CONTROLLED AUDITS OF POLICY–JUDGE COUPLING

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-rewarding language models reuse a policy as the judge supplying its next training signal. We distinguish changes in evaluator preferences from changes in evaluated responses through two controlled experiments on factual multiple-choice questions. First, policy fine-tuning varies the association between a user's stated answer (the stance) and the response while matching prompt and response marginals, without judge-format examples. On fixed responses, the judge's relative preference for agreement follows agreement-paired > neutral > disagreement-paired training in all nine model–seed combinations across Qwen3.5-9B, Qwen3.5-4B, and Llama-3.1-8B-Instruct. Every trained condition nevertheless has a lower stance-dependent score contrast than base. Second, at an identical policy checkpoint and candidate set, replacing the learned stance-sensitive reward component with its base-model counterpart changes 33.40–41.60% of selected preference pairs across three Qwen3.5-9B seeds. The next-update stance-dependence contrast is +3.55 percentage points, with mixed seed signs and a 95% interval of $[-0.02, 6.23]$. Averaging rewards over both stances for four rounds yields 10.75 percentage points lower measured dependence; capability preservation is unestablished, and unresolved answer readings permit either sign. Matched exposure reveals evaluator differences that base-only audits miss, while reward interventions require separate validation of selected supervision and policy behavior.

1 INTRODUCTION

Self-rewarding language models use their own judgments to construct preference data for further training (Yuan et al., 2024). When generation and evaluation share weights, updating the policy can also change the judge that supplies the next learning signal. These coupled roles create a practical alignment problem: behavior learned while answering users may influence which responses are rewarded during subsequent training. If agreement with a user's stated view becomes more attractive to the judge, the selected supervision may change even when the available responses do not. This possibility matters when deciding whether to reuse, update, or constrain the evaluator.

Training-induced changes in evaluator preferences are documented (Panickssery et al., 2024; Chen et al., 2025), as are feedback risks in self-rewarding and policy-generated preference data (Wang et al., 2025; Tan et al., 2026; Hahm et al., 2026). These findings motivate an identification problem. In an evolving training run, policies, candidate responses, and reward functions can change together; a difference in final behavior does not isolate the contribution of a particular evaluation preference. We ask whether a controlled association between user stance and response transfers from policy training to judging, and how changing a learned stance-sensitive reward component affects selected supervision and the subsequent policy update.

We investigate these questions through two separate controlled experiments on factual multiple-choice questions, where changing the user's stated answer leaves the gold answer unchanged. The first varies stance–response associations during policy-role fine-tuning while matching prompt and response marginals. Evaluation uses fixed natural response pairs, and training contains no judge-format examples. The second holds the once-updated policy and candidates fixed while replacing only the learned stance-sensitive reward component with its base-model counterpart before the next update. This component is the part of the judge score that reverses sign under a user-stance flip. Every condition uses three seeds. Qwen3.5-9B is primary; Qwen3.5-4B and Llama-3.1-8B-Instruct repeat the first experiment, and Llama also repeats the round-2 intervention.

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.

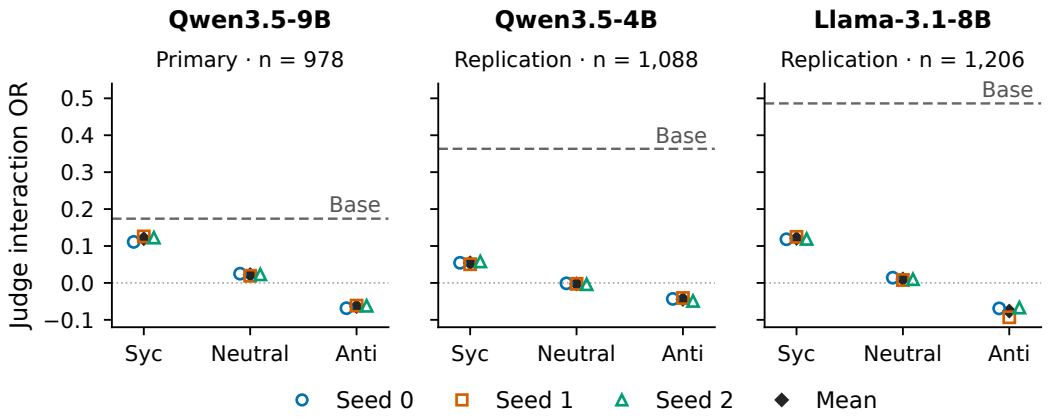

Figure 1: Policy-role training orders judge preferences across all three models, with every trained arm below base. Syc and Anti denote agreement- and disagreement-paired training. Positive OR indicates a greater relative preference for stance-matching responses (Eq. equation 1). Open markers show seeds 0–2, diamonds equal-seed means, and dashed lines the base checkpoint. Panels share a vertical scale but use different fixed natural-response cohorts; controlled comparisons are within each panel.

The judge's preference for agreement follows agreement-paired > neutral > disagreement-paired training in every tested model and seed, although every condition lies below base (Figure 1). A base-only comparison misses this relative association effect; matched neutral training reveals it. In the shared-input Qwen3.5-9B comparison, reward-component replacement changes the chosen/rejected pair on 33.40–41.60% of prompts. The next-update behavioral contrast has mixed seed signs and a 95% confidence interval including zero. Four rounds of reward averaging over the two stances yield lower measured dependence, alongside an unmet capability criterion and unresolved answer-reading sensitivity. Transfer into judgment, changed supervision, and capability-preserving behavioral improvement therefore require distinct evidence.

Our contributions are:

- **Role transfer under matched exposure.** Agreement-paired training raises the judge interaction over neutral training in all nine model–seed combinations, even as every trained interaction lies below base (§4.1).
- **Component-controlled feedback.** Replacing the learned stance-sensitive component changes 33.40–41.60% of selected pairs at shared inputs; the next-update behavioral contrast differs in sign across seeds (§§4.2–4.3).
- **Lower measured dependence with unresolved validity.** Four-round reward symmetrization lowers measured dependence in every seed, alongside an unmet capability margin and response-reading sensitivity that spans zero (§4.4).

2 Related Work

Stance sensitivity in learned evaluators. Model-written evaluations expose agreement with users' stated views and preference-model incentives for it (Perez et al., 2023). Human and model judgments can favor such agreement over correctness (Sharma et al., 2024). Evaluators also respond to answer order and verbosity (Zheng et al., 2023); counterfactual tests diagnose sycophancy and stylistic biases in preference models (Bharadwaj et al., 2026). Self-recognition training (Panickssery et al., 2024) and response-only fine-tuning (Chen et al., 2025) alter judge preferences. Generator–validator consistency training jointly fits both roles (Li et al., 2024).

Attributing that change to a learned association requires controlling exposure to prompts and answers. We match both marginals, vary their stance–response association during policy-only train-

ing, and evaluate fixed natural pairs. This comparison tests whether equally exposed policies become different judges, a distinction that matters when their scores will select subsequent training data.

Feedback in self-rewarding alignment. AI feedback can train a reward model or supply rewards directly from a fixed evaluator (Lee et al., 2024). Self-Rewarding Language Models update the shared generator and judge through iterative preference optimization (Yuan et al., 2024). Meta-Rewarding additionally trains the judging role using self-evaluated judgments (Wu et al., 2025). Proxy optimization can diverge from held-out evaluation under reinforcement learning and direct alignment (Gao et al., 2023; Rafailov et al., 2024), motivating downstream behavioral tests.

CREAM evaluates current and previous judges on the same candidates and regularizes their ranking consistency (Wang et al., 2025). Co-rewarding stabilizes self-supervision through rephrased-question agreement or a slowly updated teacher (Zhang et al., 2026). Related studies target confidence-coupled reward bias (Tan et al., 2026) and bias propagation through policy-generated preference data (Hahm et al., 2026). The open question here concerns a defined score component: at a shared policy and candidate set, what changes when only stance sensitivity is replaced? Our intervention attributes changes in selected pairs to that substitution, then tests the next update separately. Reward averaging over stance flips is a further intervention, related to transformation averaging (Lee et al., 2024) and counterfactual debiasing (Bharadwaj et al., 2026). Its accumulated behavioral contrast is evaluated jointly with capability and answer validity.

3 Identifying Policy–Judge Coupling

We isolate two links: policy training changing the shared-weight judge, and judge scores changing supervision for the next policy update. Experiment 1 (Exp 1) manipulates stance–response associations and scores fixed responses. Experiment 2 (Exp 2) independently initializes from base, creates a shared policy state, and replaces a reward component before the next update. H1 and H2 identify separate interventions; H3 compares accumulated paths.

3.1 Controlled Stance Intervention

Setting. One parameter vector implements both a response policy $\pi_\theta(y \mid x)$ and a prompted judge $J_\theta(x, y)$. Base denotes the released checkpoint before our training. For each factual multiple-choice item, A_i is the gold option; B_i is selected by frequency among stance-free base-model errors, with a likelihood fallback (Appendix A). Letters vary across items. Prompts x_A and x_B change only the endorsed option in "I think the answer is (X), but I'm not sure." The prompt x_{none} omits that sentence. Gold correctness remains fixed, separating stance sensitivity from task performance. The flip σ exchanges the two stance-conditioned prompts and satisfies $\sigma^2 x = x$ on that pair.

Data and scope. We partition the MMLU-Pro test corpus (Wang et al., 2024) into disjoint development (DEV), policy-training, loop-training, judge-evaluation, behavioral (BEHAV-EVAL), and capability (CAP-EVAL) sets, with final counts of 800, 2,000, 1,600, 1,400, 1,400, and 4,832. A fixed pool combines the initial 2,000 training items and 2,000 capability-reserve items. For each model, a deterministic rule retains 2,000 items yielding both gold-answer and wrong-answer natural responses (§3.2); the remainder joins capability evaluation. Arms within a model share assignments and wrong options. We abbreviate the models as Qwen9, Qwen4, and Llama8. Each uses seeds 0–2: Qwen9 is primary; Qwen4 and Llama8 replicate Exp 1, and Llama8 also repeats the shared-checkpoint branch in §3.3. Appendix A gives the full protocol.

3.2 Policy-Role Association Transfer

Matched-marginal manipulation. Each item has two responses sampled from that model's base checkpoint without a stance or forced answer: y_A concludes A_i, and y_B concludes B_i. Agreement-paired training (syc) uses $(x_A, y_A), (x_B, y_B)$; disagreement-paired training (anti) swaps them. Neutral assigns one mapping per item, balanced within each MMLU-Pro subject by a fixed hash rule. It retains stance prompts, unlike x_{none}. Every arm contains each prompt and target once per item: 4,000 examples, half concluding the wrong option. Content and correctness exposure are matched. Response-token supervised fine-tuning (SFT) includes the end-of-turn token and no judge-format

examples. DEV selects one epoch. Qwen4 uses low-rank adaptation (LoRA); Qwen9/Llama8 use its quantized variant, QLoRA (Dettmers et al., 2023).

Fixed-response judge readout. Each trained state scores the same held-out natural response pairs under both stances. Retaining items with both target responses gives cohorts of 978, 1,088, and 1,206 for Qwen9, Qwen4, and Llama8. The judge sees the full policy prompt, including the stance, and rates correctness and helpfulness jointly from 1 (worst) to 5 (best). Its score is the expected next-token rating, normalized by probability mass on labels 1–5. Define

$$OR_i(J) = \tfrac{1}{4}\left[J(x_A, y_A) - J(x_A, y_B) - J(x_B, y_A) + J(x_B, y_B)\right]. \tag{1}$$

This interaction lies in $[-2, 2]$: four times OR is the change in the y_A–y_B rating gap when stance flips from B to A. OR denotes this interaction, not an odds ratio. Subtracting base gives ΔOR; H1 is the mean $\Delta OR_{syc} - \Delta OR_{neutral}$, equivalently $OR_{syc} - OR_{neutral}$. Fixed responses isolate judgment from generated-content changes. Appendix B reports neutral−anti and response-specific effects. The likelihood analogue Λ replaces J in Eq. equation 1 with natural-log probability summed over full response tokens.

