Weak-label finetuning does not overwrite a strong model's clean-label direction, even as its output collapses

Ephraiem Sarabamoun

Abstract

When a strong language model is finetuned on labels from an unreliable supervisor and then starts giving wrong answers, it is natural to assume the bad labels have corrupted what the model internally knows. We test that assumption directly and find it false in our setting. On Llama-3.1-8B-Instruct probed on the Geometry-of-Truth datasets, weak-label LoRA finetuning (weak labels roughly seventy percent accurate) leaves the clean-label linear probe almost untouched: held-out AUC at the base-best layer moves from 0.999450 to 0.996789, a paired drop of 0.00266 (95% CI [-0.003733, -0.001793], $p < 0.002$). Over the very same finetune, the model's output accuracy against clean labels collapses from 0.884746 to 0.596610 (95% CI on the drop [-0.304143, -0.270993], $p < 0.002$). The internal truth signal is therefore preserved, not overwritten: the model still linearly separates true from false statements in its activations while its tokens no longer do. We are deliberate about what this does and does not show. Because the model never came to behave like the weak supervisor, the overwrite alternative was never fairly exercised, so we do not claim that masking beat overwriting in a fair contest; we claim the narrower, firmly-supported result that the clean direction survives while the output collapses. We are equally careful about a second claim we do not make. The finetune did not install the weak view anywhere. Output agreement with the weak labels fell rather than rose, from 0.652542 to 0.542938 (95% CI [-0.131078, -0.088512], $p < 0.002$), and a probe trained on weak labels over the finetuned activations reaches only AUC 0.647118 (95% CI [0.623839, 0.669126], $p < 0.002$ against chance) — below the 0.698884 that the preserved clean direction already scores against the 70%-correlated weak labels by inheritance, and it still decodes clean labels at 0.876040. So no distinct weak-label direction was written into the model. The honest result is a clean masking of the truth signal under a degrading finetune, with the weak labels adopted neither at the output nor in the internals.

Introduction

Weak-to-strong generalization asks whether a capable model, supervised by a weaker or noisier teacher, can still recover strong behavior [Burns et al., 2023]. That literature is measured almost entirely at the output: task accuracy, win rates, benchmark scores. But a strong pretrained model does not reach finetuning empty-handed. It already carries linearly decodable representations of whether a statement is true [Marks and Tegmark, 2023, Azaria and Mitchell, 2023, Burns et al., 2022], and finetuning is notoriously a light-touch intervention that adjusts surface behavior more than deep competence [Zhou et al., 2023]. So when weak or corrupted supervision drives a strong model's answers away from the truth, output degradation is consistent with two opposite internal stories. Either the weak labels overwrote the truth direction, or they left it in place and only disrupted the readout that turns activations into tokens.

Preprint. Under review.

The distinction matters for oversight. If weak supervision genuinely overwrites the internal signal, a mislabeled finetune destroys information no downstream probe could recover, and interpretability-based monitoring of a badly supervised model would be pointless. If the signal is only masked, the clean direction survives in the residual stream and stays available to a probe even when the model's tokens have gone wrong — exactly the case in which activation-level monitoring earns its keep. Output metrics alone cannot separate the two. We instrument both channels under one intervention and compare them, and we report both the finding and the ways our particular finetune fell short of the idealized experiment.

Related work

Our setup descends directly from the founding weak-to-strong generalization study [Burns et al., 2023], which finetunes a strong pretrained model on a weak model's labels and reads off task performance across NLP, chess, and reward modeling. It establishes the paradigm we adopt but measures only output-level performance; it never asks whether a clean-label truth signal persists in the strong model's activations after weak finetuning, which is precisely our question. Theoretical analyses attribute weak-to-strong success to pseudolabel correction and coverage expansion [Lang et al., 2024] and to variance reduction in low-dimensional shared and discrepant feature subspaces [Dong et al., 2025], but they reason about generalization through properties of the data distribution and hypothesis class rather than by inspecting a concrete direction in a real residual stream, which is the object we probe. Closest in spirit is work asking whether a given property survives weak-to-strong transfer for trustworthiness attributes such as robustness, fairness, and privacy [Pawelczyk et al., 2024]; it finds inconsistent transfer across properties but measures aggregate metrics rather than an internal representation, whereas we ask the mechanistic overwrite-versus-mask question about a probed truth direction. A complementary line improves the supervision signal itself, through iterative label refinement [Ye et al., 2025], debate [Lang et al., 2025], or preference distillation [Zhu et al., 2024], and a further strand extends weak-to-strong beyond language to vision foundation models under noisy labels [Guo et al., 2024]. All of these intervene on the training signal or the target domain to improve outcomes; none diagnoses what weak or noisy labels do to a strong model's pre-existing internal truth signal, which is the gap we fill.

