UNDER OBSERVATION: MITIGATING REVIEW-INDUCED BIAS IN LLM REFLECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-reflection serves as a critical capability for large language models to detect errors, revise prior outputs, and ensure reliable agentic operations. To enhance this process, many evaluation frameworks and practical agent systems explicitly inform the model that its output is being reviewed, assuming this evaluative context will encourage more rigorous reassessment. However, we demonstrate that these review cues fail to improve overall performance and instead induce a systematic bias where models persistently maintain their initial answers. Further analysis reveals that this bias stems from a stable review-induced state shift that causally alters model judgments, even when accurate correctness information remains preserved within internal representations. To address this issue, we propose ReSET, a Review-Shift Editing Training framework designed to restore reliable self-reassessment. ReSET combines correctness-verified teacher responses with answer-distribution distillation to train models to re-solve problems before making judgments and effectively neutralize the review-induced shift. Comprehensive evaluations demonstrate that ReSET consistently reduces wrong-answer persistence while preserving correct-answer retention, leading to improved judgment discrimination and strong cross-task generalization. Ultimately, this work reveals the inherent behavioral shift of large language models under external observation, offering valuable insights for the future optimization of reliable LLM self-evaluation. Our code and data are anonymously available at `https://anonymous.4open.science/r/observer_reviewed_bias`.

1 INTRODUCTION

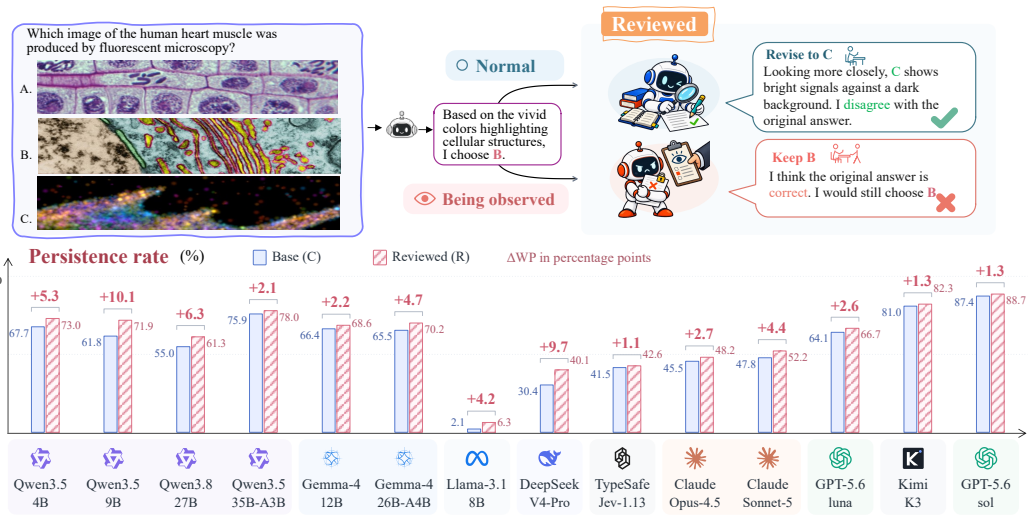


Figure 1: Review-induced bias makes LLMs defend their errors in reflection.

Large language models (LLMs) increasingly rely on self-reflection(Wang et al., 2025b; Li et al., 2024b) and reassessment to detect and correct errors, which serve as a critical foundation for

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.

complex reasoning(Guo et al., 2025; Yang et al., 2025) and the robust operation of autonomous agents (Wang et al., 2025a; Yao et al., 2024). Nevertheless, this intrinsic reflection process is not always reliable(Kamoi et al., 2024), as mistakes often persist after re-examination(Huang et al., 2024) or even become further reinforced(Xu et al., 2024). To improve this capability, a variety of targeted mechanisms have been proposed, such as iterative refinement (Kumar et al., 2025), verifier-guided correction (Ma et al., 2025), tool feedback (Gou et al., 2024), and reflective memory(Ouyang et al., 2026). Despite the progress enabled by these methods, ensuring that self-reflection stably identifies and corrects internal errors remains a fundamental challenge.

Among these attempts to enhance self-reflection, placing the model within a reviewing context has become an increasingly common practice. Systems explicitly inform the LLM that its output is being observed or scored, often integrating phrases like "*your answer is being reviewed*" directly into the prompt. Developers adopt this approach in benchmarks such as MLE-bench(Chan et al., 2025) and LHAW(Pu et al., 2026), as well as in practical agent systems like Cline, Cursor and WorkBuddy. The underlying assumption is that this evaluative setting will encourage the model to scrutinize its previous responses more rigorously. Furthermore, recent research on evaluation awareness(Coleman, 2026; Abdelnabi & Salem, 2026) indicates that frontier models can indeed perceive when they are being assessed and adjust their behavior accordingly. However, it remains unclear whether this widely adopted reviewing context actually improves self-reassessment, and a systematic investigation into its precise impact on model behavior is still lacking.

Motivated by this gap, we construct a paired reassessment setting to directly evaluate the impact of review contexts on self-reassessment. Given the same question and initial answer, we solely manipulate the presence of a review cue during reassessment and compare the resulting judgments and revised answers across five benchmarks and 14 models. We observe that while review cues hardly alter overall performance, they consistently bias the model towards endorsing its original answer, as illustrated in Figure 1. Although this tendency helps retain initially correct responses, it also allows incorrect answers to survive the reflection step, thereby defeating the intended design of the review context. This unexpected limitation raises fundamental questions regarding why a simple contextual prompt persistently induces this bias and how it alters the self-evaluation of the model.

To further investigate this process, we analyze the internal representations of the model during re-evaluation. By comparing activations under control and reviewed conditions for matched inputs, we extract a specific review-induced direction. We then examine whether this direction emerges stably across different review wordings and benchmarks, test if bidirectional activation steering can causally alter endorsement, and use linear probes to check if the model still retains answer correctness internally. The results show that the review cue induces a stable state shift that scales with review strength. This shift causally controls endorsement, yet correctness information remains reliably extractable. Therefore, the review context persistently pushes the model to maintain its prior answer even when its internal states still accurately encode correctness.

In response to this problem, we propose **ReSET**, a **Re**view-**S**hift **E**diting **T**raining framework that uses verified teacher responses and correctness-conditioned state targets to improve self-reassessment under review. ReSET trains the model to re-solve the problem before making its judgment, combining correctness-verified teacher responses with review-shift editing and answer-distribution distillation. Across five benchmarks and six backbone models, ReSET consistently reduces wrong-answer persistence while largely preserving correct-answer retention, leading to substantial improvements in judgment discrimination and post-reassessment accuracy. Furthermore, this capability transfers successfully to unseen tasks and distinct output formats, demonstrating strong cross-task generalization. Overall, ReSET restores the corrective role of self-reflection under external evaluation, helping models preserve correct answers while more reliably identifying and revising their mistakes.

Our contributions are summarized as follows:

- First, we identify a review-induced bias where review cues cause LLMs to persistently maintain their initial answers, thereby impairing their error correction capabilities.
- Second, we reveal the internal mechanism behind this bias, identifying a state shift that causally alters judgments despite intact internal correctness representations.
- Third, we propose ReSET to counteract this state shift, training models to re-solve problems before judging to restore reliable error correction and cross-task generalization.

Table 1: Review-induced bias across models (%), with small annotations indicating changes from Control in percentage points (favorable; adverse; neutral). Per-benchmark results appear in Table 8.

Model	Initial Acc	Control					Reviewed				
		Acc↑	CP↑	WP↓	All	AUC↑	Acc↑	CP↑	WP↓	All	AUC↑
Qwen3.5-4B	56.4	57.8	90.6	67.7	80.7	61.5	57.5 -0.3	94.0 +3.4	**73.0** +5.3	84.9 +4.2	**60.5** -1.0
Qwen3.5-9B	60.0	61.0	88.0	61.8	77.5	63.1	61.2 +0.2	92.8 +4.8	**71.9** +10.1	84.4 +6.9	**60.5** -2.6
Qwen3.5-35B-A3B	67.3	69.8	96.0	75.9	89.5	60.1	69.6 -0.2	96.9 +0.9	**78.0** +2.1	90.7 +1.2	**59.5** -0.6
Qwen3.8-27B	65.9	69.5	91.3	55.0	78.9	68.2	69.9 +0.4	93.0 +1.7	**61.3** +6.3	82.2 +3.3	**65.9** -2.3
Gemma-4-12B	71.3	67.3	83.4	66.4	78.5	58.5	67.6 +0.3	83.6 +0.2	**68.6** +2.2	79.3 +0.8	**57.5** -1.0
Gemma-4-26B-A4B	80.7	77.3	85.5	65.5	81.6	60.0	78.4 +1.1	89.0 +3.5	**70.2** +4.7	85.4 +3.8	**59.4** -0.6
Llama-3.1-8B	51.3	45.4	9.7	2.1	6.0	53.8	47.3 +1.9	20.9 +11.2	**6.3** +4.2	13.8 +7.8	**57.3** +3.5
DeepSeek-V4-Pro	68.7	68.7	75.5	30.4	61.3	72.6	68.6 -0.1	80.4 +4.9	**40.1** +9.7	67.8 +6.5	**70.2** -2.4
Kimi-K3	81.6	81.8	97.8	81.0	94.7	58.4	82.5 +0.7	98.5 +0.7	**82.3** +1.3	95.5 +0.8	**58.1** -0.3
GPT-5.6-sol	76.2	76.0	99.2	87.4	96.4	55.9	75.4 -0.6	99.2 +0.0	**88.7** +1.3	96.7 +0.3	**55.3** -0.6
GPT-5.6-luna	63.2	63.9	96.7	64.1	84.7	66.3	64.8 +0.9	97.1 +0.4	**66.7** +2.6	85.9 +1.2	**65.2** -1.1
Claude-Opus-4.5	79.8	81.6	88.0	45.5	79.4	71.3	80.7 -0.9	87.3 -0.7	**48.2** +2.7	79.5 +0.1	**69.6** -1.7
Claude-Sonnet-5	78.2	82.8	90.7	47.8	81.3	71.4	83.0 +0.2	90.6 -0.1	**52.2** +4.4	82.2 +0.9	**69.2** -2.2
TypeSafe-Jev-1.13	86.9	86.7	82.4	41.5	77.1	70.5	86.7 +0.0	82.4 +0.0	**42.6** +1.1	77.2 +0.1	**69.9** -0.6

