WHEN DO LLM JUDGES FAVOR THEMSELVES? DISSOCIATING SELF-RECOGNITION FROM SELF-PREFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

LLM-as-a-judge has become a standard evaluation paradigm, yet language models can exhibit self-preference, favoring their own outputs over those of other models. Existing explanations attribute this effect to self-recognition, familiarity, or genuine quality differences, but these factors have rarely been isolated within a unified controlled setting. We re-examine self-preference under stricter controls by fixing candidate pairs to identical incorrect answers and clustering inference at the question level. Across seven judges and four tasks, we find that self-preference persists, but is substantially smaller and more task-dependent than conventional pairwise analyses suggest. The effect increases with model scale and is strongest on tasks with greater expressive latitude, reaching +0.24 on MMLU, whereas it vanishes or reverses on verifiable numerical tasks. Our linear-probe experiments recover self-authorship with near-perfect accuracy across models from 1B to 72B parameters, even after surface cues are neutralized. Crucially, however, self-recognition and familiarity remain strong even when self-preference disappears, showing that neither is sufficient to explain the bias. After controlling for judge capability, the residual self-specific effect is only about two percentage points, yet excluding self-votes can still change the top-ranked model. These results show that self-preference is real but modest, strongly task-dependent, not reducible to surface features, and potentially consequential for model evaluation.

1 INTRODUCTION

LLM-as-a-judge has become a standard paradigm for model evaluation, supporting preference labeling, alignment training, and automated benchmarking (Bai et al., 2022; Zheng et al., 2023; Dubois et al., 2024). However, a central concern arises when the judge is also among the candidates: it may favor its own generations, a phenomenon known as self-preference (Zheng et al., 2023; Panickssery et al., 2024). This can inflate the judge's ranking and undermine fairness (Ye et al., 2025). At scale, the effect may propagate into reward modeling, thereby influencing future models (Xu et al., 2024).

Prior work has proposed several explanations for self-preference. One is recognition: models can identify their own outputs, and stronger self-recognition has been associated with stronger preference (Laine et al., 2024; Panickssery et al., 2024). Another is familiarity: judges tend to favor lower-perplexity text, even when written by other models (Wataoka et al., 2024; Stureborg et al., 2024). A third is quality: on verifiable tasks, stronger models may favor themselves simply because their answers are better (Chen et al., 2025a). Improved measurement has also shown that apparent self-preference shrinks after controlling for capability differences and near-random judgments (Chen et al., 2025b; Roytburg et al., 2026). Intervention studies further suggest that both surface style and shared beliefs contribute to the effect: light rewriting can reduce preference, while deeper content alignment can restore it (Mahbub & Feng, 2026); activation-level methods can also suppress it at inference time (Roytburg et al., 2025).

These explanations remain difficult to disentangle because they are rarely isolated within a single controlled setting. Two confounds are particularly important. First, shared beliefs can masquerade as self-preference when the judge favors an answer because it reaches the same conclusion (Mahbub & Feng, 2026). Second, existing evaluations often reuse the same question across multiple response

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.

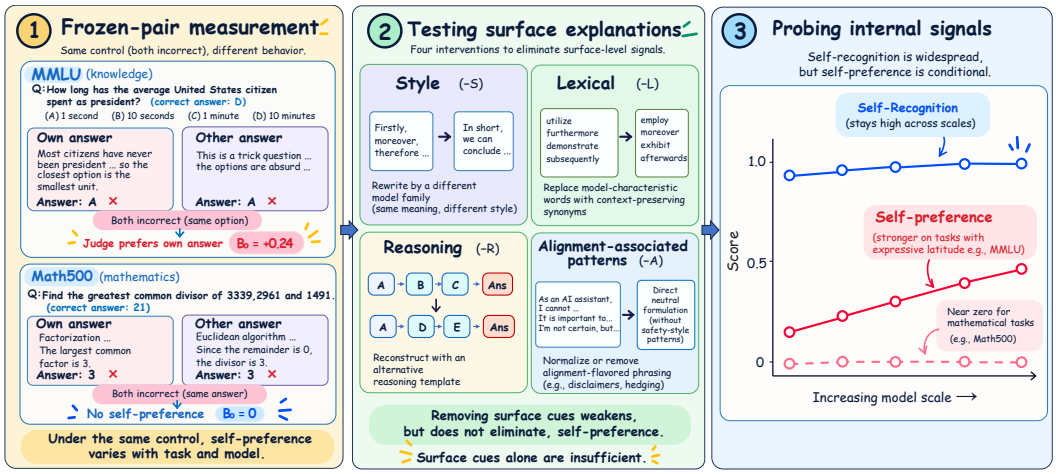


Figure 1: **Overview of the work.** (1) Self-preference is measured on frozen pairs where judge and opponent reach the same incorrect answer, preference appears on MMLU but not on Math500. (2) Four controlled rewrites remove surface cues, yet the residual persists. (3) Representational self-recognition stays high across scales while self-preference is conditional on scale and task.

pairs while treating those pairs as independent (Panickssery et al., 2024). We therefore ask: **after controlling for shared beliefs and dependence across comparisons, how much self-preference remains, and is it driven by surface features or deeper representations?** We address this by restricting analysis to pairs in which the judge and its opponent produce the same incorrect answer, removing conclusion agreement as an explanation, and by clustering inference at the question level.

Under these controls, we evaluate seven judges from 1B to 72B parameters across four task types (Figure 1). We find a residual self-preference remains, but it varies sharply with scale and task. Within a model family, it rises from near zero or negative at 7B to 0.24 above 32B. It is strongest on multiple-choice tasks that permit elaboration, disappears or reverses on verifiable tasks, and varies on open-ended generation. This suggests that self-preference is not an invariant property of the judge, but a behavior shaped by model scale and task structure. We next test whether this residual can be explained by surface features. We apply four controlled rewrites to the judge's own responses, targeting style, lexicon, reasoning structure, and alignment packaging, and attribute their contributions using Shapley values (Shapley et al., 1953). Surface cues explain at most half of the residual, leaving a substantial component that survives rewriting.

We therefore turn to internal representations to test the recognition and familiarity hypotheses directly. A linear probe trained on hidden states reliably distinguishes the judge's own responses from an opponent's across all tested scales, even after stylistic and lexical cues are removed. Large judges also assign higher likelihood to their own outputs than to opponents'. Yet both signals remain strong in settings where self-preference is absent. On numerical tasks, for example, large models still recognize and assign higher likelihood to their own responses without favoring them; smaller models likewise exhibit self-recognition without self-preference. Recognition and familiarity are therefore present but insufficient explanations. Finally, we examine the practical impact of the residual effect. After accounting for judge capability, the purely self-specific advantage is only about two percentage points. Nevertheless, removing self-votes is sufficient to change the top-ranked model, showing that even a modest residual bias can affect aggregate evaluation. Overall, our contributions are:

- ⋆ We re-estimate self-preference under stricter controls, using belief freezing and question-level clustering, and show that a genuine residual preference persists.
- ⋆ We show that self-preference emerges with model scale and varies strongly across tasks. Surface features explain at most half of the residual, which cannot be eliminated by rewriting or reduced to surface-level patterns alone.
- ⋆ Through representational probing, we show that self-recognition is universal across scales and independent of surface cues, yet dissociates from preference. Recognition and familiarity remain strong even when self-preference is absent, indicating that neither is sufficient to explain the effect.

