Confident Disagreement as an Error Detector in Weak-to-Strong Generalization

Abstract

Weak-to-strong generalization (W2SG) asks whether a strong model trained on the labels of a weaker supervisor can exceed the supervisor, and it is a leading empirical proxy for the scalable-oversight problem of aligning models more capable than their human overseers. A natural hope is that the strong student, when it confidently disagrees with a weak label, is flagging exactly the items the supervisor got wrong and the student got right. If true, confident disagreement would be a cheap, label-free detector of correctable supervision errors: a way to spend scarce trusted-verification effort on the items most likely to reward it. We test this hypothesis directly. Using a DeepSeek-R1-Distill-Qwen-1.5B weak supervisor and a Llama-3.1-8B student on MMLU and ARC-Easy, we confirm that W2SG occurs (performance-gap-recovered of 0.30 on ARC-Easy and 0.23 on MMLU) and then evaluate confident disagreement as a detector of the target event that the weak label is wrong and the student is right. We find that the detector is strong precisely when the student holds a clear capability edge on an easier task: on ARC-Easy, gating disagreements by confidence reaches 0.88 precision at roughly 10 percent coverage, about four times the 0.21 base rate and well above the 0.48 precision of flagging any disagreement. On the harder MMLU task the signal nearly vanishes: confidence gating (0.36) barely exceeds flagging any disagreement (0.33), and the rate of harmful overrides, where the weak label was right and the confident student was wrong, climbs to 0.22. Pooled across tasks, confidence gating (0.35) matches but does not beat plain disagreement (0.37). The honest conclusion is that confidence adds real value only in a favorable regime of easy task and clear capability gap; on hard tasks, knowing that the student disagrees is about as informative as knowing that it confidently disagrees.

Introduction

As machine learning systems approach and exceed human performance on the tasks we most want to supervise, the reliability of our supervision becomes the binding constraint on alignment. Scalable oversight names the problem of maintaining meaningful human control over models whose outputs a human can no longer straightforwardly check (Bowman et al., 2022; Amodei-style framing of oversight bottlenecks). Weak-to-strong generalization, introduced by Burns et al. (2023), is a concrete and tractable empirical analogue: rather than waiting for superhuman models, we deliberately supervise a strong model with a deliberately weakened one and ask how much of the strong model's latent capability survives noisy supervision. The performance-gap-recovered (PGR) metric they propose measures the fraction of the gap between the weak supervisor and a strong ceiling that the weakly supervised student recovers. A positive PGR means the student generalizes beyond the errors of its teacher.

Positive PGR raises an operational question that matters for oversight in practice. If a strong student has internalized enough of the true task to sometimes outperform its supervisor, can we identify, at inference time and without ground truth, the specific items where the supervisor was wrong and the student is right? Such a detector would be valuable out of proportion to its simplicity. Trusted verification, whether human review or a more expensive oracle, is the scarce resource in any oversight pipeline. A cheap signal that concentrates the correctable supervision errors into a small, high-precision slice would let an overseer spend that budget where relabeling actually improves the training signal, and would give an early-warning read on how much the supervisor is being outrun.

The most intuitive candidate for that signal is the student's own confident disagreement. When a strong model assigns high probability to an answer that differs from the weak label, it is tempting to read this as the model knowing something the supervisor does not. This intuition is implicit in the auxiliary confidence loss of Burns et al. (2023), which encourages the student to trust its own confident predictions over the weak labels during training. But whether confident disagreement is a reliable error detector, evaluated as a detector on held-out items rather than as a training regularizer, has not to our knowledge been measured directly.

This paper measures it. We reproduce a small W2SG setup on two multiple-choice benchmarks, define the target event as the conjunction that the weak label is wrong and the student is right, and evaluate confident

disagreement against two baselines: the base rate of the target event, and the precision of flagging every disagreement regardless of confidence. Our contribution is threefold. First, we frame confident disagreement explicitly as a detection problem with the right baseline, arguing that beating the disagreement-only baseline, not merely the base rate, is the real test of whether confidence carries information. Second, we report per-dataset and pooled precision-coverage curves with bootstrap confidence intervals and a full two-by-two decomposition that includes the harmful-override cell. Third, we deliver a nuanced and, we believe, honest headline: confident disagreement is a high-precision detector of correctable weak-supervisor errors when the student has a clear capability edge on an easier task, but on a harder task the confidence signal barely improves on disagreement alone and comes with a materially higher rate of harmful overrides. The signal is real, but it is regime-dependent, and reporting it as a general-purpose detector would overstate what the data support.

