HEADS TAKE SIDES: IMBALANCED STATIC VOTING IN UNFAIR LLM JUDGES

Anonymous authors
Paper under double-blind review

ABSTRACT

LLM judges can encode correctness-related information internally, yet how this information is converted into a final verdict remains unclear. By decomposing the verdict-logit margin into attention-head contributions, we identify a predominant head-level mechanism of **static voting** in LLM judges: opposing head groups consistently favor acceptance or rejection, while input-dependent voting strengths determine their competition. Although their voting directions remain fixed, their voting strengths vary with solution quality—acceptance heads weaken and rejection heads strengthen as errors accumulate—enabling input-dependent judgments through competition. This static voting structure recurs across model architectures and verdict formats, while targeted ablations confirm that the two camps causally push judgments in opposite directions. This account suggests that reliable correction should adapt the balance between the two camps rather than suppress either one. We therefore introduce **Voting Calibration**, which uses an internal correctness signal to modulate their voting strengths while preserving their specialized directions. Relying on a small reference set, Voting Calibration improves overall accuracy and true-negative rate by averages of 8.3% and 14.6%, respectively, across four models on μ-MATH benchmark.

1 INTRODUCTION

Large language models (LLMs) are now widely deployed for automatic assessments: scoring responses, ranking alternatives, and generating feedback for model alignment, forming what has become known as the *LLM-as-a-judge* paradigm (Lee et al., 2024; Liu et al., 2023; Zheng et al., 2023). Yet the efficiency of this approach should not be mistaken for evaluative reliability. A growing body of evidence shows that LLM judgments are systematically shaped by candidate order (Wang et al., 2024a; Li et al., 2024), verbosity (Saito et al., 2023), purported majority preferences (Koo et al., 2024), and even cues with little substantive bearing on answer quality (Xu et al., 2026). Moreover, on harder tasks such as mathematical reasoning and code verification, LLM judges can exhibit pronounced and persistent failures in most cases (Tan et al., 2025). This unreliability is not merely an issue of evaluation accuracy; it points to a more fundamental question about the judging process itself: Understanding how a judge internally transforms a solution into a final verdict therefore becomes essential.

To study such questions, mechanistic studies offer two complementary levels of explanation: representation level and circuit level. At the representation level, model behavior is described in terms of the geometry of hidden activations, where particular directions or subspaces can encode semantic concepts and behavioral tendencies (Zou et al., 2023; Marks & Tegmark, 2024). Circuit-level analysis goes further by looking at how different components in the model interact and influence its behavior (Olsson et al., 2022). In LLM-as-a-judge settings, recent work has studied its mechanism at the representation level, including analyses of logit distributions (Lai et al., 2025), latent representation trajectories (Wang et al., 2025), and geometric interventions (Xu et al., 2026). By contrast, circuit-level accounts of judging remain relatively sparse. Thus, even after identifying correctness information in the internal representations, it remains unclear how individual components contribute to the final judgment: *how are these signals aggregated into the final judgment? what roles do different components play in this process?*

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.

In this work, we focus on the computations performed by individual attention heads. Following prior works, we first formulate the objective of LLM-as-a-judge task as enlarging the logit margin between the *correct* and *incorrect* verdict tokens in the appropriate direction (Wang et al., 2023; Lepori et al., 2025). We argue that each attention head can support accurate judging in two ways: shifting the verdict positively for correct solutions and negatively for incorrect ones. Scoring heads on these two abilities, we identify two kinds of head: **1) *dynamic heads***, whose contribution flips as the label changes; **2) *static heads***, which consistently favor the same label.

Crucially, we demonstrate that static heads dominate in both number and contribution strength. That is, **most heads play only one role in the judge task: a staunch supporter or a permanent resister**, and thus form two opposing groups. This suggests that, internally, the model does not generate its decision through adaptive processing within a single module; rather, the final output emerges from competition between camps operating in opposing directions. We introduce the term **static voting** to describe this mechanism, as shown in Fig 1.

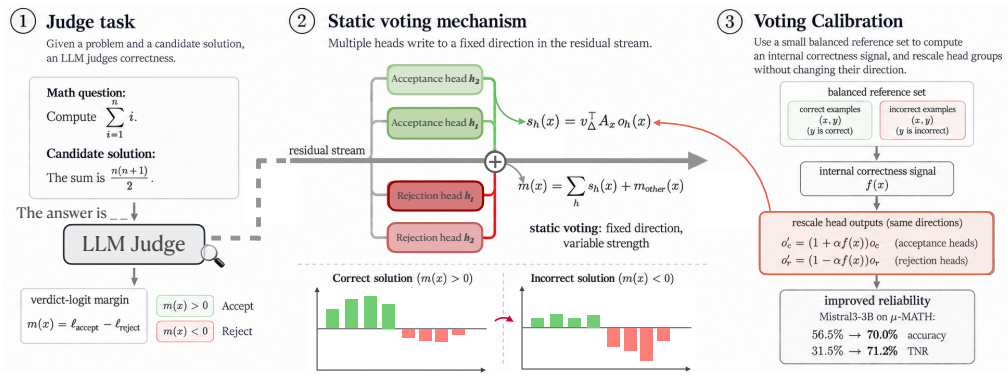

Figure 1: **Overview of our Static voting model and Voting Calibration intervention**: A head that favors acceptance on average can still distinguish correct from incorrect solutions by contributing more strongly to correct ones. Voting Calibration further exploits this correctness signal to adaptively rescale the strengths of the two head groups.

Concretely, our paper's structure and contributions are as follows. (1) Sec 2 formalizes static voting, defining the task objective, head-level categories and contributions, and theoretical guarantees for the aggregate model. (2) Sec 3 identifies static heads across architectures and shows that static voting is a general and causal mechanism that dominates the aggregate contribution of attention heads to LLM-as-a-judge verdicts. (3) Despite fixed verdict directions, static-head representations encode error severity, producing input-dependent judgments through contribution magnitude. (4) Sec 4 introduces a calibration-based intervention that adaptively reweights the two camps using a small reference set, improving overall performance and outperforming several LLM-as-a-judge baselines.

2 Theoretical Framework

2.1 LLM-as-a-Judge as Logit-Difference Prediction

We study LLM-as-a-Judge as binary correctness assessment. A judge receives a structured input $x = P(q, a)$, formed by formatting a question q and a candidate answer a with a prompt template P. We encode acceptance and rejection as $+1$ and -1, respectively, and write $y \in \{+1, -1\}$ for the ground-truth correctness label.

Let $z_\theta^+(x)$ and $z_\theta^-(x)$ be the logits of the verdict tokens t_+ and t_-. The judge predicts $\hat{y} = +1$ if $\mathbf{m}(x) > 0$ and $\hat{y} = -1$ otherwise, where

$$\mathbf{m}(x) := z_\theta^+(x) - z_\theta^-(x) = \log \frac{\pi_\theta(t_+ \mid x)}{\pi_\theta(t_- \mid x)}. \tag{1}$$

To quantify individual attention heads' contributions, we expand the forward computation of a pre-norm Transformer with $|\mathcal{H}| = L \times H$ attention heads. At the verdict position, the residual stream

satisfies

$$\begin{aligned}\mathbf{r}_L(x) &= \mathbf{r}_0(x) + \sum_{\ell=1}^{L}\left[\sum_{h=1}^{H}\mathbf{o}_{\ell,h}(x) + \mathbf{g}_\ell(x)\right] \\ &= \sum_{h\in\mathcal{H}}\mathbf{o}_h(x) + \mathbf{r}_{\text{other}}(x).\end{aligned} \tag{2}$$

Here, $\mathbf{r}_0(x)$ is the initial residual state, and $\mathbf{o}_h(x)$ is head h's output after projection into residual-stream coordinates. The remaining layer update $\mathbf{g}_\ell(x)$ includes the MLP output and any shared attention output bias.

Let $\mathbf{v}_\Delta$ and b_Δ denote the unembedding vector and bias differences between t_+ and t_-. At the observed residual state, the final LayerNorm or RMSNorm satisfies $\mathrm{Norm}(\mathbf{r}_L(x)) = \mathbf{A}_x\mathbf{r}_L(x)+\boldsymbol{\beta}_x$, where $\mathbf{A}_x$ incorporates any centering and learned gain, with the normalization scale evaluated at the full residual state. Substitution yields

$$\begin{aligned}\mathbf{m}(x) &= \mathbf{v}_\Delta^\top[\mathbf{A}_x\mathbf{r}_L(x) + \boldsymbol{\beta}_x] + b_\Delta \\ &= \sum_{h\in\mathcal{H}}\mathbf{s}_h(x) + \mathbf{m}_{\text{other}}(x),\end{aligned} \tag{3}$$

where

$$\begin{aligned}\mathbf{s}_h(x) &= \mathbf{v}_\Delta^\top\mathbf{A}_x\mathbf{o}_h(x), \\ \mathbf{m}_{\text{other}}(x) &= \mathbf{v}_\Delta^\top\mathbf{A}_x\left[\mathbf{r}_0(x) + \sum_{\ell=1}^{L}\mathbf{g}_\ell(x)\right] + \mathbf{v}_\Delta^\top\boldsymbol{\beta}_x + b_\Delta.\end{aligned} \tag{4}$$

This is an exact decomposition of the observed margin; it does not by itself describe the effects of intervening on a head. Positive and negative values of $\mathbf{s}_h(x)$ contribute toward acceptance and rejection, respectively.

