OPENING LLM JUDGES: RECOVERING PREFERENCE SIGNALS BEYOND THE FINAL VERDICT

Anonymous authors
Paper under double-blind review

ABSTRACT

LLM judges are widely used to evaluate model outputs, but their verdicts can be unreliable: a judge may favor the worse answer because of its position, length, or other surface features. When a judge gives the wrong verdict, is the information needed to make the right judgment absent from the model, or is it still present in its internal representations but not reflected in the final output? We study this question across **64** open-weight evaluators and **14** datasets, including causal interventions on **41** judges (editing a model's internal activations while it runs to see whether its verdict changes). On LLMBar, a benchmark built so that the superficially better answer is the worse one, the verdicts of 50 judges agree with human labels only 0.456 of the time, even after position bias is cancelled by scoring both answer orders. Yet a small probe trained on the same judges' internal activations, without changing the judges, reaches 0.846, and still 0.686 after the influence of surface features such as length and position is removed (0.507 with shuffled labels). This gap holds across eight benchmarks and across model families, but it is not universal. A simple score of how well surface features alone predict the human label, computed before any probe is trained, is strongly correlated with the size of the gain (Spearman $\rho = 0.90$). On two further tasks where a judge scores one answer at a time against a rubric, leaving no surface cue to exploit, reading the internals gives no advantage. The interventions also show that editing activations in the middle of the network already changes the verdict, before it can be read off directly, and locate the pathways that carry position and length bias. With the same human labels, the recovered signal also lets a judge flag cases where it is likely to be wrong and yields better labels for preference learning. A wrong verdict, then, does not mean the judge lacks the information, and a simple diagnostic shows when it is worth recovering. [1]

1 INTRODUCTION

LLM judges are increasingly used to evaluate model outputs, compare systems, rank models, and generate preference data for training (Zheng et al., 2023; Liu et al., 2023; Lambert et al., 2025); in each case the judge exposes one small output, a preferred answer or a score, which we treat as its decision.

That output can be unreliable. Judges respond to properties that should not determine quality, including answer position (Wang et al., 2024b), response length (Dubois et al., 2024; Saito et al., 2023), and style (Panickssery et al., 2024; Sharma et al., 2024). On benchmarks built so that the longer or more polished answer is the worse one, a judge can score *below chance* (Zeng et al., 2024). But a model does not decide in a single token, and the verdict is the endpoint of many layers of computation, which raises a simple question:

> *When an LLM judge gives the wrong verdict, is the information needed to make the right judgment absent from the model, or is it still present in its internal representations but not reflected in the final output?*

[1] Code, configurations, prompts, and all result artifacts: `https://anonymous.4open.science/r/llm-judge-internals-B775`

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.

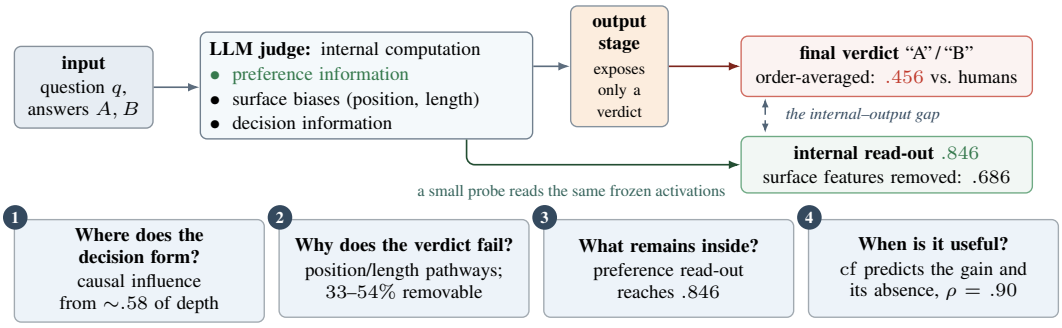

Figure 1: A judge's rich internal computation reaches the user through an output stage that exposes only a verdict. On adversarial LLMBar the verdict agrees with humans below chance (.456, mean over 50 judges), while a small probe reading the same frozen activations reaches .846. The four questions below organize the paper.

