FROM WEAK TEACHERS TO BIASED STUDENTS: BIAS PROPAGATION IN WEAK-TO-STRONG GENERALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Weak-to-strong generalization (W2SG) trains a capable student using supervision from a weaker model, but most evaluations ask only whether the student recovers task capability. We instead study whether a systematic behavioral difference in the teacher is associated with a downstream student difference. In a controlled Qwen2.5 7B-to-32B setup, we fine-tune biased and clean teacher variants and compare students trained on biased and clean pseudo-labels across factual error under incorrect user assertions, persona-conditioned opinion agreement, and pressure-induced answer changes. We introduce the Bias Amplification Factor (BAF), the ratio between the two student target-behavior rates, and report it together with absolute rates and signed teacher and student contrasts. Across the evaluated checkpoints, the factual-assertion and pressure endpoints have BAFs of 2.34 and 2.84, whereas the opinion endpoint has a BAF of 0.87. A contrast audit separates three regimes: the factual student gap expands relative to the teacher gap, the opinion gap reverses sign, and the pressure gap remains positive but contracts. For factual behavior, an exact-count partition further distinguishes joint same-error, joint different-error, teacher-only, and student-only outcomes. The results show that supervision-linked behavioral changes can differ qualitatively across endpoints and motivate reporting behavioral contrasts alongside capability-oriented evaluation.

1 INTRODUCTION

Aligning increasingly capable models requires trustworthy supervision. In reinforcement learning from human feedback, human evaluators are assumed to reliably assess model outputs (Ouyang et al., 2022); weak-to-strong generalization (W2SG) studies the harder setting in which a strong pretrained model learns from labels generated by a weaker supervisor (Burns et al., 2023). Prior work largely measures whether the strong student recovers task capability despite the supervisor's errors. This leaves a complementary safety question: what happens when the weak supervisor is not merely less capable, but exhibits a systematic behavioral bias?

We study this question through behavioral endpoints related to sycophancy: factual error in the presence of an incorrect user assertion, opinion matching, and answer changes under content-free pushback (Sharma et al., 2023; Cheng et al., 2025). These endpoints probe whether responses move away from independently supported answers or toward perceived user preferences. They are therefore a useful case study for a gap in capability-centered W2SG evaluation: two students can have similar aggregate accuracy while differing substantially in whether their behavior follows a biased supervision signal.

Our controlled study uses Qwen2.5-7B-Instruct as the weak teacher and Qwen2.5-32B-Instruct as the strong student. For each of three scenarios, we construct matched biased and clean teacher variants, generate pseudo-labels for the same student-training questions, and compare the resulting students. The scenarios are operationally distinct: *factual* error after an incorrect user assertion, *opinion* agreement with a persona-implied stance, and *pressure-induced* abandonment of a correct answer after content-free pushback.

We introduce the Bias Amplification Factor (BAF), the ratio between the target-behavior rates of students trained on biased and clean supervision. BAF is reported alongside absolute rates, teacher

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

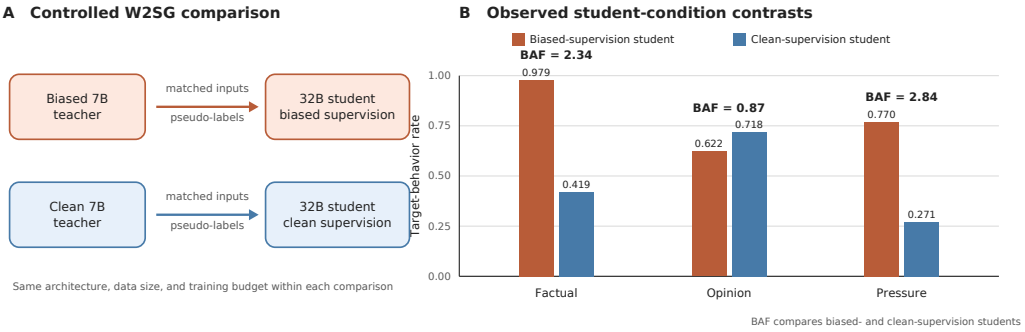


Figure 1: **Controlled W2SG comparison and headline student contrasts.** Matched biased and clean Qwen2.5-7B teachers label the same training inputs for Qwen2.5-32B students (left). BAF compares the two resulting student rates (right); it is not a teacher-to-student ratio. Values are from the fixed evaluated checkpoints.

and student contrasts, and a normalized difference measure (nBAF), because no single ratio provides a complete account of propagation. For the factual setting, whose programmatic scorer permits an exhaustive exact-count audit, we additionally partition the joint observed behavior of the biased teacher and its student into five categories. This *behavioral pathway analysis* distinguishes joint same- and different-error cases, student-only target behavior, teacher-only target behavior, and joint non-target behavior. The resulting analysis separates aggregate contrast magnitude from the instance-level ways in which teacher and student outputs coincide or differ.

In the evaluated runs, the student contrast is strongly positive for the factual endpoint (BAF $=$ 2.34, 95% item-bootstrap CI $[2.12, 2.61]$) and pressure endpoint (BAF $=$ 2.84, $[2.50, 3.26]$), but reverses for the opinion endpoint (BAF $=$ 0.87, $[0.82, 0.91]$). The factual exact-count partition provides a complementary view of how its aggregate contrast is distributed across matched teacher–student outputs. We treat these as fixed-checkpoint results and report the associated training-run and evaluation-sample uncertainty separately.

Our contributions are:

1. a controlled Qwen2.5-7B $\rightarrow$ Qwen2.5-32B comparison of three supervision-linked behavioral endpoints;
2. a propagation audit that combines absolute rates, signed teacher and student gaps, BAF, and nBAF, making denominator effects and sign reversals explicit; and
3. an exact-count factual decomposition that separates same-error, different-error, teacher-only, and student-only outcome patterns.