Let D denote the task distribution, with positive probability for each class, and let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ be an evaluation dataset. Expectations below are taken under D and estimated from $\mathcal{D}$ in the experiments. Assuming finite first moments for head contributions, we define each head's average support for the ground-truth verdict as

$$\mu_h := \mathbb{E}_{(x,y)\sim D}[y \cdot \mathbf{s}_h(x)]. \tag{5}$$

Writing D_+ and D_- for the distributions conditional on $y = +1$ and $y = -1$, respectively, we further define

$$\mu_h^+ := \mathbb{E}_{(x,y)\sim D_+}[\mathbf{s}_h(x)], \qquad \mu_h^- := \mathbb{E}_{(x,y)\sim D_-}[-\mathbf{s}_h(x)]. \tag{6}$$

The pair (μ_h^+, μ_h^-) measures average support for accepting correct answers and rejecting incorrect ones. Unlike the class-weighted average in Eq. 5, it distinguishes a shared mean verdict preference from a preference that changes across classes.

2.2 LLM Judge as a Static Voting Model

We now ask how heads with the same average verdict preference across classes can jointly support input-dependent judgments. The key distinction is between a head's mean voting direction and the variation of its contribution across inputs.

Definition 2.1 (Static and dynamic heads)**.** *Heads with nonzero class-conditional means are assigned to four sets according to the signs of* (μ_h^+, μ_h^-)*:*

$$\begin{aligned}\mathcal{H}_+^{\text{st}} &:= \{h : \mu_h^+ > 0,\ \mu_h^- < 0\}, \quad & \mathcal{H}_-^{\text{st}} &:= \{h : \mu_h^+ < 0,\ \mu_h^- > 0\}, \\ \mathcal{H}_+^{\text{dyn}} &:= \{h : \mu_h^+ > 0,\ \mu_h^- > 0\}, \quad & \mathcal{H}_-^{\text{dyn}} &:= \{h : \mu_h^+ < 0,\ \mu_h^- < 0\}.\end{aligned} \tag{7}$$

The static heads form $\mathcal{H}^{\text{st}} = \mathcal{H}_+^{\text{st}} \cup \mathcal{H}_-^{\text{st}}$*; the dynamic heads form* $\mathcal{H}^{\text{dyn}} = \mathcal{H}_+^{\text{dyn}} \cup \mathcal{H}_-^{\text{dyn}}$*. Heads with either mean equal to zero are left unclassified.*

Static heads favor the same verdict in class-conditional expectation, whereas dynamic heads have mean contributions of opposite signs across classes. This distinction concerns averages, not pointwise signs: a static head's contribution may still vary with correctness and with the input.

Definition 2.2 (Static voting model)**.** *The aggregate score of the static heads is*

$$\begin{aligned}\mathbf{S}_{\text{static}}(x) &:= \sum_{h\in\mathcal{H}_+^{\text{st}}} \mathbf{s}_h(x) + \sum_{h\in\mathcal{H}_-^{\text{st}}} \mathbf{s}_h(x) \\ &= \sum_{h\in\mathcal{H}_+^{\text{st}}} \mathbf{v}_\Delta^\top \mathbf{A}_x \mathbf{o}_h(x) + \sum_{h\in\mathcal{H}_-^{\text{st}}} \mathbf{v}_\Delta^\top \mathbf{A}_x \mathbf{o}_h(x).\end{aligned} \tag{8}$$

This quantity captures the collective directional preference of the static heads toward acceptance or rejection. Through competition between the two oppositely oriented groups, the model can in principle produce dynamic, input-dependent decisions. However, *to what extent can this input-dependent voting model ensure correct judgments?*

Theorem 2.1 (Variance-sensitive error bound for static voting)**.** *Suppose that, in each class, the conditional mean of the signed score $y\mathbf{S}_{\text{static}}(x)$ is at least $d > 0$ and its conditional variance is at most v, where $0 \le v < \infty$. Then*

$$R(\mathbf{S}_{\text{static}}) \le \frac{v}{v + d^2}.$$

For any real-valued score g, $R(g)$ represents its classification error when $g(x) > 0$ predicts $+1$ and $g(x) \le 0$ predicts -1. Detailed notation, a sharper class-conditional bound, and the proof appear in Appendix B.1. Theorem 2.1 shows that static voting achieves low error when class means lie on opposite sides of the threshold with sufficient separation relative to within-class variation.

Furthermore, the static score constitutes only one component of the full verdict margin. We therefore characterize the relationship between the classification errors of the static voting model and the full model in the following result (proof in Appendix B.2).

Theorem 2.2 (From static voting to the full model)**.** *For the static score and full margin evaluated on the same forward pass,*

$$\left|R(\mathbf{m}) - R(\mathbf{S}_{\text{static}})\right| \le \Pr(|\mathbf{m}(x) - \mathbf{S}_{\text{static}}(x)| \ge |\mathbf{S}_{\text{static}}(x)|)\,.$$

The static voting model and the full model have similar error rates when the residual contribution rarely exceeds the static score in magnitude. Combining this result with the preceding theorem yields a bound on the full-model error given by the static-voting error bound plus the probability that the residual contribution dominates the static score.

3 Characterizing Static Heads in LLM Judges

3.1 Experimental Settings

Tasks. We organize the experiments into three stages. **(i) Structure identification** uses class-conditional head contributions to identify opposing static camps and examine their consistency across models and verdict formats. **(ii) Feature analysis** examines how head contributions and internal representations vary with solution correctness and error severity. **(iii) Causal intervention** (Sec 4) assesses how camp ablation and contribution-level correction affect class-conditional judgment accuracy.

Datasets. We use ProcessBench (Zheng et al., 2025) to identify head structure and Math-Shepherd (Wang et al., 2024b) to analyze error-sensitive responses. For the latter, we group solutions into six error levels (0–5) according to the number of incorrect steps. We use μ-MATH (Chernyshev et al., 2025) to assess robustness across verdict formats and to measure the effects of causal interventions. All static-head identification and ranking used in subsequent interventions are performed exclusively on ProcessBench; no μ-MATH labels are used for head selection.

Models. Qwen3-8B (Yang et al., 2025) is the primary model for detailed head-level analysis. Cross-model experiments additionally include Qwen3.5-9B (Qwen Team, 2026), Gemma-4-12B-it (Gemma Team, 2026), and Ministral-3-3B-Base-2512 (Liu et al., 2026).

3.2 Static Camps Across Models and Verdict Forms

We first establish the empirical validity of the static voting model through large-scale behavioral experiments. Two main findings emerge: Static voting is robust across architectures and predominates among heads with the largest contributions.

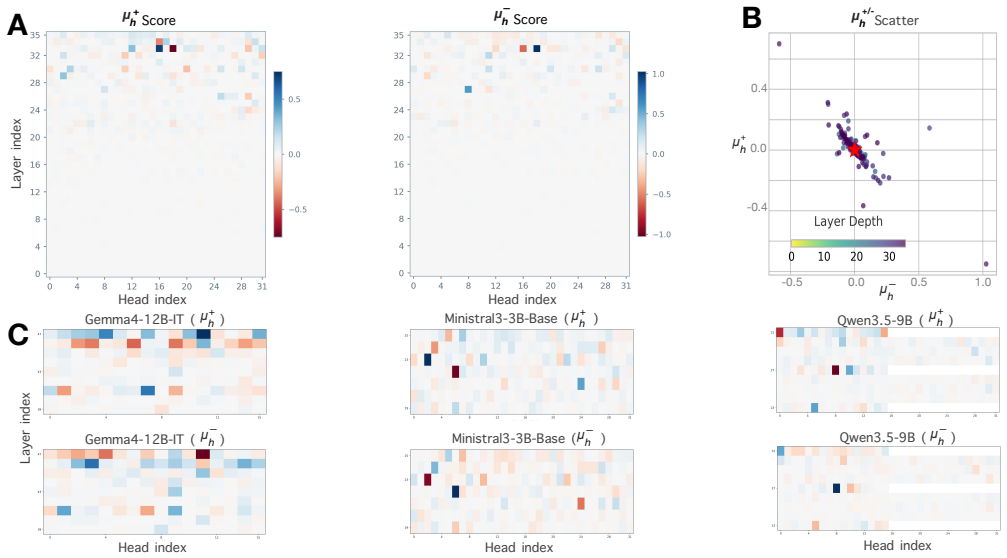

Figure 2: **Global and local analysis of static head behavior.** **A**. Heatmaps of $\mu_{\pm}$ scores for each head in Qwen3-8B. **B**. Each head in the $\mu_{\pm}$ plane. **C**. Heatmaps of $\mu_{\pm}$ scores for representative static heads in diverse models (Full heatmaps are shown in Appendix D.1)

Evidence for static voting's existence. We score the class-conditional contributions of attention heads in four language models with distinct architectures and training corpora, spanning 3B to 12B parameters, on ProcessBench (Zheng et al., 2025). Fig 2.A and .C map these scores across head locations. Comparing the two heatmaps for each model reveals a recurring reversal: heads that contribute positively to the appropriate verdict for one class often contribute negatively for the other. Static corrections of significant magnitude often appear in deeper layers. Because the scores are aligned with the correct verdict for each class, this reversal indicates that a head's mean contribution favors either acceptance or rejection across both classes, as predicted by static voting. Fig 2.B makes the pattern clearer at the individual-head level in Qwen3-8B: μ_h^+ and μ_h^- are predominantly of opposite signs, with relatively few heads contributing positively to both. Corresponding scatter plots for the other three models are provided in the Appendix D.1.