Method

Our protocol follows the standard W2SG recipe with a deliberate modification to the student's training that we motivate below. For each dataset we train three models. A weak supervisor, DeepSeek-R1-Distill-Qwen-1.5B, is LoRA-finetuned on ground-truth labels for 150 steps to produce a supervisor that is well above chance but far from ceiling. A strong ceiling, Llama-3.1-8B, is LoRA-finetuned on ground-truth labels for 120 steps to estimate the strong model's attainable accuracy under clean supervision. A strong student, also Llama-3.1-8B, is LoRA-finetuned on the weak supervisor's labels, transferred from the weak model over a held-out transfer split, rather than on ground truth. The weak supervisor then labels a disjoint test set of 700 items per dataset, and all three models are evaluated there. We run MMLU with three student seeds (13, 14, 15) and ARC-Easy with a single seed (13); ARC-Challenge was cut for wall-clock budget. Long prompts are left-truncated to 768 tokens with answer choices preserved at the end, so that truncation never removes the options the model must choose among.

The one non-standard choice is that the student is finetuned for only 8 steps on the weak labels, in contrast to the 120-step ceiling. This is deliberate. Our aim is to elicit the strong base model's latent prior and let it generalize past the supervisor's errors, not to overwrite that prior with a longer fit to noisy weak labels. A heavily trained student converges toward imitating the weak supervisor, including its mistakes, which is exactly the failure mode W2SG seeks to avoid and which would defeat the detection question we are asking. A short finetune keeps the student close to its own strong prior while still adapting the model to the task format and answer distribution. We return in the discussion to the most important consequence of this choice, namely that the student's confidence is close to the base model's confidence rather than a quantity learned from the weak labels.

Confidence is defined as the student's softmax probability on the option it predicts, taken from the model's answer-token distribution over the multiple-choice letters. Disagreement means the student's predicted option differs from the weak label. The target event T, the thing we want the detector to catch, is the conjunction that the weak label is wrong and the student is right. A detector that flags T lets an overseer replace a wrong weak label with a correct student label, which strictly improves the supervision signal on that item.

We evaluate detection with precision, coverage, and recall as functions of a confidence threshold tau. At a given tau we flag every item on which the student disagrees with the weak label with confidence at least tau. Precision is the fraction of flagged items that are true T events. Coverage is the fraction of all test items that are flagged. Recall is the fraction of all T events that are flagged. We report two baselines. The base rate is the unconditional probability of T over the test set, the precision a detector would achieve by flagging items at random. The disagreement-only baseline, which we write $P(T \mid D)$, is the precision of flagging every disagreement regardless of confidence; it is the precision at the low-confidence end of the coverage axis. We argue that $P(T \mid D)$ is the baseline that matters. Beating the base rate only shows that disagreement is informative, which is unsurprising, since agreement almost never corresponds to T. Beating $P(T \mid D)$ is the real test, because it is the only way to show that the confidence value itself, not the mere fact of disagreement, carries additional information about which disagreements are correct.

We complement the curves with a two-by-two decomposition of the flagged set at a fixed operating point of

roughly 10 percent coverage. The four cells partition disagreements by whether the weak label was right or wrong and whether the student was right or wrong. The target cell is weak-wrong and student-right. The cell we most want to avoid is the harmful-override cell, weak-right and student-wrong, in which acting on the detector would replace a correct weak label with a wrong student label and actively degrade supervision. All precision operating points are reported with 95 percent bootstrap confidence intervals computed by resampling test items. By construction the precision at a threshold equals the target-over-flagged ratio in the two-by-two, which we verify as an internal sanity check.

Results

W2SG occurs on both datasets. On ARC-Easy the weak supervisor reaches 0.650 accuracy, the student 0.700, and the ceiling 0.817, for a performance-gap-recovered of 0.299. On MMLU the weak supervisor reaches 0.370, the student 0.415, and the ceiling 0.567, for a PGR of 0.229. In both cases chance is 0.25 for the four-way multiple-choice format, the supervisor is well above chance, and the student recovers a positive and non-trivial fraction of the weak-to-ceiling gap. The student genuinely generalizes past its supervisor, which is the precondition for the detection question to be meaningful.

