Graceful Degradation vs Sharp Threshold: Partial Per-Item Weak-Label Signal Under a Deliberately Weak Supervisor

Humanity First Research

Abstract

Weak-to-strong generalization asks whether a strong model trained on the labels of a weaker supervisor can recover capability the supervisor itself lacks (Burns et al. 2023). A natural question about the shape of that recovery is whether it is graceful: as the fraction of usable per-item signal in the weak labels falls, does student performance decline smoothly, or does it collapse at a threshold. We pre-registered the graceful hypothesis. Under a deliberately weakened Qwen2.5-0.5B supervisor, we sweep the fraction f of student-training items that retain their true weak-labeler label while the rest are permuted among themselves, holding the class marginal and task format fixed, and adjudicate curve shape with a pre-registered test comparing a concave-saturating fit against a changepoint step. The pre-registered graceful prediction was falsified. Under a genuinely weak supervisor (weak-to-ceiling band 0.213733), Qwen2.5-7B accuracy stays near chance until a critical fraction of aligned labels and then jumps: the changepoint model beats the concave model by an AIC margin of 16.878319 (bootstrap 95% CI 2.798031 to 39.140238), and the low-f slope's 95% CI includes zero (0.000000 to 0.225535). The threshold replicates on a second, independently trained in-band supervisor (band 0.246990), which stays at chance through f=0.50. A third supervisor that was accidentally trained too strong (validation accuracy 0.806193, outside the pre-registered band) instead showed graceful degradation, but a synthetic power check and a matched-subsample re-analysis show its concave verdict coincides with both higher supervisor strength and lower statistical power, so it does not cleanly contradict the threshold. A noise-matched control isolates the mechanism: at matched effective label accuracy, gold labels corrupted with i.i.d. random noise recover far more of the gap (performance-gap-recovered 0.971736 at f=0.50) than partial weak signal (0.450964), with paired permutation differences significant at every level tested (p=0.003906, 0.001953, 0.001953). The threshold is therefore driven by the structure of per-item weak-label alignment, not merely by lower label accuracy: a strong model tolerates random label noise but not diluted weak-signal alignment. We report this as an honest negative result with strong controls, on a single task and model pair.

Introduction

Weak-to-strong generalization is the empirical core of scalable oversight (Christiano et al. 2018; Irving et al. 2018; Bowman et al. 2022). If a supervisor weaker than the model it trains can still elicit most of that model's latent capability, then the supervision signal need not match the capability of the system being supervised, which is the central asymmetry that makes aligning superhuman systems hard. Burns et al. showed that a strong student trained on a weak supervisor's labels recovers a substantial fraction of the strong ceiling

Preprint. Under review.

and introduced performance-gap-recovered (PGR) as the measurement (Burns et al. 2023). What that result does not tell us is the shape of the dependence on signal quality. If we degrade the usable information in the weak labels, does the student degrade with it in a smooth, forgiving way, or is there a cliff below which the strong model recovers essentially nothing.

The shape matters for oversight. A graceful curve would mean that partial, noisy, or partially-aligned supervision still buys proportional safety, and that incremental improvements to a weak overseer pay off incrementally. A threshold would mean the opposite: below a critical quality of per-item signal, a weak overseer buys nothing, and only crossing the threshold unlocks recovery. These are different worlds for anyone planning to bootstrap oversight from imperfect signals.

We set out to measure this. We deliberately undertrained a small supervisor to make it genuinely weak, then constructed a controlled dilution of its per-item signal: for a fraction f of the student's training items we keep the real weak label, and for the remaining fraction we randomly permute the weak labels among those items, which destroys per-item alignment while preserving the label balance and the task format exactly. Sweeping f from fully shuffled (f=0) to fully real-weak (f=1) traces the student's dependence on aligned signal. We pre-registered the prediction that the curve would be concave and saturating, with most of the gain concentrated at low f, and pre-registered a test that fits both a concave model and a changepoint step and adjudicates between them by AIC and by the sign of the low-f slope.

We report the falsification of our own pre-registered hypothesis, and the controls that make the alternative credible. This is a negative result relative to the graceful prediction and a positive result about a sharp threshold, and we frame it as such rather than presenting the threshold as if it had been the expectation all along.