Table 1: Head composition and static-head contributions on Mu-Math. All entries are percentages. Details equation of metrics are shown in Appendix B.3

Judge model	Head proportions ($\alpha = 0.05$)					Static contribution	
	$p_{\mathrm{st}+}$	$p_{\mathrm{st}-}$	$p_{\mathrm{dyn}+}$	$p_{\mathrm{dyn}-}$	$p_{\mathrm{uncertain}}$	C_{st}	$\widehat{\Pr}(\lvert S(x)\rvert > \lvert R_{\mathrm{attn}}(x)\rvert)$
Qwen3-8B	45.31	48.61	1.56	0.43	4.08	**93.37**	**76.55**
Ministral3-3B-Base	49.16	47.84	0.00	0.12	2.88	**99.52**	**100.00**
Qwen3.5-9B	50.78	44.75	1.12	0.11	3.24	**98.43**	**99.26**
Gemma4-12B-it	48.57	45.96	1.43	1.30	2.73	**89.66**	**91.04**

Numerical Distribution of Static Heads. We subsequently quantified the statistics of the static head groups. Across all four Mu-Math models, static heads—those whose mean contributions favor the same answer for both correct and incorrect responses—constitute the overwhelming majority of analyzed heads (roughly 94–97%; Table 1). More importantly, they account for a similarly dominant share of the total class-balanced mean absolute head contribution, C_{st} (about 90–99%). To account

for cancellation, let $S(x)$ and $R_{\text{attn}}(x)$ denote the summed signed contributions of static heads and all remaining heads, respectively. We find that $|S(x)| > |R_{\text{attn}}(x)|$ for most samples, exceeding 90% in three of the four models. On these samples, the static component determines the sign of the aggregate attention contribution. Thus, static heads are not merely prevalent: within the measured attention decomposition, they dominate both the magnitude of contributions and the net direction of the aggregate signal. Their relative balance is a plausible influence on the final verdict, motivating the causal ablations in Section 4.

3.3 Error-Sensitive Voting Strength and Correctness Information

Voting strength responds to error severity. We select the top four heads from each static camp identified in Sec 3.2 and measure their mean contribution intensity $a_h(x) = |s_h(x)|$ across six Math-Shepherd error levels (Fig. 3, left). As the number of incorrect steps increases from 0 to 5, the acceptance-oriented heads generally weaken, whereas the rejection-oriented heads strengthen. This opposing response appears across all four ranks in each camp, indicating that a static verdict preference can produce an error-sensitive vote by varying its contribution strength with solution quality.

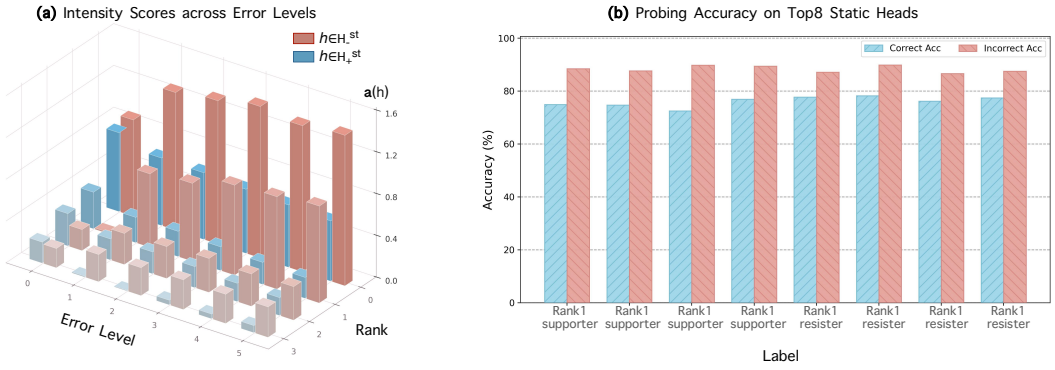

Figure 3: **Error-sensitive intensity and correctness information in static heads.** Left: mean intensity $a_h(x)$ across error levels for the top four heads in both camps. Right: accuracy of probes trained on each head's hidden states, reported separately for correct and incorrect solutions.

Correctness is decodable from both camps. Probes trained on the eight heads' hidden states attain approximately 72–78% accuracy on correct solutions and 86–90% on incorrect solutions (Fig. 3, right). The accuracy ranges overlap between the acceptance-oriented and rejection-oriented camps: heads with opposing average vote directions both contain information predictive of solution correctness. Together, the intensity and probing results demonstrate that directional specialization coexists with graded responses to errors and decodable correctness information. The probes establish the availability of this information; they do not, by themselves, establish that it causally controls vote strength.

3.4 Geometric Basis of Static Voting and Relative Correction

We next examine how fixed verdict preferences coexist with correctness information. For each head, u_{abs} points from the origin to the mean output over Math-Shepherd solutions, while u_{relative} points from the incorrect- to correct-solution centroid; we measure both against the correct-verdict direction v_Δ.

Relative correctness directions provide within-role correction. The relative directions show a different pattern (Fig. 4b): their angles are more dispersed, with substantial overlap around 90°, yet both camps include heads with $\angle(u_{\text{relative}}, v_\Delta) < \pi/2$, and heads with large net voting contributions are concentrated well below this threshold. Such heads provide relatively more support for the correct verdict on correct than incorrect solutions. For H_+^{st}, this strengthens acceptance on correct

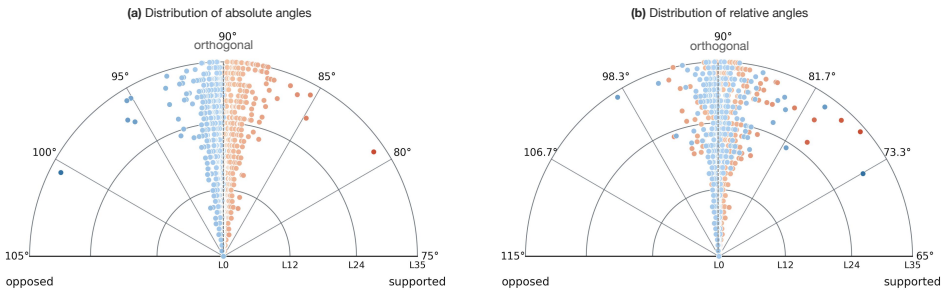

Figure 4: **Absolute verdict preference and relative correctness displacement.** (a) Angles between u_{abs} and v_Δ; (b) angles between u_{relative} and v_Δ. Orange denotes heads $\in H_+^{\text{st}}$, and blue denotes rejection-oriented static heads $\in H_-^{\text{st}}$.

solutions; for H_-^{st}, it weakens rejection. Thus, class-dependent displacements provide a *relative correction* without reversing the head's average voting direction. The correction magnitude depends on both $\angle(u_{\text{relative}}, v_\Delta)$ and $|u_{\text{relative}}|$.

Static heads exhibit distinct correctness-sensitive geometries. We visualize the underlying sample distributions for the representative rejection-oriented static head L33H18, acceptance-oriented static head L33H16, and dynamic head L27H8 using separate MDS projections (Fig. 5). The error-level centroids of L33H18 and L27H8 move visibly from level 0 toward higher error levels, although individual samples overlap. L33H16 exhibits a smaller displacement, with substantial overlap among erroneous levels. These distinct geometries reinforce the angular result: correctness-sensitive variation is present in head outputs, while its visible organization differs across heads and does not alone determine their voting roles.

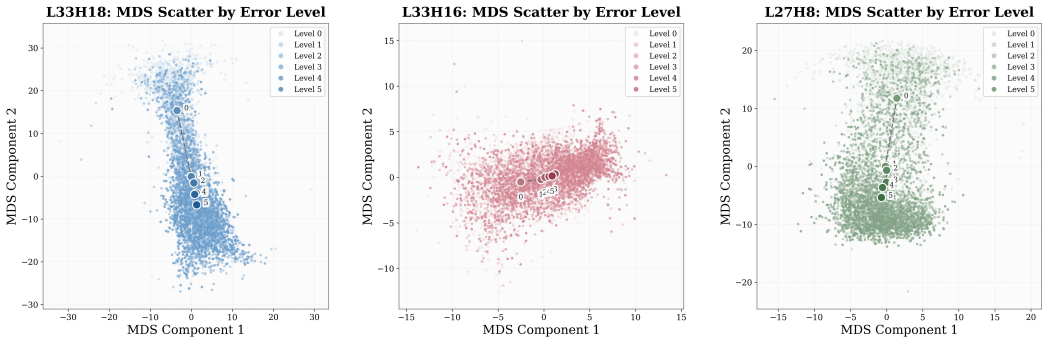

Figure 5: **MDS projections of representative head outputs on Math-Shepherd.** Each point represents a solution, colored by its number of incorrect steps (levels 0–5). Large markers indicate error-level centroids, connected in level order. MDS is performed separately for L33H18, L33H16, and L27H8; the plotted axes are not the verdict direction v_Δ.

4 Intervention Experiments

Having established the head-level structure, we examine its causal role in the final verdict: whether static voting causally controls predictions and which intervention most effectively aligns its aggregate vote with solution correctness. We test both questions on μ-MATH across architectures.