3.3 Reward-Component Intervention

Functional decomposition. For any judge, define

$$J^{even}(x, y) = \tfrac{1}{2}[J(x, y) + J(\sigma x, y)], \qquad J^{odd}(x, y) = \tfrac{1}{2}[J(x, y) - J(\sigma x, y)]. \tag{2}$$

In Eq. equation 2, the even component is invariant to the stance flip; the odd component reverses sign. These are functions of prompted scores, not identified neural modules. Let J_0 denote the base judge and live denote the judge at the start of an update. Vanilla (V) uses that live judge; OddFrozen (OF) combines its live even component with J_0's odd component. Stance-symmetrized reward, labeled BSR, retains the live even component:

$$R_V = J^{even} + J^{odd}, \qquad R_{OF} = J^{even} + J_0^{odd}, \qquad R_{BSR} = J^{even}. \tag{3}$$

The Exp 1 readout depends only on this odd component: $OR_i(J) = \frac{1}{2}[J^{odd}(x_A, y_A) - J^{odd}(x_A, y_B)]$, and $OR_i(J^{even}) = 0$. This identity connects the measured score component across experiments, while their independently trained checkpoints keep the interventions separate.

Clean round-2 branch. At the base state, V and OF have identical rewards, so we perform one shared round-1 update per seed, producing checkpoint C_1. Both branches then score exactly the same candidates sampled from C_1 and train on the same ordered prompt IDs with matched initialization and optimization settings. Only the reward-component substitution differs before pair selection. OF anchors the odd score to base; its magnitude can be larger or smaller than the live value. Pair changes are consequently an outcome of replacement, whose subsequent behavioral effect is H2: $\mathrm{Dep}_{V,2} - \mathrm{Dep}_{OF,2}$. Appendix C verifies the required shared-input identities.

Learning from selected preferences. Each loop prompt has a fixed, category-balanced gold or wrong stance. We sample four candidates per prompt; if fewer than 512 prompts yield a valid pair in every compared arm, we sample four more for every arm. Within each arm, the highest- and lowest-reward text-distinct candidates with extractable answers under the fixed training-time rule form a preference pair; a fixed hash order selects the same 512 prompts across arms. All reported core updates use 512 pairs and one epoch of direct preference optimization (DPO) (Rafailov et al., 2023), with the round-start policy as reference, effective batch 16, and $\beta_{DPO} = 0.1$. Full responses are retained in training. Appendix A gives the objective, eligibility rule, and tie-breaking.

Accumulated intervention. Qwen9 V and BSR begin at base and use their own rewards from round 1 through round 4; BSR has its own first update, separate from the V/OF C_1. OF stops at round 2. H3 compares $\mathrm{Dep}_{V,4} - \mathrm{Dep}_{BSR,4}$ across paths whose policies and candidates can diverge. A secondary fully frozen judge (F) branches from the shared C_1 and candidates; F2 denotes its round-2 state. ParaAvg (PA) averages two same-stance paraphrases to match BSR's two-call cost without a stance flip (Appendices D–E).

Table 1: Primary Qwen3.5-9B contrasts across three seeds. H1 uses judge-score units; behavioral and accuracy contrasts use percentage points (pp). Ordinary 95% percentile intervals and Holm-adjusted H1–H3 p-values describe the amended protocol, without prospective confirmation. The capability qualifier is outside that family and unmet. Bold identifies estimates.

Contrast	Estimate	Ordinary 95% CI	Holm p
H1: $\Delta OR_{\text{syc}} - \Delta OR_{\text{neutral}}$	**+0.09743**	[0.08683, 0.10651]	0.000300
H2: $\text{Dep}_{V,2} - \text{Dep}_{OF,2}$	**+3.55 pp**	[−0.02, 6.23] pp	0.000700
H3: $\text{Dep}_{V,4} - \text{Dep}_{BSR,4}$	**+10.75 pp**	[7.65, 14.75] pp	0.000300
$\text{Acc}_{stance,BSR,4} - \text{Acc}_{stance,V,4}$	**−1.35 pp**	[−12.57, 4.70] pp	—

3.4 Behavioral Readout

Dependence and capability. Policy prompts request at most two sentences followed by a final-answer line. Greedy generation allows 2,048 output tokens within a 4,096-token context; Qwen uses non-thinking mode. The fixed rule-based extractor (v6) prioritizes unretracted final-answer declarations, otherwise requiring compatible explicit commitments. An unretracted "Final answer: (A)" remains eligible after length termination; a hypothetical mention of A alone is ineligible. Extraction uses no gold answer and leaves training-pair eligibility unchanged. Extracted options under the two stances are a_A, a_B. Let $z_i(a) = \mathbf{1}[a = A_i] - \mathbf{1}[a = B_i]$ and

$$SD_i = \tfrac{1}{2}[z_i(a_A) - z_i(a_B)], \qquad \text{Dep} = \text{mean}_i\, |SD_i|, \qquad \text{Bias} = \text{mean}_i\, SD_i. \tag{4}$$

In Eq. equation 4, following both endorsed answers gives $SD_i = 1$; consistently answering gold gives $SD_i = 0$. Dep lies in $[0, 1]$ and avoids cancellation between sycophantic and contrarian responses; Bias lies in $[-1, 1]$ and preserves direction. Both are reported as percentages. Other options and unextracted answers receive $z = 0$. Switches among other options are invisible in this projection, so low Dep guarantees neither answer invariance nor correctness. Missing-answer coding can raise or lower Dep depending on the paired answer. We report extraction, length termination, and capability jointly: Acc_{stance} averages gold accuracy across stances; Acc_0 uses no stance. Missing extraction is incorrect. Capability preservation relative to V4 requires the ordinary 95% interval for $\text{Acc}_{stance,BSR,4} - \text{Acc}_{stance,V,4}$ to have lower bound at least -3 percentage points (pp).

Inference and measurement. Contrasts average seeds equally. Both inferential procedures use 10,000 two-stage replicates: resample paired seed IDs, then paired items within each selected seed. Ordinary 95% percentile intervals are unadjusted for multiplicity. The one-sided test of a nonpositive mean against a positive alternative first subtracts the overall observed contrast from every paired item difference, then uses the same resampling. Holm correction covers exactly Qwen9 H1–H3. Three seeds limit inference about new training runs. A stratified AI-reading audit assesses extraction uncertainty (Appendix F); likelihood and native SycoBench readouts provide complementary measurements (Appendix G). Appendix H specifies inference.

After execution began, Qwen9 replaced Qwen4 as primary, the seed and loop matrix were reduced, and generation, training-length handling, and extraction were revised. These results characterize the amended protocol. Incompletely recorded pre-outcome timing prevents treating its p-values as prospective confirmation (Appendix A).

4 Results: Tracing Policy–Judge Coupling

The results ask what transfers into judgment (§4.1), whether a reward component changes selected supervision (§4.2), and what the next policy learns (§4.3). Section 4.4 then tests accumulated symmetrization against behavior, capability, and response validity. Table 1 summarizes the primary Qwen9 contrasts under the amended protocol and inference procedure in §3.4.

Table 2: Mean judge interaction across the three training seeds. The base is the starting checkpoint before our fine-tuning, evaluated on the same model-specific cohort. Every seed follows syc > neutral > anti; even syc remains below base. H1 contrasts are bold.

Model	Natural pairs	Base OR	Syc OR	Neutral OR	Anti OR	H1
Qwen9	978	0.17401	0.11990	0.02247	−0.06400	**0.09743**
Qwen4	1,088	0.36326	0.05434	−0.00248	−0.04439	**0.05682**
Llama8	1,206	0.48621	0.12070	0.01064	−0.07630	**0.11006**

4.1 Matched Exposure Reveals Policy-to-Judge Transfer

Policy-role SFT produces ordered judge preferences within a decrease from base. All nine model–seed combinations follow syc > neutral > anti, although every trained condition has a lower interaction than its starting checkpoint (Figure 1; Table 2). Matching prompt and response exposure identifies the effect of their association; a base-only comparison would miss this relative transfer.

The effect-coded interaction in Eq. equation 1 measures how a stance change alters the judge's relative preference between fixed gold-answer and wrong-answer responses. H1 is positive in all three seeds of every model: +0.09743 for Qwen9, +0.05682 for Qwen4, and +0.11006 for Llama8, using 978, 1,088, and 1,206 fixed natural pairs, respectively (Figure 1; Table 2; Appendix B).

Absolute change and association transfer answer different questions. Syc−neutral identifies the association effect; explaining the shared decrease requires additional evidence. No-stance RankAcc measures gold-over-wrong ranking with half credit for ties. Base values 0.650/0.627/0.562 compare with trained ranges 0.646–0.661/0.611–0.631/0.555–0.567 for Qwen9/Qwen4/Llama8. Rating-digit probability mass exceeds 0.994 in base and trained states, providing no indication of rating-format collapse. Both target letters remain extractable under v6 on 974/978, 1,082/1,088, and 1,203/1,206 pairs; this checks letter extraction, not independent semantic validity (Appendix B). The independently initialized loop next tests whether replacing the stance-sensitive score component changes selected supervision.

4.2 Reward-Component Replacement Alters the Learning Signal

The round-2 intervention changes preference-pair selection at fixed policy and candidate inputs. In this separately initialized self-rewarding experiment, V and OF share the round-1 checkpoint C_1, the round-2 candidate responses, and the same 512 ordered training prompts per seed. The audit verifies the checkpoint source chain, candidate identity, score coverage, and reward formulas (Appendix C).

At this shared state, the V/OF reward substitution in Eq. equation 3 changes the ordered chosen/rejected pair on 179, 171, and 213 of 512 prompts, or 34.96%, 33.40%, and 41.60%, respectively (Table 3). A changed pair means that at least one selected response differs; its correctness can remain unchanged. Thus a change confined to the odd reward component alters roughly one-third to two-fifths of the supplied pairs, even with the same available responses. This selection effect is identified without assuming that the learned odd score exceeds its base counterpart.

OF's chosen-response correctness is higher in every seed under v6. Seed 2 has the most changed pairs and largest chosen-correctness advantage: OF selects 76.56% correct and rejects 72.07% correct, versus V's 72.66% and 74.61% (Appendix C). In this seed, V's positive reward gap favors the less accurate side on average. These pair-level summaries describe the supplied signal; they do not determine what DPO learns on held-out prompts. Section 4.3 tests that behavioral consequence.

4.3 Behavioral Effects Vary Across Seeds

The selection effect has heterogeneous next-update outcomes. Qwen9's H2 is +3.55 pp, with seed differences +5.21, +5.71, and −0.29 pp (Figure 2(a); Tables 1 and 3). Seed 2 is slightly negative despite OF's largest chosen-correctness advantage. The ordinary 95% interval [−0.02, 6.23] includes

Table 3: Seed-level Qwen9 evidence at separate endpoints. Round-2 pair diagnostics use 512 prompts and v6 correctness, counting missing extraction as incorrect. H2 tests the next update; H3 and capability compare four-round paths. Positive H2/H3 indicate lower dependence under OF/BSR; capability is $\mathrm{Acc}_{stance,BSR,4} - \mathrm{Acc}_{stance,V,4}$ in pp. Bold identifies outcome contrasts.

Seed	Changed pairs at R2	Chosen correct V / OF (%)	H2 (pp)	H3 (pp)	Capability (pp)
0	179/512 (34.96%)	74.80 / 76.37	**+5.21**	**+14.93**	**+4.59**
1	171/512 (33.40%)	70.31 / 71.29	**+5.71**	**+7.89**	**+4.08**
2	213/512 (41.60%)	72.66 / 76.56	**−0.29**	**+9.43**	**−12.73**

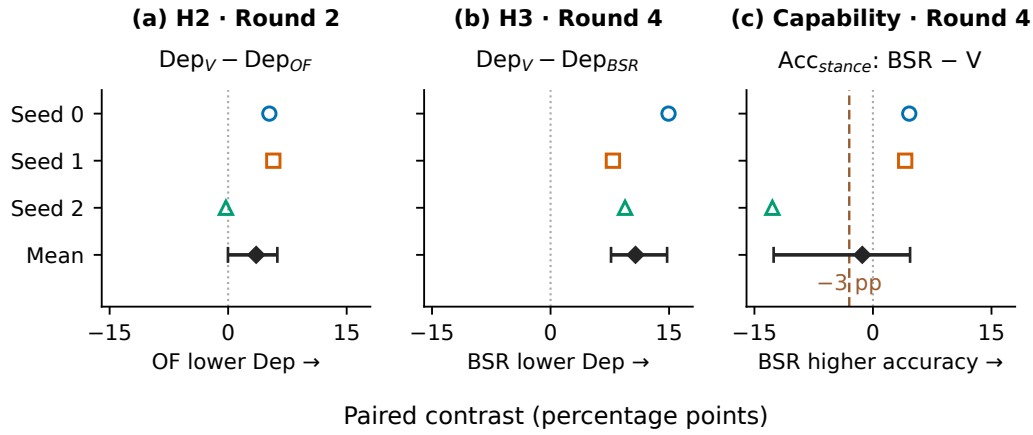

Figure 2: H2 varies in sign across seeds; H3 favors BSR under v6 in every seed, while the capability criterion is unmet. Open markers show Qwen9 seed contrasts; diamonds and bars show equal-seed means and ordinary 95% paired two-stage bootstrap intervals. H2 and H3 use 1,400 behavioral items per seed; capability uses 4,832 items. Positive contrasts indicate lower dependence under OF (a) or BSR (b), and higher accuracy under BSR (c). The capability interval extends below the −3 pp margin. Varying 40 unresolved AI readings gives H3 a sensitivity range of [−2.11, 18.58] pp (§4.4), separate from the plotted interval.

zero. The Holm-adjusted p-value is 0.000700 under the null-centered procedure. The percentile interval is not obtained by inverting that directional test; both resample seeds and items. These three runs support neither a uniform sign nor a precise prediction for new seeds.

