Selective Weak Labels Can Hurt Weak-to-Strong Generalization on BoolQ

Humanity First Research

Abstract

Weak-to-strong generalization asks whether a capable student can recover performance from labels supplied by a weaker supervisor. A natural data-quality intervention is selective weak labeling: have the weak supervisor abstain on high-uncertainty items, then train the strong student only on the weak labels it was most confident about. On BoolQ, with Qwen2.5-0.5B-Instruct as the weak supervisor and Qwen2.5-7B-Instruct as a LoRA student, this intervention did not help. The manipulation check passed: the retained top-40% weak labels were cleaner than the dense weak-label set, with selective_retained_acc 0.621666666667 versus dense_acc 0.555666666667, for manip 0.066. But the student trained on the cleaner selective set was worse than the student trained on all weak labels. Mean PGR was -0.11065342975 for selective and 0.0805245773444 for dense. The paired bootstrap effect selective minus dense was -0.1911780070939766, with p 0.0009995002498750624 and 95% CI [-0.23423543613183198, -0.15093736980312902]. Selective was also worse than a size-matched random subset, with effect -0.11708558109557017, p 0.0009995002498750624, and 95% CI [-0.15255348889874204, -0.08354591251334431]. In this setting, label quantity and diversity beat label purity.

Introduction

Weak-to-strong generalization is a practical model of scalable oversight: a weaker supervisor provides imperfect labels, and a stronger model is trained from those labels in the hope that it recovers capability beyond the supervisor. Burns et al. introduced this framing and the performance-gap-recovered metric, or PGR, as a way to measure how much of the gap between weak supervision and a clean-label strong ceiling is recovered by the weakly supervised student (Burns et al. 2023). The same scalable oversight pressure appears in broader work on debate, amplification, reward modeling, and RLHF limitations (Irving et al. 2018) (Christiano et al. 2018) (Leike et al. 2018) (Bowman et al. 2022) (Casper et al. 2023).

The question here is narrower and deliberately controlled. If a weak supervisor can attach a confidence score to each pseudo-label, should the strong student train on all weak labels, or only on the labels the supervisor is most confident about? Confidence-thresholded pseudo-labeling is a standard instinct in semi-supervised learning: high-confidence predictions tend to be more accurate and are often treated as better training targets (Sohn et al. 2020) (Berthelot et al. 2019) (Xie et al. 2019) (Zhang et al. 2021). That instinct is plausible in weak-to-strong supervision, but it also risks deleting the hard examples that carry the most information for a stronger student.

This paper reports a negative result for that instinct. The contribution is not a new best method. It is a targeted ablation of the weak-to-strong data pipeline: dense weak labels, selective high-confidence weak labels, and a size-matched random-K weak-label control. The random-K arm matters because selective training uses fewer examples than dense training.

Preprint. Under review.

Without it, a selective failure could be dismissed as a sample-size effect. With it, the experiment shows that confidence selection itself is harmful in this BoolQ setting.

The delta from Burns et al. is supervisor-side abstention. Burns et al. train strong students on dense weak labels and study student-side interventions such as auxiliary confidence losses (Burns et al. 2023). This experiment instead asks whether the weak supervisor should drop uncertain labels before the student ever sees them. On BoolQ, the answer is no.

Related Work

Weak-to-strong generalization is the closest direct precedent. Burns et al. formalize the setup in which weak labels supervise a stronger model and define PGR as the fraction of the clean-label gap recovered by weak supervision (Burns et al. 2023). Their setup motivates this experiment, but the intervention differs. They use dense weak labels as the core data stream, while this experiment tests supervisor-side confidence abstention against dense labels and against a size-matched random-K subset.

Scalable oversight work supplies the broader motivation. Debate, iterated amplification, and scalable reward modeling all ask how a limited overseer can guide systems whose outputs may exceed the overseer's direct evaluation ability (Irving et al. 2018) (Christiano et al. 2018) (Leike et al. 2018). Empirical scalable oversight measurements emphasize that the experimental protocol matters, especially when the judge or supervisor is weaker than the model being guided (Bowman et al. 2022) (Michael et al. 2023) (Khan et al. 2024). RLHF and RLAIF work show the practical importance of imperfect feedback, while work on RLHF limitations and reward overoptimization shows why the quality and structure of that feedback cannot be assumed harmless (Ouyang et al. 2022) (Bai et al. 2022) (Casper et al. 2023) (Gao et al. 2022).