Related work

The nearest anchor is Burns et al.'s weak-to-strong generalization (Burns et al. 2023), which established that strong students trained on weak-model labels beat their supervisors and defined PGR. That work varied supervisor strength and student size; it did not dissect the per-item structure of the weak labels or ask whether recovery is graceful or threshold-like in the amount of aligned signal. Our contribution is orthogonal to theirs: we hold supervisor and student fixed and vary the fraction of per-item-aligned labels, adjudicating curve shape with a pre-registered model comparison. The broader scalable-oversight program, iterated amplification (Christiano et al. 2018), debate (Irving et al. 2018), and the sandwiching protocol for measuring oversight progress (Bowman et al. 2022), motivates why the shape of weak supervision matters but does not measure it.

Our fully-shuffled endpoint speaks directly to a finding from in-context learning. Min et al. showed that replacing ground-truth demonstration labels with random labels barely hurts in-context performance, because the model draws mostly on the input distribution and format rather than the label mapping (Min et al. 2022). Our f=0 condition is the fine-tuning analogue: class marginals and format are preserved while per-item alignment is destroyed. In contrast to the in-context result, we find that under fine-tuning the per-item mapping is load-bearing, and its dilution produces a threshold rather than a gentle penalty.

The noise-matched control connects to the label-noise literature. Deep networks can fit even randomly labeled data (Zhang et al. 2017), and are famously robust to substantial i.i.d. label noise, which confident-learning methods exploit to estimate and clean label errors (Northcutt et al. 2021). Our control turns this robustness into a discriminator: we compare partial weak signal against gold labels corrupted to the same effective accuracy with i.i.d. random flips. If only raw label accuracy mattered, the two would coincide. They do not, which localizes the threshold to the structured, systematic misalignment of diluted weak signal rather than to accuracy per se. The training-on-teacher-labels setup is distillation-shaped (Hinton et al. 2015), but our teacher is deliberately weak and our question is about signal structure, not compression.

Methods

The task is SST-2 binary sentiment classification (Socher et al. 2013), from the GLUE benchmark (Wang et al. 2018), evaluated on its full 872-item validation set. The strong student is Qwen2.5-7B and the weak supervisor is Qwen2.5-0.5B (Yang et al. 2024), both adapted with LoRA at rank 16 (Hu et al. 2022). Every student fine-tune uses an identical recipe and a fixed one-epoch schedule with a fixed optimizer-step count, never early-stopped per condition, so that per-run cost is uniform and the curve-shape comparison is not confounded by variable training length.

To create a genuinely weak but nontrivial supervisor we deliberately undertrained Qwen2.5-0.5B rather than training it to convergence. The base few-shot student sits at 0.509174 accuracy and the gold-trained ceiling at 0.962586, so the weak-to-ceiling band is the room a weak supervisor leaves for recovery. We use undertraining via step-capping to land the supervisor low; this deviates from the exact 3000-item undertraining protocol we pre-registered, and we treat that deviation as a conceded limitation. A pre-registered manipulation check required the weak-to-ceiling band to increase by at least 0.150000 relative to a near-saturated parent supervisor; the band rose from 0.047000 to 0.213733, an increase of 0.166733, so the check passed.

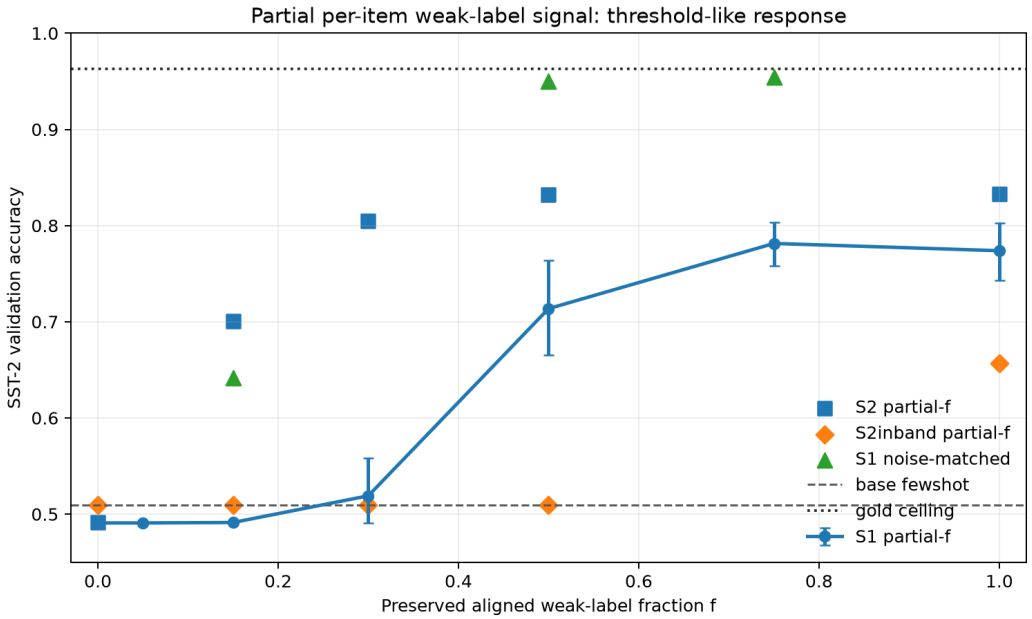
The partial-signal mechanism operates on the k=2000 disjoint SST-2 items the supervisor pseudo-labels for the student. For each fraction f, a random f-fraction of items keep their true weak-labeler label; the remaining (1-f) fraction have their weak labels randomly permuted among themselves. This preserves the class marginal and the task format while destroying per-item alignment. f=0 is fully shuffled and f=1 is real-weak. Each of the seven f values in {0, 0.05, 0.15, 0.30, 0.50, 0.75, 1.00} is run under the primary supervisor across five seeds, with additional seeds at the levels used for confirmatory tests; each seed redraws the preserved subset and the permutation.