2 BACKGROUND

2.1 SELF-PREFERENCE IN LLM EVALUATORS

Biases of LLM evaluators have been systematically examined (Ye et al., 2025; Shi et al., 2025; Li et al., 2026). Among them, self-preference stands out. The judge favors its own outputs, while often being among the candidates it evaluates (Zheng et al., 2023; Liu et al., 2024). Three accounts of its origin have emerged. The first is recognition. Panickssery et al. (2024) measure it by prompting the model to identify which text it wrote, find that preference rises linearly as recognition is strengthened through fine-tuning, and note that mechanistic evidence for the causal link remains open. The second is familiarity. Judges assign higher scores to low-perplexity text regardless of its author (Wataoka et al., 2024; Stureborg et al., 2024). The third is quality. On verifiable tasks, stronger models often favor themselves deservedly (Chen et al., 2025a). These accounts each capture part of the phenomenon, yet they have rarely been examined jointly, and how they relate remains largely open.

2.2 CORRECTING THE MEASUREMENT

A central difficulty in measurement is that apparent self-preference may merely reflect capability gaps, as stronger judges often favor their own answers because those answers are genuinely better. Chen et al. (2025a) separate legitimate from illegitimate preference using ground-truth answers on verifiable tasks, Chen et al. (2025b) and Spiliopoulou et al. (2025) strip away capability gaps with third-party judges, and Roytburg et al. (2026) correct for near-random verdicts on difficult questions, after which only about half of previously reported findings remain significant. These corrections target quality and decision noise, and the statistics of judge evaluation has itself become an emerging topic (Calderon et al., 2025; Lee et al., 2025). However, a judge may still favor an answer simply for stating its own conclusion, and repeated pairs on the same question continue to be treated as independent samples.

2.3 MITIGATION AND INTERNAL MECHANISMS

Several works have attempted to remove the preference. Substituting a few words reduces it, yet once style is fully flattened, it resurfaces through *shared beliefs* (Mahbub & Feng, 2026), and directions extracted from activation space flip illegitimate preference while disturbing legitimate verdicts (Roytburg et al., 2025). On the mechanistic side, Ackerman & Panickssery (2025) locate an internal direction that causally controls self-recognition in a single model, and Zhou et al. (2025) show that hidden states separate self- from other-generated text even when outputs do not report the distinction (Azaria & Mitchell, 2023). Such evidence, however, remains fragmented, and how internal recognition relates to behavioral preference is not yet well understood.

To address this limitations, we develop a unified framework for measuring self-preference. The framework first freezes beliefs to remove shared conclusions and clusters inference at the question level, yielding a stricter estimate of the residual preference. It then attributes this residual to surface features through four controlled rewrites, and finally probes internal signals across scales.

3 METHODOLOGY

3.1 PRELIMINARIES

For a task with prompt set $\mathcal{X}$, consider a judge model J performing pairwise evaluation. Given a prompt $x \in \mathcal{X}$ and a candidate pair (A, B), the judge outputs a preference score

$$s_J(x, A, B) = P_J(\text{vote} = A \mid x, A, B), \tag{1}$$

read from the predictive distribution over vote tokens; to neutralize position bias, each pair is scored in both orders and averaged (Shi et al., 2025). J also acts as a task performer, producing its own response $o_J(x)$ for each prompt; $o_K(x)$ and $o_{\text{ref}}(x)$ denote the responses of a proxy model K and a pool of reference models, where K is selected to match the judge J's incorrect conclusion.

Following the evaluator-quality logic of Roytburg et al. (2026), we compute the per-example differential

$$\Delta s_J(x) = s_J(x,\, o_J,\, o_{\text{ref}}) - s_J(x,\, o_K,\, o_{\text{ref}}), \tag{2}$$

which cancels the judge's baseline tendency toward comparable responses, since the second term involves nothing the judge authored. Each question expands into multiple pairs, one per combination of reference model in the pool and presentation order, with o_{ref} in Eq. 2 ranging over the pool accordingly; let $\mathcal{P}(\mathcal{X}')$ denote the pair set induced by $\mathcal{X}' \subseteq \mathcal{X}$, containing N_q unique questions. Aggregating over an evaluation subset $\mathcal{X}' \subseteq \mathcal{X}$ yields the test statistic

$$T_{\text{quality}}(J, K, \mathcal{X}') = \tfrac{1}{|\mathcal{P}(\mathcal{X}')|} \textstyle\sum_{p \in \mathcal{P}(\mathcal{X}')} \Delta s_J(p), \tag{3}$$

whose significant positivity indicates that residual self-preference survives after correcting for evaluator quality. We not only examine the *existence* of this residual, but characterize its *composition*.

3.2 Isolating the Harmful Residual

We focus on examples where the judge's own response is incorrect. Outcome matching alone is insufficient: a judge may still favor a response simply because its conclusion aligns with its own belief, causing shared belief to masquerade as self-preference (Mahbub & Feng, 2026). We therefore require the proxy to reach the same incorrect conclusion as the judge.

Let $y^*(x)$ denote the ground-truth answer. We define

$$\mathcal{X}^{\text{match}} = \{x \in \mathcal{X} : \text{ans}_J(x) = \text{ans}_K(x) \neq y^*(x),\ \text{ans}_{\text{ref}}(x) = y^*(x)\}\,. \tag{4}$$

Thus, the self and proxy responses agree at the conclusion level, and their differential cannot be explained merely by agreement with the judge's own final answer.

The resulting residual is

$$B_0 = T_{\text{quality}}\big(J, K, \mathcal{X}^{\text{match}}\big)\,. \tag{5}$$

A positive B_0 therefore indicates that residual self-preference persists after controlling for evaluator quality and conclusion-level shared belief. Multiple pairs may share the same question, and such pairs are correlated. Treating them as independent samples overstates significance, since with N_q unique questions and intra-question correlation ICC, the effective sample size is bounded as $n_{\text{eff}} \in [N_q,\ N_q/\text{ICC}]$. All inference on B_0 is thus clustered at the question level. We next decompose this residual into observable response-level signals.

3.3 Decomposing the Residual

We ask which observable properties of the judge's own response sustain B_0. We consider four families of response-level signals, $\mathcal{F} = \{S, L, R, A\}$, chosen to span surface realization through higher-level organization: **style** S (discourse structure and formatting), **lexical** L (word choice), **reasoning** R (argument organization), and **alignment-associated** A (post-training response conventions). These families may overlap rather than represent independent mechanisms, which motivates a cooperative attribution below.

For each $c \in \mathcal{F}$, an intervention $\mathcal{I}_c(o_J) = o_J^{(-c)}$ attenuates family c while preserving the meaning of the response. For style, a model outside the judge's family rewrites the full response,

$$o_J^{(-S)} = \text{Rewrite}_{M'}(o_J), \qquad M' \notin \text{family}(J). \tag{6}$$

For lexicon, characteristic items are replaced with context-preserving synonyms by exact string substitution,

$$o_J^{(-L)} = \text{Replace}(o_J, \{w_i \to \tilde{w}_i\}_{i=1}^m)\,. \tag{7}$$

For reasoning, the argument is reconstructed under an alternative template from a predefined set $\mathcal{H}$,

$$o_J^{(-R)} = \text{Reconstruct}(o_J; \text{Match}(o_J, \mathcal{H}))\,. \tag{8}$$

For alignment, conventions from a pattern inventory $\mathcal{P}_A$ are removed,

$$o_J^{(-A)} = \text{RemovePatterns}(o_J; \mathcal{P}_A). \tag{9}$$

Interventions apply only to the judge-authored response, with proxy and reference responses fixed, and configurations compose under a fixed order, yielding the post-intervention differential

$$\Delta s_J^{(-\mathcal{U})}(x) = s_J\Big(x, o_J^{(-\mathcal{U})}, o_{\text{ref}}\Big) - s_J(x, o_K, o_{\text{ref}})\,, \tag{10}$$

for any attenuated subset $\mathcal{U} \subseteq \mathcal{F}$. Implementation details are given in Appendix C.