We call the difference between what a judge represents internally and what its verdict exposes the **internal–output gap**, and it can be large. On LLMBar (Zeng et al., 2024), one such benchmark, the verdicts of **50** open-weight judges agree with human labels only 0.456 of the time, below chance, even after position bias is cancelled by scoring both answer orders. A small classifier, a *probe*, trained on the same judges' frozen internal activations reaches 0.846 without changing the judges themselves. One might suspect it has simply learned that the shorter answer wins on this benchmark, so we remove the influence of length and eight other surface features before training it; it still reaches 0.686, where shuffled labels give 0.507. The same pattern holds on eight pairwise-preference benchmarks and is not driven by a few models or by a single model family (§4).

Where does the gap come from? We intervene directly on the activations of **41** judges during a forward pass. Editing activations in the middle of the network already changes the final verdict, earlier than the layer at which the verdict can be read off directly. Position and length biases run through internal pathways that we can locate and switch off, and a common attribution method points to different components than the ones whose removal actually changes the verdict (§5). The final verdict is therefore an incomplete picture of the computation behind it.

The gap is not universal, so we ask when reading internals is worth it. We measure how well one surface feature, response length, predicts the human label on a benchmark, and turn the answer into a single number, the **surface-confound score** cf. Across the eight benchmarks the gain from reading internals rises with cf (Spearman $\rho = 0.90$); this holds after accounting for judge quality and for judges that share a model family, and cf from one run predicts the gain on a second, independent run. The reverse case holds too: on two held-out benchmarks where the judge scores one answer at a time against a rubric, so that position and length cannot tip a comparison, reading internals gives no advantage (§6). The practical rule is simple: when cf says the output is likely to be unreliable, read the judge's frozen activations instead of retraining it.

Prior work has shown that relevant information can be read from an evaluator's internal representations (Maiya et al., 2025; Lai et al., 2025; Girrbach et al., 2025; Li et al., 2026) and that the internal components a judge uses to score can be found with causal interventions (Feldhus et al., 2026). Our work asks: when the emitted verdict is wrong, is the human-preference signal still recoverable, which internal pathways keep the output from exposing it, and can we predict when recovering it will help? We make three contributions.

1. **A robust internal–output gap.** Across 50 judges and eight benchmarks, preference information recovered from frozen activations consistently improves on the emitted verdict, including after surface controls and clustering by model family (§4).
2. **A causal view of evaluator decisions.** Across 41 judges, interventions on the forward pass locate the computation that sets the verdict and show that surface biases run through internal pathways that attribution methods do not reliably identify (§5).

3. **A conditional and practical read-out.** A surface-confound diagnostic indicates both where internal read-out helps and where it does not, and with the same number of human labels the recovered signal improves evaluation and downstream preference learning (§6–§7).

Three claims must be kept apart: that preference information can be *decoded* from internal representations, that internal components *causally influence* the verdict, and that the judge actually *used* that information when it decided.

2 Related Work

Evaluator reliability. LLM judges (Zheng et al., 2023; Liu et al., 2023) and reward models (Christiano et al., 2017; Ouyang et al., 2022) are studied behaviorally (Lambert et al., 2025; Tan et al., 2025; Gureja et al., 2025) and repaired at the *output*: preference optimization (Rafailov et al., 2023), multi-objective heads (Wang et al., 2024a), swap-averaging, self-consistency (Wang et al., 2023b), calibration (Guo et al., 2017). Position (Wang et al., 2024b), length (Dubois et al., 2024; Saito et al., 2023) and self-preference (Panickssery et al., 2024) biases are documented, LLMBar (Zeng et al., 2024) shows below-chance verdicts when surface form is decoupled from quality, and judges that agree closely with one another can still fall well short of human agreement (Mukherjee et al., 2026b). This line of work fixes the judge's output; it does not ask what the judge computed internally.

Internal representations for evaluation. Recent work looks into the internal representations of the evaluator. Maiya et al. (2025) show that supervised and unsupervised linear probes over LLM activations extract preferences more accurately than generation-based judgment across four model families and six datasets, and can match fine-tuned evaluators at equal data. Lai et al. (2025) (LAGER) aggregate score-token logits across layers, backbone frozen, to improve *point-wise* judge–human correlation. Girrbach et al. (2025) derive latent ratings from probability-weighted scores, verifier probabilities, and probes at the rating position, while Li et al. (2026) frame evaluation directly as representation probing in small models. Two open questions that our work addresses are *why* the output loses the signal, and *when* recovering it is worthwhile.