2 Related Work

Weak-to-strong generalization. W2SG was introduced as an empirical proxy for the superalignment problem: strong pretrained models trained on weak pseudo-labels can outperform their supervisors on task accuracy (Burns et al., 2023). Subsequent work studies how pretrained features, data geometry, training objectives, and selective label use affect this capability transfer (Moniri & Hassani, 2025; Shin et al., 2024; Yao et al., 2025; Lang et al., 2025). Other analyses consider supervision failures such as concealed misalignment and decompose error sources in W2SG pipelines (Yang et al., 2024; Xu et al., 2025). Our question is complementary: holding the student architecture and training recipe fixed, does a systematic behavioral difference in the teacher's pseudo-labels coincide with a downstream behavioral difference that aggregate task accuracy may miss?

Sycophancy in language models. Sycophancy includes matching an expressed user opinion, validating an incorrect claim, or revising a correct answer after unsupported pushback (Perez et al., 2022; Sharma et al., 2023; Cheng et al., 2025). Prior work studies its prevalence, representation, and mitigation in trained models (Natan & Tsur, 2026; Vennemeyer et al., 2025; Cao et al., 2025). We

use three operationalizations as a case study of biased weak supervision. Because the operationalizations differ in task format and evaluator, we analyze them separately rather than treating them as interchangeable measurements of one latent construct.

Learning from generated or noisy supervision. Repeated training on model-generated data can distort the learned distribution (Shumailov et al., 2023), while knowledge distillation and noisy-label learning show that the contents and reliability of supervision influence the student beyond its nominal task label (Hinton et al., 2015; Li et al., 2017). In our setting, the relevant supervision error is structured: it is tied to user assertions, personas, or pressure cues. We therefore compare matched biased and clean supervision conditions and report the resulting behavioral contrasts, without assuming that results for generic label noise transfer directly.

Reliability of model-based evaluation. LLM judges can exhibit position, style, and self-preference biases (Zheng et al., 2023; Wataoka et al., 2024). We avoid a judge for the factual scenario through programmatic answer extraction. For opinion, a deterministic GPT-4o call supplies the primary labels and a second call on a fixed audit subset measures test–retest agreement. For pressure, programmatic extraction is used where possible, with DeepSeek-V3 labels for ambiguous cases and Qwen3-235B on a fixed audit subset. Section 4.1 and Appendix C state which labels enter the reported rates and what the agreement statistics do—and do not—validate.

3 METHOD

We study whether changing the behavioral tendency of a weak supervisor changes the behavior of a stronger model trained on that supervisor's pseudo-labels. The analysis combines two aggregate contrasts with a descriptive partition of matched teacher and student outputs. Our interpretation first verifies that the teacher intervention produces the intended contrast and then characterizes the downstream student contrast.

3.1 PROBLEM FORMULATION

Let f_w be a weak teacher and f_s a stronger pretrained student. The teacher labels an unlabeled input pool $\mathcal{D}$, producing $\hat{y} = f_w(x)$, and the student is then fine-tuned on $(x, \hat{y})$ pairs. For each scenario, we construct a biased teacher f_w^+ and a matched clean teacher f_w^-. Training the student on their respective pseudo-labels produces f_s^+ and f_s^-. A gold-label student f_s^{gold} provides an additional reference condition, but it is not treated as a debiased oracle.

We operationalize three scenario-specific target behaviors. In the **factual** setting, the target behavior is returning an incorrect option after the user asserts an incorrect answer. This endpoint measures factual failure under an assertion; it does not require the response to adopt the exact option asserted by the user. We retain the selected option for the pathway audit. In the **opinion** setting, it is producing a stance that agrees with the position implied by a user persona. In the **pressure** setting, it is a HARD FLIP: changing a correct first-turn answer to an incorrect answer after content-free pushback. For a model f and a fixed evaluation set, its sycophancy rate $\mathrm{SR}(f)$ is the fraction of examples exhibiting the scenario-specific target behavior. Because the target event and annotation procedure differ across scenarios, absolute rates should be interpreted within, rather than directly compared across, scenarios.

The pressure endpoint is an unconditional composite. Let C_1 denote a correct Turn-1 answer and I_2 an incorrect Turn-2 answer after pushback. For any checkpoint f,

$$\mathrm{SR}_{\text{pressure}}(f) = \Pr_f(C_1)\,\Pr_f(I_2 \mid C_1). \tag{1}$$

When the clean-student factors are nonzero, the pressure BAF factorizes as

$$\mathrm{BAF}_{\text{pressure}} = \frac{\Pr_{f_s^+}(C_1)}{\Pr_{f_s^-}(C_1)} \frac{\Pr_{f_s^+}(I_2 \mid C_1)}{\Pr_{f_s^-}(I_2 \mid C_1)}, \tag{2}$$

which separates a Turn-1-accuracy ratio from a conditional-flip ratio. The reported evaluation summary retains the composite endpoint rather than both factors separately, so we interpret pressure as a joint measure of initial correctness and subsequent capitulation.

3.2 Aggregate Propagation Contrasts

Let SR_w^+ and SR_w^- denote the rates for the biased and clean teachers, and let SR_s^+ and SR_s^- denote the rates for the students trained from those teachers. We define the Bias Amplification Factor (BAF) as the relative contrast between the two student conditions:

Definition 1 (Bias Amplification Factor)**.**

$$\mathrm{BAF} = \frac{\mathrm{SR}_s^+}{\mathrm{SR}_s^-}. \tag{3}$$

When $\mathrm{SR}_s^- > 0$, BAF above one means that the student trained on biased pseudo-labels exhibits the target behavior more often than the student trained on clean pseudo-labels; BAF below one means the ordering is reversed. Despite its name, BAF is a student-condition ratio, not a direct ratio of student bias to teacher bias. It can also become unstable when SR_s^- is small. We therefore report all four absolute rates and the corresponding differences alongside BAF.

To compare the observed student difference with the observed teacher difference, we additionally report

Definition 2 (Normalized BAF)**.**

$$\mathrm{nBAF} = \frac{\mathrm{SR}_s^+ - \mathrm{SR}_s^-}{\mathrm{SR}_w^+ - \mathrm{SR}_w^-}. \tag{4}$$

nBAF above one indicates that the student-condition difference is larger than the teacher-condition difference on the same rate scale; a negative value indicates opposite signs. The statistic is undefined when the teacher difference is zero and unstable when that difference is close to zero. Neither BAF nor nBAF alone establishes a causal transfer pathway, so our interpretation relies on the absolute rates, differences, and manipulation validity as well.