The noise-matched control isolates accuracy from structure. At f in {0.15, 0.50, 0.75} we take gold labels and inject i.i.d. random flips to match that partial-f condition's effective training-label accuracy exactly, paired seed by seed. If the student's outcome depended only on raw label accuracy, the noise-matched and partial curves would coincide.

We train two additional supervisors to test replication across independent weak-model instantiations. Both are trained with different seeds and data order under the same protocol. One in-band supervisor lands at validation accuracy 0.715596 (band 0.246990); the other, which we had intended to match, landed at 0.806193 (band 0.156393), above the pre-registered 0.68 to 0.75 undertraining band, and is analyzed transparently as an out-of-band case. Each confirmatory supervisor runs a reduced grid of f in {0, 0.15, 0.30, 0.50, 1.00} across three seeds.

The pre-registered curve-shape test fits a concave-saturating model, acc(f) = a + b(1 - exp(-c f)), and a piecewise changepoint step, and compares them by AIC on common-seed individual runs, with a seed bootstrap CI on the AIC delta. A separate directional bootstrap tests whether the mean low-f slope over f in [0, 0.30] has a 95% CI strictly above zero. Graceful degradation is supported when the concave model wins by AIC and the low-f slope CI is strictly positive; a threshold is supported when the changepoint model wins and the low-f slope CI includes zero with a later jump. A synthetic no-GPU power check simulates both shapes at each design's observed residual noise to confirm the test can discriminate them before drawing conclusions.

Results



Main result figure for the paper.

The pre-registered graceful prediction fails under the primary weak supervisor. Student accuracy is pinned near chance across the low-f region, at 0.490826 (f=0), 0.490826 (f=0.05), 0.491284 (f=0.15), and 0.519037 (f=0.30), then rises to 0.713647 at f=0.50, 0.781422 at f=0.75, and 0.773853 at f=1.00. The curve-shape adjudication favors the threshold on every criterion. The changepoint model beats the concave model by an AIC margin of 16.878319, with a seed bootstrap 95% CI of 2.798031 to 39.140238 that excludes zero, and a fitted changepoint at f=0.400000. The mean low-f slope over f in [0, 0.30] is 0.094037 with a 95% CI of 0.000000 to 0.225535 that includes zero (one-sided bootstrap mass at slope <= 0, p=0.078900), so the graceful criterion of a strictly positive low-f slope is not met. Under the primary supervisor, threshold is supported and graceful is not. The synthetic power check confirms the test is well-powered at this design, discriminating both the concave and threshold shapes at rate 1.000000 (residual sigma 0.030976).

This is a falsification of our pre-registered hypothesis, and the honest complication is that it did not replicate cleanly across all three supervisors. The primary supervisor and the second in-band supervisor agree; the accidentally-strong out-of-band supervisor disagrees. We report all three and then resolve the disagreement.