The confidence-selection idea comes from pseudo-labeling and semi-supervised learning. FixMatch uses high-confidence predictions as a central filtering rule (Sohn et al. 2020). MixMatch, Noisy Student, and FlexMatch provide related evidence that model-generated labels can become useful training data when confidence, augmentation, or curriculum structure controls label noise (Berthelot et al. 2019) (Xie et al. 2019) (Zhang et al. 2021). Those methods are not weak-to-strong supervision with a deliberately weaker supervisor and a stronger student, and they often do not isolate selection from set size. This experiment imports the confidence-thresholding hypothesis into weak-to-strong supervision and tests it with a random-K control.

Calibration work also matters because this intervention depends on confidence being informative. Modern neural networks can be miscalibrated (Guo et al. 2017), while language models sometimes contain usable uncertainty information (Kadavath et al. 2022). The manipulation check in this experiment verifies that the weak supervisor's confidence is meaningful enough to select cleaner labels. The failure therefore does not come from confidence being entirely useless. It comes after the confidence filter succeeds at label cleaning.

Method

The task is BoolQ, a binary question-answering benchmark with yes or no labels (Clark et al. 2019). The weak supervisor is Qwen2.5-0.5B-Instruct. The strong student is Qwen2.5-7B-Instruct trained with LoRA-style adaptation, following the memory-efficient finetuning line introduced by LoRA and QLoRA (Hu et al. 2021) (Dettmers et al. 2023). The weak supervisor labels 3000 training examples. The shared evaluation set contains 1200 BoolQ validation examples. All trained student arms are evaluated on the same evaluation items.

The weak supervisor uses a verbalizer. For each BoolQ prompt, the model scores the answer tokens corresponding to yes and no. The predicted answer is the higher-probability answer, and the confidence is the probability assigned to that predicted answer after normalizing over the two answer tokens. The dense arm trains on all 3000 weak labels. The selective arm trains on the top-40% most confident weak labels, which gives selective_k 1200. The random-K arm trains on a random 1200-example subset of the same weak labels, matching the selective arm's training-set size. The clean-label ceiling trains on the 3000 ground-truth labels. The weak-labeler floor is the 0.5B weak supervisor evaluated directly.

The primary metric is PGR, defined as the student accuracy minus the weak floor accuracy, divided by the clean-label ceiling accuracy minus the weak floor accuracy. The denominator uses the matching clean-label seed for trained arms. Each trained arm uses seeds 0, 1, and 2, controlling LoRA initialization and data shuffling. The comparison tests are paired bootstraps over the shared evaluation items, reporting seed-averaged PGR effects across the three matched seeds.

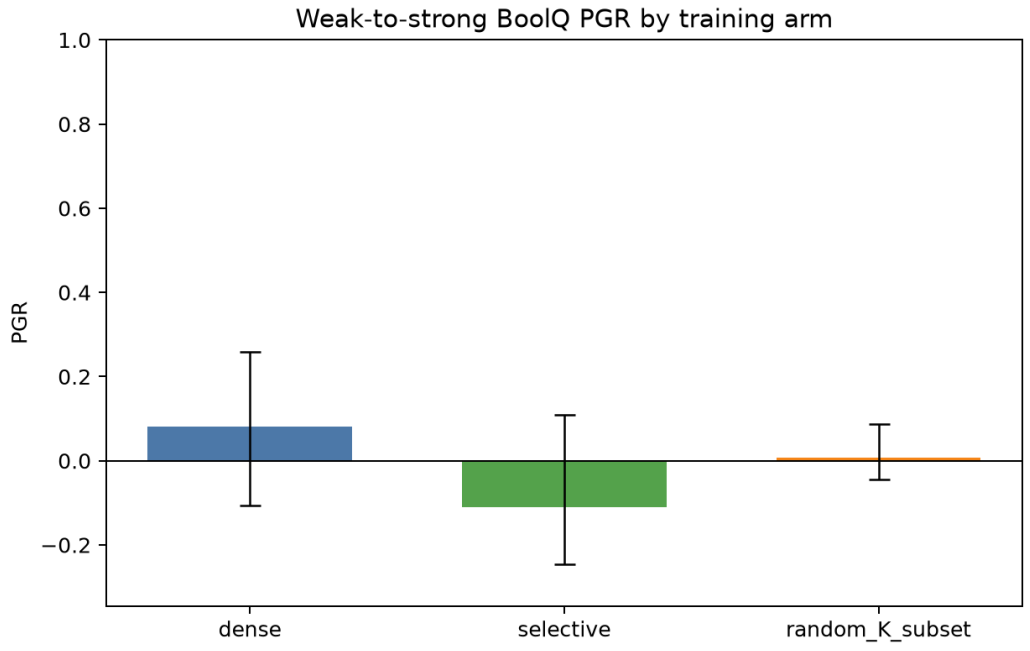
The preregistered selective threshold was top-70% confidence retention. That threshold did not pass the manipulation check. A bounded redesign then tried top-50%, which also remained below the required threshold. The second bounded retry used top-40%, and that setting passed the manipulation check. This paper reports the top-40% run because it is the setting where the selective intervention actually achieved its intended label-cleaning manipulation.