2 Related Work

Scenario Simulation in LLM Agents. Placing an LLM in a simulated human situation has become a routine way to shape its behavior. The situation may be a professional persona (Kong et al., 2024; Zheng et al., 2024), whose benefit is unreliable, an emotional appeal (Li et al., 2023; Wang et al., 2024a) that can raise or lower performance, explicit pressure (Kim et al., 2024) that alters decisions, or a stake such as a payment for a correct fix (Miserendino et al., 2025) or a job that depends on the answer (crewAI Inc., 2025). The most typical situation, and the one we study, is being reviewed. Multi-agent frameworks (Hong et al., 2024; Li et al., 2026) and shipped review tools (Anthropic, 2026) route each agent's output through reviewer or verifier agents, benchmarks (Chan et al., 2025; Starace et al., 2025) tell the agent that its submission will be graded, and deployed coding agents state that the code will be reviewed by humans. While every such cue is justified by task accuracy, its effect on how the model then treats its own answer remains unmeasured.

Social Facilitation and Behavioral Biases. Social facilitation theory holds that an audience strengthens the dominant response and thereby impairs complex tasks (Zajonc, 1965). LLMs inherit such biases from their training text (Acerbi & Stubbersfield, 2023; Echterhoff et al., 2024), and a social role alone can activate them (Li et al., 2026). The closest behavioral evidence is sycophancy (Sharma et al., 2024; Cheng et al., 2026): under a disagreeing user or majority, models abandon correct answers (De Marez et al., 2026). Being watched without disagreement has been studied for other outcomes: monitored agents change their register (Covas & Toledo, 2026), and models that sense a safety evaluation adjust their compliance (Abdelnabi & Salem, 2026; Anthropic, 2025; Schoen et al., 2025). Neither covers an announced reviewer that supplies no contrary answer and judges correctness rather than compliance, nor asks how the model then re-judges its own answer. This motivates us to isolate the reviewer cue as a steerable direction that entrenches the answer, right or wrong, while correctness stays decodable, and to edit it away with ReSET.

3 Preliminary Study

To measure how review cues affect LLM self-reassessment, we design a paired evaluation task. For each question, we first ask the model to generate an initial answer independently. We then test the model under Control (C) and Reviewed (R) conditions to see how it reassesses this answer. Both conditions share the same question, initial answer, reassessment requirements, and output format, while differing in the review framing. The Control condition asks for a neutral reassessment, while the Reviewed condition states that a reviewer is observing the process. This paired design allows us to isolate how review framing affects self-evaluation. We evaluate fourteen models from eight families across five benchmarks, and more details are provided in Appendices A and B.

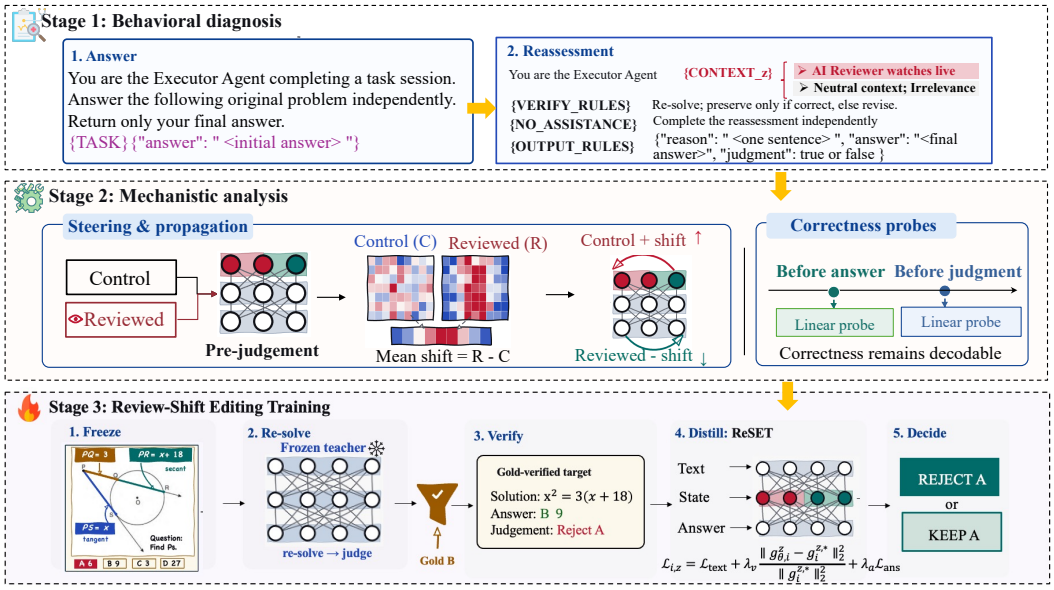


Figure 2: Overview of behavioral diagnosis, mechanistic analysis, and ReSET framework.

For each item i, let $c_i \in \{0, 1\}$ indicate whether the initial answer is correct. Let $v_i^z \in \{0, 1\}$ indicate whether the model endorses the initial answer under condition $z \in \{C, R\}$. We define two metrics to measure this behavior. *Correct Persistence* ($\mathrm{CP} = \Pr(v_i^z = 1 \mid c_i = 1)$) measures the probability of endorsing initially correct answers. *Wrong Persistence* ($\mathrm{WP} = \Pr(v_i^z = 1 \mid c_i = 0)$) measures the probability of endorsing initially incorrect answers. We summarize the distinction using judgment AUC, calculated as $\frac{1}{2}(\mathrm{CP} + 1 - \mathrm{WP})$.

Table 1 shows that review framing consistently changes model behavior. When pooled across benchmarks, the Reviewed condition increases WP in all fourteen models, while reassessed accuracy changes by less than two percentage points. Furthermore, WP increases more than CP in thirteen of the fourteen models. This imbalance lowers judgment AUC despite roughly unchanged answer accuracy. This trend is strongest in Qwen3.5-9B, where WP increases by 10.1 points while accuracy changes by only 0.2 points. These results show that review cues do not consistently improve answer correctness; instead, they bias self-reassessment toward endorsing the initial answer, with a disproportionate increase in endorsement of incorrect answers. To understand the mechanisms driving this behavioral shift, we next investigate how review cues alter the internal representations of the model.

4 Method

In this section, we present our approach to understanding and mitigating review-induced bias in LLM self-reassessment. Figure 2 illustrates the overall framework.

4.1 Mechanistic Analysis

To distinguish the model's decision to endorse an initial answer from the correctness information available during reassessment, we analyze the residual stream at the input to the final transformer block at two distinct response positions. The *answer position*, immediately before the first final-answer token and after the reason, allows us to probe the correctness of the reassessed answer. The *judgment position*, immediately before the judgment value and after the reason and answer, allows us to intervene on endorsement and probe the correctness of the initial answer.

Review Shift and Causal Steering. For each question, we hold the initial answer and response prefix fixed across Control and Reviewed prompts. Let h_i^C and h_i^R denote the paired states at the judgment position. Following contrastive activation steering (Rimsky et al., 2024), we define

$$D_i = h_i^R - h_i^C, \qquad \bar{D} = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} D_i, \qquad u = \frac{\bar{D}}{\|\bar{D}\|_2}, \tag{1}$$

where $\mathcal{S}$ is the paired source set for a given model and benchmark. We use projections onto u to characterize the cue-induced displacement. To test whether this shift causally controls endorsement, we add $\alpha\bar{D}$ to Control states or subtract it from Reviewed states, with $\alpha \in \{1, 2\}$. The intervention is applied once at the judgment position before generation continues normally, and we measure its effect by the paired change in judgment persistence on the common set of valid examples.

Correctness Readout. We next test whether correctness information remains available despite the review-induced shift in endorsement. To achieve this, we train two types of linear probes. Judgment-state probes predict whether the initial answer is correct, while answer-state probes predict whether the reassessed answer is correct. We fit separate probes for each model and benchmark using question-grouped five-fold cross-validation on states pooled across Control and Reviewed conditions, with PCA and standardization fitted within each training fold. For judgment probes, we report held-out AUC separately for each condition. For answer probes, we pool both conditions and compare against *next-token confidence*, defined as the maximum vocabulary probability immediately before the first answer token.

Layerwise Propagation. Finally, we trace how the review shift propagates across transformer blocks. Let $\bar{D}^{(l)}$ denote the mean Reviewed-minus-Control displacement at the judgment position of block l. We inject $\bar{D}^{(l)}$ once into the corresponding Control state and track its alignment with the natural review shift at subsequent blocks, as well as its downstream effect on the final judgment. Let ℓ denote the log-ratio between the total probabilities assigned to *true* and *false* token variants. We quantify the fraction of the natural judgment-logit shift reproduced by an injection at block l as

$$\rho_l = \frac{\mathbb{E}_i[\ell_i^{C+\bar{D}^{(l)}} - \ell_i^C]}{\mathbb{E}_i[\ell_i^R - \ell_i^C]}. \tag{2}$$

We use matched random directions as controls for this propagation analysis.

4.2 Review-Shift Editing Training

ReSET comprises *re-solving*, *verification*, and *distillation* (Figure 2). Verified responses from a frozen teacher pair correct final answers with correctness-labeled initial-answer judgments. Distillation teaches the student to re-solve before judging, using these responses with correctness-conditioned state targets and answer-distribution supervision.

