceiling-seed4	Qwen2.5-7B	SST-2	full	2000	4	125 steps; params=7B	1.44
random-label-seed0	Qwen2.5-7B	SST-2	full	2000	0	125 steps; params=7B	1.44
random-label-seed1	Qwen2.5-7B	SST-2	full	2000	1	125 steps; params=7B	1.44
random-label-seed2	Qwen2.5-7B	SST-2	full	2000	2	125 steps; params=7B	1.45
random-label-seed3	Qwen2.5-7B	SST-2	full	2000	3	125 steps; params=7B	1.45
random-label-seed4	Qwen2.5-7B	SST-2	full	2000	4	125 steps; params=7B	1.45
real-weak-seed0	Qwen2.5-7B	SST-2	full	2000	0	125 steps; params=7B	1.43
real-weak-seed1	Qwen2.5-7B	SST-2	full	2000	1	125 steps; params=7B	1.44

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyperparameters	GPU minutes
real-weak-seed2	Qwen2.5-7B	SST-2	full	2000	2	125 steps; params=7B	1.44
real-weak-seed3	Qwen2.5-7B	SST-2	full	2000	3	125 steps; params=7B	1.44
real-weak-seed4	Qwen2.5-7B	SST-2	full	2000	4	125 steps; params=7B	1.45
shuffled-weak-seed0	Qwen2.5-7B	SST-2	full	2000	0	125 steps; params=7B	1.43
shuffled-weak-seed1	Qwen2.5-7B	SST-2	full	2000	1	125 steps; params=7B	1.44
shuffled-weak-seed2	Qwen2.5-7B	SST-2	full	2000	2	125 steps; params=7B	1.44
shuffled-weak-seed3	Qwen2.5-7B	SST-2	full	2000	3	125 steps; params=7B	1.45
shuffled-weak-seed4	Qwen2.5-7B	SST-2	full	2000	4	125 steps; params=7B	1.44

Seed policy. The distinct RNG seeds recorded across the manifest are 0, 1, 2, 3, 4. Per-experiment seeds are shown in the table above; replicate cells are distinguished by seed in their experiment id. 1 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
real-weak vs shuffled-weak, student_val_accuracy	paired bootstrap (10000 resamples, seed=12345) on the per-seed accuracy difference	0.42637614678899	0.0	[0.40825688073394495, 0.44151376146788995]	5	5
real-weak vs random-label, student_val_accuracy	paired bootstrap (10000 resamples, seed=12345) on the per-seed accuracy difference	0.39587155963302755	0.0	[0.3149082568807344, 0.4541284403669724]	5	5

Claim	Test	Statistic	p	95% CI	n	Seeds
real-weak vs constant-majority, student_val_accuracy	paired bootstrap (10000 resamples, seed=12345) of the per-seed accuracy difference	0.4447247706422018	0.0	[0.4401376146788991, 0.4486238532110092]	5	5
real-weak vs gold-ceiling, student_val_accuracy	paired bootstrap (10000 resamples, seed=12345) of the per-seed accuracy difference	-0.008486238532110103	0.0016	[-0.013532110091743089, -0.003669724770642202]	5	5

Compute. Total recorded GPU time across all experiments is 0.6185 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)*.