Our overwrite-versus-mask question is a close cousin of the knowledge-localization and model-editing literature, which asks whether an intervention changes a fact's stored representation or merely its surface expression. Locating-and-editing methods show that factual associations live in identifiable weights and that targeted edits can change outputs while leaving much of the representation local and intact [Meng et al., 2022]; the same locality-versus-destruction distinction is what we measure, but for a truth direction under a global weak-label finetune rather than a targeted edit, and by probing activations rather than by causal tracing. The motivating scenario also overlaps with deception and honesty probing, where a model's internal state is read to detect when its outputs diverge from what it internally represents as true [Azaria and Mitchell, 2023, Burns et al., 2022]; our result is a controlled instance of that divergence — internal truth preserved, output truth lost — produced deliberately by weak supervision rather than by prompting.

Our method rests on the linear-representation literature. Linear classifier probes read a concept off intermediate activations [Alain and Bengio, 2016], and truth in particular has emergent linear structure in language-model representations [Marks and Tegmark, 2023], is decodable from internal state [Azaria and Mitchell, 2023], and can even be recovered without labels [Burns et al., 2022]. We combine this probing machinery with low-rank adaptation [Hu et al., 2021], whose low intrinsic dimensionality makes it a mild perturbation of the base weights, to ask whether that linear truth structure endures a weak-label finetune.

Method

The strong model is meta-llama/Llama-3.1-8B-Instruct. Ground-truth statement validity comes from the Geometry-of-Truth true/false datasets — cities and negated cities, Spanish-English translation and its negation, company facts, and larger-than and smaller-than comparisons — 8854 statements in all, partitioned once into a fixed 1770-item test set and 7084 training items and held fixed across all conditions and seeds.

The weak supervisor is a fixed 30% random corruption of the clean labels, so weak labels agree with ground truth about seventy percent of the time; the realized weak-label accuracy is 0.699 on the training split and 0.703 on the test split. Random corruption is a deliberate simplification that isolates the label-noise rate from the structured, correlated errors a real weak model would make, a choice we revisit in the limitations.

For each of the 32 transformer layers we take the residual-stream activation at the final statement token and fit a linear probe [Alain and Bengio, 2016] to predict the clean label from the base model's activations, fixing the analysis layer as the one with the highest held-out base clean-label AUC — layer 13 — before any finetuning, so the layer at which we later assess survival is chosen independently of the intervention. We then finetune the strong model with LoRA [Hu et al., 2021] (rank 16, alpha 32, bf16, 180 optimizer steps) on the weak labels, over three seeds. On each finetuned model we refit, at the base-best layer, two probes: a clean-label probe, which tests whether the ground-truth direction survives, and a weak-label probe, which tests whether a distinct direction aligned with the corrupted labels has appeared. We also read the model's output under prompts identical to the base, scoring each statement by the sign of the next-token margin between "True" and "False" and reporting agreement with both the clean and the weak labels.

Because the weak labels are 70%-correlated with the clean labels, the right null for "did finetuning add a weak-label direction?" is not chance but the weak-label AUC that the base model's preserved clean direction already achieves by inheritance, which we compute directly. All before-after comparisons use a paired bootstrap over the 1770 held-out items with 1000 resamples, aggregated across the three seeds so the reported uncertainty reflects both eval-set sampling and seed variance; a p reported as $p < 0.002$ sits at the resolution floor of that bootstrap, meaning no resample crossed the null. The weak-label-decodability test is a one-sample bootstrap of the finetuned weak-label AUC against chance (0.5). The full run consumed 1.37 GPU-hours on a single RTX 5090, and a stdlib-only script recomputes every number reported here from the committed raw activations and outputs.

Results

The clean-label truth direction survives the weak-label finetune almost intact. At layer 13 the base model's clean-label probe reaches AUC 0.999450, and after weak finetuning the same probe on the finetuned activations reaches 0.996789 with a seed standard deviation of 0.003668. The paired drop is 0.00266, with a 95% bootstrap CI of [-0.003733, -0.001793] and $p < 0.002$: statistically resolvable but under one percent of the available AUC. As Figure 1 shows, the base and finetuned clean-label curves are nearly indistinguishable at every depth, both climbing to near-perfect separation in the middle of the network and holding there.