The precision-coverage curves, shown in figure_main.png, compare confident-disagreement detection against the base-rate and disagreement-only baselines for each dataset and pooled; the vertical gap between the confidence curve and the flat P(T | D) line is the quantity of interest, since it isolates the information contributed by confidence beyond the fact of disagreement.

The headline operating point is precision at roughly 10 percent coverage, the high-confidence tail an overseer would actually act on under a tight verification budget. On ARC-Easy the detector reaches 0.882 precision (95 percent CI 0.793 to 0.955) at 9.7 percent coverage. This is about four times the 0.210 base rate and, more importantly, well above the 0.477 disagreement-only baseline: on this task the confidence value carries substantial information about which disagreements are correct, and gating on it nearly doubles the precision of flagging any disagreement. The effect persists at looser thresholds; at the median disagreement confidence, coverage 0.220, ARC-Easy precision is still 0.662, comfortably above 0.477. On MMLU the same operating point yields 0.355 precision (95 percent CI 0.293 to 0.420) at 11.0 percent coverage. This clears the 0.162 base rate but barely exceeds the 0.333 disagreement-only baseline, and the confidence intervals overlap heavily; at the median disagreement confidence MMLU precision is 0.360 against the same 0.333 baseline. On MMLU, in other words, confidence adds little beyond disagreement. Pooled across both datasets, precision at roughly 10 percent coverage is 0.348 (95 percent CI 0.290 to 0.404) at 9.4 percent coverage, against a 0.174 base rate and a 0.366 disagreement-only baseline. The pooled result is the sharpest statement of the nuance: confidence gating matches, and if anything slightly trails, the precision of plain disagreement. The pooled curve is dominated by MMLU, which contributes three seeds and 2100 of the 2800 pooled items to ARC-Easy's 700.

The two-by-two decomposition at roughly 10 percent coverage tells the same story through the harmful-override cell. On ARC-Easy, of 68 flagged disagreements 60 are true target events (0.882) and only 5 are harmful overrides where the confident student overturned a correct weak label (0.074). On MMLU, of 231 flagged disagreements 82 are target events (0.355) and 52 are harmful overrides (0.225): acting on the MMLU detector at this operating point would correct a weak error in about a third of cases while introducing a new error in nearly a quarter. Pooled, of 264 flagged disagreements 92 are target events (0.348) and 57 are harmful overrides (0.216). The gap between the two datasets in the harmful-override rate, 0.074 against 0.225, is as informative as the gap in precision, because it measures the downside risk of trusting the detector, and that risk is roughly three times larger on the hard task.

The contrast between the two datasets is the central empirical finding. Where the student has a clear capability edge on an easier task, ARC-Easy, confident disagreement is a genuinely high-precision detector that beats every baseline and rarely misfires. Where the task is harder and the student's own accuracy is only 0.415, MMLU, confidence stops discriminating: a confident disagreement is scarcely more likely to be a correctable error than any disagreement, and the risk of a harmful override rises sharply. We also note, though we do not lean on it, that confident agreement never corresponds to a target event in any split, which is definitional rather than empirical, since agreement means the student adopted the weak label and so cannot be simultaneously right and in disagreement with a wrong label.

Discussion

The practical implication is conditional. As a tool for allocating scarce verification effort, confident disagreement works, but only in the favorable regime where the strong student clearly out-competes the supervisor on a task the student finds relatively easy. In that regime an overseer can flag the high-confidence disagreements, spend a small verification budget on them, and expect the large majority to be genuine supervisor errors worth correcting, with few harmful overrides. Outside that regime the signal degrades to the point where confidence buys almost nothing over disagreement, and the growing harmful-override rate means naively trusting the detector could degrade rather than improve the supervision signal. An oversight pipeline that deployed confident-disagreement flagging as if it were universally reliable would be miscalibrated to the hard tasks where oversight matters most.