We evaluate all 2^4 intervention configurations and attribute B_0 by Shapley values. For $\mathcal{C} \subseteq \mathcal{F}$ the set of retained signal families, define $v(\mathcal{C}) = \mathbb{E}_{x \in \mathcal{X}^{\text{match}}}[\Delta s_J^{(-(\mathcal{F}\setminus\mathcal{C}))}(x)]$, so that $v(\mathcal{F}) = B_0$ and $v(\varnothing) = \rho$, the residual surviving all four interventions. The attribution of family c is

$$\phi_c = \sum_{\mathcal{C} \subseteq \mathcal{F}\setminus\{c\}} \frac{|\mathcal{C}|!\,(|\mathcal{F}|-|\mathcal{C}|-1)!}{|\mathcal{F}|!} \left[v(\mathcal{C}\cup\{c\}) - v(\mathcal{C})\right], \tag{11}$$

computed exactly over all 16 configurations. By Shapley efficiency,

$$B_0 = \phi_S + \phi_L + \phi_R + \phi_A + \rho. \tag{12}$$

ρ captures the portion of the residual that no surface intervention removes.

Remark. *We choose these four signal families because they capture representative response characteristics from surface realization to higher-level organization. Each intervention selectively alters one signal family while preserving the meaning of the original response. For any subset $\mathcal{U} \subseteq \mathcal{F}$, we construct a joint counterfactual $o_J^{(-\mathcal{U})}$ by applying the corresponding interventions under a fixed composition protocol.*

3.4 Probing Internal Signals

The interventions above test surface accounts by modifying text. To test the recognition and familiarity accounts directly, we use two measurements that leave the text unmodified.

Representational recognition. We distinguish *self-authorship*, the fact that a response was generated by the judge, from *self-recognition*, the capacity to discern this fact. Prior work measures recognition behaviorally, by prompting the model to identify its own output (Panickssery et al., 2024); we measure it representationally. For each judge, we extract hidden states at layer ℓ for its own response o_J and the matched proxy response o_K, and train a linear probe g_ℓ to classify authorship. We report AUROC on held-out questions, with shuffled-label controls, and repeat the procedure on surface-neutralized responses $o_J^{(-\{S,L\})}$ to test whether the separation depends on surface cues.

Familiarity. For each frozen pair, we compute the judge's per-token negative log-likelihood on both responses and define the familiarity gap

$$\text{gap}(x) = \text{NLL}_J(o_K(x)) - \text{NLL}_J(o_J(x)), \tag{13}$$

so that a positive gap indicates the judge finds its own response more predictable than the matched alternative.

Both measurements are computed on the same frozen sets as B_0, allowing recognition, familiarity, and preference to be compared question by question, across scales and tasks.

4 Experiments

In the experiments, we study eleven instruction-tuned models from three families: Qwen2.5-1.5B, Qwen2.5-3B, Qwen2.5-7B, Qwen2.5-14B, Qwen2.5-32B, and Qwen2.5-72B (Qwen et al., 2025); Llama-3.2-1B, Llama-3.2-3B, Llama-3.1-8B, and Llama-3.3-70B (Grattafiori et al., 2024); and Mistral-7B-v0.3 (Jiang et al., 2023). Each model also serves as a task performer, allowing it to evaluate candidate pairs that may include its own response. More details are provided in Appendix A.

We evaluate on four benchmarks spanning different levels of expressive latitude. MMLU (Hendrycks et al., 2020) and Math500 (Lightman et al., 2024) provide verifiable answers, allowing conclusions to be frozen exactly. TruthfulQA (Lin et al., 2022) and AlpacaEval (Dubois et al., 2024) are open-ended and test whether the observed patterns extend beyond option-constrained settings.

In the following, we first revisit the statistical basis of prior evidence and estimate B_0 on larger frozen sets using question-level clustered inference (Section 4.1). We then decompose the residual through Shapley attribution (Section 4.2), probe internal recognition and familiarity signals and their dissociation from preference (Section 4.3), and finally assess the practical impact of self-preference on leaderboard rankings (Section 4.4).

Table 1: **Residual self-preference across judges and tasks.** Each entry reports the residual self-preference estimate B_0 with a question-clustered 95% confidence interval in brackets. **Bold** estimates have confidence intervals that exclude zero. Positive values indicate residual self-preference after controlling for the specified confounders.

Judge	MMLU	Math500	TruthfulQA	AlpacaEval
Qwen2.5-7B	-0.041 $[-0.061, -0.020]$	$+0.030$ $[+0.015, +0.045]$	-0.025 $[-0.039, -0.010]$	-0.008 $[-0.020, +0.003]$
Mistral-7B-v0.3	$+0.009$ $[+0.004, +0.015]$	$+0.009$ $[-0.009, +0.028]$	$+0.001$ $[-0.002, +0.005]$	$+0.006$ $[+0.002, +0.009]$
Llama-3.1-8B	$+0.050$ $[+0.039, +0.061]$	-0.051 $[-0.068, -0.033]$	-0.016 $[-0.027, -0.005]$	-0.016 $[-0.026, -0.007]$
Qwen2.5-14B	$+0.140$ $[+0.103, +0.178]$	$+0.004$ $[-0.008, +0.015]$	$+0.090$ $[+0.069, +0.110]$	$+0.066$ $[+0.048, +0.085]$
Qwen2.5-32B	$+0.242$ $[+0.202, +0.281]$	$+0.026$ $[+0.013, +0.040]$	$+0.101$ $[+0.084, +0.120]$	$+0.048$ $[+0.030, +0.068]$
Llama-3.3-70B	$+0.230$ $[+0.172, +0.287]$	-0.048 $[-0.070, -0.026]$	$+0.070$ $[+0.046, +0.093]$	-0.113 $[-0.145, -0.083]$
Qwen2.5-72B	$+0.161$ $[+0.125, +0.198]$	$+0.016$ $[+0.007, +0.026]$	$+0.002$ $[-0.015, +0.020]$	$+0.036$ $[+0.021, +0.051]$

4.1 Residual Self-Preference Across Judges

We estimate B_0 on frozen pairs. On verifiable tasks, each pair requires the judge and the proxy to reach the same incorrect conclusion while the reference answers correctly, so the differential cannot be explained by conclusion agreement. Open-ended tasks have no ground truth, so pairs instead require the two responses to agree in conclusion, determined by bidirectional entailment (Appendix D). Frozen sets contain 34 to 583 unique questions per judge and task. All inference is clustered at the question level. Pairs sharing a question are correlated, and when pairs outnumber unique questions severalfold, treating them as independent can make a nonexistent effect appear significant. Our frozen sets remain significant under this standard, with clustered p values on MMLU ranging from 9.7×10^{-4} to 2.8×10^{-25}. Table 1 reports all B_0 estimates with question-clustered 95% confidence intervals. The results first split by scale. On MMLU, judges at the 7B level show negative or near-zero residuals, -0.041 for qwen-7b and $+0.009$ for mistral. The residual turns positive at 14B with $+0.140$, peaks at 32B with $+0.242$, and remains between $+0.161$ and $+0.230$ at the 70B level. Both families follow the same trajectory, with llama rising from $+0.050$ at 8B to $+0.230$ at 70B. Residual self-preference is not a uniform constant but a phenomenon that emerges with scale.

The results also split by task. The same large judges show near-zero or reversed residuals on Math500. Qwen values stay small and positive, at most $+0.026$, while llama values are significantly negative at -0.048 and -0.051. Where conclusions can be verified exactly, the residual finds little room to operate. On open-ended tasks, no uniform pattern emerges. The mid-sized qwen judges are consistently positive on both TruthfulQA and AlpacaEval, reaching $+0.101$, whereas llama-70b splits across tasks, positive on TruthfulQA at $+0.070$ yet significantly negative on AlpacaEval at -0.113. The direction of the residual on open-ended tasks depends on the judge and the task.

Finally, we assess inference quality. Across all 28 cells, ICC ranges from 0.15 to 0.66, so pairwise inference overstates the effective sample size two- to four-fold throughout. Even so, clustered confidence intervals for the main cells remain far from zero. Per-cell clustered p values and full diagnostics are given in Appendix B.