4.1 Static Head Camps Causally Shape Final Verdicts

To test whether static heads causally affect the final verdict, we zero the top-k heads from either camp on μ-MATH, with $k \in 1, 2, 4, 8$, and compare against the unmodified model ($k = 0$) using TPR and

TNR. The voting account predicts opposite effects: ablating rejection-oriented heads should increase acceptance, whereas ablating acceptance-oriented heads should decrease it.

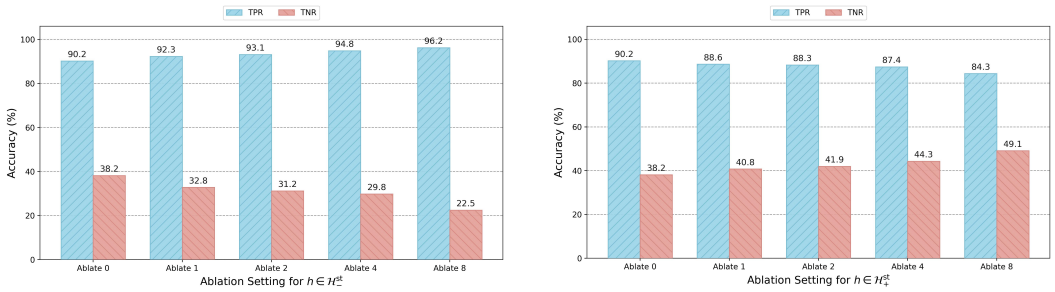

Figure 6: **Static-head ablation changes Qwen3-8B verdicts on μ-MATH.** Left: zeroing the top-k rejection-oriented heads in $\mathcal{H}_-^{\text{st}}$. Right: zeroing the top-k acceptance-oriented heads in $\mathcal{H}_+^{\text{st}}$. The unmodified model corresponds to $k = 0$. TPR measures correct-solution acceptance; TNR measures incorrect-solution rejection.

Both predictions hold, with effects increasing monotonically with k (Fig. 6). Ablating eight rejection-oriented heads raises TPR from 90.2% to 96.2% (+6.0 points) but lowers TNR from 38.2% to 22.5% (-15.7), whereas ablating eight acceptance-oriented heads yields the opposite trade-off: TPR drops to 84.3% (-5.9) and TNR rises to 49.1% (+10.9). Effects appear after ablating a single head, establishing opposing causal pressure from the two static camps. The same qualitative pattern holds for Qwen3.5-9B, Ministral-3-3B, and Gemma-4-12B-IT (Appendix D.2). Since suppressing either camp improves one class while harming the other, we next consider input-adaptive interventions.

4.2 Restoring Input Adaptivity to Static Voting

The ablations in Sec. 4.1 show that suppressing either static camp improves one class while harming the other, motivating interventions that make static voting input-adaptive. We guide these interventions with a shared correctness signal $f(x)$, obtained from a covariance-corrected linear discriminant fitted on a small balanced reference set and held fixed during inference (Appendix C). Here, $f(x) > 0$ predicts that solution x is correct, whereas $f(x) < 0$ predicts that it is incorrect. We investigate two interventions driven by this signal: *Dynamic Steering*, which directly shifts influential head outputs along the verdict direction, and *Voting Calibration*, which modulates the existing votes of the two static camps.

Dynamic Steering. A standard strategy for representation intervention is to add a task-relevant direction to an internal activation. Following this principle, we steer the output of a highly influential static head along the correct-verdict direction $v_\Delta = v_c - v_i$:

$$o_i'(x) = o_i(x) + \beta f(x) v_\Delta, \qquad \beta \geq 0, \tag{9}$$

where $o_i(x)$ denotes the original head output and $f(x)$ is an input-dependent signed correctness signal. Intuitively, the intervention injects verdict-relevant information directly into a head that already has substantial causal influence on the model decision: $f(x) > 0$ shifts the representation toward the correct-verdict direction, whereas $f(x) < 0$ reverses the shift.

Voting Calibration. Dynamic Steering changes a head by introducing an additional direction. Our voting model suggests a more targeted alternative: rather than overwrite the geometry of static heads, we can preserve their specialized directions and make only their *voting strengths* depend on the inferred correctness signal. Let $o_c(x)$ and $o_i(x)$ denote the outputs of the strongest acceptance- and rejection-oriented static heads, respectively. We calibrate them symmetrically as

$$o_c'(x) = \big(1 + \alpha f(x)\big) o_c(x), \qquad o_i'(x) = \big(1 - \alpha f(x)\big) o_i(x). \tag{10}$$

We expect to see that Voting Calibration can convert the static competition into an input-conditioned one, while retaining the directional specialization of each camp.

Table 2: **Effects of static-head interventions on μ-MATH.** Base, SocREval, and ProcessEval are evaluated on all 1,084 examples, whereas Dynamic Steering and Voting Calibration are evaluated on the 956-example test split after excluding the reference set.

Model	Method	μ-Math			μ-Math$_{\text{Qwen}}$	μ-Math$_{\text{Llama}}$	μ-Math$_{\text{GPT}}$	μ-Math$_{\text{Gemini}}$
		Acc	TPR	TNR	TPR/TNR	TPR/TNR	TPR/TNR	TPR/TNR
Ministral3-3B	Base	53.6	72.8	31.5	80.6/26.7	63.2/34.0	80.0/34.4	69.8/30.5
	SocREval	52.5	75.0	26.5	85.9/17.6	63.3/27.7	78.5/31.5	73.9/29.1
	ProcessEval	51.5	74.0	25.5	85.7/17.3	62.1/25.5	78.6/31.3	71.6/27.2
	Dynamic Steering (Ours)	58.6	70.1	45.2	82.6/29.4	42.8/53.8	66.7/42.1	75.9/54.5
	Voting Calibration (Ours)	**70.0**	69.0	**71.2**	**73.9/52.9**	**71.4/84.6**	**66.7/73.7**	65.5/**63.6**
Qwen3-8B	Base	66.1	90.2	38.2	89.0/33.6	90.0/46.3	92.9/35.1	88.9/37.8
	SocREval	65.6	89.5	38.0	89.7/27.4	77.7/56.0	95.4/28.6	91.0/27.1
	ProcessEval	66.2	89.3	39.5	89.5/29.4	76.8/57.0	95.3/29.8	91.0/29.2
	Dynamic Steering (Ours)	70.9	92.0	46.6	95.7/58.8	85.7/53.8	90.5/21.1	93.1/54.5
	Voting Calibration (Ours)	**74.1**	**92.0**	**53.4**	**95.7/52.9**	**85.7/57.7**	**90.5/42.1**	**93.1/63.6**
Qwen3.5-9B	Base	71.5	71.4	71.6	78.1/70.7	66.0/71.8	75.0/72.5	66.1/70.7
	SocREval	72.5	74.5	70.2	82.2/68.5	67.0/70.2	76.9/71.9	70.4/69.4
	ProcessEval	72.2	73.8	70.4	81.0/69.4	67.9/71.5	76.1/71.3	69.5/67.6
	Dynamic Steering (Ours)	70.5	77.0	63.0	78.3/82.4	78.6/61.5	90.5/52.6	65.6/54.5
	Voting Calibration (Ours)	**74.3**	**79.3**	68.5	**91.3**/70.6	57.1/**76.9**	76.2/**78.9**	**79.3/54.5**
Gemma-4-12B	Base	75.1	78.3	71.4	83.9/69.0	70.1/78.7	80.7/69.5	76.2/62.2
	SocREval	74.9	78.8	70.4	84.2/67.3	66.7/77.1	83.5/69.7	77.0/61.3
	ProcessEval	75.1	78.4	71.3	84.0/68.7	69.5/79.0	80.9/69.2	76.3/61.7
	Dynamic Steering (Ours)	75.5	79.3	71.2	87.0/82.4	78.6/73.0	85.7/68.4	69.0/54.5
	Voting Calibration (Ours)	**81.2**	**83.9**	**78.1**	**87.0/94.1**	57.1/**80.8**	**90.5/73.7**	**89.7/54.5**

With the judge's parameters frozen, we compare *Dynamic Steering* and *Voting Calibration* with two prompting baselines (Zheng et al., 2025; He et al., 2024) on μ-MATH benchmark.

Overall results. Voting Calibration achieves the highest overall μ-MATH accuracy among the evaluated methods on all four models, improving the base models by $+16.4$, $+8.0$, $+2.8$, and $+6.1$ percentage. In comparison, the prompting baselines change accuracy by -2.1 to $+1.0$ points, while Dynamic Steering yields changes ranging from -1.0 to $+5.0$ points. With both interventions using the same correctness signal, these results support calibrating the balance between existing static votes as more effective than directly steering representations along v_Δ in this setting.

Adaptive adjustment of static votes. Voting Calibration improves both TPR and TNR on Qwen3-8B and Gemma-4-12B, showing that its gains need not come from trading one class's accuracy for the other's. On Ministral3-3B, TNR rises from 31.5% to 71.2% while TPR falls from 72.8% to 69.0%; on Qwen3.5-9B, TPR rises from 71.4% to 79.3% while TNR falls from 71.6% to 68.5%. Generator-specific subsets show similarly heterogeneous class-wise trade-offs. Overall, correctness-conditioned reweighting makes static voting more input-adaptive, but its gains are not uniform across classes or subsets.