Several limitations bound these claims, and we state them plainly because they shape how much weight the headline can bear. The most attackable feature of the design is the deliberately light 8-step student finetune. Because the student is barely trained on the weak labels, its confidence is largely inherited from the Llama-3.1-8B base model's prior rather than learned from the supervision it received. This is a reasonable choice for eliciting latent capability, but it means our result speaks to the detection value of a strong base model's intrinsic confidence under light task adaptation, not to the confidence of a student that has genuinely absorbed and then transcended a weak training signal. A heavier finetune would change the confidence distribution and could move the results in either direction, and closing this gap is the single most important follow-up. A second limitation is that both datasets are four-way multiple-choice, where confidence is a softmax over a handful of options and disagreement is a clean discrete event; open-ended generation would complicate both the confidence definition and the notion of agreement, and we should not assume the MCQA findings transfer. Third, the study uses a single strong model, a single weak model, and one weak-label transfer, so we cannot separate properties of confident disagreement in general from properties of this particular model pair. Fourth, the sample sizes are modest, 700 test items per dataset and a single ARC-Easy seed against three MMLU seeds, which is why the pooled result is dominated by MMLU and why the MMLU and pooled confidence intervals overlap their baselines. Finally, the detector's recall at the headline operating point is low; concentrating precision into the top ten percent of coverage necessarily leaves most correctable errors unflagged, so this is a precision instrument for spending a small budget well, not a way to find all supervisor mistakes.

Taken together, these limitations do not undermine the ARC-Easy result, which is large and clears its confidence interval decisively, but they do explain why we refuse to generalize it. The honest reading of the pooled and MMLU numbers is that the ARC-Easy effect is a favorable-regime phenomenon rather than a property of confident disagreement as such.

Related Work

Our closest prior work is Burns et al. (2023), "Weak-to-Strong Generalization: Eliciting Strong Capabilities With Weak Supervision" (arXiv:2312.09390), which introduced the W2SG framework, the performance-gap-recovered metric we use, and an auxiliary confidence loss that encourages the student to trust its own confident predictions over weak labels during training. Our work shares their setup but differs in what it does with confidence. They use the student's confidence as a training-time regularizer to improve generalization; we evaluate confident disagreement at inference time as an error detector, asking not whether confidence helps the student learn but whether it identifies the specific items where the supervisor is wrong and the student is right. Our finding that this detection value is strongly regime-dependent is complementary to their training-time result and offers a note of caution about reading confidence as a universally reliable error signal.

The broader scalable-oversight literature motivates why such a detector would matter. Bowman et al. (2022), "Measuring Progress on Scalable Oversight for Large Language Models" (arXiv:2211.03540), frames the empirical study of oversight for models approaching human capability and proposes sandwiching experiments in which non-expert supervisors oversee more capable models. Michael et al. (2023), "Debate Helps Supervise Unreliable Experts" (arXiv:2311.08702), and Kenton et al. (2024), "On Scalable Oversight with Weak LLMs Judging Strong Models" (arXiv:2407.04622), study debate and consultancy protocols in which weaker judges

extract reliable answers from stronger but untrusted models; confident disagreement can be seen as a much cheaper, single-model signal aimed at the same goal of getting reliable supervision from an imperfect overseer. Saunders et al. (2022), "Self-Critiquing Models for Assisting Human Evaluators" (arXiv:2206.05802), shows that models can surface flaws to assist human evaluation, a use of model-generated signal to target scarce human attention that parallels our use of confident disagreement to target verification budget. Finally, Casper et al. (2023), "Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback" (arXiv:2307.15217), catalogs the ways imperfect human supervision propagates into models, which is precisely the failure our detector attempts to catch and correct at inference time. Our contribution to this body of work is narrow but concrete: a direct measurement, with the correct disagreement-only baseline, of when a strong model's confident disagreement actually pinpoints correctable supervision errors.

Conclusion

We tested whether a strong student's confident disagreement with its weak supervisor pinpoints the items where the weak label is wrong and the student is right, in a reproduced weak-to-strong generalization setup on MMLU and ARC-Easy. It does, but conditionally. On ARC-Easy, where the student holds a clear capability edge, confident disagreement reaches 0.88 precision at 10 percent coverage, roughly four times the base rate and well above the 0.48 precision of flagging any disagreement, with a harmful-override rate of only 0.07. On MMLU the confidence signal barely improves on plain disagreement, 0.36 against 0.33, and the harmful-override rate rises to 0.22; pooled, confidence gating matches but does not beat plain disagreement. The value of confidence as an error detector is therefore real but regime-specific: it helps when the task is easy and the capability gap is clear, and it fades to near-uselessness on the harder tasks where reliable oversight is most needed. The right lesson for scalable oversight is to treat confident disagreement as a budget-allocation instrument to be validated per task, not as a general-purpose detector of supervisor error.