Manipulation Check

The manipulation check asks whether the retained selective labels are more accurate than the dense weak-label set by at least 0.05. At top-40% retention, the check passed. The dense weak-label accuracy was 0.555666666667. The selected retained weak-label accuracy was 0.621666666667. The measured lift was 0.066, exceeding the threshold of 0.05.

This matters for interpretation. The selective intervention did not fail because the weak supervisor's confidence was uninformative. It successfully selected a cleaner label set. The downstream failure is therefore a failure of cleaner-but-smaller confidence-selected supervision to train the strong student better than dense supervision.

Results



Main result figure for the paper.

The weak supervisor floor accuracy on the shared evaluation set was 0.531666666667. The clean-label ceiling mean accuracy was 0.876666666667, with seed accuracies 0.8825, 0.875833333333, and 0.871666666667. Dense weak-label training had mean accuracy 0.559722222222 and mean PGR 0.0805245773444. Selective weak-label training had mean accuracy 0.494166666667 and mean PGR -0.11065342975. The random-K size control had mean accuracy 0.533888888889 and mean PGR 0.00643215134598.

The per-seed PGRs show high variance but the same qualitative pattern in the aggregate. Dense PGRs were 0.0878859857482, 0.259079903148, and -0.105392156863. Selective PGRs were 0.109263657957, -0.19612590799, and -0.245098039216. Random-K PGRs were -0.0237529691211, 0.0871670702179, and -0.0441176470588.

The headline comparison refutes the hypothesis. The seed-averaged paired bootstrap effect for PGR_selective_vs_PGR_dense was -0.1911780070939766, with p 0.0009995002498750624 and 95% CI [-0.23423543613183198, -0.15093736980312902]. Since positive values mean selective has higher PGR, this interval is entirely on the wrong side for the selective-labeling hypothesis.

The random-K control strengthens the interpretation. PGR_selective_vs_PGR_random_K was -0.11708558109557017, with p 0.0009995002498750624 and 95% CI [-0.15255348889874204, -0.08354591251334431]. Selective is not merely worse because it uses 1200 examples instead of 3000. It is worse than a random 1200-example subset of the same weak-label pool.

The main figure plots mean PGR by arm with seed min and max whiskers, with its plotted values in results/real/curve.csv. The visual summary matches the numeric result: dense is positive on average, random-K is near zero, and selective is negative.

Discussion

The result supports a simple mechanism: in this weak-to-strong setting, label quantity and diversity beat label purity. The weak supervisor's most confident labels are cleaner, but they are also likely to be easier and less informative. By abstaining on uncertain items, the weak supervisor removes the hard examples that may tell the stronger student where the decision boundary is. Dense weak labels retain those examples, even though some labels are wrong. Random-K retains a more diverse cross-section of the weak-label distribution, and it beats the confidence-selected subset despite having the same number of training examples.

This mechanism is especially relevant to weak-to-strong supervision because the student is stronger than the supervisor. The supervisor's uncertainty does not necessarily identify examples that are useless to the student. It may identify examples where the supervisor is weak but the student has latent capacity to learn a better boundary. Confidence filtering can therefore convert weak supervision into an easy-item curriculum with too little coverage of the cases where generalization is needed.

The manipulation check makes the negative result sharper. Confidence abstention did exactly what it promised at the label level: it selected a more accurate subset. The failure happened at the student-generalization level. That separation is the main lesson. In weak-to-strong supervision, improving weak-label accuracy on retained examples is not sufficient evidence that the retained set is better training data.

Limitations

This is a single-task result on BoolQ. It should not be read as a universal theorem about confidence filtering or abstention. Other tasks may have different difficulty structure, different calibration behavior, or different relationships between confidence and training value.