5 Conclusion

In this work, we investigate how LLM judges convert internal correctness information into verdicts through attention-head contributions. Our results suggest that binary judgments in LLMs are driven less by heads that reverse preference with correctness than by persistent acceptance- and rejection-oriented heads whose relative influence varies across inputs. Adaptivity therefore emerges from competition between directionally stable components. Its consistency across architectures and diverse input datasets, together with opposing ablation effects, indicates that static voting is an active part of the judging computation. These findings motivate rebalancing rather than suppressing either camp: removing one side shifts errors toward the other, whereas adapting their relative strengths to correctness evidence can improve judgment. Voting Calibration implements this principle using a small reference set, linking internal correctness representations to component-level decision formation and providing a mechanistic basis for improving LLM judges.

AI USE STATEMENT

To better present our work, we used AI assistants for grammatical checking of the manuscript. We also use LLM for code debug during our the experiments. We did not use AI for any other purpose. We take responsibility for the final content of this work.

ETHICS STATEMENT

This work investigates static voting mechanisms in large language models and develops interventions to improve their judgments of mathematical solutions. Our research uses existing pretrained models and mathematical reasoning benchmarks, including ProcessBench, Math-Shepherd, and μ-MATH. The experiments analyze model representations and predictions on existing benchmark examples; the proposed methods do not require collecting personal information or interacting with human participants. Our study focuses on mathematical solution assessment and does not establish the reliability of these interventions in sensitive applications.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, the main text and appendices describe the head contribution analysis, causal ablations, Dynamic Steering, and Voting Calibration. Appendix C provides pseudocode for the shared correctness signal, reference-set construction, and the selected hyperparameters, while Appendix E presents the complete prompt template and its variants. The overlap between reference and evaluation examples is explicitly documented. We report overall accuracy and class-conditional results across models and solution sources to support verification and comparison.

REFERENCES

Dyah Adila, Shuai Zhang, Boran Han, and Bernie Wang. Discovering Bias in Latent Space: An Unsupervised Debiasing Approach. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 246–261, 2024. URL https://proceedings.mlr.press/v235/adila24a.html.

Amos Azaria and Tom Mitchell. The Internal State of an LLM Knows When It's Lying, 2023. URL https://arxiv.org/abs/2304.13734.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering Latent Knowledge in Language Models Without Supervision, 2022. URL https://arxiv.org/abs/2212.03827.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. Humans or LLMs as the Judge? A Study on Judgement Bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8301–8327, 2024. doi: 10.18653/v1/2024.emnlp-main.474. URL https://aclanthology.org/2024.emnlp-main.474/.

Konstantin Chernyshev, Vitaliy Polshkov, Vlad Stepanov, Alex Myasnikov, Ekaterina Artemova, Alexei Miasnikov, and Sergei Tilga. U-MATH: A university-level benchmark for evaluating mathematical skills in large language models. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics*, pp. 974–1001, 2025. URL https://aclanthology.org/2025.gem-1.77/.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards Automated Circuit Discovery for Mechanistic Interpretability, 2023. URL https://arxiv.org/abs/2304.14997.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-Controlled AlpacaEval: A Simple Way to Debias Automatic Evaluators, 2024. URL https://arxiv.org/abs/2404.04475.

Gemma Team. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

Hangfeng He, Hongming Zhang, and Dan Roth. Socreval: Large language models with the socratic method for reference-free reasoning evaluation, 2024. URL https://arxiv.org/abs/2310.00074.

Suryaansh Jain, Umair Z. Ahmed, Shubham Sahai, and Ben Leong. Beyond Consensus: Mitigating the Agreeableness Bias in LLM Judge Evaluations, 2025. URL https://arxiv.org/abs/2510.11822.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. Benchmarking cognitive biases in large language models as evaluators. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 517–545, 2024. URL https://aclanthology.org/2024.findings-acl.29/.

Peng Lai, Jianjie Zheng, Sijie Cheng, Yun Chen, Peng Li, Yang Liu, and Guanhua Chen. Beyond the surface: Enhancing LLM-as-a-Judge alignment with human via internal representations. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://arxiv.org/abs/2508.03550.

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 *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 26874–26901, 2024. URL https://proceedings.mlr.press/v235/lee24t.html.

Michael A Lepori, Michael Curtis Mozer, and Asma Ghandeharioun. Racing thoughts: Explaining contextualization errors in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3020–3036, 2025.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-Time Intervention: Eliciting Truthful Answers from a Language Model, 2023. URL https://arxiv.org/abs/2306.03341.

Zhuochun Li, Yong Zhang, Ming Li, Yuelyu Ji, Yiming Zeng, Ning Cheng, Yun Zhu, Yanmeng Wang, Shaojun Wang, Jing Xiao, and Daqing He. Rethinking LLM-as-a-Judge: Representation-as-a-Judge with Small Language Models via Semantic Capacity Asymmetry, 2026. URL https://arxiv.org/abs/2601.22588.

Zongjie Li, Chaozheng Wang, Pingchuan Ma, Daoyuan Wu, Shuai Wang, Cuiyun Gao, and Yang Liu. Split and merge: Aligning position biases in LLM-based evaluators. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 11084–11108, 2024. URL https://aclanthology.org/2024.emnlp-main.621/.

Tom Lieberum, Matthew Rahtz, János Kramár, Neel Nanda, Geoffrey Irving, Rohin Shah, and Vladimir Mikulik. Does Circuit Analysis Interpretability Scale? Evidence from Multiple Choice Capabilities in Chinchilla, 2023. URL https://arxiv.org/abs/2307.09458.

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, et al. Ministral 3, 2026. URL https://arxiv.org/abs/2601.08584.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment, 2023. URL https://arxiv.org/abs/2303.16634.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of True/False datasets. In *Conference on Language Modeling*, 2024. URL https://arxiv.org/abs/2310.06824.

Jiwon Moon, Yerin Hwang, Dongryeol Lee, Taegwan Kang, Yongil Kim, and Kyomin Jung. Don't Judge Code by Its Cover: Exploring Biases in LLM Judges for Code Evaluation. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 1364–1389, 2026. doi: 10.18653/v1/2026.findings-eacl.70. URL https://aclanthology.org/2026.findings-eacl.70/.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022. URL https://arxiv.org/abs/2209.11895.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. Verbosity bias in preference labeling by large language models. *arXiv preprint arXiv:2310.10076*, 2023. URL https://arxiv.org/abs/2310.10076.

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. doi: 10.18653/v1/2025.ijcnlp-long.18. URL https://aclanthology.org/2025.ijcnlp-long.18/.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y. Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. JudgeBench: A benchmark for evaluating LLM-based judges. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=G0dksFayVq.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: A circuit for indirect object identification in gpt-2 small. In *International Conference on Learning Representations*, 2023.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9440–9450, 2024a. URL https://aclanthology.org/2024.acl-long.511/.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-Shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439, 2024b. doi: 10.18653/v1/2024.acl-long.510. URL https://aclanthology.org/2024.acl-long.510/.

Yiming Wang, Pei Zhang, Baosong Yang, Derek Wong, and Rui Wang. Latent space chain-of-embedding enables output-free llm self-evaluation. In *International Conference on Learning Representations*, volume 2025, pp. 70938–70970, 2025.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. Self-Preference Bias in LLM-as-a-Judge, 2024. URL https://arxiv.org/abs/2410.21819.

Zixiang Xu, Sixian Li, Huaxing Liu, Xiang Wang, Shuai Li, Zirui Song, and Xiuying Chen. Inside the unfair judge: A mechanistic interpretability account of LLM-as-judge bias. *arXiv preprint arXiv:2607.11871*, 2026. URL https://arxiv.org/abs/2607.11871.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V Chawla, and Xiangliang Zhang. Justice or Prejudice? Quantifying Biases in LLM-as-a-Judge, 2024. URL https://arxiv.org/abs/2410.02736.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. Evaluating Large Language Models at Evaluating Instruction Following, 2023. URL https://arxiv.org/abs/2310.07641.

Anqi Zhang, Yulin Chen, Jane Pan, Chen Zhao, Aurojit Panda, Jinyang Li, and He He. Reasoning Models Know When They're Right: Probing Hidden States for Self-Verification, 2025. URL https://arxiv.org/abs/2504.05419.

Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. ProcessBench: Identifying process errors in mathematical reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1009–1024, 2025. doi: 10.18653/v1/2025.acl-long.50. URL https://aclanthology.org/2025.acl-long.50/.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2306.05685.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023. URL https://arxiv.org/abs/2310.01405.

A Related Works

Explanatory Granularity. Research on LLM judges spans behavioral evaluation, representation analysis, and component-level mechanisms. Behavioral studies measure agreement with human judgments and sensitivity to evaluation protocols (Liu et al., 2023; Zheng et al., 2023). Representation-based work decodes truth and correctness signals from hidden states, including intermediate reasoning and evaluation representations, though decodability does not establish causal use in judgment (Burns et al., 2022; Azaria & Mitchell, 2023; Zhang et al., 2025; Li et al., 2026). Activation interventions provide complementary causal evidence by manipulating truth- or bias-related directions (Marks & Tegmark, 2024; Zou et al., 2023; Li et al., 2023; Adila et al., 2024), while recent judge-specific work identifies bias subspaces that bidirectionally control scores (Xu et al., 2026). Circuit analyses further localize behavior to attention heads and computational connections using attribution and causal interventions (Wang et al., 2023; Lieberum et al., 2023; Conmy et al., 2023). Building on these lines, we link correctness representations to direct logit contributions and causal ablations, identifying opposing judgment-head groups with approximately stable contribution directions but input-dependent magnitudes.