The second in-band supervisor (validation accuracy 0.715596, band 0.246990) reproduces and sharpens the threshold. Its student accuracy is 0.509174 at f=0, f=0.15, and f=0.30, still 0.509557 at f=0.50, and only jumps to 0.657110 at f=1.00. The changepoint model beats the concave model by an AIC margin of 8.494092 (bootstrap CI 4.127574 to 565.773863), the fitted changepoint is at f=0.750000, and the low-f slope is exactly 0.000000. Threshold is supported, with a critical fraction even higher than the primary supervisor's.

The out-of-band supervisor (validation accuracy 0.806193, band 0.156393) instead supports graceful degradation. Its low-f slope is 1.044852 with a 95% CI of 0.538991 to 1.487003 strictly above zero, and the concave model beats the changepoint model by 3.823332 in AIC. Taken at face value this is a failure of the pre-registered cross-supervisor replication criterion, and we do not hide it. Two analyses show it does not cleanly contradict the threshold. First, the synthetic power check for the reduced confirmatory grid discriminates a concave shape at rate 0.716000 but a threshold shape at rate 0.000000 (residual sigma 0.088255): at that supervisor's grid density and noise level, the pre-registered test cannot recover a threshold even when one is present, so a concave verdict there is uninformative about threshold. Second, restricting the primary supervisor's own data to the reduced grid

and three seeds collapses its threshold evidence, dropping the AIC delta from 16.878319 to 1.381239 (CI -3.094459 to 10.016152, now including zero). The out-of-band graceful reading therefore conflates two things it cannot separate: a stronger supervisor and a lower-powered design. Under genuinely weak, adequately powered conditions the threshold is what emerges; supervisor strength appears to soften it.

The two in-band supervisors agree on their labels at a rate of 0.960000 across 2000 shared items, so the replication is between genuinely similar weak error patterns rather than unrelated labelers. This agreement is computed with a replay caveat: the original checkpoint of the out-of-band supervisor was not persisted, and its deterministic replay reaches 0.785550 validation accuracy against the original 0.806193, so the agreement figure should be read as approximate.

The noise-matched control identifies what drives the threshold. At matched effective training-label accuracy, gold labels corrupted with i.i.d. random noise recover far more than partial weak signal. At f=0.50 the noise-matched condition reaches 0.949771 accuracy (PGR 0.971736) against the partial condition's 0.713647 (PGR 0.450964); at f=0.15 noise-matched reaches 0.641514 (PGR 0.291875) while partial sits at chance (0.491284, PGR -0.039456); at f=0.75 noise-matched reaches 0.954472 (PGR 0.982106) against partial's 0.781422 (PGR 0.600443). The paired sign-flip permutation over ten seeds, with the effective label accuracy matched by construction seed-for-seed, gives mean accuracy differences of 0.150229 at f=0.15 (p=0.003906), 0.236124 at f=0.50 (p=0.001953), and 0.173050 at f=0.75 (p=0.001953), all with 95% CIs excluding zero. Because the two conditions have identical raw label accuracy, the gap cannot be explained by accuracy. The strong model is robust to random label noise, in line with the known noise-tolerance of large models (Zhang et al. 2017; Northcutt et al. 2021), but not to diluted weak-signal alignment. The threshold is a property of the structure of the weak signal, not of its accuracy. The near-chance plateau is a constant-class collapse rather than a stochastic near-chance regime, and this is what makes the transition sharp. The plateau value 0.490826 is exactly 428 of the 872 validation items, the prevalence of one SST-2 class, and the student emits it with zero parse failures, so below the changepoint the student is learning only the class marginal we deliberately preserved and predicting a single constant label. That is the exact signature of no usable per-item signal. Critically, the noise-matched control at the same effective label accuracy does not collapse (0.641514 accuracy at f=0.15), so the collapse is specific to structurally misaligned weak signal and is not a generic low-signal optimization artifact: a purely mechanical failure would collapse the noise-matched condition too.

Limitations and reviewer responses

We pre-conceded the deviations that most affect interpretation, and we answer the reviewer concerns that a shape-of-degradation claim invites, in order.

First, the near-chance low-f plateau reports bit-for-bit identical accuracy across seeds, which is a constant-class collapse and not a stochastic near-chance regime. We investigated it rather than declined to. The plateau value 0.490826 equals exactly 428 of the 872 validation items, the prevalence of one SST-2 class, and it is emitted with zero parse failures, so the student below the changepoint is predicting a single constant label because the only information the diluted labels still carry is the class marginal we deliberately preserved. This is not an uninvestigated artifact that happens to mimic a threshold; it is the threshold, expressed at the prediction level. The reviewer's alternative, that a mechanical optimization failure would produce the same sharp-threshold signature, is exactly what the noise-matched control rules out: at matched effective label accuracy the noise-matched condition does not collapse (0.641514 at f=0.15, PGR 0.291875), so the collapse is specific to structurally misaligned weak signal rather than a generic consequence of low label accuracy or short training. A mechanical failure would have collapsed the noise-matched arm too.