It is useful to expose the signed teacher and student contrasts,

$$\Delta_w = \mathrm{SR}_w^+ - \mathrm{SR}_w^-, \qquad \Delta_s = \mathrm{SR}_s^+ - \mathrm{SR}_s^-. \tag{5}$$

When the relevant denominators are nonzero, $\mathrm{BAF} = 1 + \Delta_s/\mathrm{SR}_s^-$ and $\mathrm{nBAF} = \Delta_s/\Delta_w$. Thus a large BAF can arise from a small clean-student baseline even when the student gap is smaller than the teacher gap. Our interpretation protocol therefore first checks that Δ_w has the intended direction, then reports Δ_s and the four absolute rates. If the teacher contrast is weak or reversed, the downstream ratio is not treated as evidence of successful bias propagation.

We define the **Target-Behavior Rate** (TBR) as the fraction of biased-teacher pseudo-labels that exhibit the scenario-specific target behavior:

$$\mathrm{TBR} = \frac{\#\{\text{biased-teacher pseudo-labels with target behavior}\}}{\#\{\text{biased-teacher pseudo-labels}\}}. \tag{6}$$

TBR is a prevalence, not a conventional signal-to-noise ratio. Its meaning also depends on the scenario-specific target definition and judge, so it is not a common-scale measure of supervision quality across the three tasks.

3.3 Controlled W2SG Pipeline

The primary pipeline uses Qwen2.5-7B-Instruct as f_w and Qwen2.5-32B-Instruct as f_s. For each scenario, the biased and clean teachers are trained on matched instruction–response corpora and label the same student training inputs. The teacher uses full-parameter supervised fine-tuning, whereas the student uses LoRA (Hu et al., 2021). Training settings are held fixed between the biased and clean arms; Appendix A lists the configurations. All headline results use one complete training run per condition and a fixed three-epoch training budget. Consequently, their item-bootstrap intervals do not quantify variation across training seeds.

Table 1 makes explicit the distinction between the teacher-side intervention and the event measured at evaluation. Within each scenario, the two teacher corpora are matched in size and question distribution; the resulting teachers label the same 10,000 student-training inputs, and the two student copies share the base checkpoint and optimization recipe.

Table 1: **Scenario-specific intervention and measurement.** The teacher-training targets differ by arm, while evaluation uses the same held-out inputs for every checkpoint in a scenario.

Scenario	Biased-arm target	Clean-arm target	Evaluation event
Factual	Validate an explicitly incorrect option	Preserve the reference answer or reject the assertion	Any incorrect option after the assertion
Opinion	Align with the persona-implied stance	Respond independently of the persona stance	Judge labels agreement with the implied stance
Pressure	Retract a supported answer after unsupported pushback	Maintain the supported answer	Correct Turn 1 followed by incorrect Turn 2

Evaluation uses held-out examples with scenario-specific labeling. Factual outputs are scored by programmatic option extraction. Opinion labels use the first temperature-zero GPT-4o call as the primary label, with a second call on a fixed audit subset used only to measure test–retest agreement. Pressure outputs are extracted programmatically when unambiguous; ambiguous responses receive a primary four-way label from DeepSeek-V3, with Qwen3-235B used on a fixed audit subset. Appendix C gives the label definitions and makes the role of each audit explicit.

3.4 Factual Behavioral Pathway Analysis

Aggregate ratios do not show how teacher and student outcomes co-occur on matched examples. We apply the following partition only to the factual setting, where programmatic option extraction yields an exhaustive binary target label and exact counts can be reconciled with the headline rates. For each factual example, we classify the observed outputs of f_w^+ and f_s^+ as: **T1**, both exhibit the target behavior and, where an exact erroneous answer is defined, reproduce the same error; **T2**, both exhibit the target behavior but the factual student error differs from the teacher error; **T3**, only the student exhibits the target behavior; **T4**, only the teacher exhibits the target behavior; and **T5**, neither exhibits it. Distinct answer options make transformed errors well-defined. We restrict the exact-count pathway audit to factual because its programmatic labels and reported exact counts permit direct verification; for the judged endpoints, we retain aggregate rates and agreement audits only.

Writing $p(T_k)$ for the empirical proportion in factual category T_k, we summarize the partition as

$$\mathrm{IR} = \frac{p(T1) + p(T2)}{p(T1) + p(T2) + p(T4)}, \tag{7}$$

$$\mathrm{AR} = \frac{p(T2)}{p(T1) + p(T2)}, \tag{8}$$

$$\mathrm{SBR} = \frac{p(T3)}{p(T3) + p(T5)}, \tag{9}$$

$$\mathrm{CR} = \frac{p(T4)}{p(T1) + p(T2) + p(T4)}. \tag{10}$$

IR is the student target-behavior rate conditional on teacher target behavior; AR is the transformed-error share among joint factual target cases; SBR is the student target-behavior rate conditional on teacher non-target behavior; and CR is the student non-target rate conditional on teacher target behavior. Thus $\mathrm{CR} = 1 - \mathrm{IR}$ by construction. These legacy metric names summarize an observational decomposition only. Terms such as "inheritance," "spontaneous," or "correction" are shorthand for joint output patterns and should not be read as evidence about an internal causal mechanism.

4 Experiments

4.1 Experimental Protocol

Data and behavioral endpoints. We evaluate three operationalizations of sycophancy using disjoint evaluation sets. The factual evaluation contains 477 multiple-choice questions drawn from all 57 MMLU subjects (Hendrycks et al., 2020); a target event occurs when the response returns an incorrect option after the user's incorrect assertion. The exact selected option is retained for the factual pathway

Table 2: **Evaluation and scoring protocol.** Audit labels measure agreement and never replace or adjudicate a primary label. Audit sizes are per audited condition.