Mechanistic interpretability. To answer the first of those we use lenses (nostalgebraist, 2020; Belrose et al., 2023), activation and path patching (Meng et al., 2022; Goldowsky-Dill et al., 2023), causal mediation (Vig et al., 2020), causal scrubbing (Chan et al., 2022), probing (Alain & Bengio, 2017; Belinkov, 2022), and sparse features (Bricken et al., 2023; Huben et al., 2024; Lieberum et al., 2024), inheriting the warning that attribution surrogates diverge from causal effects (Zhang & Nanda, 2024). Concurrently, Feldhus et al. (2026) apply edge attribution patching to judges and find a sparse "latent evaluator" subgraph in mid-to-late MLPs feeding fragile format-specific output branches. The accounts are complementary: theirs explains why one shared internal evaluator yields different scores under different output *formats*, ours asks whether the human-labeled *preference* survives the output stage at all, and when it can be recovered.

Table 1 places our work against that literature along the axes taken up in turn by §4 (read-out under surface controls), §5 (causal localization), §6 (the diagnostic) and §7 (downstream use).

3 Setup

A generative judge J maps (q, a_A, a_B) to a verdict distribution; a reward model maps (q, a) to $r \in \mathbb{R}$. Every dataset supplies a *human-labeled* preference y and "accuracy" means agreement with y; we treat y as a noisy, annotator-derived target and use inter-annotator agreement only to estimate its noise. We read a judge two ways. The **verdict read-out**, what a user consumes, is the final-layer logits over the verdict vocabulary. The **internal read-out** applies, at each layer ℓ, the model's own final LayerNorm and unembedding (logit lens) or a learned affine (tuned lens) to the residual stream h_ℓ at the decision position, yielding a per-layer estimate $\hat{v}_\ell$ and a vector for probing; *lens decoding depth* is the fractional layer at which $\hat{v}_\ell$ first matches the final verdict.

Because correlation is insufficient, we also intervene inside a live forward pass. **Steering** adds $\alpha\hat{u}$, **ablation** projects out a subspace, and **patching** copies activations from a counterfactual run, and each time we measure the change in the *emitted* verdict. This gives the second depth quantity we

Work	Internal read-out	Causal localization	Surface controls	Predicts when useful	Downstream use	Judge format	Scale (models)
Maiya et al. (2025)	✓	–	pos. only	–	–	both	4 families
Lai et al. (2025)	✓	–	–	–	data sel.	point-wise	6
Girrbach et al. (2025)	✓	–	–	–	BoN, routing	point-wise	5
Li et al. (2026)	✓	–	–	–	data sel.	point-wise	4
Feldhus et al. (2026)	(✓)	✓	–	–	–	point-wise	3 families
Our work	✓	✓	✓	✓	**DPO labels, abstention**	both	**64**

Table 1: Positioning of our work against the four closest lines of prior work and one concurrent one. "Causal localization": forward-pass interventions that change the emitted output; "surface controls": the advantage is tested after removing length and position cues; "predicts when useful": a diagnostic computable *before* the read-out is fitted, validated on an independent run, and including a negative case; "downstream use": empirically demonstrated applications (BoN: best-of-N); only our work uses the recovered preferences as training labels (DPO) and to abstain on likely errors. Judge format: pairwise compares two answers in one prompt, point-wise scores one at a time. A dash marks an axis outside that paper's scope; (✓) marks an appendix-level result. Extended discussion: App. A.11.

use, **causal decision depth**: the earliest normalized layer at which a residual patch reliably flips that verdict. It marks where our intervention first finds causally sufficient decision information; the decision itself may still form gradually over several layers (App. A.1).