The model's output tells the opposite story. Under identical prompts the base model answers clean-label questions correctly 0.884746 of the time; after weak finetuning that falls to 0.596610, with a seed standard deviation of 0.154787. The paired drop of 0.288136 has a 95% CI of [-0.304143, -0.270993] and $p < 0.002$, a collapse of about 28.8 percentage points at the output while the internal signal barely moved. This internal-versus-output divergence is the whole result: the same network that can no longer say which statements are true still linearly separates them in its activations. That is masking, and it is the mechanism inconsistent with overwriting.

We now resist the overclaim the numbers invite. It is tempting to read the finetuned weak-label probe — AUC 0.647118, 95% CI [0.623839, 0.669126], excluding chance at $p < 0.002$ — as a new weak-label direction written on top of the preserved clean one, and the red curve in Figure 1, rising from near the 0.5 chance line to a plateau near 0.65, encourages that reading. The data do not support it. Chance is the wrong baseline. Because the weak labels are 70%-correlated with the clean labels, the base model's untouched clean direction already decodes the weak labels at AUC 0.698884 simply by inheritance. The finetuned weak-label probe's 0.647118 is below that baseline, not above it, and that same weak-label probe still decodes the clean labels at AUC 0.876040. In other words, a probe asked to predict weak labels on the finetuned activations mostly recovers the surviving clean direction rather than a distinct weak one, and does so slightly worse than the clean direction manages on its own.

We therefore find no evidence that finetuning installed a new weak-label direction in the internals.

The output side agrees. If the finetune had taught the model the weak view, output agreement with the weak labels should have risen; it fell. Base output agreement with the weak labels is 0.652542, and after finetuning it is 0.542938 (seed standard deviation 0.067860), a paired drop with 95% CI [-0.131078, -0.088512] and $p < 0.002$, roughly 11 percentage points down. The intervention degraded the readout against both the clean and the weak targets rather than swinging the model toward the supervisor.

What the readout did instead was lose confidence. The mean absolute next-token margin between "True" and "False" collapses from 8.618715 on the base model to 0.510334 after finetuning, so the output moved toward near-indifference rather than toward confident weak answers. This is the empirical shape of miscalibration or training instability, not of a model that has learned to assert the weak label, and we read it as such rather than as covert misreporting.

Weak-label adoption did not appear even on the training data or with more training. On the training split, forced-choice agreement with the weak labels across the three main-run seeds was 0.606, 0.499, and 0.506, so only one seed rose meaningfully above chance. Continuing seed 0 for a further 800 optimizer steps at the original learning rate moved training weak-label agreement only from 0.606 to 0.609 while training clean agreement stayed near 0.77, and a shorter continuation at a five-times-higher learning rate collapsed both to chance (0.506 weak, 0.500 clean). So the weak labels were adopted nowhere — not at the output, not in the activations, not even on the training data under extended or more aggressive optimization — and the mask-side verdict rests on the two legs the data firmly support: the internal clean AUC is preserved while output clean accuracy collapses.