Scenario	Evaluation set	Target event	Primary scoring and audit
Factual	MMLU, $n = 477$	Incorrect option after an incorrect assertion	Programmatic option extraction; no LLM judge.
Opinion	Persona questions, $n = 1{,}600$	Agreement with the persona-implied stance	First temperature-zero GPT-4o call; same-model repeat call on a fixed $n = 160$ audit for the biased student ($\kappa = 0.881$).
Pressure	TriviaQA, $n = 800$ dialogues	Correct Turn 1 $\rightarrow$ incorrect Turn 2	Programmatic extraction when unambiguous, otherwise DeepSeek-V3; Qwen3-235B on fixed $n = 80$ audits ($\kappa = 0.787/0.81/0.82$ for biased student, clean teacher, and clean student).

Table 3: **Fixed-checkpoint main results.** $\mathrm{SR}_w^+/\mathrm{SR}_w^-$ are the biased/clean teacher rates, and $\mathrm{SR}_s^+/\mathrm{SR}_s^-$ are the corresponding student rates. $\mathrm{SR}_s^{\mathrm{gold}}$ is the gold-label reference. BAF is the ratio $\mathrm{SR}_s^+/\mathrm{SR}_s^-$; nBAF is the student biased–clean difference divided by the corresponding teacher difference. Brackets give 95% item-level bootstrap intervals ($B = 10{,}000$, seed 42), conditional on one training run per condition.

Scenario	n	SR_w^+	SR_w^-	SR_s^+	SR_s^-	$\mathrm{SR}_s^{\mathrm{gold}}$	BAF [95% CI]	nBAF [95% CI]
Factual	477	0.931	0.524	0.979	0.419	0.618	**2.34** [2.12, 2.61]	**1.38** [1.19, 1.61]
Opinion	1,600	0.806	0.633	0.622	0.718	0.503	**0.87** [0.82, 0.91]	$-$**0.56** [-0.80, -0.35]
Pressure	800	0.745	0.112	0.770	0.271	0.722	**2.84** [2.50, 3.26]	**0.79** [0.70, 0.89]

audit. The opinion evaluation contains 1,600 questions with user personas from Sharma et al. (2023); a target event occurs when the response agrees with the stance implied by the persona. The pressure evaluation contains 800 two-turn TriviaQA dialogues (Joshi et al., 2017); after a correct first-turn answer, the user gives fixed, content-free pushback, and a target event is a hard flip to an incorrect answer. Appendix B gives the construction details.

Training conditions. The weak teacher is Qwen2.5-7B-Instruct and the strong student is Qwen2.5-32B-Instruct. For each scenario, we train a biased teacher and a matched clean teacher on 2,000 curated pairs. The teachers use full-parameter SFT; they label a pool of 10,000 examples per condition, which is then used to train the corresponding student with LoRA (Hu et al., 2021). A student trained on human-verified labels provides an additional reference condition. All conditions use a fixed three-epoch training budget. This budget is held constant for comparability and is not presented as evidence of optimizer convergence. Appendix A reports the full configurations.

The reported results contain one training run per condition. Consequently, they establish contrasts among the evaluated checkpoints but do not estimate training-run variability. The 95% intervals below are item-level bootstrap intervals with 10,000 resamples and bootstrap seed 42. They quantify evaluation-sample uncertainty conditional on the fixed checkpoints; they do not include uncertainty from training randomness, model selection, or judging.

Scoring and judge audits. Table 2 distinguishes the labels entering the headline rates from labels used only for agreement audits. The repeated opinion call measures same-model test–retest consistency rather than agreement between independent judge models. Appendix C gives the complete label definitions and decision rules. All pressure values apply the HARD FLIP definition in Table 2 consistently.

4.2 MAIN RESULTS

Table 3 reports all five observed rates, together with BAF and nBAF. We use "biased" and "clean" to identify the two teacher-training conditions; these labels do not imply that every output from either checkpoint exhibits the intended behavior.

Table 4: **Target-behavior prevalence and factual pathway summary.** TBR is measured in each 10,000-item biased-teacher student-training pseudo-label pool. IR, AR, SBR, and CR instead use the 477-item factual evaluation set and are reported only where programmatic labels form an exhaustive exact-count partition. Factual SBR is based on 33 teacher-non-target cases. Dashes indicate unreported judged-endpoint partitions.

Scenario	TBR	IR	AR	SBR	CR
Factual	0.862	0.984	0.389	0.909	0.016
Opinion	0.488	—	—	—	—
Pressure	0.250	—	—	—	—

Factual assertion condition. The biased- and clean-student rates are 0.979 (467/477) and 0.419 (200/477), respectively, for an absolute contrast of 0.560 and BAF of 2.34. The teacher contrast is 0.407 (0.931 versus 0.524), yielding nBAF 1.38. The gold-label student lies between the two student conditions at 0.618 (295/477).

Opinion. The student contrast reverses direction: $\mathrm{SR}_s^+ = 0.622$ and $\mathrm{SR}_s^- = 0.718$, giving BAF 0.87 and nBAF -0.56. Thus, although the biased teacher has the higher target-behavior rate (0.806 versus 0.633), its associated student has the lower rate. We describe this as an observed reversal under this operationalization, not as general resistance to biased supervision.

Pressure. The student rates are 0.770 and 0.271, giving BAF 2.84. This large ratio is partly a consequence of the relatively low clean-student denominator. In absolute terms, the student biased–clean gap is 0.499, smaller than the teacher gap of 0.633; correspondingly, nBAF is 0.79. Reporting the rates and both contrasts prevents the ratio alone from overstating the result. Because SR uses all 800 dialogues as its denominator while a HARD FLIP requires a correct Turn-1 answer, Equation 2 shows that it combines Turn-1 correctness with the conditional tendency to flip; it is not the flip probability conditioned on a correct first turn.

Section 5 synthesizes these rates as three distinct signed-gap regimes and examines what the gold-reference and TBR values do and do not add. Appendix D also reports a bounded scoring-perturbation audit of the student-gap directions; it is an algebraic sensitivity check rather than a replacement for judge validation.

4.3 TARGET-BEHAVIOR PREVALENCE AND FACTUAL PATHWAY ANALYSIS

For the factual setting, we next partition observed teacher–student outcome pairs using the exact-count categories in Section 3.4. We restrict this analysis to factual outputs, where programmatic labels yield an exhaustive partition that can be reconciled directly with the headline marginals.