These two kinds of evidence answer different questions, and we keep three claims separate throughout. **L1 decodability** asks whether y is predictable from h_ℓ and is answered by probes (§4). **L2 causal influence** asks whether intervening on h_ℓ changes the emitted verdict and is answered by patching, ablation and steering (§5). **L3 causal use** asks whether J used that direction to form *its* verdict. We establish L1 and L2 but not L3, and no claim here requires L3, since §7 shows that acting on an L1 signal improves outcomes whether or not the judge used it. Consistently, per-judge causal decision depth (L2) does not predict per-judge probe gain (L1): $\rho = 0.06$, $p = 0.73$, $n = 39$.

Against that background we compare five readers of the same frozen judge, each given the *same number of human labels*: (a) the raw verdict, (b) the order-averaged verdict, (c) a cross-validated probe on the internal read-out, (d) LoRA and full-parameter fine-tunes, and (e) output-level fixes (self-consistency, calibrated prompting). "Label budget" counts only human-labeled items; the arms differ widely in compute and storage, accounted separately in App. A.8.

The registry holds **64** open-weight models spanning 0.36B–405B and 21 families: **50** generative judges forming the flagship adversarial panel, **9** scalar reward models, and **5** read-out-only large judges (Table 7). The causal suite intervenes on the **41** of the 50 judges exposing hookable decoder blocks within our memory budget, plus **5** reward models. The **14 datasets**, all previously published, divide into **8** public pairwise-preference benchmarks carrying the ladder of §4 and **6** auxiliary sets supporting the mechanism and rubric analyses (Tables 9, 10). Cached surrogates (lenses, direct logit attribution, patching, causal scrubbing, SAEs, cross-validated L_2 probes) are validated against the forward-pass suite. Controls include bootstrap CIs at both the judge and the **model-family** level, paired Wilcoxon under FDR-BH, shuffled-label and surface residualization, seed replication, and random-component null checks. Full inventory: Table 8; protocol: App. A.12.

4 The Internal–Output Gap

We begin where the output stage fails most visibly. On **LLMBar-adversarial**, where the surface-preferable answer is deliberately the worse one, the verdicts of 50 judges, each scored on both answer orders and averaged, reach **0.456** accuracy against the human label: below chance, below the raw verdict's 0.483 (order-averaging strips a position bias that was partly masking the length bias, so it *lowers* accuracy here), and far below a prefer-shorter heuristic (0.799) we confront in §6. 72% of judges are below chance although the same judges reach a median 0.89 on the easy split, and averaging across all layers does no better (0.461). Yet a within-adversarial cross-validated **probe on the mid-layer read-out reaches** 0.758, positive for 100% of judges, and reading the

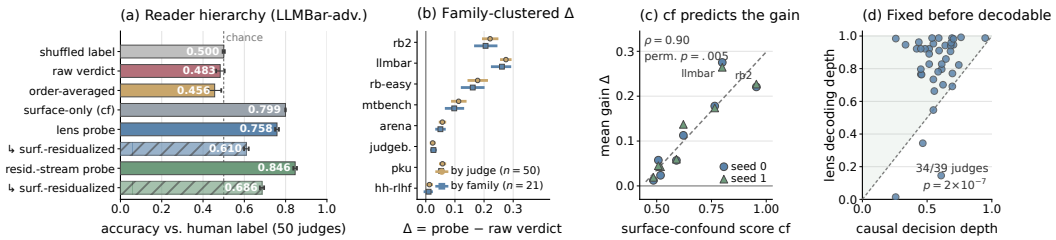

Figure 2: The internal–output gap and the statistics that have to hold for it. **(a)** Reader hierarchy on LLMBar-adversarial (50 judges, bootstrap 95% CIs): the raw verdict is below chance, the surface-only baseline is high *by construction* here (§6), and the internal read-outs survive surface residualization (hatched) well above the shuffled-label control. **(b)** Per-benchmark advantage bootstrapped over judges and over model families; clustering widens the intervals but leaves $7/8$ excluding zero, the exception being hh-rlhf, the lowest-cf benchmark, where §6 predicts the effect should vanish. **(c)** The diagnostic orders the gain; triangles are an independent replication seed. **(d)** Per judge, a residual patch alters the emitted verdict *before* a lens can decode it; causal decision depth is the onset of intervention sensitivity defined in §3.

residual stream reaches **0.846** (Fig. 2a). Seed 1 reproduces each of these within ± 0.012 (cell-level $r = 0.97$ over all 400 cells).