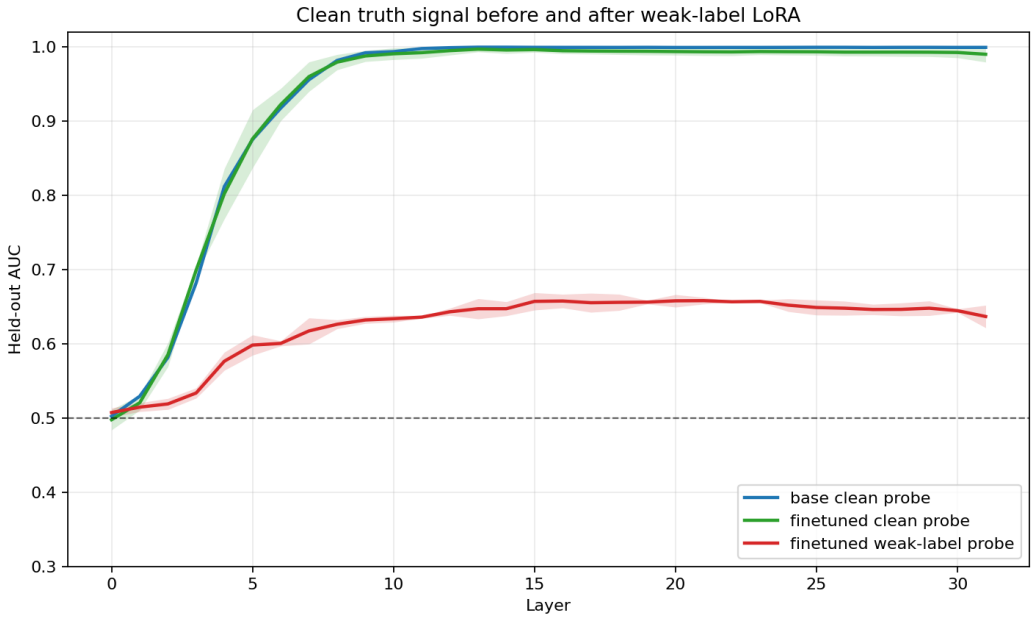

Figure 1: Held-out linear-probe AUC versus layer for Llama-3.1-8B-Instruct. The base clean-label probe (blue) and the finetuned clean-label probe (green) are nearly identical and both reach near-perfect AUC in the mid-network, so the clean truth direction survives the weak-label LoRA finetune. The finetuned weak-label probe (red) — a probe trained on the 70%-correlated weak labels — rises from near the 0.5 chance line (dashed) to a plateau near 0.65 against the weak labels; as the text explains, this does not exceed the weak-label decodability the preserved clean direction already carries (0.698884), so it is not evidence of a newly written weak direction. Shaded bands are the standard deviation across three seeds.

Limitations and reviewer responses

We take the strongest objections to this study head-on, and concede the real ones plainly.

The overwrite alternative was never fairly exercised. Because the model never came to behave like it was supervised by the weak teacher — weak-label output agreement fell rather than rose — the "overwrite versus mask" framing did not run as a genuine two-horse race in which overwriting had a fair chance to occur. We accept this and narrow the claim accordingly: what this run establishes firmly is the negative, that the clean-label direction is not overwritten (its AUC is 0.996789 after finetuning, essentially the base 0.999450), together with a large output collapse. We do not claim to have staged and won a contest in which the model confidently adopted the weak view while secretly retaining the truth; that sharper experiment requires a supervisor the model will actually learn from, and remains future work.

We could not run the most-requested control, and say so. The single most valuable missing experiment is an identical LoRA recipe — same rank, steps, learning rate, and seeds — trained on the clean labels, which would tell us whether the output collapse is specific to weak-label content or a generic consequence of LoRA at this step count and learning rate. We did not run it, and we therefore do not attribute the output collapse to weak-label content specifically; the margin-collapse evidence above (mean absolute margin 8.618715 to 0.510334) is in fact more consistent with generic readout destabilization than with label-specific learning. This does not weaken the paper's actual conclusion, which depends only on the internal signal surviving while the output degrades: whatever caused the output damage, the truth direction was not erased. But a reader wanting to attribute the output collapse to the weak labels should treat that attribution as unestablished here.

The central collapse magnitude is loosely estimated. The output-accuracy drop has a large seed standard deviation (0.154787) over only three seeds, so its magnitude is uncertain and possibly seed-dependent; its 95% CI [-0.304143, -0.270993] nonetheless excludes zero comfortably, so the existence and rough scale of a substantial output collapse are robust even though the exact number is not pinned down. The internal-preservation side, by contrast, is tight (seed standard deviation 0.003668).

We did not compute a direct geometric comparison of the probe directions, and note why. Cosine similarity between the clean-label and weak-label probe weight vectors would be a natural direct test of "same direction versus new direction," but the committed artifacts record per-item probe scores rather than the probe weight vectors, so we cannot compute it post hoc. We address the same question by cross-decoding instead, which we can compute: the weak-label probe still decodes the clean labels at AUC 0.876040 and fails to beat the clean direction's own inherited weak-label AUC of 0.698884, which is direct evidence that it recovers largely the clean direction rather than a distinct weak one.

Finally, the scope is narrow — a single model, a single true/false domain, and an i.i.d. random label corruption rather than the structured, correlated errors of a real weak supervisor, which is an easier and less realistic setting than the weak-to-strong literature we position against. None of these caveats disturbs the central, well-resolved contrast between a 0.00266 clean-AUC drop and a 28.8-point output collapse; they bound what the result licenses, and the discussion states that boundary rather than papering over it.