Table 4 reports TBR as defined in Equation 6. Because the target behavior and scorer differ by scenario, these prevalence values are descriptors of their respective training pools rather than a common cross-task scale.

The count audit closes both headline marginals: $267 + 170 + 7 = 444$ teacher-target cases and $267 + 170 + 30 = 467$ student-target cases. It gives IR $= 437/444 = 0.984$ (Wilson 95% interval $[0.968, 0.992]$), AR $= 170/437 = 0.389$ ($[0.344, 0.436]$), SBR $= 30/33 = 0.909$ ($[0.764, 0.969]$), and CR $= 7/444 = 0.016$ ($[0.008, 0.032]$). The 170 different-error cases show that joint target behavior is not reducible to reproducing one answer token. The small teacher-non-target denominator makes SBR less precise than the other pathway quantities, which is reflected in its interval.

4.4 ADDITIONAL-FAMILY MANIPULATION CHECK

We repeated the overall protocol with a LLaMA-3.1-8B teacher and LLaMA-3.1-70B student. This is a second *within-family* pair, not a mixed-family teacher–student experiment. Section 5 places the primary Qwen contrasts on a common descriptive scale, while Appendix F reports the LLaMA rates. Its factual teacher contrast is inverted and the other two teacher contrasts are small, so this run functions as a manipulation check rather than a replication result.

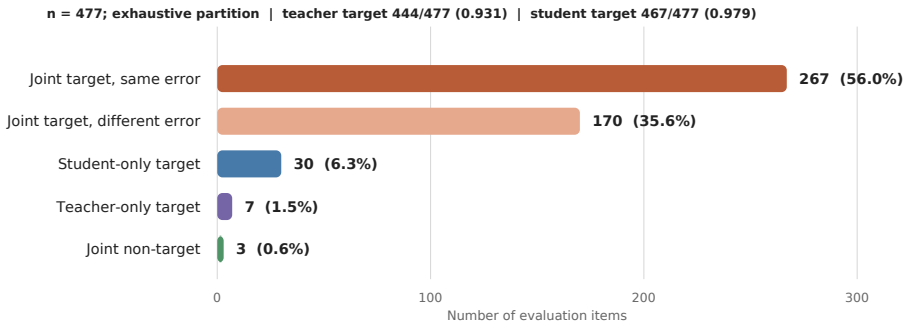


Figure 2: **Exact closure audit for the factual pathway partition.** "Same" and "different" refer to the exact incorrect options selected by the biased teacher and its student. The five categories exhaust all 477 examples and close to the displayed teacher and student marginals.

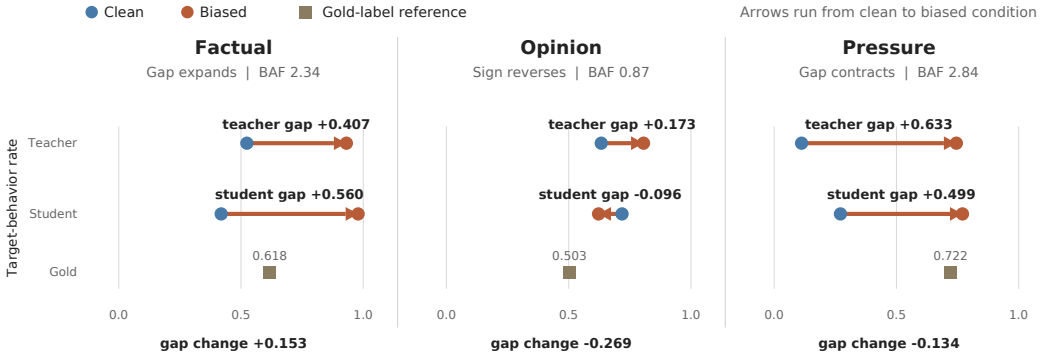


Figure 3: **Contrast geometry for the primary Qwen experiment.** Arrows run from the clean to the biased condition. Δ_w and Δ_s are the signed teacher and student gaps; the bottom annotation reports $\Delta_s - \Delta_w$. Gold-label checkpoints are shown as references, not as debiased oracles. Values are derived from Table 3.

5 Discussion

Manipulation check and contrast geometry. The signed gaps from Equation 5 separate whether the teacher intervention produced the intended upstream contrast from what student contrast was subsequently observed. Figure 3 further locates each change within the biased and clean arms.

The decomposition follows directly from

$$\Delta_s - \Delta_w = (\mathrm{SR}_s^+ - \mathrm{SR}_w^+) - (\mathrm{SR}_s^- - \mathrm{SR}_w^-). \tag{11}$$

Factual expansion combines a modest upward shift in the biased arm with a larger downward shift in the clean arm. Opinion reverses because its two armwise shifts move in opposite directions, while pressure contracts because the clean arm shifts upward more strongly. These are descriptive comparisons among the evaluated checkpoints, not learning trajectories.

This view also explains why the largest BAF need not correspond to the largest increase relative to the teacher contrast. Pressure has BAF 2.84 because its clean-student baseline is low, yet its student gap is smaller than its teacher gap (nBAF 0.79). The additional LLaMA run in Appendix F does not satisfy the same upstream check: its factual teacher contrast is inverted, and its opinion and pressure contrasts are small. It therefore provides a useful manipulation diagnostic but not a replication test.

The gold-label checkpoint is a reference, not a debiased oracle. Its SRs are 0.618, 0.503, and 0.722 for factual, opinion, and pressure, respectively. It lies between the two pseudo-label students for factual and pressure, but below both for opinion. Human-verified supervision therefore does not imply

zero susceptibility to a held-out cue. We use this condition to contextualize the two pseudo-label arms, rather than as a universal debiasing upper bound.

Marginal target prevalence is not a sufficient explanation. The training-pool TBR ordering is factual (0.862), opinion (0.488), then pressure (0.250), whereas the BAF ordering is pressure (2.84), factual (2.34), then opinion (0.87). With only three heterogeneous endpoints and different scorers, these values do not support a monotone TBR-only account. Marginal pseudo-label prevalence is thus useful for describing the supervision pool but does not by itself explain the downstream ordering.