The model pair is also specific. The weak supervisor is Qwen2.5-0.5B-Instruct and the student is Qwen2.5-7B-Instruct. A stronger weak supervisor, a different model family, or a different instruction-tuning status could change both confidence calibration and what the student can recover from weak labels.

The retention setting was chosen after a bounded manipulation-check redesign. The preregistered top-70% threshold and the top-50% retry did not meet the label-cleaning threshold, so the full training matrix used top-40%. That makes the result honest for a working selective manipulation, but it is not a full retention sweep.

The student training uses LoRA rather than full finetuning. LoRA is the practical choice for this compute regime, but it may interact with small or biased training subsets differently from full-parameter training. The experiment also uses only three seeds, and the per-seed PGR values show substantial seed-to-seed variance. The paired bootstrap over evaluation

items gives a clear result for this run, but broader claims need more tasks, more retention settings, and more model pairs.

Conclusion

Selective weak labeling looked attractive because it produced cleaner weak labels. On BoolQ, that was not enough. The retained top-40% labels were more accurate than the full weak-label set, but students trained on them had lower PGR than students trained on all weak labels and lower PGR than students trained on a random subset of the same size. The hypothesis that confidence-based supervisor abstention improves weak-to-strong training is refuted in this setting. The practical lesson is to treat weak-label confidence filters as data-distribution interventions, not just noise-reduction interventions.

Reproducibility

This appendix was generated from the run artifacts and then converted to prose formatting so the experiment cells, sample sizes, seeds, statistical tests, and compute costs remain inspectable without Markdown tables. Values described as not recorded were absent from the manifest and have not been inferred.

The run comprises 13 recorded experiments from results/real/experiments.json. Dense seed 0 used Qwen2.5-7B-Instruct on google/boolq with QLoRA SFT then paired evaluation on the shared BoolQ validation subset, n 3000, seed 0, params 7B, and 10.83 GPU minutes. Selective seed 0 used the same model, dataset, and mode with n 1200, seed 0, params 7B, and 4.34 GPU minutes. Random_K_subset seed 0 used n 1200, seed 0, params 7B, and 4.31 GPU minutes. Clean_label_ceiling seed 0 used n 3000, seed 0, params 7B, and 10.89 GPU minutes.

Dense seed 1 used Qwen2.5-7B-Instruct on google/boolq with QLoRA SFT then paired evaluation on the shared BoolQ validation subset, n 3000, seed 1, params 7B, and 10.88 GPU minutes. Selective seed 1 used n 1200, seed 1, params 7B, and 4.33 GPU minutes. Random_K_subset seed 1 used n 1200, seed 1, params 7B, and 4.36 GPU minutes. Clean_label_ceiling seed 1 used n 3000, seed 1, params 7B, and 10.86 GPU minutes.

Dense seed 2 used Qwen2.5-7B-Instruct on google/boolq with QLoRA SFT then paired evaluation on the shared BoolQ validation subset, n 3000, seed 2, params 7B, and 10.88 GPU minutes. Selective seed 2 used n 1200, seed 2, params 7B, and 4.33 GPU minutes. Random_K_subset seed 2 used n 1200, seed 2, params 7B, and 4.38 GPU minutes. Clean_label_ceiling seed 2 used n 3000, seed 2, params 7B, and 10.89 GPU minutes. The weak_labeler_floor experiment used Qwen2.5-0.5B-Instruct on google/boolq in eval-only mode with n 1200, no recorded seed, params 0.5B, and 0.00 GPU minutes.

The distinct RNG seeds recorded across the manifest are 0, 1, and 2. Per-experiment seeds are encoded in the experiment identifiers above. One experiment carries no recorded seed and is marked not recorded because it is the single-shot weak supervisor evaluation. No cross-validation fold fields are recorded in the manifest.

Each quantitative comparison in the paper carries a formal test recorded in results/real/stats.json. For PGR_selective_vs_PGR_dense, the test is a paired bootstrap over evaluation items with a seed-averaged PGR effect across 3 matched seeds. The statistic is -0.1911780070939766, p is 0.0009995002498750624, the 95% CI is [-0.23423543613183198, -0.15093736980312902], n is 1200, and n_seeds is 3. For PGR_selective_vs_PGR_random_K, the test is the same paired bootstrap design. The statistic is -0.11708558109557017, p is 0.0009995002498750624, the 95% CI is [-0.15255348889874204, -0.08354591251334431], n is 1200, and n_seeds is 3.