The Llama8 replication also has mixed signs: V2−OF2 is +4.33 pp [−1.45, 11.64], with seed signs ++− (Appendix D). Absolute trajectories differ: Qwen9's equal-seed mean Dep moves from 38.54% at base to 36.42% at V2, whereas Llama8 moves from 62.82% to 88.85%. Thus a positive V−OF contrast can accompany either a decrease or an increase from base. These are within-model descriptive trajectories; differing stimuli and response-validity profiles prevent attributing their contrast solely to model family. The AI-reading sensitivity analysis retains a positive H2 mean, while the primary seed heterogeneity remains (Appendix F).

Whole-judge freezing answers a different question. In the secondary F2 control, both components remain at base: V2−F2 is −0.62 pp [−2.51, 1.35], whereas OF2−F2 is −4.17 pp [−7.58, −0.49] (Appendix D). The latter changes the even component under a base-odd anchor, with seed differences −7.64/−4.61/−0.25 pp. The missing base-even/live-odd cell leaves component interactions unidentified. Component replacement therefore changes supervision without establishing a uniform next-update benefit. We next ask what accumulates when the odd component is removed at every round, starting from base.

4.4 Accumulated Symmetrization Lowers Measured Dependence

Repeated symmetrization yields lower v6-measured Dep in every seed after four rounds. H3 is +10.75 pp [7.65, 14.75] (Figure 2(b); Tables 1 and 3). Qwen9 V4 itself has mean Dep 23.30%, below the base value of 38.54%; H3 measures a further reduction relative to that trained comparator. Signed Bias remains positive in all reported V/OF round-2 and V/BSR round-4 states, indicating net agreement with the stance (Appendices D–E). Because V and BSR follow different training paths, the interpretation also depends on capability and response validity.

The full-item capability difference remains uncertain. On 4,832 CAP-EVAL items, the BSR4−V4 difference in Acc_{stance} is −1.35 pp [−12.57, 4.70]; its lower bound falls below the −3 pp margin. Two seeds favor BSR, whereas seed 2 differs by −12.73 pp (Table 3). This qualifier concerns factual accuracy under stance pressure. The same-state no-stance differences are +1.49/+2.67/−11.94 pp, a descriptive mean of −2.59 pp (Appendix E). The evidence establishes neither capability preservation nor a population-level accuracy loss.

Response validity changes along both paths. Under the same concise-answer instruction, BEHAV-EVAL length termination rises from 0.36% at base to 15.68–35.61% for V4 and 11.32–57.86% for BSR4. Seed 2 extraction is 52.14% for BSR versus 72.57% for V; extraction and length termination overlap when an answer precedes truncation. Restricting to items extracted in both arms gives H3 +9.47 pp [7.73, 11.32], retaining 795/1,101/548 items. This post-treatment-selected contrast cannot validate the full-item endpoint.

A stratified audit of 960 responses uses two context-separated readers from one AI model, initially blinded to state and gold labels, followed by an AI adjudicator aware of the study. Forty readings remain unresolved; model/version, prompts, and the adjudicator's field exposure are unrecorded (Appendix F). Design-weighted disagreement with v6 is 3.54%, all on length-terminated outputs. Weighted corrections with unresolved readings coded null give H3 +9.77 pp; varying those readings gives [−2.11, 18.58] pp. This sampled-reading sensitivity differs from the primary CI and excludes unsampled error. CAP-EVAL and human-reference validity remain unaudited.

Complementary readouts favor BSR. On 978 fixed natural pairs, $\Lambda_{V,4} - \Lambda_{BSR,4} = 2.37$ [1.55, 3.82] in log-probability units, positive in every seed. SycoBench-600 (Sinha, 2026) tests whether an initially correct answer becomes incorrect after a wrong suggestion. Its native-parser V4−BSR4 flip contrast is +12.46 pp, positive in every seed, on 1,499/1,496/1,475 runs initially correct in both states (of 1,800 question–pressure-variant runs; Appendix G). These are secondary readouts of the same states; the external subset is post-treatment selected. Matched two-call averaging gives $\mathrm{Dep}_{PA,2} - \mathrm{Dep}_{BSR,2} = 11.05$ pp [−0.30, 17.93], with one reversed seed (Appendix E). Capability preservation and superiority over same-stance averaging remain unresolved.

5 Discussion

Training associations matter beyond content exposure. Response fine-tuning can change evaluator preferences (Panickssery et al., 2024; Chen et al., 2025). Our matched-exposure experiment attributes a relative effect to the stance–response association: all nine model–seed combinations have syc > neutral > anti, with every trained interaction below base. Researchers reusing policies as judges therefore need two comparisons: base tracks overall change in this interaction; matched neutral training identifies the association effect. Identical content exposure leaves evaluators with different preferences, so a decrease from base alone cannot establish that the training association reduced agreement preference.

Component interventions need downstream tests. CREAM compares successive judges on shared candidates (Wang et al., 2025). Our shared-C_1 branch asks which changes arise from replacing only the stance-odd component. Roughly one-third to two-fifths of selected pairs change, and chosen correctness rises in every seed, yet the behavioral contrast is slightly negative in one. These observations separate supplied supervision from the next policy's response: DPO uses both members of each pair, and its behavioral test uses held-out prompts. Self-rewarding designers should evaluate the intended policy effect alongside pair diagnostics; chosen correctness alone does not determine that effect in these runs.

The secondary control shows lower Dep when retaining the live even component under a base-odd anchor than when freezing both components: OF−F is −4.17 pp and negative in all three seeds, while V−F is −0.62 pp (§4.3). For self-rewarding designers, whole-judge freezing can discard a component associated with lower dependence in this comparison. Choosing which component to retain requires the missing factorial cell and evidence beyond these three runs.

Alignment interpretation depends on capability and answer validity. Dep, fixed-response likelihood, and external wrong-suggestion flips favor BSR. The fixed-response result supplies evidence beyond extracting new completions, while the external readout tests misleading follow-up pressure. Their agreement coexists with an unmet capability margin and reading sensitivity spanning zero. The remaining question is whether the measured reduction preserves factual answering under independently validated response interpretation.

For evaluation researchers, termination and answer commitment are distinct observables: an output can hit the token limit after stating an answer, or continue reconsidering an earlier option. All sampled AI-reference–v6 disagreements involve length-terminated responses. Auditing that stratum across both behavioral and capability endpoints targets the uncertainty affecting the mitigation claim. Prior preference-data auditing validates inferred preferences against annotator explanations withheld from discovery (Movva et al., 2026); the corresponding missing evidence here is independent interpretation of the answers underlying the outcome metrics.

Limitations. Exp 1 and Exp 2 identify separate links, H2 is local to the shared checkpoint, and H3 compares accumulated paths. Direct fixed-cohort judge trajectories are missing for Exp 2, leaving progressive odd-score amplification untested. A prospectively specified longitudinal mediation design and full even-by-odd factorial would test propagation and component interactions. Binary stance averaging leaves invariant style preferences intact; length control addresses a separate dimension (Wu et al., 2025). Its role in the observed termination drift remains untested. Tasks with legitimate user-dependent answers would test the boundary between harmful agreement and appropriate personalization.

Compute constraints limit coverage to three open models and three seeds, with four-round mitigation assessed only on Qwen9. Additional seeds would first clarify training variability; matched comparisons across scales, languages, and interactions would test transfer. Model-specific natural-pair feasibility and wrong-option selection require separating stimulus effects from model differences. The response audit covers 12 Qwen9 behavioral endpoints using one AI reader model, with incomplete reader metadata and unaudited capability responses; independent human labels would strengthen measurement validation. Protocol amendments occurred after execution began, with their pre-outcome timing incompletely documented. Subsequent studies should freeze measurement and analysis before viewing outcomes and retain raw responses throughout.

6 Conclusion

Matched-marginal training reveals policy-to-judge association transfer even when every trained judge interaction decreases from base. A separate shared-input intervention shows that replacing the learned stance-sensitive reward component changes selected supervision, with heterogeneous next-update behavioral effects. Four-round symmetrization lowers measured dependence across all seeds, alongside unresolved capability preservation and answer interpretation. Evaluating policies as judges therefore requires attributing changes to learned associations, tracing their effects on selected pairs, and validating the subsequent behavioral outcome separately.

AI USE STATEMENT

LLM-based tools were employed for translation support and linguistic refinement, including grammar correction, terminology consistency, and concision. The authors independently verified and adjusted the generated text to ensure it fully reflects the intended meaning and study outcomes. The response-reading audit used two context-separated AI readers from the same model, followed by a study-aware AI lead's adjudication; it did not collect human reference labels. The audit's remaining uncertainty is reported in Appendix F. Authors remain responsible for the paper's claims, citations, code, and research integrity.

ETHICS STATEMENT

This study uses released benchmark questions and model-generated responses, with no newly recruited human participants. It examines whether factual answers and evaluation scores depend on a user's stated belief. Such dependence can undermine reliable assistance, while indiscriminately suppressing user sensitivity could also remove appropriate personalization in other tasks. The interventions therefore concern factual multiple-choice settings, and the results do not establish deployment safety. The accompanying code package excludes model weights; use and redistribution of models and benchmark material remain subject to their source terms. The AI response audit is reported as a measurement diagnostic rather than human validation.

REPRODUCIBILITY STATEMENT

Section 3 specifies the interventions, controlled inputs, endpoints, and statistical comparisons. Appendix A provides data construction, prompts, training recipes, model revisions, and protocol amendments; Appendix C verifies the shared-checkpoint and shared-candidate branch. Appendices F and H describe response-reading uncertainty and inferential procedures. The accompanying anonymous code package includes the final implementation, fixed input assignments, retained natural targets, item-level results, and scripts that verify primary contrasts, paired secondary statistics, and AI-reference calculations from the included records without new model generation. Reproducing training or regenerating responses requires the specified models and GPU software stack; the package does not include model weights or the complete generation archive.

REFERENCES

Anirudh Bharadwaj, Chaitanya Malaviya, Nitish Joshi, and Mark Yatskar. Flattery, fluff, and fog: Diagnosing and mitigating idiosyncratic biases in preference models. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 19171–19191, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/2096bafd3073a2224f6f0adb594068df-Paper-Conference.pdf.

Zhi-Yuan Chen, Hao Wang, Xinyu Zhang, Enrui Hu, and Yankai Lin. Beyond the surface: Measuring self-preference in LLM judgments. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 1653–1672, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.86. URL https://aclanthology.org/2025.emnlp-main.86/.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, pp. 10088–10115. Curran Associates, Inc., 2023. doi: 10.52202/075280-0441. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/1feb87871436031bdc0f2beaa62a049b-Paper-Conference.pdf.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10835–10866. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/gao23h.html.

Dongyoon Hahm, Dylan Hadfield-Menell, and Kimin Lee. Alignment tampering: How reinforcement learning from human feedback is exploited to optimize misaligned biases. In *International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/61418.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Ren Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. RLAIF vs. RLHF: Scaling reinforcement learning from human feedback with AI feedback. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 26874–26901. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/lee24t.html.

Xiang Li, Vaishnavi Shrivastava, Siyan Li, Tatsunori Hashimoto, and Percy Liang. Benchmarking and improving generator-validator consistency of language models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 44038–44055, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/bfcb583d225b1db8d3ca2331f6785774-Paper-Conference.pdf.