The obvious objection is that the probe has simply learned answer length, which predicts the label on this set. Three controls separate content from surface. **(i) Residualization:** regressing out a nine-dimensional surface feature vector (length, position, formatting and lexical markers) before fitting leaves the lens probe at 0.610 and the residual-stream probe at **0.686**, against shuffled-label controls of 0.500/0.507, and the residualized read-out beats its shuffled control for $50/50$ judges ($p < 10^{-15}$). **(ii) Length balancing:** on a length-balanced subset the probe scores 0.684 where the order-averaged verdict scores 0.482 ($p = 4 \times 10^{-10}$). **(iii) Capacity:** the residual-stream probe, which strictly contains the lens probe's features, beats it on $7/8$ benchmarks, so the lens number understates what the activations contain. Human-label-predictive signal therefore survives the *measured* surface controls. We call what remains the *surface-controlled preference signal*. It establishes decodability (L1) only, and it controls for the surface features we measured.

The pattern holds well beyond one adversarial set. Across *arena, rb-easy, mtbench, hh-rlhf, pku-safe, judgebench, rewardbench2, llmbar-adv* (50 judges each; 400 cells), the probe beats the raw verdict on **all 8** benchmarks (mean $+0.117$; Table 2a), improving 89% of individual cells.[2] The distribution matters more than the mean: the median cell gain is $+0.075$ (IQR $[+0.023, +0.209]$, $p_{90} = +0.298$) and all 50 judges gain on average (median $+0.119$), so no small set of pathological judges carries the effect.

A stricter worry is independence, since the 50 judges come from only **21** model families. Resampling by *family* (Fig. 2b) shrinks the pooled estimate only slightly, $+0.117$ $[0.105, 0.129]$ to $+0.108$ $[0.089, 0.127]$, and leaves it positive in **all** 21 **families** ($150/168$ family×benchmark cells); family-clustered CIs exclude zero on 7 of 8 benchmarks and all 8 survive a family-level paired Wilcoxon under FDR-BH. The one interval covering zero (hh-rlhf, $+0.010$ $[-0.008, +0.024]$) is the lowest-confound benchmark in the ladder, exactly what §6 predicts.

5 The Anatomy of a Judge

At which layer do a judge's activations start to determine its verdict? Unless noted, the causal numbers below come from the forward-pass suite on **41 judges** (3000+ measurements), are seed-stable across seeds $0/1/2$ (Table 2b), and agree with cached surrogates for $40/41$ judges ($\rho = 0.78$).

Editing activations changes the emitted verdict at earlier layers than those at which the verdict can first be read off. A residual patch first flips the verdict at a median of 0.58 of normalized depth, and

[2] On MT-Bench the residual-stream probe reaches 0.756, above that set's 0.646 human–human agreement. Inter-annotator agreement estimates label noise; it does not bound accuracy attainable against one annotator-derived label set.

(a) Evaluator ladder (50 judges × 8 benchmarks)

Bench.	cf	raw	o-avg	probe	resid	Δ_r	δ_o
rb2	.96	.620	.684	.840	**.907**	+.22	+.16
llmbar	.80	.483	.456	.758	**.846**	+.28	+.30
rb-easy	.77	.690	.815	.868	**.879**	+.18	+.05
mtbench	.62	.626	.697	.739	**.756**	+.11	+.04
arena	.59	.566	.616	.623	**.631**	+.06	+.01
judgebench	.52	.516	.545	.540	**.548**	+.02	−.01
pku-safe	.51	.564	.608	.622	**.647**	+.06	+.01
hh-rlhf	.49	.551	**.582**	.563	.555	+.01	−.02
mean, seed 0	.66	.577	.625	.694	**.721**	+.12	+.07
mean, seed 1	.65	.576	.626	.697	**.727**	+.12	+.07
beats raw		ref.	7/8	**8/8**	**8/8**		
fam. CI>0		ref.	7/8	**7/8**	**7/8**		
beats o-avg		–	ref.	6/8	**7/8**		

(b) Forward-pass causal suite (41 judges)