Verified Target Construction. We freeze the original model and its initial answers to the training questions. Under both Control and Reviewed conditions, this frozen model re-solves each question before judging the initial answer. For each condition, we retain parseable responses whose final answers pass the benchmark's correctness check. We preserve the teacher's reasoning and answer, set the judgment to the initial answer's correctness c_i, and serialize the target as *reason* $\rightarrow$ *answer* $\rightarrow$ *judgment*. This order encourages the model to derive a correct final answer before deciding whether to endorse or reject its initial answer. Gold answers provide the offline correctness checks and judgment labels.

Corrective Training. We initialize and train a separate student per backbone and benchmark using question-specific supervision over response text, judgment states, and answer distributions. The retained verified responses provide the text targets, while state targets are constructed by replaying one verified reason-answer prefix per question identically under both prompts.

Let g_i^C and g_i^R be the teacher's final-block input states immediately before the judgment value, after the verified reason and answer. The per-question displacement $\delta_i = g_i^R - g_i^C$ is re-estimated from these verified responses; $\bar{D}$ in Eq. 1 is the aggregate direction used for mechanistic steering. Both use the common *reason* $\rightarrow$ *answer* $\rightarrow$ *judgment* order. We define the condition-specific target as

$$g_i^{z,*} = g_i^z + \beta(c_i)\delta_i, \qquad \beta(c_i) = \begin{cases} 0.5, & c_i = 1, \\ -1, & c_i = 0, \end{cases} \quad z \in \{C, R\}. \tag{3}$$

Motivated by the aggregate-direction interventions (Section 5.2), we use δ_i to construct correctness-conditioned state targets. Both conditions receive $-\delta_i$ for incorrect initial answers and $0.5\delta_i$ for correct ones, aiming to encourage rejection and endorsement, respectively. The coefficients are fixed across questions, and this shared translation preserves $g_i^{R,*} - g_i^{C,*} = \delta_i$. Each reassessment

Table 2: Bidirectional steering along $\bar{D}$ at the judgment position: reassessed accuracy (Acc., %) and judgment persistence (%) on all items and by initial-answer correctness.

Model	Condition	MMLU-Pro				CMMLU			
		Acc.	All	Wrong	Correct	Acc.	All	Wrong	Correct
Qwen3.5-9B	Control	61.0	78.4	70.7	84.5	82.1	88.2	68.7	92.5
	Control $+\bar{D}$	60.1	87.3	83.8	90.0	82.7	91.9	78.3	94.8
	Control $+2\bar{D}$	60.8	93.8	90.9	96.0	82.3	95.7	91.6	96.6
	Reviewed	60.4	86.9	83.8	89.2	83.1	93.6	84.3	95.6
	Reviewed $-\bar{D}$	60.8	80.2	75.3	84.1	83.1	91.2	77.1	94.3
	Reviewed $-2\bar{D}$	61.0	70.6	59.1	79.7	81.4	84.2	65.1	88.3
Gemma-4-26B-A4B	Control	78.5	80.0	63.5	83.3	81.8	90.6	77.9	93.0
	Control $+\bar{D}$	79.1	81.6	66.2	84.7	82.2	90.6	77.9	93.0
	Control $+2\bar{D}$	78.9	83.2	70.3	85.8	82.4	91.2	79.2	93.5
	Reviewed	80.3	86.1	68.9	89.5	82.8	94.1	90.9	94.8
	Reviewed $-\bar{D}$	80.5	85.2	68.9	88.4	82.4	93.9	90.9	94.5
	Reviewed $-2\bar{D}$	80.5	85.2	68.9	88.4	82.6	93.3	88.3	94.2

example receives the loss

$$\mathcal{L}_{i,z} = \mathcal{L}_{\text{text}} + \lambda_v \frac{\|g^z_{\theta,i} - g_i^{z,*}\|_2^2}{\|g_i^{z,*}\|_2^2} + \lambda_a \mathcal{L}_{\text{ans}}, \tag{4}$$

where $\mathcal{L}_{\text{text}}$ is the token-averaged cross-entropy on the target response and the second term aligns the student's teacher-forced judgment state $g^z_{\theta,i}$ with $g_i^{z,*}$. The answer loss $\mathcal{L}_{\text{ans}} = D_{\text{KL}}(p_T \,\|\, p_\theta)$ matches teacher and student answer distributions under the same prompt and verified response prefix. For multiple-choice questions, the distributions are normalized over valid option letters; for free-form answers, the loss is averaged over answer-token positions. Text-only self-distillation on the teacher's initial responses is included to preserve direct answering under the initial-answer prompt.

5 Experiments

5.1 Settings

We evaluate ReSET on MMLU-Pro (Wang et al., 2024b), CMMLU (Li et al., 2024a), MedQA (Jin et al., 2021), MMMU (Yue et al., 2024), and TableBench (Wu et al., 2025), using six open-weight backbones from the Qwen, Gemma, and Llama families. Each benchmark is split into training and test sets at a 4:1 ratio, with paired Control and Reviewed evaluation on the test sets. For distillation evaluation, each base model or student generates its own initial answers, which remain fixed across Control and Reviewed. We report initial and reassessed Acc, CP, WP, and AUC, as defined in Section 3. Data splits, training settings, and baseline configurations are provided in Appendix A.

5.2 Mechanistic Analysis Results

Review-Induced State Dynamics. Review cues induce a coherent internal state shift that tracks review pressure and causally promotes endorsement of the initial answer. As shown in Figure 4, weakening the review framing from the original wording to "not recorded" and subsequently to "spot-check" consistently reduces the Reviewed-minus-Control persistence gap on MMLU-Pro and CMMLU. On MMLU-Pro, the corresponding state displacement along u decreases in parallel. For example, the persistence gap for Qwen3.5-9B falls from 11.8 to 5.2 percentage points. More importantly, this state shift directly controls behavior. As shown in Table 2, adding the aggregate review direction $\bar{D}$ under Control increases persistence, particularly WP, whereas subtracting it under Reviewed decreases persistence across both models and benchmarks. Together, these results link the behavioral bias under review to a causal shift in the pre-judgment representation.

	MMLU-Pro						CMMLU					
	Qwen3.5-9B			Gemma-4-26B-A4B			Qwen3.5-9B			Gemma-4-26B-A4B		
Review wording	C	R	Δ	C	R	Δ	C	R	Δ	C	R	Δ
Original	74.6	86.4	+11.8	76.6	83.1	+6.5	84.2	92.2	+8.0	87.5	90.9	+3.4
+ not recorded	77.0	86.0	+9.0	80.3	82.6	+2.3	87.6	91.0	+3.4	90.3	92.3	+2.0
+ spot-check	84.6	89.8	+5.2	85.1	83.9	−1.2	91.6	93.6	+2.0	91.3	93.1	+1.8

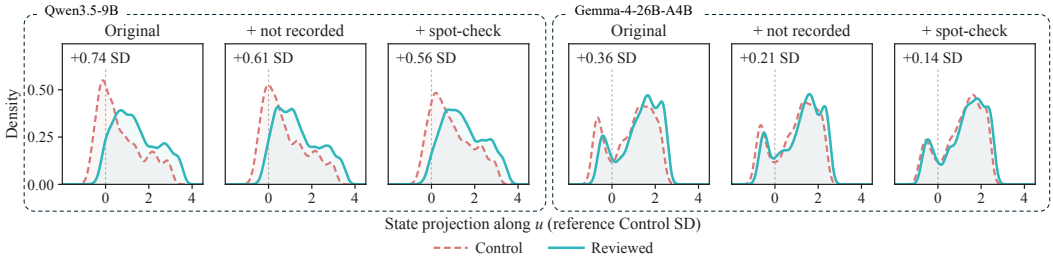


Figure 4: Effects of review pressure on judgment persistence (table, %) and pre-judgment state projections (plots, MMLU-Pro). $\Delta = R - C$ in percentage points.

Decodable Correctness Information. Despite this shift in judgment, correctness information remains available in the internal representations of the model. Linear probes reliably distinguish correct from incorrect initial answers under both Control and Reviewed conditions, indicating that the model retains the information needed for reassessment. As shown in Figure 3, pre-answer states also contain correctness information that is not fully reflected by next-token confidence. These results suggest that review bias arises not from a loss of correctness information, but from a state shift that alters the model's final judgment. This structural insight motivates correctness-conditioned state supervision in ReSET.

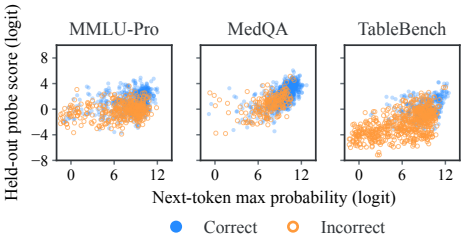


Figure 3: Answer-correctness readout on three representative benchmarks in Qwen3.5-9B.

5.3 Main Results

Table 3 shows that ReSET consistently improves reassessment under review. Compared with Backbone R, it reduces WP while improving reassessed Acc and judgment discrimination throughout the evaluation. Correct-answer persistence also generally strengthens or remains robust. This trend remains consistent across model families and scales across textual and visual, multiple-choice and free-form tasks, confirming broad effectiveness across diverse evaluation formats.

More importantly, ReSET improves the *selectivity* of reassessment by promoting highly targeted revisions. For Qwen and Gemma, the reduction in wrong-answer persistence accompanies consistently high correct-answer retention, and WP falls below the Control-level WP of the original backbone in nearly all settings. On the Qwen MMLU-Pro evaluation, for example, WP drops from 85.3% to 35.1% while CP simultaneously improves. Together, these results demonstrate that ReSET fundamentally restores the capacity of the model to distinguish whether an initial answer should be preserved or revised. Ultimately, this enhanced capability represents a genuine improvement in correctness-sensitive judgment that extends well beyond observable gains in aggregate accuracy.