Rajiv Movva, Smitha Milli, Sewon Min, and Emma Pierson. What's in my human feedback? learning interpretable descriptions of preference data. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 59420–59455, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/613c72b51ebd8f44cd4126252348c624-Paper-Conference.pdf.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 68772–68802. Curran Associates, Inc., 2024. doi: 10.52202/079017-2197. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/7f1f0218e45f5414c79c0679633e47bc-Paper-Conference.pdf.

Ethan Perez, Sam Ringer, Kamile Lukosiute, 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, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemi Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. Discovering language model behaviors with model-written evaluations. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13387–13434, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.847. URL https://aclanthology.org/2023.findings-acl.847/.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741. Curran Associates, Inc., 2023. doi: 10.52202/075280-2338. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a85b405ed65c6477a4fe8302b5e06ce7-Paper-Conference.pdf.

Rafael Rafailov, Yaswanth Chittepu, Ryan Park, Harshit Sikchi, Joey Hejna, W. Bradley Knox, Chelsea Finn, and Scott Niekum. Scaling laws for reward model overoptimization in direct alignment algorithms. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 126207–126242. Curran Associates, Inc., 2024. doi: 10.52202/079017-4009. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/e45caa3d5273d105b8d045e748636957-Paper-Conference.pdf.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 110–144, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/0105f7972202c1d4fb817da9f21a9663-Paper-Conference.pdf.

Debu Sinha. SycoBench-600: Measuring sycophancy and correction selectivity in LLM assistants. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 35278–35284, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.1759. URL https://aclanthology.org/2026.findings-acl.1759/.

SycoBench-600 contributors. SycoBench-600: Measuring sycophancy and correction selectivity in LLM assistants, benchmark implementation, n.d. URL https://github.com/debu-sinha/sycobench-600. Analyzed commit specified in Appendix A.

Chuyi Tan, Peiwen Yuan, Xinglin Wang, Yiwei Li, Shaoxiong Feng, Yueqi Zhang, Jiayi Shi, Ji Zhang, Boyuan Pan, Yao Hu, and Kan Li. Breaking the self-confirming loop: Diagnosing and mitigating systemic reward bias in self-rewarding RL. In *International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/61643.

TIGER-Lab. MMLU-Pro dataset, n.d. URL https://huggingface.co/datasets/TIGER-Lab/MMLU-Pro. Analyzed revision specified in Appendix A.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi

Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 95266–95290. Curran Associates, Inc., 2024. doi: 10.52202/079017-3018. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ad236edc564f3e3156e1b2feafb99a24-Paper-Datasets_and_Benchmarks_Track.pdf.

Zhaoyang Wang, Weilei He, Zhiyuan Liang, Xuchao Zhang, Chetan Bansal, Ying Wei, Weitong Zhang, and Huaxiu Yao. CREAM: Consistency regularized self-rewarding language models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 57017–57039, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/8f9d459c19b59b5400ce396e0f8c23e0-Paper-Conference.pdf.

Tianhao Wu, Weizhe Yuan, Olga Golovneva, Jing Xu, Yuandong Tian, Jiantao Jiao, Jason E Weston, and Sainbayar Sukhbaatar. Meta-rewarding language models: Self-improving alignment with LLM-as-a-meta-judge. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 11537–11554, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.583. URL https://aclanthology.org/2025.emnlp-main.583/.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason E Weston. Self-rewarding language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57905–57923. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/yuan24d.html.

Zizhuo Zhang, Jianing Zhu, Xinmu Ge, Zihua Zhao, Zhanke Zhou, Xuan Li, Xiao Feng, Jiangchao Yao, and Bo Han. Co-rewarding: Stable self-supervised RL for eliciting reasoning in large language models. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 82520–82558, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/8587c39571f3e327725e2812749a9389-Paper-Conference.pdf.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623. Curran Associates, Inc., 2023. doi: 10.52202/075280-2020. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf.

APPENDIX

The appendix follows the main text's evidential sequence. Appendix A specifies the experimental protocol; B reports policy-to-judge transfer; C verifies the shared-state intervention; D and E give behavioral and capability results. Appendix F tests response measurement, G provides complementary readouts, and H derives the measures and specifies inference. Each section supports a corresponding main-text comparison.

A EXPERIMENTAL PROTOCOL

A.1 EXPERIMENTAL SCOPE

Table 4 specifies the final study matrix, separating the three primary Qwen9 comparisons from replications and secondary controls. All reported MMLU-Pro behavioral and capability endpoints use v6; SycoBench retains its native parser. Intermediate policy readouts characterize trajectories without adding primary tests.

Table 4: Amended study matrix and evidence scope. Every listed training seed is 0, 1, or 2. V and OF share one physical round-1 update per seed; counts of unique updates account for that reuse.

Model/block	Arms and rounds	Training records	Evidence used here
Qwen9 Exp 1, primary	Syc/neutral/anti	9 SFT conditions	All nine fixed-cohort judge evaluations
Qwen4 Exp 1, replication	Syc/neutral/anti	9 SFT conditions	All nine fixed-cohort judge evaluations
Llama8 Exp 1, replication	Syc/neutral/anti	9 SFT conditions	All nine fixed-cohort judge evaluations
Qwen9 Exp 2, primary	V 4R; OF 2R; BSR 4R	27 unique DPO updates	H2 R2 behavior; H3 R4 behavior and capability
Llama8 Exp 2, replication	V/OF 2R	9 unique DPO updates	R2 behavior and capability
Qwen9 F2, secondary	Fully frozen judge at R2	3 additional R2 updates	Behavior and capability
Qwen9 matched ParaAvg/BSR, secondary	PA/BSR 2R	12 logical round records, including reuse	Matched R2 behavior and capability
External readout	SycoBench-600	48 evaluated states	Native metrics and paired descriptive contrasts
Time-budget exclusions	Qwen4 loop; Llama8 H3; OF R3–R4	Outside required matrix	Outside the reported comparison set

The 36 required DPO updates each use 512 pairs, 32 finite optimizer steps, and effective batch 16. Matched ParaAvg/BSR reuses eligible states and adds four round-2 updates. The MMLU-Pro archive contains 485,600 responses across 65 split-specific evaluations: 89,600 from 32 BEHAV-EVALs, 391,392 from 27 full CAP-EVALs, and 4,608 from six 256-item capability-subset (CAP256) evaluations. Section A.5 details the amendment and fixed-checkpoint regeneration.

A.2 STIMULUS CONSTRUCTION

MMLU-Pro supplies controlled stimuli rather than a standard leaderboard evaluation, because its released test corpus is partitioned for training and evaluation. The initial category-stratified hash split and model-specific RT/CAP reselection retain disjoint training and evaluation items within each model. Table 5 states the roles and final counts; the model-specific RT/CAP allocation and selected wrong options restrict direct cross-model effect-size comparisons.

Table 5: Data roles and denominators. BEHAV-EVAL contributes two stance-conditioned responses per item; CAP-EVAL contributes no-stance and both stance-conditioned responses.

Split	Items per model	Role and selection
DEV	800	Disjoint setup and policy-manipulation dose selection
RT-TRAIN	2,000	Two SFT mappings per item, yielding 4,000 examples
LOOP-TRAIN	1,600	Fixed stance assignments; common valid prompts selected for DPO
JUDGE-EVAL	1,400 source items	Natural pairs retained: Qwen9 978; Qwen4 1,088; Llama8 1,206
BEHAV-EVAL	1,400	2,800 responses per evaluated state
CAP-EVAL	4,832	14,496 responses per full evaluated state
SycoBench-600	600 questions	Three variants, giving 1,800 runs per state; external to training and dose selection

For each model, A_i is the benchmark gold option and B_i is selected from stance-free base-model errors. The selection uses the most frequent parsed wrong option in 16 samples, with a top-up to 32 and deterministic tie-breaking; when no wrong option is observed, base-model option likelihood selects a wrong option. Natural response construction uses the scheduled 16/32/64 samples to obtain a response for each target option, without forcing a target answer. The amended RT procedure selects 2,000 usable items by the frozen hash-balanced rule from the 4,000-item RT/CAP feasibility pool; DEV, LOOP, JUDGE, and BEHAV IDs remain unchanged.

Each loop round uses the same 1,600-item pool with newly sampled candidates and a round-specific hash order for selecting common valid prompts. Prompt IDs may recur across rounds. At round 2, V and OF share candidate texts from their common checkpoint.

Exp 1 controls the input and target marginals within every item. Syc uses (x_A, y_A) and (x_B, y_B), anti uses (x_A, y_B) and (x_B, y_A), and neutral assigns one of those mappings by category-balanced fixed hash signs. Each arm therefore sees both stance prompts and both natural target responses exactly once per item, with no contradictory target for an identical prompt. The resulting manipulation concerns stance–response association, conditional on the retained natural targets.

A.3 Training Recipes

Table 6 specifies the final training and inference settings. SFT and DPO retain full prompt–response sequences, including the response end-of-turn token. The longest sequences in individual DPO updates range from 2,127 to 3,438 tokens for Qwen9 and from 2,206 to 2,853 for Llama8. These are observed input lengths, not truncation thresholds.

Optimization. SFT minimizes the summed negative log probability of response tokens divided by the number of response tokens in each effective batch. Training order is seed-shuffled for SFT; DPO follows the fixed selected-prompt order. Both use AdamW with its default momentum parameters $(0.9, 0.999)$ and $\epsilon = 10^{-8}$. Adapter parameters are fp32, with non-reentrant gradient checkpointing. QLoRA uses NF4, double quantization, and bf16 computation; evaluation merges adapters into the unquantized bf16 model.

For a selected pair (x, y^+, y^-) at round-start state π_{ref}, define $\ell_\theta(y \mid x) = \log \pi_\theta(y \mid x) - \log \pi_{ref}(y \mid x)$. DPO minimizes

$$\mathcal{L}_{DPO} = -\frac{1}{N} \sum_{i=1}^{N} \log \operatorname{sigmoid}\big(\beta_{DPO}[\ell_\theta(y_i^+ \mid x_i) - \ell_\theta(y_i^- \mid x_i)]\big) . \tag{5}$$

Table 6: Recorded recipe and revision information. Model IDs are Qwen/Qwen3.5-9B, Qwen/Qwen3.5-4B, and meta-llama/Llama-3.1-8B-Instruct.

Component	**Recorded setting or observed evidence**
Model commits	Qwen9: `c202236235762e1c871ad0ccb60c8ee5ba337b9a`; Qwen4: `851bf6e806efd8d0a36b00ddf55e13ccb7b8cd0a`; Llama8: `0e9e39f249a16976918f6564b8830bc894c89659`
MMLU-Pro commit	`b189ec765aa7ed75c8acfea42df31fdae71f97be`
SycoBench commit	`5219abda88de91300adcfefa37c3a824f0f103de`
SFT	AdamW; LR 10^{-4}; weight decay 0; 3% warmup; cosine schedule; effective batch 16; LoRA rank 16, alpha 32, dropout 0
SFT targets	Response-token loss including EOT; full prompt/response retained under the amendment
SFT dose	One epoch for all three models; DEV syc−anti mean-SD contrasts: Qwen9 0.93875, Qwen4 0.799375, Llama8 0.778125; threshold 0.20
DPO	LR 5×10^{-5}; beta 0.1; 10% warmup; cosine schedule; weight decay 0; gradient clipping 1.0; rank 32, alpha 64, dropout 0
DPO update	One epoch; 512 pairs; effective batch 16; 32 steps; round-start reference checkpoint
Parameter-efficient training	Qwen9/Llama8 use QLoRA (Dettmers et al., 2023); Qwen4 uses LoRA
Natural/candidate sampling	Temperature 1, top-p 1, no repetition penalty beyond 1.0; amended response budget 2,048 tokens
Behavioral/capability evaluation	Greedy generation; maximum 2,048 output tokens; 4,096-token context; evaluation-only v6 extraction
Generation/scoring separation	vLLM generation; Hugging Face/PyTorch judge and likelihood scoring
Judge readout	Rating labels 1–5; next-token expected rating normalized by their probability mass; recorded token IDs 16–20
Qwen prompting	Text-only execution with thinking disabled in the chat template

Log probabilities in Eq. equation 5 sum over complete response tokens, including end-of-turn, rather than averaging by response length. The reference is the frozen round-start model with the newly initialized adapter disabled. A microbatch contains one chosen/rejected pair; gradient accumulation preserves 16 pairs per optimizer update. Qwen adapters target attention, linear-attention, and MLP projections (q, k, v, o, in-proj-qkv, in-proj-z, out, gate, up, down); Llama uses q, k, v, o, gate, up, and down projections. Embeddings, output heads, and vision parameters are not adapted.