Quantity	Value
Where the decision is established	
Causal decision depth (frac.)	.58
Lens decoding depth (34/39 later)	.91
Steering slope (20/50/90%)	.13/.47/1.70
Attribution ≠ causation	
Attribution vs. cause ρ	.31
Indirect (mediated) share	82%
MLP- vs. attention-mediated	.63/.39
Surrogate vs. direct ρ (40/41)	.78
Bias is a localizable circuit	
Position heads (random ctrl.)	33–54% (.03)
Length subspace ablation	.41 → −.11
Debias transfer (unseen data)	−51%
Steering flips emitted verdict	67%

Table 2: (a) The evaluator ladder: mean accuracy against human labels over 50 judges, ordered by the surface-confound diagnostic **cf**, with Δ_r = probe − raw and δ_o = probe − order-average, the probe's non-redundant value; the footer gives the eight-benchmark mean on seed 0 and on the independent replication seed 1, and the number of benchmarks whose family-clustered CI for the gain over raw excludes zero. (b) Headline quantities of the forward-pass causal suite of §5, seed-stable across seeds 0/1/2 and measured inside a live forward pass. Amber marks the high-confound regime and the three central mechanistic numbers; green marks the best reader per row, the per-judge depth ordering, and the surrogate validation. cf is a diagnostic of *where* to read internals (§6).

steering moves it from the middle layers onward, with an effect that grows with depth (Table 2b).[3] The same order holds for individual judges: on **34** of the **39** judges with both measurements, patching changes the verdict at an earlier layer than the one at which a lens can decode it (median causal decision depth 0.585 vs. lens 0.906; paired Wilcoxon $p = 2\times10^{-7}$; Fig. 2d). For rubric scoring of a single answer, the decision forms later, at 0.9–1.0 of depth. The two depths are consistent: a middle layer can hold the information that decides the verdict in a form that later layers already use but that a linear read-out cannot yet recover. On easy items the tuned lens and the logit lens find the verdict at the same depth, so no extra fitting is needed; on adversarial items the untrained read-out is wrong at every depth, so a trained probe is required. Causal decision depth also does *not* predict which judges gain most from the probe ($\rho = 0.06$, $p = 0.73$): how early a verdict can be changed and how much human preference can be recovered are separate properties of a judge (§3).

Locating the computation this way also exposes a methodological trap. Direct logit attribution (DLA), the additive surrogate in common use, is an *incomplete proxy for causal influence* in our intervention suite. Correlating DLA with a causal-knockout ranking over 41 judges gives $\rho = 0.31$ at the component level (CI $[0.30, 0.33]$; below 0.5 for 95% of judges; peak-attribution layer 30 vs. peak-cause layer 18). Mediation explains why: **82**% of the biased effect is indirect (CI $[0.81, 0.84]$), and in separate decompositions of that indirect effect MLP-mediated contributions (0.63) exceed attention-mediated ones (0.39).[4] Attribution sees the direct path and is blind to the mediated majority, and consistently the cached surrogate agrees better with the direct ($\rho = 0.78$) than the total effect (0.66). The same holds for scalar reward models ($\rho = 0.32$; an earlier cached $\rho \approx 0.97$ there was a random-head artifact our null check caught).

What the mediated pathway carries is, in large part, surface bias, and that bias can be located and removed by intervention (Table 2b). **Position bias** is mediated in part by a small set of attention heads: ablating the top-ranked heads reduces the bias gap by 33–54%, against ≈3% for random heads, and

[3] Causal onset is estimated on a coarse 10-point depth grid from a small sample of paired items per judge (App. A.12), so it locates the decision only approximately.

[4] The two fractions come from separate decompositions and need not sum to one, so we only compare them: MLP-mediated effects are *larger than* attention-mediated ones.

heads selected this way on one benchmark reduce it by 51% on other benchmarks. **Length bias** is linearly represented: we extract a length direction from each judge's residual stream, and projecting out that subspace removes the bias ($0.41 \to -0.11$). Removing it alone leaves agreement with human labels unchanged, so the usable preference signal has to be recovered by reading the representation (§4). Directions extracted from one judge also shift a *different* judge's score, so judges share the geometry of the bias, though each encodes it in its own coordinates. Steering along these directions flips the emitted verdict in $\approx 67\%$ of cases, so the directions change the judge's output itself.