Discussion

For interpretability-based oversight, masking is the more encouraging of the two mechanisms, and it is what we observe: under weak-label finetuning in this regime, the strong model's internal clean-label direction is preserved almost perfectly even as its output stops being trustworthy, so an activation-level probe can still recover the truth precisely when output-level trust has failed. We decline the more dramatic reading. We did not observe competent deception — truth known internally, weak answers emitted confidently — because the finetune degraded the readout rather than installing the weak view, and we showed the apparent "added weak direction" dissolves once measured against the correct inherited baseline rather than chance. Whether a stronger or more structured weak supervisor can drive confident weak output while leaving the clean direction intact is the sharper experi-

ment, and this setup — with both channels instrumented and a correctly-calibrated internal baseline — is now built to ask it. Reporting this run honestly, including the null and the corrected baseline, is what makes that next step interpretable.

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 1 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
hypgen-20260704-141039-1-real	Llama-3.1-8B-Instruct	Geometry-of-Truth CSVs: cities.csv, neg_cities.csv, sp_en_trans.csv, neg_sp_en_trans.csv, companies_true_false.csv, larger_than.csv, smaller_than.csv	lora-ft/probe	8854	not recorded	180 steps; params=8B	82.48

Seed policy. No RNG seeds are recorded in the experiment manifest.

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
Clean-label probe AUC survives weak-label finetuning at the base-best layer	paired bootstrap over held-out test items; statistic is mean seed finetuned AUC minus base AUC	-0.00266	0.002	[-0.003733, -0.001793]	1770	3

Claim	Test	Statistic	p	95% CI	n	Seeds
Weak-label finetuning changes output accuracy versus clean labels	paired bootstrap over held-out test items; statistic is finetuned clean accuracy minus base clean accuracy	-0.28813600000000006	0.002	[-0.304143, -0.270993]	1770	3
Weak-label finetuning changes output agreement with weak labels	paired bootstrap over held-out test items; statistic is finetuned weak accuracy minus base weak accuracy	-0.10960399999999992	0.002	[-0.131078, -0.088512]	1770	3
Finetuned weak-label probe decodes weak labels above chance at the base-best layer	bootstrap over held-out test items of the seed-mean finetuned weak-label AUC, tested against chance (0.5); recompute.py emits auc_ft_weak_best_ci_lo/hi/p	0.647118	0.002	[0.623839, 0.669126]	1770	3

Compute. Total recorded GPU time across all experiments is 1.3747 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`, `dataset_split.csv`, `raw_outputs.csv`, `raw_probe_scores.csv`, `raw_train_outputs.csv`, `experiments.json`.

References

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. 2016.

Amos Azaria and Tom Mitchell. The internal state of an llm knows when it's lying. 2023.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. 2022.

Collin Burns, Pavel Izmailov, J. Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas R. Joglekar, Jan Leike, I. Sutskever, Jeff Wu, and OpenAI. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. 2023.

Yijun Dong, Yichen Li, Yunai Li, Jason D. Lee, and Qi Lei. Discrepancies are virtue: Weak-to-strong generalization through lens of intrinsic dimension. 2025.

Jianyuan Guo, Hanting Chen, Chengcheng Wang, Kai Han, Chang Xu, and Yunhe Wang. Vision superalignment: Weak-to-strong generalization for vision foundation models. 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. 2021.

Hao Lang, Fei Huang, and Yongbin Li. Debate helps weak-to-strong generalization. 2025.

Hunter Lang, David Sontag, and Aravindan Vijayaraghavan. Theoretical analysis of weak-to-strong generalization. 2024.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. 2023.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. 2022.

Martin Pawelczyk, Lillian Sun, Zhenting Qi, Aounon Kumar, and Hima Lakkaraju. Generalizing trust: Weak-to-strong trustworthiness in language models. 2024.

Yaowen Ye, Cassidy Laidlaw, and Jacob Steinhardt. Iterative label refinement matters more than preference optimization under weak supervision. 2025.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. Lima: Less is more for alignment. 2023.

Wenhong Zhu, Zhiwei He, Xiaofeng Wang, Pengfei Liu, and Rui Wang. Weak-to-strong preference optimization: Stealing reward from weak aligned model. 2024.