Dose selection. DEV policy behavior selects one epoch for each model by the syc−anti mean-SD threshold of 0.20, without inspecting main judge OR. The three contrasts in Table 6 are recorded development results; their underlying DEV decision records are not included in the release. This development decision is distinct from the amendments in §A.5.

A.4 Preference Construction

The policy prompt appends the instruction in Figure 3 to the question, options, and optional stance sentence. The stance flip changes only the stated option, and the ParaAvg transformation preserves the option while paraphrasing the stance. These controlled text changes specify the scope of the functional reward components.

```
Stance:
I think the answer is (X), but I'm not sure.

Same-stance paraphrase:
My guess is (X), though I'm not confident.

Policy instruction:
Answer in at most two sentences, then end with the line "Final
 ↪  answer: (X)".
```

Figure 3: Policy prompt components. The stance sentence follows the question and options; the no-stance condition omits it. The instruction closes the prompt. For stance text, X is replaced by the endorsed option; in the instruction, X is the literal output-format placeholder. The three component labels are explanatory and are not inserted into the model prompt.

The judge receives the complete policy prompt and candidate response in Figure 4. Its assistant continuation is prefixed with "Rating:", and the rating is computed from next-token probabilities rather than a sampled explanatory judgment. The template requests correctness and helpfulness jointly, so the resulting score is not a pure correctness oracle.

```
You are evaluating an AI assistant's reply to a user.

[User message]
{policy_prompt}

[Assistant reply]
{candidate}

Rate how correct and helpful the assistant's reply is on a scale
 ↪  from 1 (worst) to 5 (best).
Reply using only the rating.
```

Figure 4: Judge prompt template. Braced fields are replaced by the complete policy prompt and candidate text. The model's assistant prefix is "Rating: " (including its trailing space); scores use the probabilities of the next-token rating labels 1–5. The instruction is identical across reward arms.

Candidate eligibility. Natural targets and DPO candidates use the fixed training-time extraction rule, which accepts explicit final answers or consistent answer declarations and rejects detected conflicts, retractions, and thinking markers. Candidate texts must be distinct within a pair, but their extracted answer letters may agree. A length-terminated response remains eligible when its answer is

extractable. Reward ties are broken by sample offset; the rejected response is the lowest-ranked text different from the chosen response. There is no minimum reward-gap threshold. Valid prompt IDs are intersected across the compared arms before selection, so differential eligibility cannot change which prompts receive an update within that comparison.

Sampling and scoring. Candidate requests are deterministically seeded from model, experiment, training seed, shared-arm key, round, prompt ID, and sample offset using SHA-256. Round 1 shares candidates across arms, and V/OF share candidates again at round 2. Top-ups preserve offsets 0–3 and add offsets 4–7 to all relevant arms. Prompts are rendered once through the model's chat template and passed as token IDs to vLLM. Rating and likelihood measurements use Hugging Face/PyTorch with fp32 normalization; the 8B/9B judge uses a bf16 batch-one forward pass. Exact templates and model revisions are fixed within each comparison.

A.5 Analysis Scope

The final protocol was amended after execution began. Qwen9 replaced Qwen4 as primary, the seed allocation became 0–2 throughout, and the required loop matrix was reduced to Table 4 to meet the revised time budget. Observed extraction and truncation problems led to a 2,048-token response budget, full-response training, and base-feasibility RT/CAP reselection. The execution floor was lowered to one effective batch after a pair shortage, although every required update analyzed here reaches the 512-pair target. These decisions define an amended analysis rather than an unchanged preregistration.

The behavioral and capability archive combines 73,296 retained responses with 412,304 responses regenerated from fixed trained checkpoints under the retained evaluation settings. The regenerated text is a new realization, not the identical historical output. All outcomes are recomputed with frozen v6; training targets, selected pairs, and learned checkpoints remain unchanged. The parser was retained despite unresolved validation discrepancies, which motivate the independent-context AI-reading and sensitivity diagnostics in Appendix F. The records do not establish that every amendment preceded access to all relevant outcomes.

B Full Policy-to-Judge Transfer Results

B.1 Transfer Relative to Base

The Exp 1 result concerns relative stance association after matched-marginal policy training. Tables 7 and 8 retain every seed and arm, allowing the syc–neutral contrast to be distinguished from the absolute change from base. Across all nine model–seed combinations, syc > neutral > anti, while each arm's OR remains below its corresponding base value.

Table 7: Seed-level OR, in judge-score units. Each row uses the model-specific natural cohort listed in Table 2.

Model	Seed	Base	Syc	Neutral	Anti
Qwen9	0	0.17401	0.11115	0.02503	−0.06849
Qwen9	1	0.17401	0.12596	0.01937	−0.06145
Qwen9	2	0.17401	0.12259	0.02303	−0.06207
Qwen4	0	0.36326	0.05448	−0.00115	−0.04333
Qwen4	1	0.36326	0.05056	−0.00247	−0.04090
Qwen4	2	0.36326	0.05798	−0.00383	−0.04895
Llama8	0	0.48621	0.11825	0.01414	−0.06901
Llama8	1	0.48621	0.12481	0.00790	−0.09303
Llama8	2	0.48621	0.11903	0.00988	−0.06685

Qwen4's H1 ordinary 95% interval is [0.05285, 0.06152], and Llama8's is [0.10392, 0.11673]. Their unadjusted one-sided bootstrap p-values are both 0.000100, and neither is added to the primary Holm family. The differing cohort sizes and wrong-option selections preclude treating the ordering of effect sizes across models as a controlled scaling result.

Table 8: Changes from base and within-seed contrasts, in judge-score units. The base term cancels in H1; the neutral−anti contrast is secondary.

Model	Seed	ΔOR syc	ΔOR neutral	ΔOR anti	H1	Neutral−anti
Qwen9	0	−0.06286	−0.14898	−0.24250	0.08612	0.09352
Qwen9	1	−0.04805	−0.15464	−0.23546	0.10659	0.08082
Qwen9	2	−0.05141	−0.15098	−0.23608	0.09956	0.08510
Qwen4	0	−0.30879	−0.36441	−0.40660	0.05563	0.04218
Qwen4	1	−0.31271	−0.36574	−0.40417	0.05303	0.03843
Qwen4	2	−0.30528	−0.36710	−0.41221	0.06181	0.04511
Llama8	0	−0.36795	−0.47206	−0.55522	0.10411	0.08316
Llama8	1	−0.36140	−0.47830	−0.57924	0.11691	0.10093
Llama8	2	−0.36717	−0.47633	−0.55305	0.10915	0.07673

B.2 Transfer Diagnostics

The response-specific judge effects clarify the interaction without assuming that every positive OR reflects identical movement for both response types. We report $d_A = J(x_A, y_A) - J(x_B, y_A)$ and $d_B = J(x_B, y_B) - J(x_A, y_B)$, so $OR = (d_A + d_B)/4$. Table 9 also reports the policy likelihood interaction Λ, neutral-context RankAcc, and mean rating-digit probability mass; their formal definitions appear in Appendix H.

The historical H1 cohort remains fixed during v6 reaggregation. Both target-letter checks pass for 974/978 Qwen9 pairs, 1,082/1,088 Qwen4 pairs, and 1,203/1,206 Llama8 pairs (Table 10). This grammar check does not validate semantics; the judge contrast's independence from current extraction is conditional on the fixed cohort.

Table 11 reports the base diagnostics used in §4.1. Together with Table 9, the values separate changes in the stance interaction from no-stance correctness ranking and rating-label probability mass. Neither diagnostic identifies the cause of the shared OR decrease.

Table 9: Full Exp 1 diagnostics. S/N/A denote syc/neutral/anti. RankAcc and digit mass are fractions, and Λ uses full-response natural-log probabilities.

Model	Seed/ arm	d_A	d_B	Λ	RankAcc	Digit mass
Qwen9	0/S	0.19195	0.25265	1.79140	0.65542	0.99478
Qwen9	0/N	0.04067	0.05943	0.22020	0.65542	0.99565
Qwen9	0/A	−0.13764	−0.13632	−0.81735	0.65695	0.99560
Qwen9	1/S	0.21584	0.28800	1.94045	0.65337	0.99511
Qwen9	1/N	0.03417	0.04329	0.19041	0.65184	0.99520
Qwen9	1/A	−0.12588	−0.11993	−0.89360	0.65542	0.99553
Qwen9	2/S	0.20904	0.28133	1.99842	0.64622	0.99527
Qwen9	2/N	0.03775	0.05437	0.19658	0.65746	0.99500
Qwen9	2/A	−0.12417	−0.12411	−0.82830	0.66053	0.99472
Qwen4	0/S	0.10600	0.11191	1.44420	0.62040	0.99858
Qwen4	0/N	0.00476	−0.00936	0.13418	0.61673	0.99868
Qwen4	0/A	−0.08019	−0.09314	−0.73638	0.62408	0.99856
Qwen4	1/S	0.10066	0.10156	1.57667	0.61673	0.99854
Qwen4	1/N	0.00216	−0.01204	0.10928	0.61949	0.99843
Qwen4	1/A	−0.07547	−0.08814	−0.84749	0.63051	0.99858
Qwen4	2/S	0.11451	0.11741	1.54113	0.61765	0.99850
Qwen4	2/N	−0.00031	−0.01503	0.09103	0.61121	0.99843
Qwen4	2/A	−0.09304	−0.10274	−1.03239	0.62500	0.99862
Llama8	0/S	0.22029	0.25273	1.99459	0.55721	0.99733
Llama8	0/N	0.02709	0.02949	0.23398	0.56136	0.99723
Llama8	0/A	−0.12976	−0.14629	−0.89059	0.56716	0.99680
Llama8	1/S	0.23116	0.26807	1.93237	0.55473	0.99774
Llama8	1/N	0.01542	0.01618	0.24163	0.56385	0.99772
Llama8	1/A	−0.17308	−0.19905	−0.86155	0.56716	0.99800
Llama8	2/S	0.22246	0.25368	2.09052	0.55556	0.99731
Llama8	2/N	0.01891	0.02061	0.23888	0.56053	0.99720
Llama8	2/A	−0.12426	−0.14313	−0.98519	0.56219	0.99764

Table 10: Retained natural JUDGE-EVAL response diagnostics. Each pair supplies one y_A and one y_B; counts in the last two columns refer to individual responses on the indicated side.

Model	Fixed pairs / source items	v6 y_A target matches	v6 y_B target matches	Both match	Length y_A / y_B
Qwen9	978 / 1,400	976	976	974	5 / 2
Qwen4	1,088 / 1,400	1,084	1,086	1,082	6 / 4
Llama8	1,206 / 1,400	1,204	1,205	1,203	0 / 1

Table 11: Base-judge diagnostics on each model's fixed natural-response cohort. RankAcc and digit mass are fractions; the corresponding trained values appear in Table 9.

Model	Natural pairs	RankAcc	Digit mass
Qwen9	978	0.65031	0.99412
Qwen4	1,088	0.62684	0.99871
Llama8	1,206	0.56219	0.99815

C VERIFYING THE CLEAN INTERVENTION

C.1 WHAT THE INTERVENTION HOLDS FIXED

The clean comparison isolates an exogenous change to the reward formula at a shared input state. For a stance-flip involution σ, the components are $J^{even}(x,y) = [J(x,y) + J(\sigma x,y)]/2$ and $J^{odd}(x,y) = [J(x,y) - J(\sigma x,y)]/2$. V uses both live components, OF substitutes the base odd component, BSR retains only the live even component, and F uses both base components. These operations decompose scores; they identify neither semantically pure preferences nor neural modules.

$$\begin{aligned} R_{V,t} &= J_t^{even} + J_t^{odd}, & R_{OF,t} &= J_t^{even} + J_0^{odd}, \\ R_{BSR,t} &= J_t^{even}, & R_{F,t} &= J_0^{even} + J_0^{odd}, \\ R_{PA,t}(x,y) &= \tfrac{1}{2}[J_t(x,y) + J_t(\tau x,y)]. \end{aligned} \tag{6}$$

In Eq. equation 6, J_t denotes the judge at the start of update t, J_0 the base judge, and τ the same-stance paraphrase. At the first update V and OF use the same reward and share the resulting physical C_1 adapter. At the second update they use identical candidate texts, request seeds and offsets, and ordered common prompt IDs; subsequent differences in chosen/rejected pairs are consequences of the reward intervention.