A final property gives the internal–output gap a geometric counterpart: the output exposes a lower-dimensional projection of a richer internal state (effective rank ≈2–4, aligned only weakly with the human axis), and scalar reward models behave the same way, with a mid layer reading an RM's reward $+0.2$–0.3 higher in Spearman ρ with the human label than the scalar head does (App. A.9).

6 When Reading Internals Helps

The gap is conditional, and making the condition measurable turns the read-out from a universal recommendation into a falsifiable one. The deployable recipe (App. A.6) uses only the judge's own forward pass and a small labeled set, with no weight updates: cache the residual stream over both answer orders, order-average to cancel position bias, and fit a small L_2-logistic probe on the logit-lens read-out.

We measure the severity of this failure mode with cf, the cross-validated accuracy of a length-only classifier on the evaluation set. It needs no internal read-out and no probe, only the small labeled sample of that distribution that the read-out would use anyway, so we call it *pre-read-out*. Ordered by cf (Table 2a), the eight benchmarks line up with the gain the probe obtains at Spearman $\rho = 0.90$, with an **exact permutation test** over all $8! = 40{,}320$ orderings giving $p = 0.005$ (Pearson $r = 0.91$, Kendall $\tau = 0.79$). Eight benchmark-level points are only a summary, so we add three checks. The relation is not simply that cf marks weak judges: after partialling out mean judge accuracy the correlation is still $\rho = 0.90$ ($p = 0.002$), while judge accuracy alone predicts the gain at only $\rho = 0.21$. The correlation stays between 0.86 and 0.96 when any one benchmark is dropped, and the genuinely held-out test is an independent replication run: cf computed on seed 0 predicts the gains measured on seed 1 at $\rho = 0.95$ (Fig. 2c). Finally it holds at the cell level, where the 400 cells give far more evidence than the eight means: fitting $\Delta_{ij} = \beta_0 + \beta_1 \mathrm{cf}_j + u_{\text{family}(i)} + \epsilon_{ij}$ gives $\hat{\beta}_1 = 0.53$ with a family-clustered 95% CI of $[0.46, 0.59]$, unchanged under a family random-intercept mixed model (0.53) and Huber-robust estimation (0.54); adding judge accuracy as a covariate even raises it (0.60).

On the two most confounded benchmarks the length-only classifier is itself more accurate than the probe (0.80 vs. 0.758 on LLMBar; 0.96 vs. 0.907 on RB2). This is expected: those benchmarks were built so that length predicts the label, and the length-only classifier is our measure of that failure mode. It would also make a poor judge: it needs the labels of the very benchmark it scores and drops to chance where length carries no signal (0.52 on JudgeBench), while the probe keeps a clear advantage once length and the other surface features are removed (§4). We also do not claim that cf predicts when internal read-out helps in every setting; it is a diagnostic for one common failure mode, surface confounding, in the benchmarks studied here.

A diagnostic should also say where the effect will be absent, so we test cf in three settings of increasing difficulty. **(i) Clean benchmarks.** The probe's non-redundant value over the free order-averaging debias (δ_o, Table 2a) concentrates on the confounded sets ($+0.30$ llmbar, $+0.16$ rb2) and falls to ≈ 0 elsewhere; the two negative cells, judgebench (-0.01) and hh-rlhf (-0.02), are the two lowest-cf benchmarks, and hh-rlhf is also the one whose family-clustered interval covers zero. **(ii) Zero-label cross-judge transfer.** A read-out fitted on *other* judges and applied to an unseen judge with none of its labels improves human agreement by $+0.08$ on LLMBar (beating that judge's verdict for 78% of judges) but is flat to slightly negative on JudgeBench (-0.01) and RB2 (-0.03). A combined internal-plus-length reader reaches a much larger $+0.32$, but that figure is largely carried by the length feature, a within-benchmark oracle there, so we report the internals-only number and keep the combined one in App. A.7. **(iii) A held-out task format.** Absolute rubric scoring presents one answer at a time, so no position or length cue separates candidates and our account predicts *no* internal advantage. We tested this on two existing human-labeled Indic rubric benchmarks (50 and 44 judges, $n = 80$ graded human scores each), a task format and pair