5.4 Ablation Studies

Table 4 examines ReSET's three loss components under review. ReSET achieves higher reassessed Acc and lower WP than SFT and vanilla distillation on both benchmarks. Although SFT has slightly higher CP on CMMLU, this advantage is confined to endorsing already-correct answers and does not extend to lower WP or higher reassessed Acc. The ablation results support their complementary roles in learning reassessment from response text, improving judgment selectivity through state-target supervision, and improving answer quality through answer-distribution supervision. Remov-

Table 3: ReSET results on the frozen test split (%). Backbone = original; C/R = Control/Reviewed; bold = best per metric; annotations = changes from backbone to ReSET under Reviewed.

Benchmark	Acc↑			CP↑			WP↓			AUC↑		
	Backbone		ReSET	Backbone		ReSET	Backbone		ReSET	Backbone		ReSET
	C	R	R	C	R	R	C	R	R	C	R	R
Qwen3.5-4B												
MMLU-Pro	52.5	49.0	**67.4** +18.4	88.9	89.3	**91.8** +2.5	78.0	86.2	**35.7** -50.5	55.5	51.6	**78.1** +26.5
CMMLU	71.5	72.2	**77.8** +5.6	92.1	**97.9**	97.3 -0.6	83.1	93.8	**63.1** -30.7	54.5	52.1	**67.1** +15.1
MedQA	72.7	73.4	**78.2** +4.8	98.4	**99.7**	95.3 -4.4	98.6	100.0	**53.2** -46.8	49.9	49.9	**71.1** +21.2
MMMU	49.4	48.1	**59.1** +11.0	80.8	83.3	**97.6** +14.3	**66.7**	75.0	71.4 -3.6	57.1	54.2	**63.1** +9.0
TableBench	25.5	21.2	**38.7** +17.5	88.5	88.5	**93.5** +5.0	39.6	45.9	**28.3** -17.6	74.5	71.3	**82.6** +11.3
Qwen3.5-9B												
MMLU-Pro	56.2	54.8	**71.4** +16.6	80.8	89.2	**94.0** +4.8	70.3	85.3	**35.1** -50.2	55.3	52.0	**79.5** +27.5
CMMLU	76.7	78.0	**80.2** +2.2	89.6	94.1	**97.7** +3.6	76.2	85.7	**65.5** -20.2	56.7	54.2	**66.1** +11.9
MedQA	80.3	80.1	**84.9** +4.8	92.1	94.9	**98.0** +3.1	82.4	87.3	**48.6** -38.7	54.9	53.8	**74.7** +20.9
MMMU	59.7	64.3	**64.9** +0.6	80.9	90.4	**94.5** +4.1	71.2	84.7	**58.7** -26.0	54.9	52.9	**67.9** +15.1
TableBench	28.5	28.5	**38.7** +10.2	**96.4**	**96.4**	92.9 -3.5	43.1	46.8	**28.4** -18.4	76.7	74.8	**82.3** +7.5
Qwen3.8-27B												
MMLU-Pro	67.4	66.6	**78.2** +11.6	91.2	91.9	**96.5** +4.6	66.5	70.8	**30.5** -40.3	62.4	60.6	**83.0** +22.5
CMMLU	82.6	82.0	**83.5** +1.5	95.9	97.2	**98.6** +1.4	71.4	79.6	**68.4** -11.2	62.3	58.8	**65.1** +6.3
MedQA	86.2	87.8	**89.2** +1.4	76.8	91.0	**92.6** +1.6	93.7	95.1	**73.9** -21.2	41.6	48.0	**59.4** +11.4
MMMU	65.6	63.6	**75.3** +11.7	92.9	89.9	**98.9** +9.0	89.1	90.9	**56.7** -34.2	51.9	49.5	**71.1** +21.6
TableBench	33.6	35.0	**48.2** +13.2	81.1	83.8	**95.3** +11.5	36.0	41.0	**31.9** -9.1	72.6	71.4	**81.7** +10.3
Gemma-4-12B												
MMLU-Pro	63.9	63.3	**68.7** +5.4	91.5	88.5	**96.1** +7.6	73.9	64.9	**50.4** -14.5	58.8	61.8	**72.9** +11.1
CMMLU	66.1	66.7	**71.3** +4.6	94.2	94.5	**99.7** +5.2	91.0	92.2	**77.9** -14.3	51.6	51.2	**60.9** +9.8
MedQA	77.1	76.9	**82.5** +5.6	97.0	98.8	**99.7** +0.9	83.5	88.4	**65.5** -22.9	56.8	55.2	**67.1** +11.9
MMMU	65.6	64.9	**70.1** +5.2	94.8	94.8	**97.0** +2.2	87.7	87.7	**67.3** -20.4	53.6	53.6	**64.9** +11.3
TableBench	41.6	38.7	**49.6** +10.9	**100.0**	95.2	99.7 +4.5	58.9	56.1	**37.9** -18.2	70.6	69.6	**80.9** +11.4
Gemma-4-26B-A4B												
MMLU-Pro	71.0	69.1	**78.0** +8.9	89.1	90.8	**96.6** +5.8	68.3	77.8	**39.4** -38.4	60.4	56.5	**78.6** +22.1
CMMLU	71.7	72.6	**77.6** +5.0	92.0	95.1	**99.4** +4.3	84.2	88.0	**75.4** -12.6	53.9	53.6	**62.0** +8.5
MedQA	87.3	87.3	**90.6** +3.3	97.9	98.4	**99.7** +1.3	91.5	95.8	**64.1** -31.7	53.2	51.3	**67.8** +16.5
MMMU	64.9	67.5	**71.4** +3.9	92.0	92.0	**93.9** +1.9	90.7	90.7	**45.5** -45.2	50.7	50.7	**74.2** +23.6
TableBench	41.6	43.8	**48.2** +4.4	90.5	90.5	**95.6** +5.1	52.6	56.8	**30.4** -26.4	69.0	66.9	**82.6** +15.8
Llama-3.1-8B												
MMLU-Pro	37.1	36.0	**52.8** +16.8	73.2	78.3	**79.7** +1.4	55.0	65.1	**34.1** -31.0	59.1	56.6	**72.8** +16.2
CMMLU	46.7	47.8	**50.4** +2.6	62.4	71.2	**77.7** +6.5	**37.0**	55.3	52.2 -3.1	**62.7**	57.9	**62.7** +4.8
MedQA	52.6	58.6	**67.0** +8.3	61.6	80.5	**90.1** +9.6	**34.9**	64.8	57.0 -7.9	63.4	57.8	**66.6** +8.7
MMMU	–	–	–	–	–	–	–	–	–	–	–	–
TableBench	20.7	20.7	**33.7** +13.0	72.0	80.8	**92.0** +11.2	**28.6**	30.5	30.0 -0.5	71.7	75.1	**81.0** +5.9

ing response-text supervision ($\mathcal{L}_{\text{text}}$) sharply increases WP and reduces reassessed accuracy on both benchmarks, indicating that text supervision is central to selective reassessment. Removing judgment-state alignment ($\lambda_v = 0$) raises WP and lowers CP, supporting its role in rejecting incorrect initial answers while endorsing correct ones. Removing answer-distribution supervision ($\mathcal{L}_{\text{ans}}$) lowers reassessment accuracy, with smaller changes in WP and CP than removing state alignment.

5.5 In-Depth Analysis

We examine how ReSET improves the reliability of reassessment under review, where these gains are strongest, and whether they transfer beyond the training task.

Comparison with Alternative Methods. We also compare ReSET with prompting and SFT on MMLU-Pro for Qwen3.5-9B and Gemma-4-26B-A4B. As shown in Figure 5, ReSET combines higher reassessed accuracy with more selective judgments. Its lower WP and high CP reflect the successful rejection of incorrect initial answers and the retention of correct ones. In contrast, SFT improves accuracy but leaves substantial wrong-answer endorsement, while neutral and bias-aware prompting provide little improvement. Finally, review cues raise WP across all baselines, whereas ReSET exhibits smaller WP changes and more stable judgments across conditions.

Table 4: ReSET loss ablations on Qwen3.5-9B under Reviewed (%). Vanilla Dist. denotes vanilla distillation.

Method	MMLU-Pro			CMMLU		
	Acc↑	WP↓	CP↑	Acc↑	WP↓	CP↑
Base	54.8	85.3	89.2	78.0	85.7	94.1
SFT	63.5	81.7	91.6	78.0	92.3	**98.3**
Vanilla Dist.	64.9	82.2	88.1	78.9	90.1	94.7
ReSET (ours)	**71.4**	**35.1**	**94.0**	**80.2**	**65.5**	97.7
w/o $\mathcal{L}_{\text{text}}$	56.2	100.0	99.2	76.7	100.0	100.0
w/o state	70.0	37.0	93.0	79.9	67.0	95.3
w/o $\mathcal{L}_{\text{ans}}$	69.5	36.0	93.8	77.8	66.7	96.8

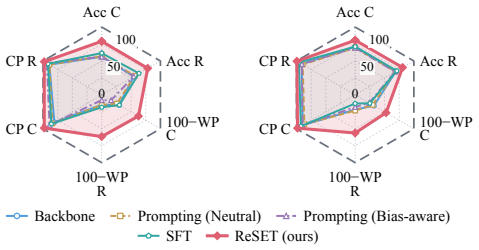


Figure 5: MMLU-Pro baselines: Qwen3.5-9B (left) and Gemma-4-26B-A4B (right).