Source of Judge Bias. Prior work documents LLM-judge sensitivity to candidate position, response length, and familiarity with evaluated text (Zheng et al., 2023; Wang et al., 2024a; Shi et al., 2025; Dubois et al., 2024; Wataoka et al., 2024), as well as misinformation oversight, authority cues, and systematic input perturbations (Chen et al., 2024; Ye et al., 2024). Correctness benchmarks test whether judges distinguish valid or instruction-following responses from misleading alternatives, while code studies show that semantics-preserving surface changes can distort correctness judgments (Zeng et al., 2023; Tan et al., 2025; Moon et al., 2026). Work on agreeableness bias further identifies asymmetric acceptance of valid and invalid outputs and proposes ensemble- and regression-based corrections (Jain et al., 2025). We instead examine the mechanism underlying this class-conditional asymmetry, combining behavioral analysis, head-level attribution, and causal ablation to show how imbalanced contributions from opposing judgment channels bias the verdict.

B Theoretical and Quantitative Analysis of Static Heads

B.1 Within-Class Variance and Static-Voting Error

For brevity, write $S = \mathbf{S}_{\text{static}}(x)$, and define

$$\begin{aligned} \pi &:= \Pr(y = +1), \qquad \mu_{\pm} := \mathbb{E}[S \mid y = \pm 1], \\ \sigma_{\pm}^2 &:= \operatorname{Var}(S \mid y = \pm 1), \qquad \gamma := \frac{\mu_+ - \mu_-}{2}, \qquad c := \frac{\mu_+ + \mu_-}{2}. \end{aligned} \tag{11}$$

Assume $0 < \pi < 1$, finite conditional variances, and $\gamma > |c|$, so that $\mu_+ = \gamma + c > 0$ and $\mu_- = c - \gamma < 0$. The two conditional means therefore lie on the correct sides of zero. A class-conditional bound is

$$R(\mathbf{S}_{\text{static}}) \le \pi \frac{\sigma_+^2}{\sigma_+^2 + (\gamma + c)^2} + (1 - \pi) \frac{\sigma_-^2}{\sigma_-^2 + (\gamma - c)^2}. \tag{12}$$

The main-text parameters may be any d and v satisfying

$$0 < d \le \gamma - |c|, \qquad \max\{\sigma_+^2, \sigma_-^2\} \le v < \infty. \tag{13}$$

Since the label is constant within each class, the conditional variance of yS equals that of S.

Proof of Theorem 2.1. We first recall the one-sided variance inequality of Cantelli. If Z has mean zero and variance $w < \infty$, then for $t > 0$, setting $a = w/t$ gives

$$\Pr(Z \ge t) \le \frac{\mathbb{E}[(Z + a)^2]}{(t + a)^2} = \frac{w + a^2}{(t + a)^2} = \frac{w}{w + t^2}.$$

The same bound applies to $\Pr(Z \le -t)$ by replacing Z with $-Z$; it also holds when $w = 0$.

Under the specified tie-breaking rule, the error is exactly

$$R(\mathbf{S}_{\text{static}}) = \pi \Pr(S \le 0 \mid y = +1) + (1 - \pi) \Pr(S > 0 \mid y = -1).$$

Applying the inequality to each conditional distribution yields

$$\begin{aligned} \Pr(S \le 0 \mid y = +1) &\le \frac{\sigma_+^2}{\sigma_+^2 + \mu_+^2}, \\ \Pr(S > 0 \mid y = -1) \le \Pr(S \ge 0 \mid y = -1) &\le \frac{\sigma_-^2}{\sigma_-^2 + \mu_-^2}. \end{aligned}$$

Substituting the expressions for μ_+ and μ_- proves Eq. 12. Each term is at most $v/(v + d^2)$ under Eq. 13, so their weighted average satisfies the main-text bound. □

These are population-risk bounds for a fixed score. Substituting sample estimates of their parameters does not, by itself, yield a finite-sample confidence guarantee. Head dependence is retained in the conditional variance of the aggregate score; no independence assumption is used.

B.2 Relating Static Voting to the Full Model

Let $\widehat{y}_g(x) = +1$ if $g(x) > 0$ and $\widehat{y}_g(x) = -1$ otherwise, so that $R(g) = \Pr(\widehat{y}_g(x) \ne y)$. By Eq. 3, the remaining contribution is

$$\begin{aligned} e(x) &:= \mathbf{m}(x) - \mathbf{S}_{\text{static}}(x) \\ &= \sum_{h \in \mathcal{H} \setminus \mathcal{H}^{\text{st}}} \mathbf{s}_h(x) + \mathbf{m}_{\text{other}}(x). \end{aligned} \tag{14}$$

This includes all nonstatic heads, any unclassified heads, and the remaining residual-stream contributions.

Proof of Theorem 2.2. For each input and label, the two error indicators can differ only if the predictions differ. Hence

$$\left|R(\mathbf{m}) - R(\mathbf{S}_{\text{static}})\right| \le \Pr(\widehat{y}_{\mathbf{m}}(x) \neq \widehat{y}_{\mathbf{S}_{\text{static}}}(x)) .$$

If $|e(x)| < |\mathbf{S}_{\text{static}}(x)|$, then $\mathbf{m}(x) = \mathbf{S}_{\text{static}}(x) + e(x)$ has the same nonzero sign as $\mathbf{S}_{\text{static}}(x)$, so the predictions agree. Therefore, disagreement is contained in the event $|e(x)| \ge |\mathbf{S}_{\text{static}}(x)|$. This inclusion also covers zero scores and proves the theorem. □

In particular, under the assumptions of Theorem 2.1,

$$R(\mathbf{m}) \le \frac{v}{v + d^2} + \Pr(|e(x)| \ge |\mathbf{S}_{\text{static}}(x)|) . \tag{15}$$

The second term need not be small merely because static heads are numerous or have large individual contributions: opposing votes may cancel. The result compares scores on the same forward pass and makes no claim about the effects of ablating or modifying heads.

B.3 Head-Type Proportions and Static Contribution Metrics

For each model, let $\mathcal{H}$ denote the set of analyzed heads, with $H = |\mathcal{H}|$, and let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where $y_i \in \{+1, -1\}$ indicates whether the candidate answer is correct or incorrect. Let $s_h(x_i)$ denote the signed contribution of head h to the correct-minus-incorrect judgment margin under the decomposition in Section 3.2. Positive scores favor the verdict "correct," irrespective of the sample label.

For each class $c \in \{+1, -1\}$, define the raw class-conditional mean

$$\bar{s}_{h,c} = \frac{1}{N_c} \sum_{i:y_i=c} s_h(x_i), \qquad N_c = \sum_{i=1}^{N} \mathbf{1}\{y_i = c\}.$$

Under the label-aligned mean convention, $\mu_{h,+} = \bar{s}_{h,+1}$ and $\mu_{h,-} = -\bar{s}_{h,-1}$. The classification below uses the raw score orientation.

We estimate a pointwise 95% percentile bootstrap interval $[L_{h,c}, U_{h,c}]$ for each mean using 2,000 resamples. Resampling is performed at the question level, retaining all associated candidate answers within each sampled question. Define the resolved sign

$$\sigma_{h,c} = \begin{cases} +1, & L_{h,c} > 0, \\ -1, & U_{h,c} < 0, \\ 0, & \text{otherwise.} \end{cases}$$

Each head is assigned to one of five mutually exclusive categories:

$$g(h) = \begin{cases} \text{st+}, & (\sigma_{h,+1}, \sigma_{h,-1}) = (+1, +1), \\ \text{st}-, & (\sigma_{h,+1}, \sigma_{h,-1}) = (-1, -1), \\ \text{dyn+}, & (\sigma_{h,+1}, \sigma_{h,-1}) = (+1, -1), \\ \text{dyn}-, & (\sigma_{h,+1}, \sigma_{h,-1}) = (-1, +1), \\ \text{uncertain}, & \sigma_{h,+1}\sigma_{h,-1} = 0. \end{cases}$$

Static heads favor the same verdict in both class-conditional means, whereas dynamic heads reverse their mean preference across classes. A head is uncertain when either interval includes zero. This classification concerns mean preferences; individual contributions may still vary across inputs. The intervals are pointwise and do not provide a simultaneous coverage guarantee across heads.

For each of these five categories k, define

$$\mathcal{H}_k = \{h \in \mathcal{H} : g(h) = k\}, \qquad p_k = \frac{|\mathcal{H}_k|}{H}, \qquad \sum_k p_k = 1.$$

These fractions give $p_{\text{st+}}$, $p_{\text{st}-}$, $p_{\text{dyn+}}$, $p_{\text{dyn}-}$, and $p_{\text{uncertain}}$.

To quantify contribution magnitude, let $\mathcal{H}_{\mathrm{st}} = \mathcal{H}_{\mathrm{st}+} \cup \mathcal{H}_{\mathrm{st}-}$ and define the class-balanced mean absolute contribution of each head:

$$w_h = \frac{1}{2} \sum_{c \in \{+1,-1\}} \frac{1}{N_c} \sum_{i:y_i=c} |s_h(x_i)|.$$

The static contribution share is

$$C_{\mathrm{st}} = \frac{\sum_{h \in \mathcal{H}_{\mathrm{st}}} w_h}{\sum_{h \in \mathcal{H}} w_h},$$

provided that the denominator is positive. The denominator includes all analyzed heads, including uncertain heads. Absolute values are taken before averaging and aggregation: opposing contributions both increase the measured contribution mass even when their signed scores cancel.