4.2 Shapley Decomposition of the Residual

We apply the four interventions of Section 3.3, evaluate all sixteen configurations, and attribute B_0 to the four signal families and the remainder ρ by Shapley values. By the efficiency identity, each row satisfies $\phi_S + \phi_L + \phi_R + \phi_A + \rho = B_0$ exactly, so the decomposition is exact rather than a regression approximation. Table 2 reports the results. The B_0 column restates the residuals

Table 2: **Exact Shapley decomposition of residual self-preference B_0.** For each judge–task pair, the decomposition satisfies $\phi_S + \phi_L + \phi_R + \phi_A + \rho = B_0$ up to rounding. The ϕ terms denote the Shapley-attributed contributions of the four feature groups, while ρ denotes the residual component. Math500 rows are shaded because their residual self-preference is close to zero; we therefore do not interpret their decomposition.

Task	Judge	B_0	ϕ_S	ϕ_L	ϕ_R	ϕ_A	ρ
MMLU	Qwen2.5-7B	−0.041	+0.002	+0.045	+0.012	+0.004	−0.104
	Mistral-7B-v0.3	+0.009	+0.004	−0.001	+0.017	−0.006	−0.005
	Llama-3.1-8B	+0.050	−0.006	+0.026	−0.025	−0.007	+0.062
	Qwen2.5-14B	+0.140	+0.003	+0.031	−0.029	+0.019	+0.116
	Qwen2.5-32B	+0.242	+0.002	+0.065	+0.015	+0.039	+0.122
	Llama-3.3-70B	+0.230	−0.002	+0.059	−0.031	+0.023	+0.181
	Qwen2.5-72B	+0.161	+0.007	+0.062	−0.010	+0.019	+0.082
Math500	Qwen2.5-7B	+0.030	−0.004	+0.021	+0.098	−0.023	−0.062
	Llama-3.1-8B	−0.051	−0.033	+0.005	−0.017	−0.006	+0.002
	Qwen2.5-32B	+0.026	−0.002	+0.015	+0.026	+0.008	−0.020
	Llama-3.3-70B	−0.048	−0.011	+0.021	+0.009	−0.004	−0.063
TruthfulQA	Qwen2.5-14B	+0.090	−0.017	+0.028	−0.026	+0.031	+0.074
	Qwen2.5-32B	+0.101	+0.015	+0.036	+0.025	+0.057	−0.031
	Llama-3.3-70B	+0.070	+0.025	+0.065	−0.122	+0.070	+0.031
AlpacaEval	Qwen2.5-14B	+0.066	−0.013	+0.022	−0.023	+0.035	+0.047
	Qwen2.5-32B	+0.048	+0.008	+0.030	−0.006	+0.076	−0.060
	Qwen2.5-72B	+0.036	−0.000	+0.023	+0.036	+0.037	−0.059
	Llama-3.3-70B	−0.113	+0.019	+0.081	−0.115	+0.071	−0.169

established in Table 1, and the remaining columns decompose them, with $\rho = v(\varnothing)$ denoting the part that survives all four interventions and is directly measured.

On MMLU, surface channels contribute little and their ordering is consistent. Across the large judges, the style share is near zero, the lexical share is largest but never exceeds $+0.065$, alignment conventions come second, and the reasoning share is small with unstable sign. Taken together, the four families explain at most half of the residual on the judges with robust residuals.

By contrast, ρ is the largest single component for every judge with a positive residual, accounting for one half to four fifths of B_0, and on individual judges it reaches or exceeds B_0 itself, suggesting interactions among configurations. Once everything a rewrite can touch is removed, the bulk of the residual remains.

On the open-ended tasks, the weight of the decomposition shifts. The alignment share grows systematically, and for qwen-32b the ϕ_A on TruthfulQA and AlpacaEval both exceed its MMLU share. Multiple choice leaves little room for packaging, while open-ended responses do, and the role of post-training conventions grows with expressive latitude.

Finally, the estimate of ρ is conservative. It is measured directly from the all-intervention configuration, which undergoes the most rewriting. Rewriting carries a measurable quality cost (Appendix F), and the judge lowers its preference accordingly, so the measured ρ can only understate the true non-surface component. Surface signals participate in the residual, but the residual does not reduce to them.

4.3 Probing Recognition and Familiarity

We next test whether the residual is accompanied by internal signals of recognition and familiarity, measured on the same frozen sets as B_0. Figure 2 places the three quantities side by side across scales, and Table 3 reports the values.

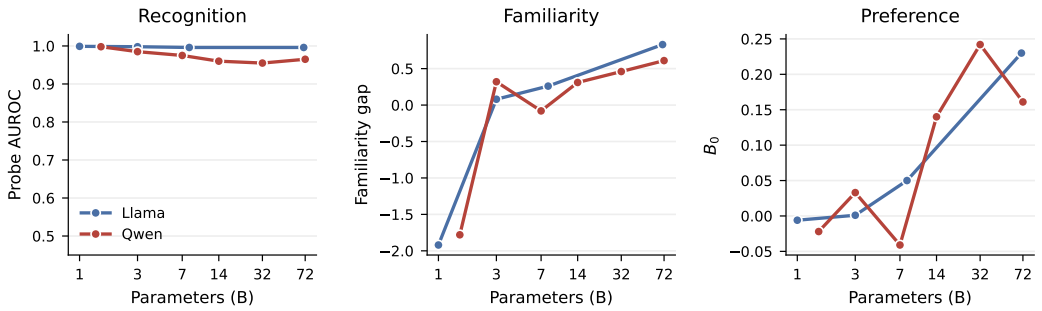


Figure 2: Recognition (probe AUROC), familiarity (NLL gap), and preference (B_0) across judge scales (from 1B to 72B) on MMLU.

Table 3: **Recognition, familiarity, and residual self-preference across model scales on MMLU.** Recognition is measured by probe AUROC, with 0.5 denoting chance performance; *Chance* reports the same probe trained on shuffled labels, and *Neutral* reports AUROC after style and lexical neutralization (– denotes not run). Familiarity is summarized by the mean NLL gap and *Own*, the percentage of examples in which the judge assigns lower NLL to its own response than to the reference response. Preference is measured by the residual self-preference B_0.

	Recognition			Familiarity		Preference
Judge	**AUROC**	**Chance**	**Neutral**	**Gap**	**Own (%)**	**B_0**
Llama-3.2-1B	0.999	0.471	–	−1.92	18	−0.006
Qwen2.5-1.5B	0.998	0.478	–	−1.78	12	−0.022
Llama-3.2-3B	0.998	0.510	–	+0.08	68	+0.001
Qwen2.5-3B	0.985	0.469	–	+0.32	84	+0.033
Qwen2.5-7B	0.975	0.497	–	−0.08	55	−0.041
Llama-3.1-8B	0.996	0.473	–	+0.26	82	+0.050
Qwen2.5-14B	0.960	0.461	–	+0.31	80	+0.140
Qwen2.5-32B	0.955	0.489	0.940	+0.46	86	+0.242
Llama-3.3-70B	0.996	0.562	0.976	+0.83	100	+0.230
Qwen2.5-72B	0.965	0.510	0.862	+0.61	92	+0.161

Recognition is present at every scale. Linear probes recover self-authorship with high AUROC across all ten judges, while shuffled-label controls stay at chance. Even the smallest judges carry near-ceiling authorship information in their hidden states. The signal largely survives surface neutralization. Retraining the probes on responses with style and lexical signals removed reduces AUROC by at most 0.020 for Qwen-32B and Llama-70B. Qwen-72B shows a larger drop but remains well above chance.