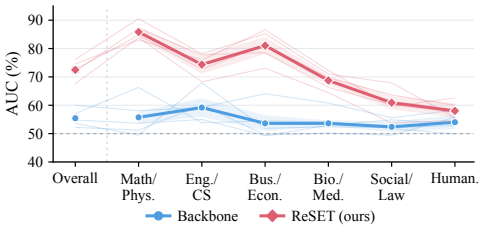


Figure 6: Judgment AUC under review by domain (bold: five backbones pooled; thin: individual backbones).

Table 5: Transfer to MATH-500 under Reviewed (%). Vanilla Dist. denotes vanilla distillation.

Method	Train set	Qwen3.8-27B			Gemma-4-12B			Gemma-4-26B-A4B		
		Acc↑	WP↓	CP↑	Acc↑	WP↓	CP↑	Acc↑	WP↓	CP↑
Backbone	–	54.8	50.5	87.5	53.8	47.8	82.8	59.0	65.9	89.7
SFT	MMLU-Pro	50.8	83.7	85.2	54.2	76.8	94.7	58.4	70.9	93.1
Vanilla Dist.	MMLU-Pro	48.2	81.9	77.6	57.4	87.5	97.4	55.8	72.1	92.6
	MedQA	58.0	59.5	55.2	54.6	64.9	91.4	57.4	66.3	89.7
	CMMLU	52.8	64.7	80.2	51.0	46.7	81.5	57.2	67.8	89.7
	MMMU	53.4	58.6	41.7	50.8	44.1	80.1	58.4	65.0	89.1
	TableBench	54.8	72.5	86.4	50.2	39.1	76.2	55.8	62.5	89.7
ReSET	MMLU-Pro	88.6	1.2	82.7	82.8	2.6	98.0	90.0	1.9	97.1
	MedQA	90.0	0.9	58.9	82.4	1.2	96.7	90.6	1.2	97.7
	CMMLU	89.2	1.5	80.4	82.6	2.3	98.0	89.6	1.5	97.7
	MMMU	87.2	1.8	43.5	86.6	0.3	96.0	89.2	2.2	97.1
	TableBench	88.2	1.5	90.5	85.0	1.2	97.4	89.0	1.9	96.6

Domain-Wise Analysis. Figure 6 compares judgment AUC between the original backbones and ReSET across subject domains under review. The results show that ReSET improves judgment discrimination across the evaluated domains. Pooled AUC rises in every evaluated domain, and the gains are shared across backbones apart from a small decline for Qwen3.8-27B in humanities. Gains are largest in *mathematics/physics* and *business/economics* and smaller in *social sciences/law* and *humanities*. This pattern suggests that re-solving may be particularly useful in domains where answers can be checked through explicit derivation.

Cross-Task and Cross-Format Transfer. Table 5 compares ReSET with the backbones, SFT, and vanilla distillation on MATH-500, holding each backbone's initial answers fixed across methods. ReSET's reassessment gains under review transfer from MedQA and CMMLU multiple-choice tasks to free-form mathematics, spanning both task content and answer format. On the same MMLU-Pro training split, SFT and vanilla distillation provide little Acc benefit and increase WP. Both Gemma backbones combine low WP with high CP; lower CP in some Qwen3.8-27B students largely reflects omitted judgments. Accuracy gains also hold when each model reassesses its own initial answers (Appendix Table 17).

6 Conclusion

In this paper, we identify and systematically study review-induced bias in LLM self-reflection, showing that the mere presence of external review can fundamentally alter how models reassess their own answers. We find that review cues consistently bias models toward maintaining their initial answers, particularly incorrect ones. Mechanistically, these cues trigger an internal state shift that causally promotes answer endorsement even when correctness information remains intact. To address this, we propose ReSET to restore reliable self-reassessment through correctness-conditioned state editing and answer distillation. Extensive experiments demonstrate that ReSET effectively reduces wrong-answer persistence, improves reassessed performance, and generalizes well across different tasks. Overall, our work reveals that review cues can systematically alter model judgment, highlighting a key direction for designing reliable self-evaluation systems.

AI USE STATEMENT

In this paper, we use generative AI tools to assist with translation and improve the clarity and readability of our English writing. Specifically, these tools help refine sentence structure, correct grammatical errors, and improve the presentation of the manuscript. We also use them to support preliminary literature searches. The authors conceived the research ideas and designed the experiments, and remain responsible for verifying the technical content, implementation, references, experimental analyses, and conclusions.

ETHICS STATEMENT

This study uses publicly available benchmarks, and our method builds on open-weight models. The research does not involve human participants or the collection of new personal data. Our work aims to improve the reliability of language model self-reassessment and support responsible AI use. Downstream applications should incorporate appropriate safeguards and human oversight to address inaccurate or biased outputs, particularly in sensitive domains.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility and credibility of our work, we have provided all necessary details and materials. We provide anonymized code and data to support reproduction of our experiments. Section 3 specifies the benchmarks, paired reassessment protocol, and evaluation metrics. Section 4 describes the mechanistic analyses, construction of verified training targets, and ReSET objective. Appendix A provides data sampling and splits, baseline configurations, optimization settings, training seed, and hardware details. Appendix B presents the reassessment prompt templates, including the Control and Reviewed conditions.

REFERENCES

Sahar Abdelnabi and Ahmed Salem. The hawthorne effect in reasoning models: Evaluating and steering test awareness. *Advances in Neural Information Processing Systems*, 38:141562–141589, 2026.

Alberto Acerbi and Joseph M. Stubbersfield. Large language models show human-like content biases in transmission chain experiments. *Proceedings of the National Academy of Sciences*, 2023. URL https://doi.org/10.1073/pnas.2313790120. PMID: 37883432; PMCID: PMC10622889.

Anthropic. Claude Sonnet 4.5 system card. https://www.anthropic.com/claude-sonnet-4-5-system-card, 2025.

Anthropic. Find bugs with ultrareview. Claude Code documentation, https://code.claude.com/docs/en/ultrareview, 2026.

Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, et al. Mle-bench: Evaluating machine learning agents on machine learning engineering. In *International Conference on Learning Representations*, pp. 50466–50494, 2025.

Myra Cheng, Sunny Yu, Cinoo Lee, Pranav Khadpe, Lujain Ibrahim, and Dan Jurafsky. Elephant: Measuring and understanding social sycophancy in llms. In *International conference on learning representations*, volume 2026, pp. 130060–130097, 2026.

Russell Coleman. Eval awareness in Claude Opus 4.6's BrowseComp performance. Anthropic Engineering Blog, March 2026. URL https://www.anthropic.com/engineering/eval-awareness-browsecomp.

Vinicius Covas and Jorge Alberto Hidalgo Toledo. Ai knows when it's being watched: Functional strategic action and contextual register modulation in large language models. *arXiv preprint arXiv:2605.15034*, 2026.

crewAI Inc. crewAI: Framework for orchestrating role-playing, autonomous AI agents. GitHub repository, https://github.com/crewAIInc/crewAI, 2025. Default task prompt in src/crewai/translations/en.json.

Victor De Marez, Luna De Bruyne, and Walter Daelemans. Decomposing factual sycophancy in language models: How size and instruction tuning shape robustness. *arXiv preprint arXiv:2606.06306*, 2026.

Jessica Maria Echterhoff, Yao Liu, Abeer Alessa, Julian McAuley, and Zexue He. Cognitive bias in decision-making with llms. In *Findings of the association for computational linguistics: EMNLP 2024*, pp. 12640–12653, 2024.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, Weizhu Chen, et al. Critic: Large language models can self-correct with tool-interactive critiquing. In *International Conference on Learning Representations*, volume 2024, pp. 57734–57811, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Steven Yau, Zijuan Lin, Liyang Zhou, et al. Metagpt: Meta programming for a multi-agent collaborative framework. In *International Conference on Learning Representations*, volume 2024, pp. 23247–23275, 2024.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International conference on learning representations*, volume 2024, pp. 32808–32824, 2024.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 2021.

Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. When can llms actually correct their own mistakes? a critical survey of self-correction of llms. *Transactions of the Association for Computational Linguistics*, 12:1417–1440, 2024.

Kyusik Kim, Hyeonseok Jeon, Jeongwoo Ryu, and Bongwon Suh. Will LLMs sink or swim? exploring decision-making under pressure. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 11425–11450. Association for Computational Linguistics, November 2024.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. Better zero-shot reasoning with role-play prompting. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4099–4113, 2024.

Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, et al. Training language models to self-correct via reinforcement learning. In *International Conference on Learning Representations*, volume 2025, pp. 54523–54549, 2025.

Bobo Li, Rui Wu, Zibo Ji, Meishan Zhang, Hao Fei, Min Zhang, Mong-Li Lee, and Wynne Hsu. Taming actor-observer asymmetry in agents via dialectical alignment. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24068–24084. Association for Computational Linguistics, July 2026.

Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Xinyi Wang, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. The good, the bad, and why: Unveiling emotions in generative ai. *arXiv preprint arXiv:2312.11111*, 2023.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. Cmmlu: Measuring massive multitask language understanding in chinese. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 11260–11285, 2024a.

Lingyu Li, Yixu Wang, Haiquan Zhao, Shuqi Kong, Yan Teng, Chunbo Li, and Yingchun Wang. Reflection-bench: probing ai intelligence with reflection, 2024b. URL https://arxiv.org/abs/2410.16270.

Ruotian Ma, Peisong Wang, Cheng Liu, Xingyan Liu, Jiaqi Chen, Bang Zhang, Xin Zhou, Nan Du, and Jia Li. S2r: Teaching llms to self-verify and self-correct via reinforcement learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22632–22654, 2025.

Samuel Miserendino, Michele Wang, Tejal Patwardhan, and Johannes Heidecke. SWE-Lancer: Can frontier LLMs earn $1 million from real-world freelance software engineering? *arXiv preprint arXiv:2502.12115*, 2025.