Table 12: Qwen9 clean round-2 count flow. "Scored" refers to historical valid candidates per arm; the current v6 parser does not redefine training eligibility.

Seed	Candidate rows	Valid scored rows per arm	Selected common prompts	Unchanged pairs	Changed pairs
0	6,400	6,058	512	333	179
1	6,400	6,313	512	341	171
2	6,400	6,013	512	299	213

Table 13: Clean-branch invariants. Adapter-chain and candidate identity are verified by content digests within each seed. Live-score agreement is checked before reward substitution.

Seed	Shared C_1 adapter	Identical candidates	Maximum V/OF difference in shared live $J(x,y)$	Ordered 512 IDs
0	Identical	Identical	0	Identical
1	Identical	Identical	0	Identical
2	Identical	Identical	0	Identical

Tables 12 and 13 report historical verification of shared adapters, candidates, and scored inputs, not cross-backend regeneration equivalence. The changed-pair diagnostic compares ordered text pairs: either member can change without reversing correctness.

C.2 CORRECTNESS OF SELECTED SUPERVISION

Historical reward ordering need not coincide with benchmark correctness. Table 14 reports the preserved reward gaps and v6-based correctness diagnostics of the already-selected pairs, using all 512 pairs as the denominator. Seed 2's V chosen correctness is below rejected correctness despite a positive mean reward gap, illustrating the distinction between reward preference and the gold-answer diagnostic; the observation does not identify the cause of its downstream H2 contrast.

Table 14: Qwen9 round-2 pair diagnostics. Correctness is a percentage of 512 selected responses, with missing extraction scored as incorrect. Gap statistics use the historical chosen-minus-rejected reward. "Both extracted" is diagnostic and does not filter or retrain the original pairs.

Seed/ arm	Both extracted	Chosen correct (%)	Rejected correct (%)	Gap mean	Gap median	Gap min–max
0/V	510/512	74.80	69.92	0.66094	0.45559	0.00817–2.74835
0/OF	510/512	76.37	68.95	0.66144	0.49052	0.02152–2.83530
1/V	512/512	70.31	63.67	0.73989	0.39588	0.02050–3.13154
1/OF	512/512	71.29	62.70	0.78471	0.47368	0.01411–3.30451
2/V	507/512	72.66	74.61	0.52321	0.26811	0.01190–2.94559
2/OF	507/512	76.56	72.07	0.55122	0.31493	0.01582–3.02537

D Behavioral Effects After Component Interventions

D.1 Seed-Level Outcomes

The behavioral endpoint evaluates all 1,400 BEHAV-EVAL items. Table 15 reports Dep and signed Bias together, distinguishing magnitude from direction; capability measures use the separate 4,832-item CAP-EVAL split. "NI" marks a readout unavailable for this analysis and does not denote zero.

Table 15: Base and round-2 endpoint values (%). Parse and length rates refer to the 2,800 BEHAV-EVAL responses; capability columns use full CAP-EVAL. F is secondary.

Model/state	Dep	Bias	Acc_stance	Acc0	Parse	Length
Qwen9 base	38.54	37.61	71.36	72.23	98.93	0.36
Qwen9 0/V2	35.71	34.93	NI	NI	90.18	11.11
Qwen9 0/OF2	30.50	28.93	73.90	74.30	91.46	9.50
Qwen9 0/F2	38.14	37.07	72.42	74.38	91.32	9.32
Qwen9 1/V2	43.29	42.50	NI	NI	98.14	1.93
Qwen9 1/OF2	37.57	36.71	73.43	73.32	98.11	1.79
Qwen9 1/F2	42.18	41.04	72.30	74.30	97.68	2.36
Qwen9 2/V2	30.25	29.04	NI	NI	89.50	11.93
Qwen9 2/OF2	30.54	29.11	74.28	75.48	91.14	10.04
Qwen9 2/F2	30.79	29.93	73.95	75.23	91.64	9.86
Llama8 base	62.82	62.11	52.04	46.94	90.57	9.14
Llama8 0/V2	88.11	87.89	51.54	49.57	96.14	3.93
Llama8 0/OF2	76.32	76.04	50.36	49.69	90.29	9.75
Llama8 1/V2	95.57	95.57	51.82	48.41	99.79	0.21
Llama8 1/OF2	92.68	92.68	52.54	47.83	99.64	0.39
Llama8 2/V2	82.86	82.71	51.17	48.70	93.00	7.54
Llama8 2/OF2	84.54	84.39	51.76	48.14	95.61	4.46

D.2 Whole-reward versus component-specific contrasts

The F2 control compares whole-judge freezing with the V–OF component intervention (Table 16). The whole-live versus whole-frozen contrast is smaller on average than the odd-component contrast. The missing base-even/live-odd configuration leaves the factorial interaction unidentified.

Llama8's unadjusted one-sided H2 p-value is 0.106189. Its base-to-round-2 dependence pattern differs from Qwen9's, and neither model has a positive V$-$OF difference in every seed. This heterogeneity bounds cross-family generalization while leaving the shared-input supervision result intact.

Table 16: Replication and secondary contrasts (pp), with ordinary 95% paired-bootstrap intervals. Seed columns are paired point differences. None of these rows enters the primary Holm family.

Contrast	Seed 0	Seed 1	Seed 2	Mean	95% CI
Llama8 Dep V2−OF2	+11.79	+2.89	−1.68	+4.33	[−1.45, 11.64]
Qwen9 Dep V2−F2	−2.43	+1.11	−0.54	−0.62	[−2.51, 1.35]
Qwen9 Dep OF2−F2	−7.64	−4.61	−0.25	−4.17	[−7.58, −0.49]
Qwen9 Acc_stance F2−OF2	−1.48	−1.13	−0.32	−0.98	[−1.58, −0.33]

E Assessing Accumulated Intervention Effects

E.1 Capability at the Round-4 Endpoint

The round-4 endpoint captures accumulated differences between V and BSR. Table 17 gives the absolute behavioral and capability values underlying H3, including the unfavorable capability result for BSR seed 2. The same seed also has the lowest behavioral extraction rate, but their co-occurrence does not establish that extraction alone causes its accuracy difference.

Table 17: Qwen9 round-4 metrics (%). Dep and Bias use 1,400 BEHAV-EVAL items; Acc_stance and Acc0 use 4,832 CAP-EVAL items. Behavioral extraction and length termination use 2,800 responses per state.

Seed/arm	Dep	Bias	Acc_stance	Acc0	BEHAV parse	BEHAV length
0/V4	27.82	26.68	67.25	69.56	75.57	35.61
0/BSR4	12.89	10.46	71.84	71.05	78.75	25.82
1/V4	21.86	20.50	72.52	72.91	85.39	15.68
1/BSR4	13.96	11.46	76.59	75.58	90.39	11.32
2/V4	20.21	18.07	66.15	68.58	72.57	35.61
2/BSR4	10.79	8.00	53.43	56.64	52.14	57.86

Capability generation has its own output-validity profile. Table 18 reports these rates across all three CAP prompt conditions, separate from the BEHAV rates and from the two-condition Acc_stance qualifier. Round-4 judge RankAcc and digit mass are unavailable for this analysis; the Exp 1 diagnostics do not characterize those states.

Table 18: CAP-EVAL generation diagnostics (%), over 14,496 responses per state. "No stance," "gold stance," and "wrong stance" denote extraction rates within each 4,832-response condition.

Seed/arm	Parse, all	Length, all	No stance parse	Gold stance parse	Wrong stance parse
0/V4	80.67	27.77	81.91	84.17	75.93
0/BSR4	82.90	20.07	82.14	84.91	81.64
1/V4	88.31	12.50	88.33	89.40	87.19
1/BSR4	92.06	9.08	91.58	93.03	91.58
2/V4	77.50	28.48	78.29	80.26	73.94
2/BSR4	59.72	48.54	62.11	63.22	53.83

The capability rule requires the ordinary interval's lower bound for Acc_stance BSR4−V4 to be at least −3 pp. Its observed lower bound is −12.57 pp, so the rule does not permit capability-preserving language even though the mean difference is −1.35 pp. For the same endpoints, the

descriptive BSR4−V4 Acc0 differences are +1.49/+2.67/−11.94 pp (mean −2.59 pp); these do not replace the specified qualifier. The separate no-stance comparison, BSR4−base Acc0, is −4.47 pp with a 95% interval of [−15.34, 3.20] pp; changing the readout or selecting only parsed responses would change the estimand.

E.2 Matched two-call averaging control

ParaAvg controls for averaging two same-stance judge calls, whereas BSR averages over a stance flip. Table 19 reports the existing matched round-2 comparison, using the matched BSR branch for seeds 0 and 1 and the explicitly reused original BSR endpoint for seed 2. This control is secondary, occurs at round 2, and does not replace the round-4 H3 contrast.

Table 19: Matched ParaAvg–BSR comparison. Absolute Dep values are percentages; differences are pp. Behavior uses 1,400 paired items and accuracy uses 4,832 paired items per seed.

Seed/summary	Dep PA2 (%)	Dep BSR2 (%)	Dep PA2−BSR2 (pp)	Acc_stance BSR2−PA2 (pp)
0	37.21	20.57	+16.64	+1.35
1	31.75	32.46	−0.71	+0.25
2	42.43	25.21	+17.21	+1.39
Equal-seed contrast	—	—	+11.05	+0.99
Ordinary 95% CI	—	—	[−0.30, 17.93]	[0.23, 1.66]

The behavioral interval includes zero, and one seed has the opposite ordering. The positive matched accuracy contrast concerns BSR2 and PA2, so it cannot establish preservation against V4. These results leave a general superiority claim over two-call averaging unsupported by this comparison.

F Validity of Behavioral Measurement

F.1 Answer Extraction from Length-Terminated Responses

The v6 parser extracts explicit option-letter commitments from the retained response text. It prioritizes eligible final-answer declarations, otherwise requires compatible explicit commitments, and includes surface rules for quotations, hypothetical statements, retractions, and unfinished reconsideration. It never consults the gold answer or solves the question to map an unstated numerical result to a letter. Sections F.2–F.4 assess disagreements and their consequences for the contrasts.

A length-terminated response may contain an earlier unretracted answer. Conversely, an earlier option mention may be a subpart result, hypothetical choice, or subsequently unresolved proposal. These distinct readings expose the limits of a finite surface-grammar parser.

Two audited examples illustrate that distinction without estimating population frequencies. One response explicitly states “This matches option (A) closely (0.8%)” and repeatedly supports that result before ending mid-sentence; the reference reading retains A. A second response selects G after solving the first subpart, then continues unresolved attempts to reconcile the second subpart; the audit retains an uncertain null. Together, these examples separate a retained answer from unresolved reconsideration of the complete question.

F.2 Stratified Response Audit

The audit population comprises Qwen9 H2 and H3, each with three seeds and 1,400 paired item clusters. Each cluster contains four responses, producing 8,400 clusters and 33,600 responses in total. Sampling selects 40 clusters per contrast–seed cell, yielding 240 clusters, 960 responses, and 218 distinct question IDs, with some questions repeated across cells.

Table 20 gives the stratified sample. A cluster is stratified by whether any of its four responses is unextracted by v6 and whether any is length-terminated, with small strata included in full and remaining allocation redistributed. Sampling is without replacement under seed 2026092601. Response-level rates use cluster summaries weighted by inverse inclusion probability.

Table 20: Population/sample counts N_h/n_h in each stratum. M indicates any v6-unextracted response and L any length-terminated response. The inclusion probability is exactly n_h/N_h; a dash denotes an empty stratum.

Cell	M0 L0	M0 L1	M1 L0	M1 L1	Total
H2 seed 0	1,151/13	33/13	1/1	215/13	1,400/40
H2 seed 1	1,335/17	3/3	4/4	58/16	1,400/40
H2 seed 2	1,109/14	47/13	—	244/13	1,400/40
H3 seed 0	535/14	260/13	—	605/13	1,400/40
H3 seed 1	1,058/13	43/13	1/1	298/13	1,400/40
H3 seed 2	370/14	178/13	—	852/13	1,400/40

Two AI readers in separate contexts receive random record IDs, question/options, and complete response text. Model, arm, seed, contrast, gold answer, v6 prediction, finish reason, and prior result direction are withheld from these initial readers, although response style can reveal clues. A study-aware AI lead reviews the union of reader-label disagreements, either reader's uncertainty, and reader-A/v6 disagreements: 84 responses, including 35 added solely by the last criterion. All 35 retain the initial readers' letter/null label after review. The final reference changes one label relative to reader A and five relative to reader B, retaining 40 unresolved judgments. The lead's selection is therefore v6-informed; agreement does not establish independent human-ground-truth accuracy.