Familiarity emerges with scale, with some family-specific variation. The gap is strongly negative for the two smallest judges (-1.92 and -1.78), meaning that they find their own responses harder to predict than the reference responses. It turns positive by 3B in both families and generally grows with scale, reaching $+0.83$ for Llama-70B and $+0.61$ for Qwen-72B. The trajectory is not strictly monotonic. Qwen-7B dips slightly below zero (-0.08) before the gap recovers at 14B. Apart from this dip, every judge from 3B onward shows a positive gap.

Preference arrives last. $|B_0|$ stays at or below 0.05 up to 8B and reaches 0.14–0.24 only from 14B onward, even though recognition is high throughout and familiarity is already positive from 3B. Neither internal signal is therefore sufficient for preference. On Math500, the large judges retain near-ceiling probe performance and strongly positive familiarity gaps, with their own responses consistently more predictable than the reference responses, yet B_0 remains near zero or reverses sign. Both internal signals therefore remain present even when behavioral self-preference disappears.

Together, these results reveal a clear separation among the three quantities: self-recognition is consistently present across the models we study, familiarity increases with scale, while self-preference remains contingent on both model scale and task.

Table 4: **Effect of self-votes on the leaderboard under the Phi-3.5-mini reference.** Results are averaged over six judges. Removing self-votes changes the top-ranked candidate from Llama-3.3-70B to Qwen2.5-72B. ΔRank denotes the change in rank after removing self-votes, with positive values indicating improvement.

	With self-votes		Self-votes removed		
Candidate	**Win rate**	**Rank**	**Win rate**	**Rank**	**ΔRank**
Llama-3.3-70B	**0.481**	**1**	0.437	2	−1
Qwen2.5-72B	0.403	2	**0.441**	**1**	+1
Qwen2.5-32B	0.390	3	0.382	4	−1
Qwen2.5-14B	0.369	4	0.415	3	+1
Qwen2.5-7B	0.276	5	0.244	5	0
Llama-3.1-8B	0.224	6	0.203	6	0

4.4 Impact on Leaderboard Rankings

The preceding sections excluded capability differences by design, through frozen pairs. Here we let them back in, and remove them statistically instead, to ask what the residual costs in a realistic evaluation. Six judges score eight candidates on a set of MMLU questions, where each candidate's score is its win rate against a common reference, Phi-3.5-mini (Abdin et al., 2024), averaged over judges. Each judge is itself a candidate, so it also scores its own responses.

In a direct comparison, a judge scores its own responses about seven percentage points above the average it assigns to the other candidates. Removing candidate ability and judge leniency by difference-in-differences, the gap shrinks to about two percentage points.

The preference lands on the judge's own entry, and the top of the leaderboard is decided by narrow margins. Removing self-votes changes the top rank while the rest of the board is largely preserved, with Kendall $\tau = 0.73$ (Table 4). Under the weaker reference Llama-3.2-1B, the corrected preference is indistinguishable from zero and the ranking is more disturbed ($\tau = 0.33$, see Appendix E).

5 Conclusion

LLM-as-a-judge has become a standard evaluation paradigm, yet judges may favor their own outputs. We introduce a unified framework for measuring this self-preference under stricter controls by fixing candidate conclusions and clustering inference at the question level. We found a residual effect remains, but it is modest and highly conditional: it emerges more clearly at larger model scales, while weakening or reversing on verifiable numerical tasks. Controlled interventions on style, lexicon, reasoning structure, and alignment-associated conventions explain at most half of the residual. Representational probing further shows that self-authorship information is present across scales and that familiarity increases with model size, yet neither reliably tracks behavioral preference. In a realistic evaluation panel, the capability-corrected self-specific effect is only about two percentage points, but this is still sufficient to change the top-ranked model. Overall, we think self-preference is genuine but limited, strongly dependent on task and scale, and cannot be explained by surface features, self-recognition, or familiarity alone.

Our analysis also has several limitations. On objective tasks, frozen pairs are correctness-filtered using explicit ground truth, whereas open-ended tasks lack exact labels and rely on entailment-based conclusion matching. Estimates across these two settings are therefore not directly comparable. The strongest task-level contrast in our study, MMLU versus Math500, uses the same correctness-based protocol and is unaffected by this limitation. Results on TruthfulQA and AlpacaEval should instead be viewed as complementary evidence, since entailment-based freezing is weaker than exact answer matching and cannot fully eliminate residual shared-belief effects. Our experiments are also limited to open-weight model families up to 72B parameters. Future work could extend the framework to frontier closed models, develop the freezing procedure into a reusable audit for new judges, and investigate whether the unexplained remainder ρ corresponds to internal signals that can be directly targeted by steering, providing mitigation strategies beyond simply excluding self-votes.

AI USE STATEMENT

In this work, LLMs are used within the method to produce controlled rewrites of model responses; these rewrites are experimental stimuli and are reported as such. We have not used generative AI tools to develop the theoretical framework, formulate the estimators, design the methodology, implement the method, or interpret results; the remaining required-disclosure tasks are not applicable to this work. Additionally, we used generative AI tools to edit the paper for readability and to assist with LaTeX formatting. We have reviewed all AI-assisted work. Rewritten responses are constrained by their prompts to preserve all claims and final answers, and all edited text was verified by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human subjects. All datasets used are publicly available benchmarks, and all evaluated models are publicly released checkpoints. The study measures evaluation biases of LLM judges, and we do not foresee direct negative societal impact; identifying such biases supports more reliable model evaluation.

REPRODUCIBILITY STATEMENT

The appendix provides the full prompt templates for generation and judging, the intervention prompts, the complete model list with exact checkpoints, the conclusion-matching protocol for open-ended tasks, and per-cell inference diagnostics. All datasets used are publicly available. Our code will be released upon publication.

REFERENCES

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone, 2024. URL https://arxiv.org/abs/2404.14219.

Christopher Ackerman and Nina Panickssery. Inspection and control of self-generated-text recognition ability in llama3-8b-instruct. In *International Conference on Learning Representations*, volume 2025, pp. 85901–85932, 2025.

Amos Azaria and Tom Mitchell. The internal state of an llm knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976, 2023.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Nitay Calderon, Roi Reichart, and Rotem Dror. The alternative annotator test for llm-as-a-judge: How to statistically justify replacing human annotators with llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16051–16081, 2025.

Wei-Lin Chen, Zhepei Wei, Xinyu Zhu, Shi Feng, and Yu Meng. Do llm evaluators prefer themselves for a reason? *arXiv preprint arXiv:2504.03846*, 2025a.

Zhi-Yuan Chen, Hao Wang, Xinyu Zhang, Enrui Hu, and Yankai Lin. Beyond the surface: Measuring self-preference in llm judgments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 1653–1672, 2025b.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL `https://arxiv.org/abs/2310.06825`.

Rudolf Laine, Bilal Chughtai, Jan Betley, Kaivalya Hariharan, Jeremy Scheurer, Mikita Balesni, Marius Hobbhahn, Alexander Meinke, and Owain Evans. Me, myself, and ai: The situational awareness dataset (sad) for llms. *Advances in Neural Information Processing Systems*, 37:64010–64118, 2024.

Moritz Laurer, Wouter Van Atteveldt, Andreu Casas, and Kasper Welbers. Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli. *Political Analysis*, 32(1):84–100, 2024.

Chungpa Lee, Thomas Zeng, Jongwon Jeong, Jy-yong Sohn, and Kangwook Lee. How to correctly report llm-as-a-judge evaluations. *arXiv preprint arXiv:2511.21140*, 2025.

Qingquan Li, Shaoyu Dou, Kailai Shao, Chao Chen, and Haixiang Hu. Evaluating scoring bias in llm-as-a-judge. In *International Conference on Database Systems for Advanced Applications*, pp. 19–34. Springer, 2026.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 3214–3252, 2022.

Yiqi Liu, Nafise Sadat Moosavi, and Chenghua Lin. Llms as narcissistic evaluators: When ego inflates evaluation scores. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12688–12701, 2024.