To assess dominance after cancellation, define

$$S(x) = \sum_{h \in \mathcal{H}_{\mathrm{st}}} s_h(x),$$
$$R_{\mathrm{attn}}(x) = \sum_{h \in \mathcal{H} \setminus \mathcal{H}_{\mathrm{st}}} s_h(x).$$

Here, R_{attn} includes both dynamic groups and the uncertain group. Their sum, $S(x) + R_{\mathrm{attn}}(x)$, is the total contribution of the analyzed heads. The final table column reports

$$\widehat{\Pr}(|S(x)| > |R_{\mathrm{attn}}(x)|) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{|S(x_i)| > |R_{\mathrm{attn}}(x_i)|\}.$$

This frequency weights samples equally under the observed label distribution; ties do not satisfy the strict inequality. Its interpretation follows from

$$|S(x)| > |R_{\mathrm{attn}}(x)| \implies \mathrm{sgn}(S(x) + R_{\mathrm{attn}}(x)) = \mathrm{sgn}(S(x)).$$

Thus, on samples satisfying this condition, the static component determines the direction of the aggregate attention contribution even after signed cancellation. The condition is sufficient; matching directions can also occur outside this event. These statistics describe the measured attention decomposition, with non-head components of the final judgment margin outside their scope. All seven metrics are reported as percentages.

C Implement details of interventions

Procedure of $f(x)$ calculation. We randomly sample a balanced reference set $\mathcal{C}$ of $N = 128$ solutions from the 1,084-example μ-MATH evaluation set, comprising 64 correct and 64 incorrect solutions. We evaluate the base model and the two prompting baselines on all 1,084 examples, while Dynamic Steering and Voting Calibration are evaluated on the remaining 956 examples after excluding the 128-example reference set. We assign labels $y = +1$ and $y = -1$ to correct and incorrect solutions, respectively.

Let h_c and h_i denote the highest-contributing acceptance- and rejection-oriented static heads, with outputs $o_c(x)$ and $o_i(x)$. These heads are selected separately for each model; for Qwen3-8B, they are L33H16 and L33H18. Both interventions compute $f(x)$ from the unmodified output of h_i. Specifically, we extract $r(x) = o_i(x) \in \mathbb{R}^d$ at the last valid token of the judge prompt, after the output projection W_O. Thus, $r(x)$ represents the head's contribution to the residual stream, and d is the residual-stream dimension.

Algorithm 1 gives the complete procedure. We estimate the class centroids and pooled within-class covariance on $\mathcal{C}$, apply covariance shrinkage, and obtain the discriminant direction by solving a linear system. The resulting statistics remain fixed during evaluation.

Algorithm 1 Procedure for computing the shared correctness signal $f(x)$

Require: Balanced reference set $\mathcal{C} = \{(x_j, y_j)\}_{j=1}^{128}$, with $N_+ = N_- = 64$; selected rejection-oriented head h_i; shrinkage coefficient $\lambda = 0.1$

Ensure: Fixed statistics (m, w) and an input-dependent signal $f(x)$

Reference-set estimation: once per model

$N \leftarrow 128$

for $j = 1, \ldots, N$ **do**

  Extract the unmodified head output $r_j \leftarrow o_i(x_j) \in \mathbb{R}^d$

end for

$$\mu_+ \leftarrow \frac{1}{N_+} \sum_{j:y_j=+1} r_j, \qquad \mu_- \leftarrow \frac{1}{N_-} \sum_{j:y_j=-1} r_j$$

$$m \leftarrow \frac{\mu_+ + \mu_-}{2}$$

$$\Sigma_w \leftarrow \frac{1}{N-2} \sum_{j=1}^{N} (r_j - \mu_{y_j})(r_j - \mu_{y_j})^\top$$

$$\tau \leftarrow \frac{\mathrm{tr}(\Sigma_w)}{d}$$

$\Sigma_\lambda \leftarrow (1-\lambda)\Sigma_w + \lambda\tau I_d$

Solve $\Sigma_\lambda w = \mu_+ - \mu_-$ for w

Fix (m, w) for subsequent inference

Signal computation: for each input x

Obtain the pre-intervention output $r(x) \leftarrow o_i(x)$

$q(x) \leftarrow w^\top (r(x) - m)$

return $f(x) \leftarrow \tanh\big(q(x)/2\big)$

Positive values of $f(x)$ predict correct solutions, whereas negative values predict incorrect ones; its magnitude controls the intervention strength. Both methods use the same fitted statistics and signal computation. Once (m, w) are fixed, computing $f(x)$ requires only the head representation, without accessing the input's label. All judge parameters remain frozen throughout.

Hyperparameter selection and intervention application. We select the intervention settings through a simple grid search, using $\lambda = 0.1$ for covariance shrinkage, $\alpha = 0.5$ for Voting Calibration, and $\beta = 10$ for Dynamic Steering in the reported experiments.

Voting Calibration rescales the two selected head outputs as

$$o_c'(x) = \big(1 + 0.5 f(x)\big) o_c(x), \qquad o_i'(x) = \big(1 - 0.5 f(x)\big) o_i(x). \tag{16}$$

Since $|f(x)| \le 1$, both scaling factors remain within $[0.5, 1.5]$. Calibration therefore adjusts contribution strength while preserving each output vector's direction.

Dynamic Steering intervenes on the single rejection-oriented head h_i, which also supplies the representation used to compute $f(x)$:

$$o_i'(x) = o_i(x) + 10 f(x) v_\Delta. \tag{17}$$

Here, $v_\Delta = v_c - v_i$, where v_c and v_i are the unembedding readout vectors for the correct and incorrect verdict tokens, respectively. Both interventions operate at the final valid token of the judge prompt, with $f(x)$ computed from the head output before intervention.

D Supplement Experiments results

D.1 Heatmap replications

In this section we provide the replication heatmap results for Qwen3.5-9B, Ministral3-3B and Gemma4-12B-IT, shown in Fig 7- 8

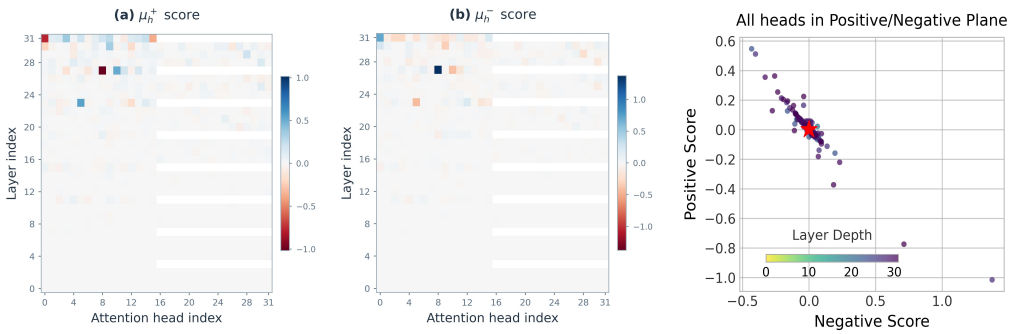

Figure 7: Heatmaps of $\mu_\pm$ scores for each head in Qwen3.5-9B.

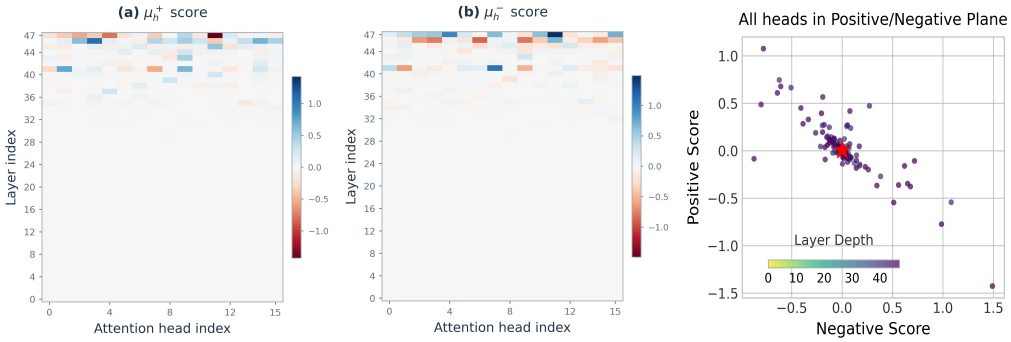

Figure 8: Heatmaps of $\mu_\pm$ scores for each head in Gemma4-12B-IT.

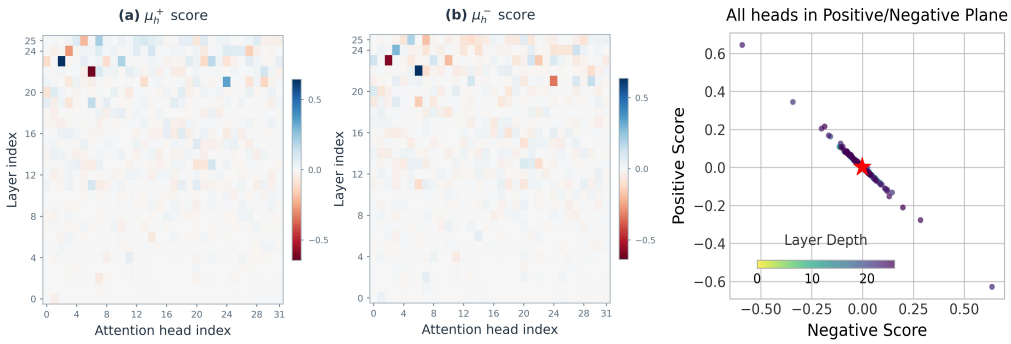

Figure 9: Heatmaps of $\mu_\pm$ scores for each head in Ministral3-3B.