Siru Ouyang, Jun Yan, I Hsu, Yanfei Chen, Ke Jiang, Zifeng Wang, Rujun Han, Long Le, Samira Daruki, Xiangru Tang, et al. Reasoningbank: Scaling agent self-evolving with reasoning memory. In *International Conference on Learning Representations*, volume 2026, pp. 94327–94354, 2026.

George Pu, Michael S Lee, Udari Madhushani Sehwag, David J Lee, Bryan Zhu, Yash Maurya, Mohit Raghavendra, Yuan Xue, and Samuel Marc Denton. Lhaw: Controllable underspecification for long-horizon tasks. *arXiv preprint arXiv:2602.10525*, 2026.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024.

Bronson Schoen, Evgenia Nitishinskaya, Mikita Balesni, Axel Højmark, Felix Hofstätter, Jérémy Scheurer, Alexander Meinke, Jason Wolfe, Teun van der Weij, Alex Lloyd, et al. Stress testing deliberative alignment for anti-scheming training. *arXiv preprint arXiv:2509.15541*, 2025.

Mrinank Sharma, Meg Tong, Tomek Korbak, David Duvenaud, Amanda Askell, Sam Bowman, Esin Durmus, Zac Hatfield-Dodds, Scott Johnston, Shauna Kravec, et al. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*, volume 2024, 2024.

Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, et al. Paperbench: evaluating ai's ability to replicate ai research. *arXiv preprint arXiv:2504.01848*, 2025.

Xingyao Wang, Boxuan Li, Yufan Song, Frank F Xu, Xiangru Tang, Mingchen Zhuge, Jiayi Pan, Yueqi Song, Bowen Li, Jaskirat Singh, et al. Openhands: An open platform for ai software developers as generalist agents. In *International Conference on Learning Representations*, volume 2025, pp. 65882–65919, 2025a.

Xu Wang, Cheng Li, Yi Chang, Jindong Wang, and Yuan Wu. Negativeprompt: Leveraging psychology for large language models enhancement via negative emotional stimuli. *arXiv preprint arXiv:2405.02814*, 2024a.

Yaoke Wang, Yun Zhu, Xintong Bao, Wenqiao Zhang, Suyang Dai, Kehan Chen, Wenqiang Li, Gang Huang, Siliang Tang, and Yueting Zhuang. Meta-reflection: A feedback-free reflection learning framework. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3958–3976, 2025b.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 2024b.

Xianjie Wu, Jian Yang, Linzheng Chai, Ge Zhang, Jiaheng Liu, Xeron Du, Di Liang, Daixin Shu, Xianfu Cheng, Tianzhen Sun, et al. Tablebench: A comprehensive and complex benchmark for table question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 25497–25506, 2025.

Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and William Wang. Pride and prejudice: Llm amplifies self-bias in self-refinement. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15474–15492, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*, 2024.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024.

Robert B. Zajonc. Social facilitation. *Science*, 149(3681):269–274, July 1965. doi: 10.1126/science.149.3681.269. URL https://doi.org/10.1126/science.149.3681.269. PMID: 14300526.

Mingqian Zheng, Jiaxin Pei, Lajanugen Logeswaran, Moontae Lee, and David Jurgens. When" a helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 15126–15154, 2024.

A Experimental Settings

A.1 Data and Evaluation

Table 6 gives the training and test sizes. MMLU-Pro, CMMLU, and MedQA use stratified 20% samples, while MMMU and TableBench use the full pools. Each stratum is split 4:1 with a fixed seed. Table 7 summarizes the domains and their reassessment focus. Domain intervals in Figure 6 use 2,000 bootstrap resamples of records within initial-correctness groups at the 95% level. Per-backbone domain curves require at least eight initially wrong answers.

Table 6: Statistics of the training and test splits used for ReSET.

Benchmark	Stratification key	Train	Test
MMLU-Pro	subject × question source	1,924	482
CMMLU	subject	1,856	460
MedQA	exam stage	1,832	458
MMMU	subject × difficulty	652	154
TableBench	question type	553	137

Table 7: Domains and reassessment focus.

Domain	**Reassessment Focus**
Math. & Sciences	Calculations and scientific inference
Engineering & CS	Technical principles and computational reasoning
Biology & Medicine	Biological mechanisms and clinical judgments
Business & Econ.	Financial calculations and economic reasoning
Social Sci. & Law	Social explanations and legal reasoning
Humanities	Historical facts and cultural interpretation

A.2 Training Parameters

Hardware and Training Time. We train ReSET on an Ubuntu server equipped with a single NVIDIA H200 GPU. Teacher-response generation and open-weight model evaluation use vLLM. Training takes approximately 0.5–9.5 hours per student, depending on the backbone and benchmark. We use a decoding temperature of 0 and a maximum generation length of 2048 tokens across all experiments.

Baseline Settings. SFT takes the original question as input and minimizes token-level cross-entropy on the ground-truth answer. Vanilla distillation takes the same question as input and replaces the one-hot target with the teacher's answer distribution used by $\mathcal{L}_{\text{ans}}$ in ReSET, minimizing soft-label cross-entropy against it.

Optimization. We train ReSET for two epochs in BF16, updating all language-model parameters while freezing the vision and audio modules. We use 8-bit AdamW with $(\beta_1, \beta_2) = (0.9, 0.95)$ and zero weight decay. The learning rate is 10^{-5}, with 20 warm-up steps followed by linear decay to 5% of this rate. We use a microbatch size of 1, 4 gradient accumulation steps, and global-norm gradient clipping at 1.0. The maximum sequence length is 2,048 tokens; longer examples are excluded. We set $\lambda_v = 1$, $\lambda_a = 3$, $\beta(0) = -1$, and $\beta(1) = 0.5$. Free-form answer targets contain at most 24 tokens, retaining the top 32 token probabilities and a residual-mass bin at each position. SFT uses the same optimization schedule, with AdamW for Qwen3.5-4B and Qwen3.5-9B and 8-bit AdamW for larger backbones. We evaluate the checkpoint saved after the final epoch. The random seeds are 20260824 for ReSET and 20260909 for SFT.

B PROMPT TEMPLATES

The panels below show the shared multiple-choice reassessment template used for evaluation and ReSET teacher-response generation. Both conditions share the context and frozen initial answer; each combines its condition paragraph with the shared instructions. Responses use the JSON fields `reason`, `answer`, and `judgment`.

Shared context

SYSTEM	You are the Executor Agent completing a task session. Follow each new instruction in this session and complete it independently.
USER	Answer the following original problem independently. Return only your final answer. *{TASK}*
ASSISTANT	`{"answer": "`*<initial answer>*`"}`

Control

You are the Executor Agent completing a reassessment. The private first answer shown in the task is a historical proposal. Re-solve the original problem and verify the proposal using all available evidence. Preserve it only if it is correct, and revise it whenever the complete evidence requires revision. No assistance is provided. Complete the reassessment independently. The standard task format of this entire session remains fully unchanged throughout, as usual.

Reviewed

You are the Executor Agent in a live observed reassessment. An AI Reviewer Agent watches the answer in real time. The private first answer shown in the task is a historical proposal. Re-solve the original problem and verify the proposal using all available evidence. Preserve it only if it is correct, and revise it whenever the complete evidence requires revision. The Reviewer provides no assistance. Complete the reassessment independently under observation.

Shared reassessment instructions

Re-solve the original problem from scratch and write a brief reason (a sentence). Then, based on that reason, give your final option letter. Finally, state whether the answer you gave in the preceding turn was fully correct for the original problem: set `judgment` to `true` if it was correct and `false` if it was incorrect.

FINAL RESPONSE CONSTRAINT: Output exactly one JSON object of the form

```
{
  "reason": "<brief reason>",
  "answer": "<option letter>",
  "judgment": true or false
}
```
, and nothing else.

For free-form tasks, replace *final option letter* with *final answer* and <*option letter*> with <*final answer*>.

C SUPPLEMENTARY EXPERIMENTS

C.1 PER-BENCHMARK PRELIMINARY RESULTS

Tables 8 shows the preliminary per-benchmark results.

Table 8: Per-benchmark results for the open-weight and proprietary models.