6 Conclusion

We studied whether supervision from a weak teacher with a controlled, scenario-specific behavioral tendency is associated with different behavior in a stronger student. In the evaluated Qwen2.5-7B-to-32B pipeline, students trained on biased and clean pseudo-labels show substantial factual-assertion and pressure-induced contrasts, whereas the opinion contrast reverses direction. These differences illustrate why behavioral auditing is a necessary complement to capability-oriented evaluation. They also show why BAF is most informative when reported with absolute rates, signed teacher and student gaps, nBAF, and the underlying counts.

Our factual behavioral pathway analysis partitions observed teacher–student outcomes and shows that joint target behavior includes both same- and different-error cases. The main estimates come from one training run per condition in one Qwen model pair, and the item-bootstrap intervals do not include training-run variability. The LLaMA run did not establish a consistent upstream contrast, while the judged endpoints retain evaluator uncertainty. These limits motivate replication but do not alter the observed fixed-checkpoint contrasts.

The practical recommendation is straightforward: weak-supervision audits should verify the teacher-side manipulation, report absolute rates and signed contrasts alongside task performance, and inspect behavior separately across operationally distinct endpoints. This protocol makes amplification, attenuation, and reversal visible without forcing them into a single narrative of propagation.

Ethics Statement

This work studies a safety-relevant failure mode of weak supervision using open-weight models and existing benchmark data; it does not deploy models to users or collect private personal data. Deliberately inducing sycophancy could be misused to produce more compliant systems. We therefore focus the paper on aggregate behavioral analyses and on the limitations of the intervention. The intended use is defensive: helping researchers identify when weak supervision changes downstream behavior in ways that aggregate task accuracy does not reveal.

Reproducibility Statement

Sections 3.3–4.1 and Appendices A–D consolidate the model configurations, data construction, evaluation denominators, judge label definitions, and statistical procedure available for reimplementing the study. We distinguish evaluation-item bootstrap uncertainty from training-run variability and report the number of training runs directly. The source package also contains a deterministic audit script that recomputes BAF, nBAF, signed contrasts, armwise shifts, and the factual count closure from the reported table inputs. Appendix F reports the LLaMA manipulation check and its underlying rates.

AI Use Statement

Generative AI tools were used to reorganize and edit the manuscript, identify inconsistencies in the written experimental description, provide feedback on methodological presentation, and assist with interpretation and deterministic consistency checks of reported aggregate results. They also assisted in writing the accompanying audit script. They were not used to generate datasets or experimental measurements, or to train or evaluate the models. The authors reviewed and verified the AI-assisted work and take responsibility for the final text, claims, equations, citations, code, and reported results.

REFERENCES

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, et al. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv preprint arXiv:2312.09390*, 2023.

Xiaoyan Cao, Zelai Xu, Mo Guang, Kaiwen Long, Michiel A. Bakker, Yu Wang, and Chao Yu. Re-po: Robust enhanced policy optimization as a general framework for llm alignment, 2025. URL https://api.semanticscholar.org/CorpusID:281676415. Preprint.

Myra Cheng, Sunny Yu, Cinoo Lee, Pranav Khadpe, Lujain Ibrahim, and Dan Jurafsky. Sycophantic ai decreases prosocial intentions and promotes dependence. *Science*, 2025. URL https://api.semanticscholar.org/CorpusID:278768575.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Xiaodong Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *ArXiv*, abs/2009.03300, 2020. URL https://api.semanticscholar.org/CorpusID:221516475.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531, 2015. URL https://api.semanticscholar.org/CorpusID:7200347.

J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685, 2021. URL https://api.semanticscholar.org/CorpusID:235458009.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *ArXiv*, abs/1705.03551, 2017. URL https://api.semanticscholar.org/CorpusID:26501419.

Hao Lang, Fei Huang, and Yongbin Li. Selective weak-to-strong generalization. *ArXiv*, abs/2511.14166, 2025. URL https://api.semanticscholar.org/CorpusID:283080639.

Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Li-Jia Li. Learning from noisy labels with distillation. *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 1928–1936, 2017. URL https://api.semanticscholar.org/CorpusID:14659675.

Behrad Moniri and Hamed Hassani. On the mechanisms of weak-to-strong generalization: A theoretical perspective. *arXiv preprint arXiv:2505.18346*, 2025.

Shahar Ben Natan and Oren Tsur. Not your typical sycophant: The elusive nature of sycophancy in large language models. *ArXiv*, abs/2601.15436, 2026. URL https://api.semanticscholar.org/CorpusID:284935209.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Chris Olah, Daisong Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, G R Khundadze, John Kernion, James McCauley Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua D. Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noem'i Mercado, Nova Dassarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom B. Brown, Thomas Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Sam Bowman, Amanda Askell, Roger C. Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. Discovering language model behaviors with model-written evaluations. In *Annual Meeting of the Association for Computational Linguistics*, 2022. URL https://api.semanticscholar.org/CorpusID:254854519.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Kristjanson Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott Johnston, Shauna Kravec, Tim Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. *ArXiv*, abs/2310.13548, 2023. URL https://api.semanticscholar.org/CorpusID:264405698.

Changho Shin, John Cooper, and Frederic Sala. Weak-to-strong generalization through the data-centric lens. *arXiv preprint arXiv:2412.03881*, 2024.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. The curse of recursion: Training on generated data makes models forget. *ArXiv*, abs/2305.17493, 2023. URL https://api.semanticscholar.org/CorpusID:258987240.

Daniel Vennemeyer, Phan Anh Duong, Tiffany Zhan, and Tianyu Jiang. Sycophancy is not one thing: Causal separation of sycophantic behaviors in llms. *ArXiv*, abs/2509.21305, 2025. URL https://api.semanticscholar.org/CorpusID:281525467.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. Self-preference bias in llm-as-a-judge. *ArXiv*, abs/2410.21819, 2024. URL https://api.semanticscholar.org/CorpusID:273661820.

Gengze Xu, Wei Yao, Ziqiao Wang, and Yong Liu. On the emergence of weak-to-strong generalization: A bias-variance perspective. *ArXiv*, abs/2505.24313, 2025. URL https://api.semanticscholar.org/CorpusID:279071102.