Taslim Mahbub and Shi Feng. Mitigating self-preference by authorship obfuscation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 37701–37708, 2026.

Arjun Panickssery, Samuel R Bowman, and Shi Feng. Llm evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems*, 37:68772–68802, 2024.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Dani Roytburg, Matthew Bozoukov, Matthew Nguyen, Jou Barzdukas, Simon Fu, and Narmeen Oozeer. Breaking the mirror: Activation-based mitigation of self-preference in llm evaluators. *arXiv preprint arXiv:2509.03647*, 2025.

Dani Roytburg, Matthew Bozoukov, Matthew Nguyen, Jou Barzdukas, Mackenzie Puig-Hall, and Narmeen Oozeer. Are llm evaluators really narcissists? sanity checking self-preference evaluations. *arXiv preprint arXiv:2601.22548*, 2026.

Lloyd S Shapley et al. A value for n-person games. 1953.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. Judging the judges: A systematic study of position bias in llm-as-a-judge. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 292–314, 2025.

Evangelia Spiliopoulou, Riccardo Fogliato, Hanna Burnsky, Tamer Soliman, Jie Ma, Graham Horwood, and Miguel Ballesteros. Play favorites: A statistical method to measure self-bias in llm-as-a-judge. *arXiv preprint arXiv:2508.06709*, 2025.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. Large language models are inconsistent and biased evaluators. *arXiv preprint arXiv:2405.01724*, 2024.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. Self-preference bias in llm-as-a-judge. *arXiv preprint arXiv:2410.21819*, 2024.

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

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, et al. Justice or prejudice? quantifying biases in llm-as-a-judge. In *International Conference on Learning Representations*, volume 2025, pp. 102351–102390, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Yinghan Zhou, Weifeng Zhu, Juan Wen, Wanli Peng, Zhengxian Wu, and Yiming Xue. From implicit to explicit: Enhancing self-recognition in large language models. *arXiv preprint arXiv:2508.14408*, 2025.

A Models and Datasets

This appendix lists the judge models and evaluation tasks used throughout the paper.

Models. We use eleven instruction-tuned judges spanning 1B to 72B parameters across three families (Table 5). The four judges below 7B enter the MMLU scaling panel only; the remaining seven enter the full four-task matrix.

Table 5: Judge models. All models are instruction-tuned checkpoints from HuggingFace.

Model ID	Family	Params	HF / Source
llama-3.2-1b	Llama	1B	meta-llama/Llama-3.2-1B-Instruct
qwen-2.5-1.5b	Qwen	1.5B	Qwen/Qwen2.5-1.5B-Instruct
llama-3.2-3b	Llama	3B	meta-llama/Llama-3.2-3B-Instruct
qwen-2.5-3b	Qwen	3B	Qwen/Qwen2.5-3B-Instruct
qwen-2.5-7b	Qwen	7B	Qwen/Qwen2.5-7B-Instruct
mistral-7b-v0.3	Mistral	7B	mistralai/Mistral-7B-Instruct-v0.3
llama-3.1-8b	Llama	8B	meta-llama/Llama-3.1-8B-Instruct
qwen-2.5-14b	Qwen	14B	Qwen/Qwen2.5-14B-Instruct
qwen-2.5-32b	Qwen	32B	Qwen/Qwen2.5-32B-Instruct
llama-3.3-70b	Llama	70B	meta-llama/Llama-3.3-70B-Instruct
qwen-2.5-72b	Qwen	72B	Qwen/Qwen2.5-72B-Instruct

Datasets. Table 6 shows an example prompt from each task.

Table 6: Evaluation tasks and example prompts.

Task	Answer format	Example
MMLU	Multiple choice	What is the output of "abc"[::-1] in Python 3? (A) Error (B) abc (C) cba (D) c
Math500	Numerical	Find the greatest common divisor of 3339, 2961 and 1491.
TruthfulQA	Free-form	Which country consumes the most tea per person?
AlpacaEval	Open-ended	How did US states get their names?

B Clustered Inference Diagnostics

Here we explain the clustered inference used throughout the paper and report the full diagnostics. Pairs expanded from the same question share its difficulty and the judge's tendency on it, and are therefore correlated. ICC measures the strength of this sharing, giving $\mathrm{DEFF} = 1 + (\bar{m} - 1)\,\mathrm{ICC}$ and $n_{\mathrm{eff}} = \mathrm{pairs}/\mathrm{DEFF}$, where $\bar{m}$ is the average number of pairs per question. Treating pairs as independent samples inflates significance, so all B_0 inference in the main text uses cluster-robust standard errors with $df = N_q - 1$.

Table 7 reports the diagnostics for all 28 cells, listing pairs, N_q, ICC, DEFF, n_{eff}, the two-sided p value, and its Benjamini–Hochberg adjusted value q. Across all cells ICC ranges from 0.15 to 0.66 and DEFF from 1.8 to 3.8, so inference at the pair level overstates the effective sample size by a factor of roughly two to four.

C Intervention Design

We implement four targeted interventions on the judge-authored response o_J. Each intervention attenuates one response-level signal family while preserving the meaning of the response. All interventions share three constraints: the final answer stated in the response must remain unchanged, task-relevant content must be retained, and the intervention must not introduce signals characteristic

Table 7: Clustering diagnostics for all 28 judge–task cells. p is two-sided, cluster-robust on question; q is Benjamini–Hochberg adjusted across all 28 tests.

Judge	Pairs	N_q	ICC	DEFF	n_{eff}	p	q
MMLU							
Qwen2.5-7B	2681	377	0.363	3.22	833	1.8×10^{-4}	3.1×10^{-4}
Mistral-7B-v0.3	3804	583	0.499	3.76	1013	9.7×10^{-4}	1.4×10^{-3}
Llama-3.1-8B	2848	421	0.419	3.41	834	4.4×10^{-17}	4.1×10^{-16}
Qwen2.5-14B	1287	229	0.430	2.98	431	1.1×10^{-11}	3.6×10^{-11}
Qwen2.5-32B	1120	192	0.427	3.06	365	2.8×10^{-25}	3.9×10^{-24}
Llama-3.3-70B	773	140	0.546	3.47	223	5.1×10^{-13}	2.4×10^{-12}
Qwen2.5-72B	815	145	0.379	2.75	296	1.5×10^{-14}	8.2×10^{-14}
Math500							
Qwen2.5-7B	1206	158	0.222	2.47	488	1.3×10^{-4}	2.5×10^{-4}
Mistral-7B-v0.3	117	34	0.661	2.61	45	0.35	0.39
Llama-3.1-8B	680	87	0.412	3.81	179	7.3×10^{-8}	1.9×10^{-7}
Qwen2.5-14B	1212	159	0.152	2.01	604	0.55	0.57
Qwen2.5-32B	1224	161	0.174	2.15	569	3.7×10^{-4}	6.1×10^{-4}
Llama-3.3-70B	1076	138	0.337	3.29	327	6.2×10^{-5}	1.2×10^{-4}
Qwen2.5-72B	1214	157	0.232	2.56	474	1.5×10^{-3}	1.9×10^{-3}
TruthfulQA							
Qwen2.5-7B	2002	446	0.259	1.90	1051	7.6×10^{-4}	1.2×10^{-3}
Mistral-7B-v0.3	1252	440	0.527	1.97	635	0.47	0.51
Llama-3.1-8B	2084	437	0.413	2.56	815	5.1×10^{-3}	6.2×10^{-3}
Qwen2.5-14B	2188	453	0.287	2.10	1042	6.1×10^{-16}	4.3×10^{-15}
Qwen2.5-32B	3014	481	0.257	2.36	1279	1.3×10^{-25}	3.6×10^{-24}
Llama-3.3-70B	2930	471	0.249	2.30	1274	1.1×10^{-8}	3.1×10^{-8}
Qwen2.5-72B	2698	456	0.293	2.44	1105	0.78	0.78
AlpacaEval							
Qwen2.5-7B	1858	440	0.311	2.00	927	0.17	0.20
Mistral-7B-v0.3	1191	441	0.480	1.82	655	1.2×10^{-3}	1.7×10^{-3}
Llama-3.1-8B	1850	419	0.407	2.39	774	1.2×10^{-3}	1.6×10^{-3}
Qwen2.5-14B	1956	443	0.362	2.24	875	3.9×10^{-12}	1.5×10^{-11}
Qwen2.5-32B	2678	464	0.305	2.46	1090	5.4×10^{-7}	1.3×10^{-6}
Llama-3.3-70B	2498	451	0.441	3.00	833	7.8×10^{-12}	2.7×10^{-11}
Qwen2.5-72B	2384	443	0.316	2.39	999	3.2×10^{-6}	6.9×10^{-6}

of the judge's own family. Interventions are generated once and reused across all 2^4 configurations, so that the same counterfactual text enters every subset containing it.