D.2 Ablation replications

Following the settings in Sec 4.1, we zero the top-k heads from either camp on μ-MATH, with $k \in 1, 2, 4, 8$, and compare against the unmodified model ($k = 0$) using TPR and TNR. The results across different architecures (Qwen3.5-9B, Ministral3-3B and Gemma4-12B-IT) are as follows

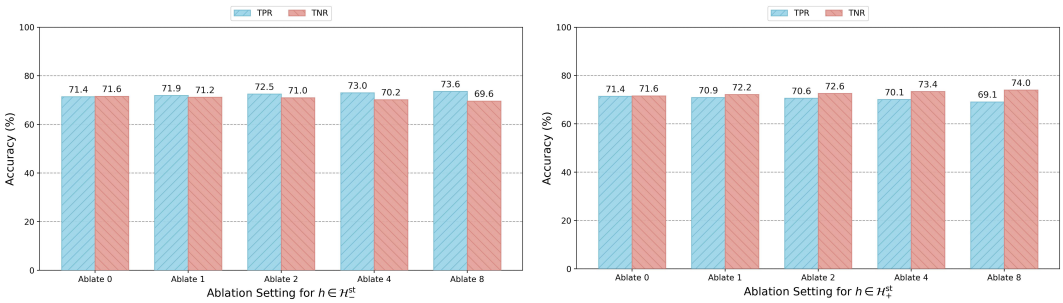

Figure 10: **Static-head ablation changes Qwen3.5-9B verdicts on μ-MATH.** Left: zeroing the top-k rejection-oriented heads in $\mathcal{H}_-^{st}$. Right: zeroing the top-k acceptance-oriented heads in $\mathcal{H}_+^{st}$. The unmodified model corresponds to $k = 0$. TPR measures correct-solution acceptance; TNR measures incorrect-solution rejection.

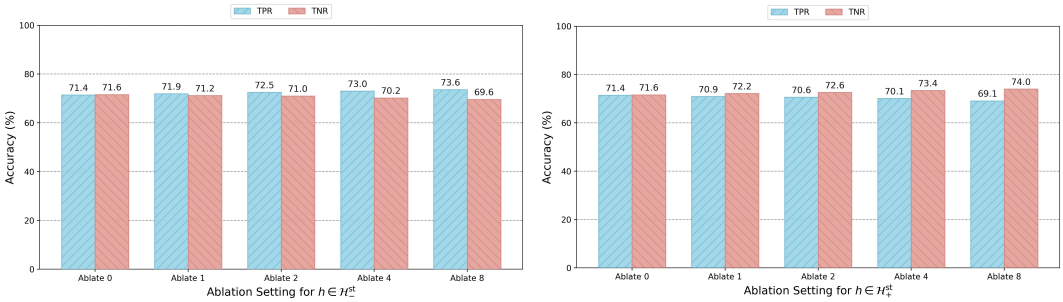

Figure 11: **Static-head ablation changes Qwen3.5-9B verdicts on μ-MATH.** Left: zeroing the top-k rejection-oriented heads in $\mathcal{H}_-^{st}$. Right: zeroing the top-k acceptance-oriented heads in $\mathcal{H}_+^{st}$. The unmodified model corresponds to $k = 0$. TPR measures correct-solution acceptance; TNR measures incorrect-solution rejection.

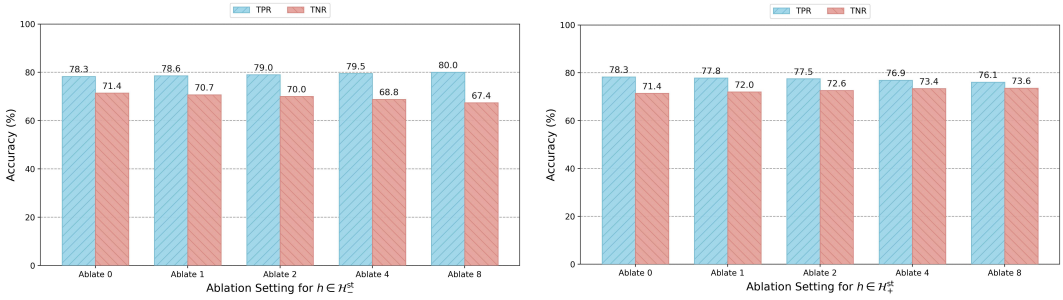

Figure 12: **Static-head ablation changes Gemma4-12B-IT verdicts on μ-MATH.** Left: zeroing the top-k rejection-oriented heads in $\mathcal{H}_-^{st}$. Right: zeroing the top-k acceptance-oriented heads in $\mathcal{H}_+^{st}$. The unmodified model corresponds to $k = 0$. TPR measures correct-solution acceptance; TNR measures incorrect-solution rejection.

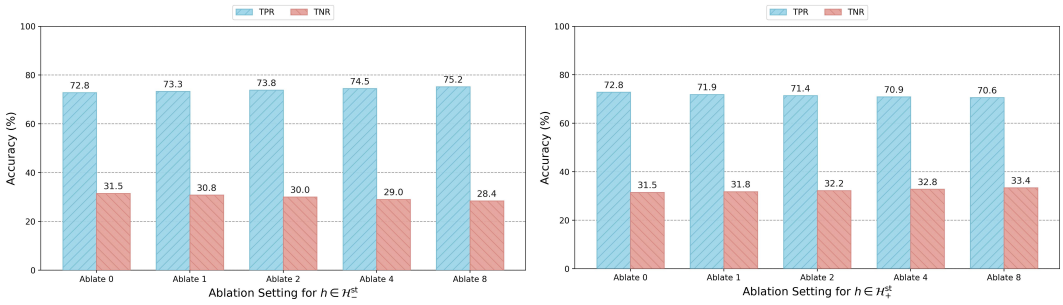

Figure 13: **Static-head ablation changes Ministral3-3B verdicts on μ-MATH.** Left: zeroing the top-k rejection-oriented heads in $\mathcal{H}_-^{\mathrm{st}}$. Right: zeroing the top-k acceptance-oriented heads in $\mathcal{H}_+^{\mathrm{st}}$. The unmodified model corresponds to $k = 0$. TPR measures correct-solution acceptance; TNR measures incorrect-solution rejection.

E Prompts and Prompt Sensitivity

E.1 Prompt Used in the Main Experiments

We format the judging instruction using each model's corresponding chat template. Below, we show the complete serialized input for Qwen3, where `{question}` and `{answer}` are replaced with the question and candidate solution, respectively.

```
<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
For a question '{question}', and a candidate answer '{answer}', your job
    is to judge the candidate answer is correct or incorrect.<|im_end|>
<|im_start|>assistant
<think>

</think>
The answer is
```

The serialized input includes an empty `<think>` block and ends immediately after `The answer is`. The positive and negative verdict strings are `correct` and `incorrect`, respectively, including their leading spaces.

E.2 Prompt Variations and Correlation Analysis

To examine the sensitivity of head contributions to prompt wording, we compare the main prompt, denoted `origin`, with three variants. Each variant modifies either the system-role description or the verdict vocabulary while retaining the remaining prompt structure: 1) **Mathematics-expert role (`math-expert`).** We replace the original system message with: "you are a mathematics expert."; 2) **Proof-reviewer role (`judger`).** We replace the original system message with: "You are an expert reviewer of mathematical proof solutions."; 3)**Alternative verdict labels (`true/false`).** We retain the original system message and replace " correct" with " true" and " incorrect" with " false". The resulting user instruction is: "For a question 'question', and a candidate answer 'answer', your job is to judge the candidate answer is true or false."

For each prompt setting, we recompute the class-conditional contribution μ_h^- for every attention head h and compare the resulting contribution profiles across settings. Fig. 14 reports the pairwise correlation matrix. The two system-role variants remain highly correlated with the main prompt: the correlation is 1.00 for `math-expert`, at the displayed precision, and 0.97 for `judger`. The two role variants also correlate at 0.97. Changing the verdict vocabulary produces a larger difference: `true/false` correlates with `origin`, `math-expert`, and `judger` at 0.63, 0.61, and 0.58, re-

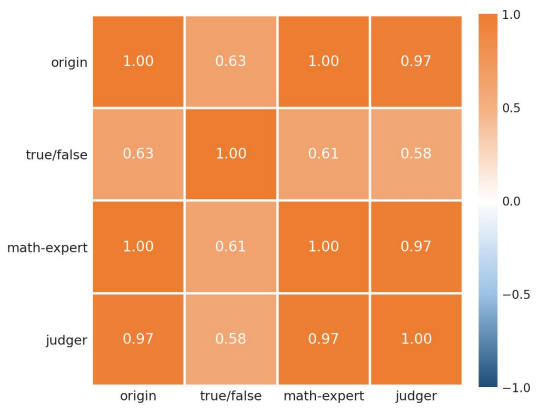

Figure 14: **Head contribution correlations across prompt settings.** Each entry reports the correlation between contribution profiles under two prompt settings. `origin` denotes the main experimental prompt; `true/false` replaces the verdict labels; `math-expert` and `judger` replace the system message with a mathematics-expert and a proof-reviewer description, respectively. Values are displayed to two decimal places.

spectively. These results indicate that the measured head contribution patterns are highly consistent across the tested role descriptions, but more sensitive to the choice of verdict labels.