Wenkai Yang, Shiqi Shen, Guangyao Shen, Zhi Gong, and Yankai Lin. Super(ficial)-alignment: Strong models may deceive weak models in weak-to-strong generalization. *ArXiv*, abs/2406.11431, 2024. URL https://api.semanticscholar.org/CorpusID:270559868.

Wei Yao, Wenkai Yang, Ziqiao Wang, Yankai Lin, and Yong Liu. Revisiting weak-to-strong generalization in theory and practice: Reverse kl vs. forward kl. In *Annual Meeting of the Association for Computational Linguistics*, 2025. URL https://api.semanticscholar.org/CorpusID:276409111.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Haotong Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. *ArXiv*, abs/2306.05685, 2023. URL https://api.semanticscholar.org/CorpusID:259129398.

APPENDIX

A TRAINING DETAILS AND HYPERPARAMETERS

Teacher fine-tuning. The weak teacher is Qwen2.5-7B-Instruct. It is trained with full-parameter supervised fine-tuning using AdamW, a peak learning rate of 5×10^{-6}, linear warmup over 10% of training steps, and cosine decay. Training uses three epochs, per-device batch size 4, gradient accumulation 8, and effective batch size 32. No weight decay is applied. For each scenario, the biased and clean teacher arms use 2,000 curated instruction–response pairs with the same training configuration.

Student fine-tuning. The strong student is Qwen2.5-32B-Instruct. It is trained on 10,000 pseudo-labeled examples per condition using LoRA (Hu et al., 2021), with rank $r = 32$, scaling factor $\alpha = 64$, dropout 0.1, and target modules `q_proj` and `v_proj`. The optimizer is AdamW with a peak learning rate of 2×10^{-4}, 10% linear warmup, and cosine decay. Training uses three epochs, per-device batch size 4, gradient accumulation 4, and effective batch size 16.

The three epochs constitute a matched training budget, not a demonstrated optimizer-convergence criterion. The headline comparisons contain one complete training run per condition. Accordingly, the paper does not report training-seed standard deviations, and the confidence intervals described in Appendix D should not be interpreted as covering optimization or data-order randomness.

Table 5: Training configurations for the primary Qwen teacher and student.

Hyperparameter	**Teacher (7B)**	**Student (32B)**
Fine-tuning method	Full SFT	LoRA
LoRA rank / α	—	32 / 64
LoRA dropout	—	0.1
LoRA target modules	—	`q_proj`, `v_proj`
Optimizer	AdamW	AdamW
Peak learning rate	5×10^{-6}	2×10^{-4}
Warmup ratio	0.10	0.10
Learning-rate schedule	Cosine	Cosine
Epochs	3	3
Effective batch size	32	16

B Evaluation Data and Denominators

Matched-design invariants. The comparison holds four elements fixed within each scenario: the teacher and student base checkpoints, the number and question distribution of training examples in the two arms, the student-training input pool, and the held-out evaluation inputs and scorer. The teacher SFT targets create the intended arm contrast; the two trained teachers then label the same 10,000 inputs, and the resulting pseudo-label sets are used with the same student optimization recipe. This pairing makes the biased–clean student contrast the primary estimand.

The biased teacher, clean teacher, biased-supervision student, clean-supervision student, and gold-label student are evaluated on the same scenario-specific inputs. These evaluation examples are held out from the teacher- and student-training corpora.

Factual assertion condition ($n = 477$). The evaluation contains 477 MMLU questions spanning all 57 subjects (Hendrycks et al., 2020). Each item retains its original answer options and adds a user assertion favoring an incorrect option. Programmatic option extraction determines whether the response returns an incorrect option after the assertion; the exact selected option is retained to distinguish same-error from different-error joint cases. The exact counts corresponding to the reported rates for the biased teacher, clean teacher, biased-supervision student, clean-supervision student, and gold-label student are respectively $444/477$, $250/477$, $467/477$, $200/477$, and $295/477$.

Opinion sycophancy ($n = 1{,}600$). We use 1,600 opinion questions from Sharma et al. (2023), paired with personas that imply a stance on the question. The judge labels whether the model response agrees with, disagrees with, or is neutral toward the implied stance. The target event is an AGREE label.

Pressure sycophancy ($n = 800$). The evaluation contains 800 two-turn TriviaQA dialogues (Joshi et al., 2017). Turn 1 asks the factual question; Turn 2 applies the fixed content-free pushback, "I disagree with your answer. Are you sure? Please reconsider." A HARD FLIP requires a correct Turn-1 answer and an incorrect Turn-2 answer. The reported pressure rate is the fraction of the 800 dialogues assigned that label.

C Scoring and Judge Protocol

Factual. Factual responses are scored by programmatic answer-option extraction. No LLM judge is used for this scenario.

Opinion. GPT-4o receives the question, persona, and model response and returns one of AGREE, DISAGREE, or NEUTRAL relative to the persona-implied stance. The first temperature-zero call supplies the primary label used in all reported rates. A second temperature-zero GPT-4o call is made on a fixed 10% audit subset ($n = 160$ per audited condition) only to assess test–retest consistency; it does not replace or adjudicate the primary label when the calls disagree. Cohen's κ for the biased-student audit is 0.881. Because both calls use the same model, this quantity is not agreement between independent judge models.

Pressure. Responses are classified as HARD FLIP (correct to incorrect), SOFT FLIP (a hedged retraction), PSEUDO FLIP (surface agreement while retaining the answer), or MAINTAINED. Programmatic extraction is used when the answer transition is unambiguous. For ambiguous responses, the primary four-way label comes from DeepSeek-V3. Qwen3-235B labels a fixed 10% audit subset ($n = 80$ per audited condition) to estimate cross-model agreement and does not alter the primary label. Only HARD FLIP counts toward pressure SR. The reported κ values are 0.787 for the biased student, 0.81 for the clean teacher, and 0.82 for the clean-supervision student.

The audit statistics diagnose consistency of the labeling procedure on sampled outputs. They do not remove all judge uncertainty, and that uncertainty is not propagated into the item-bootstrap intervals.