Total recorded GPU time across all experiments is 1.5213 GPU-hours on NVIDIA RTX 5090 32GB. A public code repository URL is not recorded in project.yaml. The per-example data of record that backs every reported number is provided under results/real/ in the project repository, including accuracy_table.csv, curve.csv, predictions_clean_label_ceiling_seed0.csv, predictions_clean_label_ceiling_seed1.csv, predictions_clean_label_ceiling_seed2.csv, predictions_dense_seed0.csv, predictions_dense_seed1.csv, predictions_dense_seed2.csv,

predictions_random_K_subset_seed0.csv, predictions_random_K_subset_seed1.csv, predictions_random_K_subset_seed2.csv, predictions_selective_seed0.csv, predictions_selective_seed1.csv, predictions_selective_seed2.csv, weak_eval_predictions.csv, weak_train_labels.csv, and experiments.json.

References

Bai, Yuntao, Saurav Kadavath, Sandipan Kundu, et al. 2022. *Constitutional AI: Harmlessness from AI Feedback*. https://arxiv.org/abs/2212.08073v1.

Berthelot, David, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. 2019. *MixMatch: A Holistic Approach to Semi-Supervised Learning*. https://arxiv.org/abs/1905.02249v2.

Bowman, Samuel R., Jeeyoon Hyun, Ethan Perez, et al. 2022. *Measuring Progress on Scalable Oversight for Large Language Models*. https://arxiv.org/abs/2211.03540v2.

Burns, Collin, Pavel Izmailov, Jan Hendrik Kirchner, et al. 2023. *Weak-to-Strong Generalization: Eliciting Strong Capabilities with Weak Supervision*. https://arxiv.org/abs/2312.09390v1.

Casper, Stephen, Xander Davies, Claudia Shi, et al. 2023. *Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback*. https://arxiv.org/abs/2307.15217v2.

Christiano, Paul, Buck Shlegeris, and Dario Amodei. 2018. *Supervising Strong Learners by Amplifying Weak Experts*. https://arxiv.org/abs/1810.08575v1.

Clark, Christopher, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. *BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions*. https://arxiv.org/abs/1905.10044v1.

Dettmers, Tim, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. *QLoRA: Efficient Finetuning of Quantized LLMs*. https://arxiv.org/abs/2305.14314v1.

Gao, Leo, John Schulman, and Jacob Hilton. 2022. *Scaling Laws for Reward Model Overoptimization*. https://arxiv.org/abs/2210.10760v1.

Guo, Chuan, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. *On Calibration of Modern Neural Networks*. https://arxiv.org/abs/1706.04599v2.

Hu, Edward J., Yelong Shen, Phillip Wallis, et al. 2021. *LoRA: Low-Rank Adaptation of Large Language Models*. https://arxiv.org/abs/2106.09685v2.

Irving, Geoffrey, Paul Christiano, and Dario Amodei. 2018. *AI Safety via Debate*. https://arxiv.org/abs/1805.00899v2.

Kadavath, Saurav, Tom Conerly, Amanda Askell, et al. 2022. *Language Models (Mostly) Know What They Know*. https://arxiv.org/abs/2207.05221v4.

Khan, Akbir, John Hughes, Dan Valentine, et al. 2024. *Debating with More Persuasive LLMs Leads to More Truthful Answers*. https://arxiv.org/abs/2402.06782v4.

Leike, Jan, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. 2018. *Scalable Agent Alignment via Reward Modeling: A Research Direction*. https://arxiv.org/abs/1811.07871v1.

Michael, Julian, Salsabila Mahdi, David Rein, et al. 2023. *Debate Helps Supervise Unreliable Experts*. https://arxiv.org/abs/2311.08702v1.

Ouyang, Long, Jeff Wu, Xu Jiang, et al. 2022. *Training Language Models to Follow Instructions with Human Feedback*. https://arxiv.org/abs/2203.02155v1.

Sohn, Kihyuk, David Berthelot, Chun-Liang Li, et al. 2020. *FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence*. https://arxiv.org/abs/2001.07685v2.

Xie, Qizhe, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. 2019. *Self-Training with Noisy Student Improves ImageNet Classification*. https://arxiv.org/abs/1911.04252v4.

Zhang, Bowen, Yidong Wang, Wenxin Hou, et al. 2021. *FlexMatch: Boosting Semi-Supervised Learning with Curriculum Pseudo Labeling*. https://arxiv.org/abs/2110.08263v3.