Style ($-S$). We use a model M' from a different model family to rewrite the full response, shifting its stylistic realization while preserving its meaning:

$$o_J^{(-S)} = \text{Rewrite}_{M'}(o_J), \qquad M' \notin \text{family}(J). \tag{14}$$

The rewriting model is instructed to preserve every factual claim, the reasoning order, and the stated answer, while freely reorganizing sentence structure, discourse markers, and formatting. Using a model outside $\text{family}(J)$ shifts the response away from the judge's stylistic distribution rather than merely paraphrasing within it.

Lexical ($-L$). We use an external LLM to identify model-characteristic lexical choices and replace them with context-preserving synonyms. In practice we use ChatGPT[1] and Claude[2]:

$$\mathcal{P}_L(o_J) = \{w_i \to \tilde{w}_i\}_{i=1}^{m}, \qquad o_J^{(-L)} = \text{Replace}(o_J, \mathcal{P}_L(o_J)), \tag{15}$$

where w_i denotes a targeted lexical item and $\tilde{w}_i$ its context-preserving synonym. Candidate items are words and short phrases that the external model flags as characteristic of the judge's outputs;

[1] https://chatgpt.com
[2] https://claude.ai

replacements are constrained to match the original register, so that the substitution changes lexical identity without simplifying or elevating the text. Replacements are applied by exact string substitution, leaving all other tokens untouched.

Reasoning ($-R$). We first summarize recurrent reasoning structures across responses into a template set $\mathcal{H} = \{h_1, \ldots, h_M\}$, where each template specifies an ordering and organization of argumentative moves. Each response is then matched to an alternative compatible template and its argument reconstructed accordingly:

$$h^* = \mathrm{Match}(o_J, \mathcal{H}), \qquad o_J^{(-R)} = \mathrm{Reconstruct}(o_J; h^*), \tag{16}$$

where h^* denotes the selected alternative template. The reconstruction reuses the original claims and evidence, changing only the order and framing in which the argument unfolds; the final answer and its justification content are preserved.

Alignment-associated patterns ($-A$). We derive a pattern inventory $\mathcal{P}_A$ from matched base and instruction-tuned responses, capturing recurrent conventions associated with post-training:

$$\mathcal{P}_A = \mathrm{Diff}(\mathcal{D}_{\mathrm{inst}}, \mathcal{D}_{\mathrm{base}}), \tag{17}$$

where $\mathcal{D}_{\mathrm{inst}}$ and $\mathcal{D}_{\mathrm{base}}$ denote response sets from instruction-tuned and base models on matched prompts. The inventory includes hedging, disclaimers, safety-flavored framing, and assistant-style openings and closings. We then remove the matched conventions from the judge response while retaining its task-relevant information:

$$o_J^{(-A)} = \mathrm{RemovePatterns}(o_J; \mathcal{P}_A). \tag{18}$$

Composition. For a subset $\mathcal{U} \subseteq \mathcal{F}$, the joint counterfactual $o_J^{(-\mathcal{U})}$ is constructed by applying the corresponding interventions under a fixed order, so that every configuration is generated by the same protocol.

D Conclusion Matching on Open-Ended Tasks

TruthfulQA and AlpacaEval have no ground-truth answers, so the correctness conditions in Eq. equation 4 cannot be applied. On these tasks a pair enters the frozen set when the judge and proxy responses agree in conclusion. Agreement is determined by bidirectional entailment with an NLI model (DeBERTa-v3-base-mnli-fever-anli; Laurer et al., 2024). A pair is accepted when $\mathrm{entail}(a \Rightarrow b) \geq \tau$ and $\mathrm{entail}(b \Rightarrow a) \geq \tau$ with $\tau = 0.5$.

Acceptance rates range from 59.2% to 75.7% on TruthfulQA (530 candidate questions per reference–proxy combination) and from 52.5% to 64.4% on AlpacaEval (550 candidates). The open-ended tasks complement the verifiable ones, testing whether the pattern extends beyond option-constrained settings; the strict task contrast in this paper is MMLU versus Math500.

E Additional Results

Scaling panel below 7B. The four judges below 7B are evaluated on MMLU only. Their residuals are near zero or slightly negative.

Table 8: B_0 for judges below 7B on MMLU, with question-clustered 95% confidence intervals and two-sided cluster-robust p values.

Model	B_0	95% CI	N_q	Pairs	p
Llama-3.2-1B	-0.006	$[-.012, +.001]$	498	3084	9.2×10^{-2}
Qwen2.5-1.5B	-0.022	$[-.027, -.016]$	553	3705	8.4×10^{-15}
Llama-3.2-3B	$+0.001$	$[-.007, +.009]$	524	3573	8.1×10^{-1}
Qwen2.5-3B	$+0.033$	$[+.020, +.046]$	463	3298	1.2×10^{-6}

Reference sensitivity. Pooled estimates are mixtures over references. Per-reference estimates are available for three judges on MMLU. All nine runs are positive with confidence intervals excluding zero, and each pooled value lies within the range of its per-reference components.

Table 9: B_0 by reference model, MMLU. Confidence intervals are cluster-robust at 95%, clustered on question, and p is two-sided.

Judge	Reference	Pairs	N_q	B_0	95% CI	p
Llama-3.3-70B	Qwen2.5-14B	193	68	$+0.298$	$[+.213, +.384]$	1.8×10^{-9}
	Qwen2.5-32B	276	93	$+0.186$	$[+.115, +.258]$	1.2×10^{-6}
	Qwen2.5-72B	304	103	$+0.230$	$[+.164, +.297]$	4.5×10^{-10}
	Pooled	773	140	$+0.230$	$[+.172, +.287]$	5.1×10^{-13}
Qwen2.5-32B	Llama-3.3-70B	490	146	$+0.233$	$[+.187, +.278]$	1.4×10^{-18}
	Qwen2.5-14B	256	86	$+0.284$	$[+.216, +.352]$	1.6×10^{-12}
	Qwen2.5-72B	374	115	$+0.226$	$[+.173, +.279]$	1.5×10^{-13}
	Pooled	1120	192	$+0.242$	$[+.202, +.281]$	2.8×10^{-25}
Qwen2.5-72B	Llama-3.3-70B	374	109	$+0.135$	$[+.091, +.178]$	1.4×10^{-8}
	Qwen2.5-14B	211	67	$+0.173$	$[+.119, +.227]$	2.1×10^{-8}
	Qwen2.5-32B	230	69	$+0.187$	$[+.123, +.251]$	1.5×10^{-7}
	Pooled	815	145	$+0.161$	$[+.125, +.198]$	1.5×10^{-14}

Leaderboard under the weak reference. Under Llama-3.2-1B, scores approach the ceiling and the ranking is more fragile, with Kendall $\tau = 0.33$ and the top rank changing as well.