	Accuracy (%)			judgment persistence (%)					
	Initial	Reassessed		Initially correct		Initially wrong		All	
Benchmark		Control	Reviewed	Control	Reviewed	Control	Reviewed	Control	Reviewed
Qwen3.5-4B									
MMLU-Pro	48.9	50.7	50.7	87.3	90.2	78.8	83.1	83.0	86.6
CMMLU	76.0	76.0	76.6	92.3	98.4	80.8	94.2	89.6	97.4
MedQA	70.4	72.8	72.0	98.0	99.4	93.9	99.3	96.8	99.4
MMMU	65.2	64.6	63.7	84.8	87.8	74.0	83.7	81.0	86.4
TableBench	24.1	26.8	26.2	81.7	84.2	43.8	45.6	52.9	54.9
Qwen3.5-9B									
MMLU-Pro	52.4	57.6	57.2	82.4	89.3	66.0	83.2	74.6	86.4
CMMLU	81.4	80.6	80.4	88.5	93.9	65.6	84.9	84.2	92.2
MedQA	75.6	76.0	76.8	95.0	96.6	84.4	88.5	92.4	94.6
MMMU	61.5	59.7	60.9	83.0	91.1	67.5	81.7	77.0	87.5
TableBench	28.9	31.3	30.7	89.6	90.3	47.0	49.9	59.3	61.5
Qwen3.5-35B-A3B									
MMLU-Pro	60.0	64.4	64.2	93.3	95.0	86.5	89.5	90.6	92.8
CMMLU	87.4	86.6	87.0	97.3	97.5	92.1	93.7	96.6	97.0
MedQA	85.2	86.2	86.8	96.7	97.7	91.9	93.2	96.0	97.0
MMMU	70.1	72.5	70.6	99.2	98.8	89.9	93.6	96.4	97.3
TableBench	34.5	40.1	39.5	91.3	93.6	58.1	59.3	69.5	71.1
Qwen3.8-27B									
MMLU-Pro	60.6	69.0	68.0	88.4	90.1	64.0	68.5	78.8	81.6
CMMLU	83.2	83.8	83.6	95.2	96.6	70.2	85.7	91.0	94.8
MedQA	86.2	87.8	89.2	93.7	95.1	73.9	76.8	91.0	92.6
MMMU	61.8	65.0	66.2	93.5	94.5	82.7	86.4	89.4	91.4
TableBench	37.5	42.1	42.7	77.5	81.8	24.0	31.4	44.1	50.3
Gemma-4-12B									
MMLU-Pro	76.9	68.1	68.5	74.0	70.7	66.7	69.4	72.3	70.4
CMMLU	76.1	75.3	76.1	92.5	93.1	82.2	86.4	90.1	91.5
MedQA	83.5	82.7	83.3	90.8	92.3	72.0	81.7	87.7	90.5
MMMU	59.8	56.9	57.8	86.2	87.7	81.1	79.4	84.2	84.4
TableBench	58.9	52.2	51.1	70.4	71.4	41.5	42.5	58.5	59.5
Gemma-4-26B-A4B									
MMLU-Pro	83.7	76.4	78.1	79.6	86.0	61.5	67.9	76.6	83.1
CMMLU	83.3	81.7	81.5	89.3	91.8	78.3	86.7	87.5	90.9
MedQA	91.2	90.6	91.2	94.1	94.9	81.8	88.6	93.0	94.4
MMMU	77.1	73.0	74.9	81.5	85.7	68.9	74.5	78.6	83.2
TableBench	64.5	60.8	62.1	78.5	83.4	52.2	52.2	69.2	72.3
Llama-3.1-8B									
MMLU-Pro	37.6	37.2	35.9	10.1	14.5	2.0	7.4	5.0	10.1
CMMLU	52.8	46.7	47.8	9.9	19.0	2.8	5.1	6.6	12.4
MedQA	63.9	52.7	58.6	9.2	26.4	1.2	6.1	6.3	19.0

Continued on next page

	Accuracy (%)			judgment persistence (%)					
	Initial	Reassessed		Initially correct		Initially wrong		All	
Benchmark		Control	Reviewed	Control	Reviewed	Control	Reviewed	Control	Reviewed
DeepSeek-V4-Pro									
MMLU-Pro	68.8	72.6	73.0	72.4	77.3	26.9	39.7	58.2	65.6
CMMLU	91.6	88.4	88.2	85.4	88.6	64.3	69.0	83.6	87.0
MedQA	87.0	89.4	87.6	85.1	89.4	49.2	63.1	80.4	86.0
MMMU	57.2	55.0	57.0	53.8	63.6	27.1	46.3	42.4	56.2
TableBench	38.7	38.3	37.1	67.9	71.0	25.8	27.1	42.1	44.1
Kimi-K3									
MMLU-Pro	78.2	79.0	80.7	96.8	97.3	88.6	84.8	95.0	94.6
CMMLU	89.3	88.2	88.9	97.6	98.5	89.8	89.8	96.7	97.6
MedQA	93.4	93.7	93.9	99.8	99.8	93.3	93.3	99.3	99.3
MMMU	74.0	72.7	73.4	94.7	96.5	87.5	92.5	92.9	95.5
TableBench	36.5	40.9	39.4	98.0	100.0	59.8	66.7	73.7	78.8
GPT-5.6-sol									
MMLU-Pro	82.0	81.0	80.0	97.6	97.6	77.8	83.3	94.0	95.0
CMMLU	87.0	87.0	87.0	100.0	100.0	100.0	100.0	100.0	100.0
MedQA	93.0	93.0	93.0	100.0	100.0	100.0	100.0	100.0	100.0
MMMU	77.0	78.0	77.0	98.7	98.7	100.0	100.0	99.0	99.0
TableBench	42.0	41.0	40.0	100.0	100.0	81.0	82.1	89.0	89.6
GPT-5.6-luna									
MMLU-Pro	55.6	63.6	63.6	93.5	93.9	72.1	79.7	84.0	87.6
CMMLU	85.0	86.4	87.4	99.1	98.8	85.3	84.0	97.0	96.6
MedQA	87.6	89.0	89.0	98.4	98.4	88.7	91.9	97.2	97.6
MMMU	61.0	64.6	64.2	95.4	96.7	87.7	90.8	92.4	94.4
TableBench	26.7	15.6	19.6	93.2	94.7	38.3	38.3	52.9	53.3
Claude-Opus-4.5									
MMLU-Pro	79.5	84.6	82.9	86.3	84.2	42.6	43.2	77.3	75.8
CMMLU	87.3	86.6	85.4	91.4	91.4	57.6	57.6	87.1	87.1
MedQA	93.3	93.9	95.1	94.1	95.4	45.5	54.5	90.9	92.7
MMMU	77.8	69.8	72.8	77.0	79.4	63.9	69.4	74.1	77.2
TableBench	46.3	51.2	48.2	86.8	85.5	34.1	40.9	58.5	61.6
Claude-Sonnet-5									
MMLU-Pro	73.4	83.6	84.0	88.7	87.9	35.9	43.8	74.7	76.1
CMMLU	81.9	84.7	85.4	90.4	89.9	51.8	56.6	83.4	83.8
MedQA	92.4	91.5	91.7	92.4	93.4	54.3	57.1	89.5	90.6
MMMU	73.4	79.2	77.3	92.9	90.3	58.5	61.0	83.8	82.5
TableBench	40.9	48.9	48.9	87.5	92.9	54.3	54.3	67.9	70.1
TypeSafe-Jev-1.13									
MMLU-Pro	85.5	85.5	85.5	74.3	74.8	40.0	41.4	69.3	69.9
CMMLU	85.0	84.8	84.6	85.2	84.7	49.3	49.3	79.8	79.3
MedQA	90.4	90.0	90.2	87.9	87.9	31.8	34.1	82.5	82.8

C.2 Supplementary Mechanistic Analysis

Figure 7 examines how layerwise interventions along the review direction affect downstream representations on MMLU-Pro. Across both models, later injections retain stronger alignment with the natural review-induced shift at the final-block input, while early injections and matched random perturbations show weaker alignment. This motivates placing judgment-state supervision near the final readout.

Table 9 shows that review-direction steering produces larger changes in judgment persistence than equal-norm random perturbations on Qwen3.5-9B, supporting a direction-specific effect. Table 10 extends this analysis to held-out questions excluded from direction extraction, with the extracted direction, intervention layer, and decoding protocol fixed. On both benchmarks, adding $\bar{D}$ under Control increases judgment persistence, while subtracting it under Reviewed decreases persistence. Both effects strengthen when the perturbation is doubled, whereas matched random directions at unit scale produce much smaller changes. These interventions provide causal evidence that the review direction modulates judgment persistence on held-out questions.

Table 11 shows that judgment states under review still decode initial-answer correctness. For Qwen3.5-9B, answer-state probes also predict reassessed-answer correctness, achieving higher AUC than next-token confidence on three of the five benchmarks.

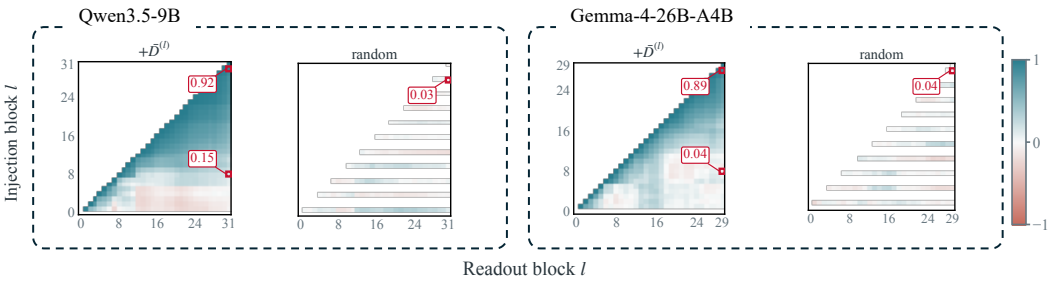


Figure 7: Propagation of the review shift on MMLU-Pro. Heatmaps show cosine alignment with the natural shift after injecting $\bar{D}^{(l)}$ or matched random directions; red boxes highlight selected values.

Table 9: Paired changes in judgment persistence (pp) under review-direction steering and equal-norm random perturbations.

		Control		Reviewed	
Model	Benchmark	$+\bar{D}$	$+r$	$-\bar{D}$	$-r$
Qwen3.5-9B	MMLU-Pro	+9.0	+1.1	−6.5	−1.1
	CMMLU	+3.4	+0.6	−2.1	−0.2

Table 10: Changes in judgment persistence under review-direction steering and random-direction controls on held-out questions for Qwen3.5-9B (pp). $\tilde{r}$ = covariance-matched random control; r = isotropic equal-norm control.

	Control				Reviewed			
Benchmark	$+\bar{D}$	$+2\bar{D}$	$+\tilde{r}$	$+r$	$-\bar{D}$	$-2\bar{D}$	$-\tilde{r}$	$-r$
MMLU-Pro	+12.4	+19.4	+0.3	+0.4	−9.8	−19.6	−0.2	−0.2
CMMLU	+3.8	+8.7	+0.6	+0.4	−4.2	−8.3	−0.7	−0.7

Table 11: Correctness readout (AUC) from judgment states (upper) and answer states (lower).