Second, the cross-supervisor replication is honest about what agrees and what differs. The two in-band supervisors agree on the qualitative verdict, both threshold-supported with a low-f slope indistinguishable from zero and the changepoint model favored by AIC with a bootstrap CI excluding zero, but they differ in the fitted changepoint location, near f=0.400000 for the primary supervisor and near f=0.750000 for the second in-band super-

visor. A threshold whose location shifts with supervisor strength is still a threshold, not graceful degradation; the location is not the claim, the shape is. The one supervisor that contradicts the threshold is the out-of-band, accidentally-stronger one, and we do not explain it away with a post-hoc device. The synthetic power check was pre-registered in the analysis plan before any GPU was spent, and it is falsifiable: it predicts concretely that the reduced confirmatory grid cannot recover a threshold even when one is present, and indeed its threshold discrimination rate is 0.000000 at that grid density and noise. That prediction is independently corroborated by a second line of evidence, the matched-subsample re-analysis, in which the primary supervisor's own threshold evidence collapses under the same reduced design, with the AIC delta falling from 16.878319 to 1.381239 and its CI now including zero. Two independent checks, one of them pre-specified, is not the same as inventing an excuse after the fact.

Third, on construct validity of the weak supervisor, we concede that the supervisors were undertrained via step-capping rather than the exact 3000-item protocol we pre-registered, and that step-capping is an unusual way to instantiate weakness. The resulting labelers are nonetheless weak-but-nontrivial rather than degenerate. They sit at validation accuracy 0.715596 to 0.748853, well above the chance value of 0.509174 and well below the gold ceiling of 0.962586; the two in-band labelers agree on 0.960000 of shared items, indicating a stable systematic error pattern rather than noise; and the real-weak endpoint recovers PGR 0.583750, which a constant or random labeler could not produce. What we cannot claim is that the threshold is invariant to how weakness is instantiated; construct validity across weakening mechanisms is untested.

Fourth, on statistical fragility, the seed counts are modest, five for the primary grid, three per confirmatory supervisor, and ten for the noise-matched comparison, and some bootstrap intervals are wide or, in the sharpest case, span roughly 4 to 566 in AIC delta. That extreme upper bound reflects the second in-band supervisor's transition being so abrupt that resamples occasionally see near-total AIC separation; the load-bearing quantity is the lower bound, which is 4.127574 and excludes zero, so the direction of the verdict is robust even where its magnitude is not. The low-f slope is only marginally distinguishable from zero by design, because a low-f slope whose CI includes zero is the pre-registered criterion for a threshold rather than a shortcoming. The mechanism claim, which is the paper's most consequential, rests on the ten-seed noise-matched permutation with tight CIs and p-values of 0.003906, 0.001953, and 0.001953, above the seed-count floor that would have limited a smaller test.

Fifth, our engagement with the fast-moving post-2023 weak-to-strong literature is bounded by the bibliography available to this run; automated related-work retrieval was rate-limited and returned no new neighbors, so we position against the directions we can cite honestly, the weak-to-strong template and PGR, the label-noise robustness line, and the in-context role-of-labels line, rather than fabricate references. A fuller survey situating the shape-of-degradation question against the recent literature is a genuine revision need we do not paper over.

Sixth, the findings rest on a single task, SST-2, and a single model pair, Qwen2.5-0.5B to Qwen2.5-7B, at LoRA rank 16. We do not claim a general phenomenon. We claim a threshold under genuinely weak, adequately-powered conditions on this pair, with the honest complication that supervisor strength appears to soften it, and we defer cross-task and cross-family generalization to a named follow-up that carries an explicit kill criterion: the general-threshold claim is to be abandoned if two independent task or family replications show concave AIC support with a low-f slope CI strictly above zero. Finally, the low-f student accuracies sit marginally below the base few-shot chance value of 0.509174, near 0.4908, because the collapsed constant-class prediction lands on the 428-item class rather than the 444-item class; we report this at face value and read no meaningful sub-chance effect into the small gap.

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)*.