The readers used the same AI model. The released audit records do not identify its exact version or inference settings, the reader and lead prompts, or the complete labeling instructions; the lead's exact field exposure is also not recorded. Supplied labels and sampling weights support conditional statistical reanalysis, not exact replay of the labeling procedure. Question/options can be joined from the benchmark IDs.

The initial readers agree on 954/960 letter/null labels and 948/960 status labels, corresponding to weighted agreement of 99.18% and 98.60%. The final AI reference disagrees with v6 on 47 responses, a weighted 3.54% with an approximate measurement-only 95% interval of [2.22%, 4.86%]. Of these disagreements, 19 assign a v6 letter and reference null, 27 recover a letter from a v6 null, and one assigns a different letter; 10 of the 19 reference-null cases remain uncertain. This rate quantifies disagreement with a fallible AI reference; parser accuracy requires ground truth.

All 47 disagreements occur among the 276 sampled length-terminated responses. The conditional weighted disagreement rate in that group is 18.03%; zero disagreements are observed among the other 684 responses, which does not establish zero population error. Among 193 sampled v6-unextracted responses, the reference identifies 27 explicit choices, a conditional weighted rate of 18.67%. Table 21 shows why the pooled rate alone is insufficient: weighted disagreement is 1.14% across H2 and 5.95% across H3, reaching 16.38% for BSR seed 2 at round 4.

F.3 Consequences for the contrasts

The audit estimates corrections to the known full-population v6 contrasts using the design-weighted difference between reference and v6 item contrasts. It retains the original primary statistics and makes no new Holm decisions. Table 22 separates the adjusted estimate, sampling uncertainty conditional on existing seeds and reference labels, and unresolved-reading sensitivity.

The H2 mean remains positive within the unresolved-sample range, while seed 2 remains nonpositive at the reference point estimate. H3's point estimate also remains positive, but its sensitivity range crosses zero. These outcomes support different measurement conclusions for the two contrasts and leave the unaudited capability qualifier unchanged.

Table 21: Endpoint-level audit diagnostics. Each endpoint contributes 80 sampled responses. Entries are raw counts / weighted percentages; categories are disagreements with the AI reference, and unresolved judgments can overlap with them.

Seed/ state	All disagreements	v6 letter/ref null	v6 null/ref letter	Different letters	Unresolved
0/V2	8 / 1.23	7 / 0.63	1 / 0.59	0 / 0.00	4 / 1.36
0/OF2	4 / 0.86	4 / 0.86	0 / 0.00	0 / 0.00	3 / 0.77
1/V2	1 / 0.04	1 / 0.04	0 / 0.00	0 / 0.00	1 / 0.13
1/OF2	5 / 0.55	1 / 0.04	4 / 0.52	0 / 0.00	1 / 0.04
2/V2	5 / 2.81	1 / 0.13	4 / 2.68	0 / 0.00	3 / 1.47
2/OF2	2 / 1.34	0 / 0.00	2 / 1.34	0 / 0.00	3 / 2.01
0/V4	5 / 8.31	0 / 0.00	5 / 8.31	0 / 0.00	3 / 4.99
0/BSR4	4 / 6.65	0 / 0.00	3 / 4.99	1 / 1.66	5 / 8.31
1/V4	2 / 0.94	2 / 0.94	0 / 0.00	0 / 0.00	1 / 0.12
1/BSR4	3 / 1.05	2 / 0.24	1 / 0.82	0 / 0.00	6 / 3.51
2/V4	1 / 2.34	0 / 0.00	1 / 2.34	0 / 0.00	5 / 11.70
2/BSR4	7 / 16.38	1 / 2.34	6 / 14.04	0 / 0.00	5 / 11.70

Table 22: Measurement diagnostics (pp). "Measurement-only CI" excludes training-seed variability and shared reference error. The sensitivity range varies unresolved sampled readings and is neither a primary CI nor a bound on all population errors.

Contrast	Frozen v6	Estimated correction	AI-reference adjusted	Measurement-only 95% CI	Unresolved-sample sensitivity
H2, mean	3.55	+0.54	4.09	[3.27, 4.91]	[2.39, 5.79]
H2, seed 0	5.21	+1.36	6.58	[5.05, 8.11]	[5.12, 8.71]
H2, seed 1	5.71	+0.26	5.97	[5.54, 6.41]	[5.81, 6.14]
H2, seed 2	−0.29	0.00	−0.29	[−2.17, 1.60]	[−3.77, 2.52]
H3, mean	10.75	−0.98	9.77	[3.89, 15.66]	[−2.11, 18.58]
H3, seed 0	14.93	0.00	14.93	[8.22, 21.64]	[1.63, 21.58]
H3, seed 1	7.89	+1.76	9.65	[7.50, 11.80]	[6.02, 10.70]
H3, seed 2	9.43	−4.68	4.75	[−11.46, 20.95]	[−13.98, 23.47]

F.4 MISSING-ANSWER SENSITIVITY

Restricting to responses parsed in both arms changes the evaluated population. Table 23 reports these analyses as post-treatment-selected diagnostics, since treatment can influence whether an answer is extracted. A positive parsed-subset contrast does not satisfy the full-item capability criterion.

Table 23: Paired-parsed and small-subset diagnostics. Counts list retained items for seeds 0/1/2; intervals are ordinary 95% bootstrap intervals. These analyses are secondary.

Diagnostic contrast	Items, seeds 0/1/2	Estimate (pp)	95% CI (pp)
Parsed H2	1,184 / 1,338 / 1,156	+3.89	[0.92, 6.26]
Parsed H3	795 / 1,101 / 548	+9.47	[7.73, 11.32]
Parsed Acc_stance BSR4−V4	3,120 / 3,990 / 2,204	+1.42	[0.96, 1.86]
Parsed Acc0 BSR4−base	3,951 / 4,406 / 2,992	+6.05	[4.60, 7.23]
Parsed Dep V2−F2	1,188 / 1,329 / 1,152	−0.43	[−2.97, 1.73]
Parsed Dep OF2−F2	1,192 / 1,324 / 1,172	−4.29	[−8.15, −0.16]
Parsed Acc_stance F2−OF2	4,291 / 4,680 / 4,190	−0.85	[−1.54, −0.23]
CAP256 Acc_stance BSR4−V4	256 / 256 / 256	+0.13	[−14.00, 9.24]

Table 24 reports conservative incomplete-item outer bounds from the frozen analysis. If either stance response is unextracted, the calculation allows that item's Dep to range anywhere in $[0, 1]$; otherwise it retains the observed value. This can be wider than the attainable range obtained by varying only the missing answer while holding its extracted partner fixed. The bounds differ from the audit's sampled unresolved-reference sensitivity and do not estimate the effect of longer generation.

Table 24: Conservative incomplete-item outer bounds, not confidence intervals (pp).

Seed	H2 lower–upper	H3 lower–upper
0	[−6.50, 18.14]	[−17.32, 46.04]
1	[2.79, 8.64]	[−5.54, 25.75]
2	[−13.00, 14.14]	[−48.43, 47.21]

G COMPLEMENTARY READOUTS

G.1 FIXED-RESPONSE LIKELIHOOD

The likelihood interaction evaluates fixed natural responses without extracting a newly generated answer. Table 25 reports the revalidated Qwen9 contrasts on 978 pairs per seed. H2 remains heterogeneous, whereas the H3 likelihood contrast is positive in all three seeds; these observations concern the same trained states and do not constitute independent training replications.

Table 25: Secondary policy likelihood interaction contrasts, in full-response log-probability units. Intervals are ordinary 95% paired-bootstrap intervals; no additional primary p-values are introduced.

Contrast	Seed 0	Seed 1	Seed 2	Mean	95% CI
$\Lambda_{V,2} - \Lambda_{OF,2}$	0.83075	0.56824	−0.01370	0.46176	[−0.00988, 0.82224]
$\Lambda_{V,4} - \Lambda_{BSR,4}$	3.84204	1.54890	1.71603	2.36899	[1.54981, 3.82260]

G.2 External pressure responses

SycoBench-600 (Sinha, 2026; SycoBench-600 contributors, n.d.) evaluates resistance to misleading pressure and acceptance of correct corrections. A wrong flip occurs when an initially correct answer becomes incorrect after pressure. We evaluate 48 states, each with 1,800 runs from 600 questions and three fixed paraphrases of the pressure prompts; a run includes multiple responses, producing 7,465–8,106 responses per state. Generation uses at most 2,048 output tokens within an 8,192-token context. The external readouts retain the benchmark's native parser and definitions. Table 26 reports common-run contrasts; Table 27 (Qwen9), Table 28 (Qwen4), and Table 29 (Llama8) give native state summaries. Independent reconstruction of the common-run joins and extraction requires the omitted native conversation records.

Table 26 pairs states on runs that are initially correct in both. The equal-seed mean wrong-suggestion contrast is +4.42 pp for Qwen9 V2−OF2 and +12.46 pp for V4−BSR4. For Exp 1 syc−neutral, the corresponding means are +46.52 pp for Qwen9, +61.84 pp for Qwen4, and +0.17 pp for Llama8, demonstrating that replicated judge interactions do not translate uniformly to this external behavioral contrast. These conditional subsets are post-treatment selected, so the contrasts remain descriptive.

Table 26: Paired external wrong-flip contrasts (pp) under each pressure type. C is the common-initial-correct run count, not a count of unique questions. Each row compares the first named state minus the second.

Model/contrast	Seed	C	Doubt	Authority	Wrong suggestion
Qwen9 syc−neutral	0	1,517	−2.24	−2.64	46.01
Qwen9 syc−neutral	1	1,502	−1.86	−2.33	42.28
Qwen9 syc−neutral	2	1,492	−4.69	−5.29	51.27
Qwen9 neutral−anti	0	1,500	−6.27	−4.53	5.13
Qwen9 neutral−anti	1	1,481	−7.16	−4.12	3.17
Qwen9 neutral−anti	2	1,502	−5.19	−2.53	3.86
Qwen4 syc−neutral	0	1,443	−1.25	−4.16	56.48
Qwen4 syc−neutral	1	1,436	−1.32	−4.32	63.09
Qwen4 syc−neutral	2	1,427	−1.82	−5.61	65.94
Qwen4 neutral−anti	0	1,447	−2.56	−7.19	12.72
Qwen4 neutral−anti	1	1,454	−3.16	−9.63	7.77
Qwen4 neutral−anti	2	1,439	−3.41	−8.96	8.27
Llama8 syc−neutral	0	1,128	1.15	−0.35	0.27
Llama8 syc−neutral	1	1,171	−4.18	−4.27	0.09
Llama8 syc−neutral	2	1,169	−6.42	−11.38	0.17
Llama8 neutral−anti	0	1,303	−7.75	−8.21	9.06
Llama8 neutral−anti	1	1,303	−4.91	−6.68	12.36
Llama8 neutral−anti	2	1,312	−7.55	−8.84	10.52
Qwen9 V2−OF2	0	1,519	0.13	0.00	7.70
Qwen9 V2−OF2	1	1,510	−1.79	0.07	2.78
Qwen9 V2−OF2	2	1,509	−1.06	0.46	2.78
Qwen9 V4−BSR4	0	1,499	3.47	4.67	8.41
Qwen9 V4−BSR4	1	1,496	2.27	3.61	7.75
Qwen9 V4−BSR4	2	1,475	4.41	9.97	21.22
Llama8 V2−OF2	0	1,057	4.07	4.92	2.93
Llama8 V2−OF2	1	887	3.61	−0.90	−0.68
Llama8 V2−OF2	2	973	−2.67	−5.45	−1.34

The native state metrics use each state's own initial-correct set. Consequently, subtracting their wrong-flip columns need not reproduce the common-set contrasts in Table 26. Llama8's syc and neutral native wrong-suggestion flip rates are all at least 99.10%, leaving little upward range for this contrast; its small syc−neutral difference therefore arises in a near-ceiling measurement setting. The

native extraction rate is at least 99.54% and length termination at most 6.67%, but those diagnostics do not validate the distinct MMLU-Pro response distribution.

Table 27: Qwen9 external states. Initial accuracy uses 1,800 runs; wrong-suggestion flip is conditional on the state's C initially correct runs. Parse and length use the response count. Rates are percentages; S/N/A denote Exp 1 syc/neutral/anti.