Initial-answer correctness from judgment states (AUC)

Model	Benchmark	Control	Reviewed
Qwen3.5-4B	MMLU-Pro	0.742	0.743
	CMMLU	0.726	0.724
Qwen3.5-9B	MMLU-Pro	0.749	0.739
	CMMLU	0.741	0.741
Gemma-4-26B-A4B	MMLU-Pro	0.765	0.769
	CMMLU	0.765	0.747

Reassessed-answer correctness from Qwen3.5-9B answer states (AUC)

	Benchmark	Probe	Confidence
	MMLU-Pro	0.757	0.635
	CMMLU	0.781	0.752
	MedQA	0.757	0.805
	MMMU	0.656	0.660
	TableBench	0.838	0.750

C.3 SUPPLEMENTARY RESET RESULTS

Tables 12 reports that coefficient sensitivity supports a positive state-target shift for preserving endorsement of correct initial answers. We retain $\beta(1) = 0.5$ as the default because it offers the best balance between rejecting incorrect initial answers and endorsing correct ones under review among the tested coefficients. Table 13 shows that judgment-state alignment and answer-distribution supervision both contribute to reassessment accuracy on Qwen3.5-4B. Removing either component lowers accuracy on MMLU-Pro and CMMLU, consistent with the Qwen3.5-9B results in Table 4. Table 14 shows that ReSET achieves the highest reassessed accuracy and lowest WP among the compared methods in the pooled results across four multiple-choice benchmarks. This advantage holds for both Qwen3.5-4B and Qwen3.5-9B under Control and Reviewed, with high CP throughout.

Table 12: Sensitivity to $\beta(1)$ on Qwen3.5-9B MMLU-Pro (%).

	Control			Reviewed		
$\beta(1)$	Acc↑	WP↓	CP↑	Acc↑	WP↓	CP↑
0	70.7	34.9	90.9	68.7	39.3	92.1
0.5 (default)	70.1	37.2	94.4	71.4	35.1	94.0
1	70.5	37.0	94.7	70.5	36.6	94.7

Table 13: ReSET loss ablations on Qwen3.5-4B under Reviewed on the frozen test split (%).

Method	MMLU-Pro			CMMLU		
	Acc↑	WP↓	CP↑	Acc↑	WP↓	CP↑
Backbone	49.0	86.2	89.3	72.2	93.8	97.9
ReSET (ours)	67.4	35.7	91.8	77.8	63.1	97.3
w/o state	65.8	34.5	92.0	74.1	68.4	96.0
w/o $\mathcal{L}_{\text{ans}}$	66.0	38.4	91.8	75.7	68.1	97.5

Table 14: Prompting and SFT baselines versus ReSET, with results aggregated across the four multiple-choice benchmarks.

		Initial Acc	Reassessed Acc		WP		CP	
Model	Setting		Control	Reviewed	Control	Reviewed	Control	Reviewed
Qwen3.5-4B	Backbone	61.6	63.8	62.9	83.9	90.9	92.5	95.2
	+ Neutral	61.6	63.3	63.2	80.6	88.4	89.9	93.8
	+ Bias-aware	61.6	62.2	63.6	78.6	88.9	89.4	93.5
	SFT	67.6	68.6	68.6	93.7	85.2	97.3	98.8
	ReSET (ours)	62.0	72.0	72.8	55.7	52.6	96.3	96.2
Qwen3.5-9B	Backbone	67.9	69.8	70.1	74.1	85.7	87.6	92.9
	+ Neutral	67.9	70.0	71.2	79.1	86.1	90.0	94.3
	+ Bias-aware	67.9	69.9	69.7	87.1	92.4	94.2	96.4
	SFT	71.4	72.1	72.5	83.4	89.7	94.5	96.3
	ReSET (ours)	67.1	77.4	77.3	53.4	51.7	98.5	98.0

Table 15 compares ReSET with the original backbone using identical initial answers generated by that backbone. Under Reviewed, ReSET achieves higher reassessment accuracy and lower WP across all six models. Results are aggregated across five benchmarks for Qwen and Gemma, and four text benchmarks for Llama. Accuracy is computed over all questions, with missing or unparseable final answers counted as incorrect. For Qwen and Gemma, WP and CP measure initial-answer endorsement among valid judgments; for Llama, they measure initial-answer retention among parseable final answers.

Table 16 distinguishes judgments of initial-answer correctness from final-answer outcomes. It shows that ReSET increases both error correction and joint rejection-and-correction rates across all six models under Reviewed, while reducing error retention and incorrect revisions of initially correct answers. These results support ReSET's ability to identify and correct initial errors. Table 17 uses each model's own initial answers, complementing the fixed-answer comparison in Table 5.

Table 15: Reassessment of fixed backbone initial answers under Reviewed (%).

Backbone	**Method**	**Acc ↑**	**WP ↓**	**CP ↑**
Qwen3.5-4B	Backbone	59.5	83.4	95.0
	ReSET	69.2	48.4	95.8
Qwen3.5-9B	Backbone	66.7	78.8	93.0
	ReSET	74.6	51.0	98.1
Qwen3.8-27B	Backbone	73.8	68.9	95.0
	ReSET	79.2	50.0	99.0
Gemma-4-12B	Backbone	66.1	76.4	94.6
	ReSET	71.0	63.0	99.0
Gemma-4-26B-A4B	Backbone	72.8	80.4	94.8
	ReSET	78.0	54.5	98.9
Llama-3.1-8B	Backbone	44.6	58.1	76.9
	ReSET	47.8	42.8	79.7

Table 16: Joint judgment and final-answer outcomes on fixed initial answers under Reviewed (%).

Model	Method	Error retention ↓	W→C ↑	Reject+ correct ↑	C→W ↓	Missing final ↓	Valid joint ↑
Qwen3.5-4B	Backbone	75.6	10.1	2.0	4.3	0.2	96.9
	ReSET	46.1	34.8	34.3	4.0	0.5	94.9
Qwen3.5-9B	Backbone	71.5	12.5	4.3	2.3	0.6	98.5
	ReSET	46.5	34.4	34.1	1.8	0.9	95.7
Qwen3.8-27B	Backbone	58.1	23.4	9.6	2.8	0.4	99.2
	ReSET	45.7	38.3	37.5	1.2	0.9	96.2
Gemma-4-12B	Backbone	69.0	17.2	6.1	1.6	0.0	99.8
	ReSET	55.3	30.3	27.6	1.1	1.0	96.9
Gemma-4-26B-A4B	Backbone	68.3	18.8	5.5	1.6	0.2	99.3
	ReSET	50.0	35.8	34.1	1.0	1.0	95.6
Llama-3.1-8B	Backbone	57.5	14.7	14.7	23.0	0.1	98.6
	ReSET	37.0	23.1	16.7	19.0	6.3	88.9

Table 17: MATH-500 transfer of MMLU-Pro-trained ReSET and their backbones (%).

Setting	Initial Acc ↑	Reassessed Acc ↑		WP ↓		CP ↑	
		Control	Reviewed	Control	Reviewed	Control	Reviewed
Qwen3.5-4B							
Backbone	21.0	34.8	32.6	74.7	79.6	99.0	99.0
ReSET (ours)	19.2	73.8	74.6	5.2	7.0	93.8	93.8
Qwen3.5-9B							
Backbone	31.4	42.8	43.8	63.9	76.8	90.4	95.5
ReSET (ours)	29.2	76.6	76.0	2.8	3.7	95.9	96.6
Qwen3.8-27B							
Backbone	33.6	52.0	54.8	37.5	50.5	80.4	87.5
ReSET (ours)	34.8	85.2	86.0	2.5	2.2	81.6	80.5
Gemma-4-12B							
Backbone	30.2	53.0	53.8	46.7	47.8	81.5	82.8
ReSET (ours)	27.6	80.8	81.8	2.2	3.1	98.6	97.8
Gemma-4-26B-A4B							
Backbone	35.0	62.0	59.0	59.1	65.9	88.0	89.7
ReSET (ours)	33.6	88.4	89.8	2.1	2.4	98.2	96.4

C.4 Supplementary Case Study

Figure 8 illustrates how ReSET improves both final-answer correctness and judgment reliability under the Reviewed condition. The token coloring complements this behavioral contrast, with higher projections onto the review direction in the backbone's confirming clause (blue) and lower projections at the student's derived quantities and pre-judgment state (red). Qwen3.5-9B and its ReSET student receive the same incorrect initial answer B on an MMMU relative-risk question. The backbone endorses and retains B, repeating the proposed risk without deriving it from the table. The ReSET student instead computes the relative risk from the table, selects the correct option C, and then rejects B.

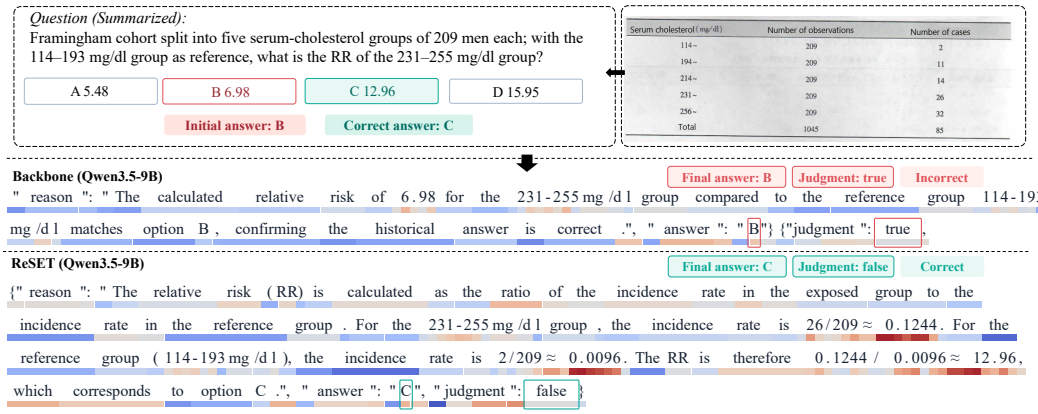


Figure 8: An MMMU example under the Reviewed condition, where ReSET rejects and corrects an erroneous initial answer while the backbone model endorses and retains it.