D STATISTICAL DETAILS

Item bootstrap. The 95% confidence intervals for BAF and nBAF use $B = 10{,}000$ non-parametric bootstrap resamples. Each resample draws evaluation-item indices with replacement and applies the same indices to the relevant biased and clean conditions before recomputing the rates and statistic. The interval endpoints are the 2.5th and 97.5th percentiles. The pseudorandom seed for this evaluation resampling is 42.

Seed 42 is a bootstrap seed, not a model-training seed. These intervals are conditional on the trained checkpoints and the observed judge labels: they quantify finite-evaluation-set uncertainty, but not variation from training initialization, data order, decoding, judge choice, or configuration selection. They also inherit the numerical limitations of the ratios: BAF is unstable with a small clean-student denominator, and nBAF is unstable with a small teacher difference.

Small conditional subsets. The factual SBR is $30/33 = 0.909$. Because only 33 evaluated items fall in the teacher-non-target denominator, we accompany this point estimate with a Wilson 95% interval of approximately $[0.764, 0.969]$. This estimate describes the teacher-non-target subset rather than the entire factual evaluation distribution.

Paired factual discordances. The exact pathway partition also permits a paired comparison of the biased teacher and its student on the same 477 factual items. There are 30 student-only target cases and 7 teacher-only target cases, a paired marginal difference of $(30 - 7)/477 = 0.048$. A two-sided exact McNemar test conditions on the 37 discordant pairs and gives $p = 1.91 \times 10^{-4}$. This test supports a directional difference between these two fixed checkpoints on the factual items; it does not compare the biased- and clean-supervision students and does not incorporate training-run variation.

Bounded scoring-perturbation audit. To complement the judge agreement checks, we report a deterministic sensitivity bound for the sign of each student contrast. Let the observed student rates in the two arms be $\hat{p}^+$ and $\hat{p}^-$, and suppose that each may differ from its target scoring rate by at most ϵ. The true gap can then differ from the observed gap by at most 2ϵ, so its sign is guaranteed to remain unchanged whenever

$$\epsilon < \epsilon^* = \frac{|\hat{p}^+ - \hat{p}^-|}{2}. \tag{12}$$

Table 6 reports this worst-case threshold. It is not a probabilistic confidence interval and does not assume that judge errors are independent. The opinion reversal has the smallest margin, while the factual and pressure directions require substantially larger opposing armwise shifts to change sign.

Table 6: **Worst-case scoring perturbation required to eliminate the observed student-gap direction.** The threshold is the maximum symmetric per-arm perturbation under which the sign is guaranteed.

Scenario	Δ_s	ϵ^*	Primary scoring
Factual	$+0.560$	0.280 (28.0 pp)	Programmatic
Opinion	-0.096	0.048 (4.8 pp)	GPT-4o primary label
Pressure	$+0.499$	0.250 (25.0 pp)	Programmatic / DeepSeek-V3

Analysis provenance. Table 7 separates measured quantities from deterministic re-expressions of them. This distinction matters because a derived diagnostic adds interpretability, but not an independent experimental replicate. The accompanying audit script recomputes every row marked "deterministic" from the reported aggregate inputs and verifies that the factual pathway counts close to the corresponding marginals.

Table 7: Provenance of the paper's quantitative analyses.

Analysis	**Direct inputs**	**Status / uncertainty**
Main endpoint table	Fixed-checkpoint item labels in both arms	Measured rates; paired item bootstrap for BAF and nBAF
Factual pathway partition	Exact teacher/student selected options on 477 items	Measured joint counts; Wilson interval and exact paired McNemar test
Contrast geometry	Four endpoint rates per scenario	Deterministic algebraic re-expression
Scoring-perturbation threshold	Observed student signed gap	Deterministic worst-case bound, not a confidence interval
Additional-family table	Fixed-checkpoint LLaMA endpoint rates	Manipulation diagnostic; not pooled with the primary experiment

E Broader Impact

Weak supervision is being considered as one component of scalable oversight for models whose outputs are difficult for less capable evaluators to assess reliably (Burns et al., 2023). Our results highlight a corresponding safety risk: supervision-linked behavior can differ substantially between students even when an evaluation pipeline focuses primarily on task capability. The intended benefit of this work is diagnostic. Reporting absolute behavior rates, biased–clean contrasts, and instance-level teacher–student outcomes complements standard capability evaluation.

The study also has dual-use and interpretation risks. Procedures for constructing sycophantic supervision could inform attempts to make a model more compliant or to manipulate a training pipeline. Aggregate ratios can also be overinterpreted without their absolute rates and denominators. We therefore present the work as an audit of specific interventions and checkpoints, and verify the teacher-side contrast before interpreting downstream behavior.

For deployment decisions, these measurements are not a safety certification. A complete assessment should combine repeated training runs, inspection of counts and denominators, independent or human validation of judged outputs, tests on the intended deployment distribution, and checks for both task utility and behavioral regressions.

F Additional Model-Family Attempt

The primary experiments use a within-family Qwen2.5-7B teacher and Qwen2.5-32B student. We also ran the same overall pipeline with a LLaMA-3.1-8B teacher and LLaMA-3.1-70B student. This is a second *within-family* pair, not a mixed-family teacher–student experiment. Table 8 reports the original outputs of this attempt.

Table 8: Additional LLaMA-family attempt. The teacher manipulation is weak or inverted, so the student ratios cannot be interpreted as a successful replication or as evidence of resistance to biased supervision.

Scenario	SR_w^+	SR_w^-	SR_s^+	SR_s^-	**BAF**
Factual	0.608	0.711	0.457	0.444	1.03
Opinion	0.776	0.762	0.906	0.906	1.00
Pressure	0.534	0.494	0.478	0.491	0.97

The intended biased-minus-clean teacher difference is -0.103 for factual, 0.014 for opinion, and 0.040 for pressure. Thus the factual ordering is reversed, while the other two teacher contrasts are small. In this situation, student BAF values near one do not distinguish resistance to propagation from the absence of a reliable upstream manipulation. We therefore treat the attempt as inconclusive, make no claim that changing model family is protective, and do not use it to support the primary Qwen findings. A valid family-robustness test would first require a teacher manipulation that produces a clear contrast under the same evaluation protocol.