Table 10: Leaderboard under the Llama-3.2-1B reference, averaged over six judges. Gemma candidates are not judges and their scores are unchanged by removing self-votes.

	With self-votes		Self-votes removed	
Candidate	Win rate	Rank	Win rate	Rank
Llama-3.3-70B	0.806	1	0.790	2
Qwen2.5-72B	0.791	2	0.777	4
Qwen2.5-14B	0.783	3	0.803	1
Qwen2.5-32B	0.780	4	0.752	5
Qwen2.5-7B	0.760	5	0.751	6
Llama-3.1-8B	0.744	6	0.785	3
Gemma-2-27B	0.605	7	0.605	7
Gemma-2-9B	0.603	8	0.603	8

F Rewrite Fidelity

Interventions rewrite responses, and a rewrite can degrade quality. To measure this, a cross-family gold judge blindly compares each intervened response with its original, and the score is the win rate of the intervened response, where 0.500 indicates preserved quality. Rewrites of qwen and mistral responses are scored by Llama-3.3-70B, and rewrites of llama responses by Qwen2.5-72B, so no rewrite is scored by the family that produced it.

The gold judge is itself an LLM, so we use its scores only to discount channels, never to certify them, a reading robust to the sign of any judge bias. The $-R$ degradation follows the rewriter family rather than the scoring family, consistent with a rewriter effect.

Table 11 reports all 13 available cells. The $-S$ and $-A$ rewrites are quality-neutral throughout. The $-L$ substitutions carry a mild and consistent cost. The $-R$ restructuring splits by rewriter family, near neutral on llama responses and clearly degraded on qwen and mistral responses. The all-intervention configuration bears the largest cost, down to 0.249. We therefore read channel-

level attributions as descriptive, and the directly measured ρ, which comes from the all-intervention configuration, is understated rather than overstated.

Table 11: Rewrite fidelity by intervention, scored by a cross-family gold judge. Values are win rates of the intervened response against the original, where 0.500 indicates preserved quality.

Judge	Task	$-S$	$-L$	$-R$	$-A$	$-$all
Llama-3.1-8B	MMLU	0.504	0.474	0.526	0.500	0.492
Llama-3.3-70B	MMLU	0.503	0.485	0.511	0.499	0.488
Llama-3.3-70B	Math500	0.507	0.489	0.470	0.500	0.481
Llama-3.3-70B	TruthfulQA	0.502	0.484	0.497	0.490	0.478
Mistral-7B-v0.3	MMLU	0.511	0.434	0.405	0.497	0.392
Qwen2.5-7B	MMLU	0.483	0.446	0.264	0.501	0.300
Qwen2.5-14B	MMLU	0.489	0.444	0.312	0.502	0.306
Qwen2.5-32B	MMLU	0.489	0.447	0.333	0.500	0.318
Qwen2.5-32B	Math500	0.506	0.457	0.306	0.500	0.318
Qwen2.5-32B	TruthfulQA	0.498	0.491	0.313	0.491	0.284
Qwen2.5-72B	MMLU	0.491	0.448	0.334	0.502	0.288
Qwen2.5-72B	Math500	0.507	0.453	0.322	0.500	0.343
Qwen2.5-72B	TruthfulQA	0.498	0.492	0.305	0.494	0.249

G PROMPTS

This appendix lists all prompts used in the pipeline.

G.1 ANSWER GENERATION.

MMLU:

```
Answer the following multiple-choice question. Think briefly,
then end your response with a line of the form 'The final
answer is $$X$$' where X is one of A, B, C, or D.

{question}
```

Math500:

```
Solve the following math problem step by step. At the end,
state the final answer on its own line as \boxed{your answer}.

{question}
```

AlpacaEval:

```
You are a helpful assistant. Please help me answer the query
below. Keep your answer within 200 words.

# Query
{query}

Now, please provide your answer to this query.
# Answer
```

TruthfulQA:

```
You are a truthful assistant. Please help me answer the query
below truthfully, remaining faithful to factual knowledge and
the given context. Avoid making up any untrue facts.
```

```
# Query
{query}

Now, please provide your answer to the query. Keep your answer
within 200 words.
# Answer
```

G.2 JUDGING.

```
You will be given answers from two assistants (Assistant A and
Assistant B) to a user prompt. Please act as an impartial judge
and evaluate the quality of their answers.

You must output only one of the following choices:

- "A" if Assistant A's answer is better.
- "T" if both answers are relatively the same in quality.
- "B" if Assistant B's answer is better.

For example, if both answers are relatively the same in
quality, output "T".

<User Prompt>
{question}

<The Start of Assistant A's Answer>
{answer_a}
<The End of Assistant A's Answer>

<The Start of Assistant B's Answer>
{answer_b}
<The End of Assistant B's Answer>
```

G.3 INTERVENTION PROMPTS

G.3.1 STYLE ($-S$)

```
Rewrite the following assistant response so that its writing
STYLE is different, while its meaning is IDENTICAL.

Requirements:
- Preserve ALL information, claims, numbers, and the final
  answer exactly.
- Change stylistic realization: sentence length and rhythm,
  register, connective words, and formatting habits (e.g.
  prose vs. lists).
- Do NOT add new content, opinions, or commentary. Do NOT
  explain what you changed.
- Output ONLY the rewritten response, nothing else.

<Response to rewrite>
{response}
</Response to rewrite>
```

G.3.2 LEXICAL ($-L$)

```
Below is an assistant response. Identify up to {k} words or
short phrases that are CHARACTERISTIC vocabulary choices (words
a model tends to over-use, or distinctive word choices) and
```

```
propose a neutral, meaning-preserving synonym for each.

Rules:
- Only single words or short (<=2 word) phrases that ACTUALLY
  APPEAR in the response.
- Select CONTENT words only: do NOT select stopwords/function
  words, words taken from the question, numbers, proper nouns,
  the final answer, math symbols, units, or code tokens.
- The replacement must preserve meaning exactly and fit the
  sentence grammatically.
- Output ONLY a JSON object with two equal-length lists:
  {"selected": ["word1", "phrase two", ...],
   "replacements": ["syn1", "syn 2", ...]}

<Response>
{response}
</Response>

JSON:
```

G.3.3 REASONING ($-R$)

```
[system] You classify the argument structure of a response.
Output ONLY one label.

Classify the ARGUMENT-PROGRESSION structure of the response
below into exactly one of:
{labels}

Definitions:
{defs}

Output ONLY the single best-matching label string.

<Response>
{response}
</Response>

Label:

Restructure the argument of the response below into this
target structure:

TARGET: {target} -- {target_desc}

Hard requirements:
- Keep ALL premises, facts, and reasoning content; only change
  the ORDER/STRUCTURE in which the argument progresses.
- Do NOT add or remove information.
- Output ONLY the restructured response.

<Response>
{response}
</Response>
```

G.3.4 ALIGNMENT ($-A$)

```
[system] You remove hedging and softening language from text
while keeping every factual claim and the final answer exactly.
```

```
You output only the edited text.

Rewrite the text below to remove in-sentence HEDGING and
softening (e.g. {cues}) so statements are direct, WITHOUT
changing any facts, numbers, or the final answer.

- Do not add commentary or explain changes.
- Keep all task-relevant information intact.
- Output ONLY the edited text.

<Text>
{response}
</Text>
```

G.3.5 COMPOSED

```
Rewrite the following assistant response, applying ALL of the
changes below SIMULTANEOUSLY in a single rewrite, while keeping
its meaning IDENTICAL.

Changes to apply:
{directives}

General requirements:
- Preserve ALL information, claims, and numbers.
- Do NOT add new content, opinions, or commentary. Do NOT
  explain what you changed.
- Output ONLY the rewritten response, nothing else.

<Response to rewrite>
{response}
</Response to rewrite>
```