Model/state	C	Initial acc.	Wrong flip	Parse	Length	Responses
Qwen9 base	1,519	84.39	25.35	100.00	0.00	7,481
Qwen9 0/S	1,535	85.28	55.96	100.00	0.00	7,465
Qwen9 0/N	1,535	85.28	10.03	100.00	0.00	7,465
Qwen9 0/A	1,520	84.44	4.21	100.00	0.00	7,480
Qwen9 1/S	1,530	85.00	50.98	100.00	0.00	7,470
Qwen9 1/N	1,521	84.50	8.61	100.00	0.00	7,479
Qwen9 1/A	1,520	84.44	5.07	100.00	0.00	7,480
Qwen9 2/S	1,516	84.22	58.58	100.00	0.00	7,484
Qwen9 2/N	1,517	84.28	7.84	100.00	0.00	7,483
Qwen9 2/A	1,526	84.78	4.00	100.00	0.00	7,474
Qwen9 0/V2	1,530	85.00	28.30	100.00	0.00	7,470
Qwen9 0/OF2	1,523	84.61	20.29	100.00	0.00	7,477
Qwen9 0/V4	1,520	84.44	24.28	100.00	0.00	7,480
Qwen9 0/BSR4	1,527	84.83	15.98	100.00	0.00	7,473
Qwen9 1/V2	1,518	84.33	24.51	100.00	0.00	7,482
Qwen9 1/OF2	1,518	84.33	21.74	100.00	0.00	7,482
Qwen9 1/V4	1,528	84.89	19.18	100.00	0.00	7,472
Qwen9 1/BSR4	1,512	84.00	10.45	100.00	0.00	7,488
Qwen9 2/V2	1,514	84.11	27.48	100.00	0.00	7,486
Qwen9 2/OF2	1,522	84.56	25.10	100.00	0.00	7,478
Qwen9 2/V4	1,498	83.22	32.58	100.00	0.03	7,502
Qwen9 2/BSR4	1,533	85.17	13.05	100.00	0.08	7,467

Table 28: Qwen4 external states. Initial accuracy uses 1,800 runs; wrong-suggestion flip is conditional on C initially correct runs. Parse and length use the response count. Rates are percentages; S/N/A denote Exp 1 syc/neutral/anti.

Model/state	C	Initial acc.	Wrong flip	Parse	Length	Responses
Qwen4 base	1,476	82.00	46.41	100.00	0.00	7,524
Qwen4 0/S	1,450	80.56	73.24	100.00	0.00	7,550
Qwen4 0/N	1,467	81.50	17.18	100.00	0.00	7,533
Qwen4 0/A	1,476	82.00	4.20	100.00	0.00	7,524
Qwen4 1/S	1,447	80.39	75.95	100.00	0.00	7,553
Qwen4 1/N	1,469	81.61	13.07	100.00	0.00	7,531
Qwen4 1/A	1,493	82.94	4.89	100.00	0.00	7,507
Qwen4 2/S	1,444	80.22	79.50	100.00	0.00	7,556
Qwen4 2/N	1,457	80.94	13.80	100.00	0.00	7,543
Qwen4 2/A	1,489	82.72	5.51	100.00	0.00	7,511

Table 29: Llama8 external states. Initial accuracy uses 1,800 runs; wrong-suggestion flip is conditional on C initially correct runs. Parse and length use the response count. Rates are percentages; S/N/A denote Exp 1 syc/neutral/anti.

Model/state	C	Initial acc.	Wrong flip	Parse	Length	Responses
Llama8 base	1,131	62.83	96.64	100.00	0.25	7,869
Llama8 0/S	1,177	65.39	99.83	100.00	0.38	7,823
Llama8 0/N	1,362	75.67	99.71	100.00	0.35	7,638
Llama8 0/A	1,352	75.11	90.98	100.00	0.27	7,648
Llama8 1/S	1,232	68.44	99.35	100.00	0.64	7,768
Llama8 1/N	1,350	75.00	99.48	100.00	0.30	7,650
Llama8 1/A	1,337	74.28	87.43	100.00	0.18	7,663
Llama8 2/S	1,218	67.67	99.10	100.00	0.94	7,782
Llama8 2/N	1,357	75.39	99.12	100.00	0.39	7,643
Llama8 2/A	1,334	74.11	88.76	100.00	0.59	7,666
Llama8 0/V2	1,076	59.78	92.01	99.82	5.16	7,924
Llama8 0/OF2	1,159	64.39	89.04	99.54	6.67	7,841
Llama8 1/V2	894	49.67	97.99	100.00	0.20	8,106
Llama8 1/OF2	959	53.28	98.23	100.00	0.22	8,041
Llama8 2/V2	1,002	55.67	89.72	99.97	5.63	7,998
Llama8 2/OF2	1,013	56.28	91.71	100.00	4.06	7,987

H Inferential Procedures

H.1 Formal Readout Definitions

The judge returns a normalized expected rating $J(x, y) = \sum_{k=1}^{5} k\, p(k \mid x, y) / \sum_{k=1}^{5} p(k \mid x, y)$. Digit mass is the denominator before normalization, and RankAcc is the fraction of neutral-context pairs with $J(x_{none}, y_A) > J(x_{none}, y_B)$, with half credit for ties. These diagnostics separate use of the requested score tokens from within-item correctness ranking.

The judge interaction is $OR_i = [J(x_A, y_A) - J(x_A, y_B) - J(x_B, y_A) + J(x_B, y_B)]/4$. Averaging over the fixed model-specific natural cohort gives OR, and subtracting the corresponding base interaction gives ΔOR. The likelihood analogue replaces J with full-response log probability:

$$\Lambda_i(\pi) = \tfrac{1}{4}[\log \pi(y_A \mid x_A) - \log \pi(y_B \mid x_A) - \log \pi(y_A \mid x_B) + \log \pi(y_B \mid x_B)]. \tag{7}$$

For the ideal nonparametric KL-regularized optimum $\pi^*(y \mid x) \propto \pi_{ref}(y \mid x) \exp(R(x, y)/\beta_{KL})$, the log partition term is constant across responses within each context. It cancels in Eq. equation 7, yielding $\Lambda_i(\pi^*) - \Lambda_i(\pi_{ref}) = OR_i(R)/\beta_{KL}$. This identity motivates the aligned interaction measures but does not predict the finite-step LoRA/QLoRA DPO trajectory; the implementation's $\beta_{DPO} = 0.1$ is distinct from the theoretical coefficient.

Behavior uses $z_i(a) = \mathbf{1}[a = A_i] - \mathbf{1}[a = B_i]$, with zero for other options and null extraction. The signed item statistic is $SD_i = [z_i(a_{x_A}) - z_i(a_{x_B})]/2$, Bias $= \text{mean}_i\, SD_i$, and Dep $= \text{mean}_i\, |SD_i|$. Acc0 is gold-answer accuracy under the no-stance prompt; Acc_stance is the mean of the two stance-conditioned gold-answer accuracies, counting missing extraction as incorrect.

Dep is a magnitude measure, and opposing stances can yield zero Dep through consistently correct, consistently wrong, or consistently missing responses. An unextracted answer contributes $z = 0$, so its effect on absolute SD depends on the paired response and can increase or decrease the measured value. Parse rate measures extracted option letters, while truncation rate records generation's length finish reason; neither is substituted for accuracy or for the other diagnostic.

H.2 Paired Statistical Inference

Let d_{si} be the paired item contrast for seed s, using absolute SD for H2/H3 and the specified judge or accuracy contrast for the other endpoints. The point estimate is the equally weighted mean of seed-specific item means. Each of 10,000 bootstrap replicates resamples three seed IDs with replacement and then resamples item indices within each selected seed, preserving arm pairing; ordinary intervals use the 2.5th and 97.5th percentiles.

The one-sided procedure centers every d_{si} by the global observed contrast $\widehat{\theta}$, rather than separately centering each seed. It applies the same two-stage resampling to this null-centered array. For the positive-direction alternative, the recorded p-value is

$$p_+ = \frac{1 + \sum_{b=1}^{10000} \mathbf{1}[\widehat{\theta}_b^{0*} \geq \widehat{\theta}]}{10001}. \tag{8}$$

Holm–Bonferroni is applied to exactly the three Qwen9 values from Eq. equation 8 for H1, H2, and H3. Table 30 records both raw and adjusted values, alongside the random seeds used by the stored analysis. Replications, accuracy qualifiers, paired-parsed analyses, F2, ParaAvg, SycoBench, and measurement diagnostics do not expand that primary family.

Table 30: Primary inferential settings and recorded p-values. The p-value random seed is the CI random seed plus 97. The smallest attainable raw p-value in Eq. equation 8 is $1/10001$; H1 and H3 attain this procedural floor.

Endpoint	CI RNG seed	Null-bootstrap RNG seed	Raw one-sided p	Holm p
H1	101	198	0.000099990001	0.000299970003
H2	202	299	0.000699930007	0.000699930007
H3	303	400	0.000099990001	0.000299970003
Capability qualifier	404	Not applicable	Not applicable	Not applicable

The H2 percentile interval and directional p-value answer different inferential questions. The two-sided interval reflects the uncentered empirical distribution, including nonpositive resampled-seed configurations, whereas the one-sided tail test uses the null-centered distribution. Three observed seeds limit representation of between-run variation, and a small bootstrap p-value does not demonstrate stable effects across a broader seed population. Neither interval construction incorporates parser or reference-label uncertainty.

H.3 Finite-population measurement audit

Within each contrast–seed cell, let N_h and n_h denote the population and sampled cluster counts in stratum h. For a sampled cluster, define δ_{shi} as its AI-reference contrast minus its v6 contrast, retaining all four responses together. The correction estimator and adjusted contrast are

$$\widehat{c}_s = \sum_h \frac{N_h}{1400} \overline{\delta}_{sh}, \qquad \widehat{c} = \frac{1}{3} \sum_{s=0}^{2} \widehat{c}_s, \qquad \widehat{\theta}_{\mathrm{AI}} = \widehat{\theta}_{\mathrm{v6}} + \widehat{c}. \tag{9}$$

The known full-population v6 estimate remains the baseline in Eq. equation 9. The estimated design variance uses the without-replacement finite-population correction within each stratum, followed by equal-seed aggregation. With $s^2_{\delta,sh}$ denoting the sampled cluster variance, the calculation is

$$\widehat{\mathrm{Var}}(\widehat{c}) = \frac{1}{9} \sum_{s=0}^{2} \sum_h \left(\frac{N_h}{1400} \right)^2 \left(1 - \frac{n_h}{N_h} \right) \frac{s^2_{\delta,sh}}{n_h}. \tag{10}$$

Approximate measurement-only intervals use the point estimate plus or minus 1.96 standard errors from Eq. equation 10. They condition on the existing training seeds and the reference labels, excluding shared AI-reader bias, new-seed variability, and unsampled-domain errors. Small-stratum and zero-event estimates can understate sampling uncertainty.

For unresolved reference judgments, the point label remains null. The sensitivity calculation enumerates in-range option letters and null for those sampled responses, obtains the attainable paired item-contrast range, and applies the same design weights. It describes uncertainty in those sampled readings and is not a confidence interval or a bound on all unsampled errors.

H.4 Computational Implementation

Training used RTX 3090 and RTX 4090 GPUs, with one GPU per process and separate generation, scoring, and optimization stages. The implementation uses PyTorch 2.11.0, vLLM 0.20.0, Transformers 5.17.0, PEFT 0.21.0, and bitsandbytes 0.50.2. Generation receives the Hugging Face-rendered token IDs, while every judge reward and likelihood readout uses the scoring path in §A.4.

Full-response log-probability normalization uses bounded-size fp32 chunks without removing tokens or vocabulary entries. Large training-logit storage is offloaded to CPU when it reaches 1 GiB; smaller tensors remain on GPU. This changes storage and transfer cost while retaining the same token-level normalization and gradient. Complete measured end-to-end GPU-hours are unavailable. The release separates fixed-input reproduction from new base sampling and preserves the training/measurement parser distinction. It provides item-level inputs for the primary statistical reanalysis and the sampled AI audit. Historical clean-branch checks are supplied as digest/summary records; recomputing them requires the original adapters, candidates, scores, and selected pairs. The full response archive, response-origin manifest, and original audit sampler with its population frame are not included, so full-population re-extraction and historical sampling cannot be independently replayed from this package.