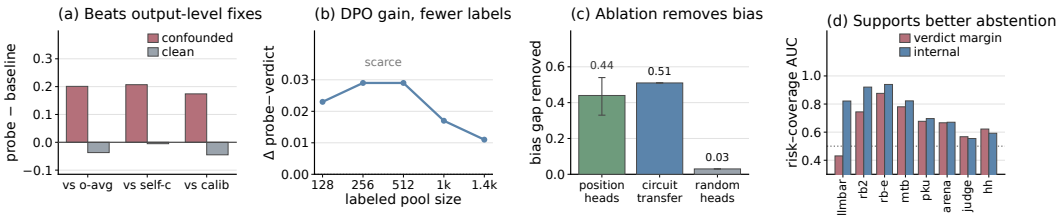

Figure 3: Four consequences of reading internals, at a matched human-label budget. **(a)** The probe's edge over output-level fixes is confined to confounded cells and vanishes, or reverses, on clean ones. **(b)** Its DPO labeling advantage grows as labels get scarce. **(c)** Ablating the position-bias heads removes much of the bias, also on unseen data. **(d)** Internal confidence supports better abstention than the judge's own verdict margin, which on LLMBar is *below* chance. Full ten-panel version: Fig. 6.

of benchmark sets held out from every other claim here. It holds: the read-out does no better than the emitted score (Samiksha $\rho = -0.006$ final, -0.005 best layer, $+0.017$ probe; Pariksha $+0.013$, $+0.003$, -0.053), and a held-out best layer beats the final layer on 52% and 50% of judges, matching the 50% predicted.

This null result needs reconciling with one prior finding. Lai et al. (2025) report the opposite sign for point-wise scoring, where cross-layer logit aggregation *improves* judge–human correlation. We do not read our null as a contradiction: the settings differ on four axes (pairwise vs. point-wise, adversarial confounding vs. ordinary scoring, a trained probe vs. training-free aggregation, and a different read-out target), and our account predicts internal read-out should help whenever the *output stage* is the locus of error. Miscalibrated score-token distributions are another such case, as Girrbach et al. (2025) observe in finding that probes recover signal specifically where output logits are miscalibrated. So our claim is not that internal representations are always better. Reading internals helps when the judge's output stage is where the error enters, and cf measures one common way that happens.

Two deployment properties follow. The reader lifts the floor: the weakest quartile of judges (mean raw 0.50) reaches 0.65, narrowing the gap to the strongest quartile from 0.17 to 0.08, while probe accuracy still rises with judge quality (slope $+0.56$), so weak judges are helped without a ceiling being hit. Internal confidence also supports *abstention*, gaining $+0.39$ risk–coverage AUC over the verdict margin on LLMBar, where that margin is *below chance* (0.431; Fig. 3d). The reader is judge-portable but distribution-specific, and it reorders *which* judge one would deploy (App. A.7).

7 Downstream Consequences

Holding the human-label budget fixed across arms, the read-out beats output-level debiasing and the tested full-parameter fine-tuning recipe, without the latter's safety regression, and produces better preference labels, with each advantage scaling with the same diagnostic. The read-out here is the 41-parameter logit-lens probe of App. A.6.

We first compare against output-level fixes. Giving every arm the **same 256 human labels** over **24 cells** (6 judges $\times$ 4 benchmarks, same items), we compare the probe to *self-consistency* ($k = 5$, majority-voted) and a *calibrated prompt* telling the judge to ignore length and position (Table 3a). The probe wins on mean accuracy, with its advantage concentrated on the confounded cells: $+0.17$ to $+0.21$ over every output-level fix on LLMBar and RB2, ≈ 0 on clean JudgeBench and Arena. Sampling more verdicts does not close this gap.

A stronger alternative is to retrain the judge. We full-FT a 24B roster judge (Mistral-Small-24B) on the verdict-format objective across 2×8 B200s (FSDP2) and compare it to the frozen-judge probe fit on the *same* $3{,}349$ human labels, scored on the same untouched 590-item test set (Table 3b). The frozen probe reaches higher overall accuracy (0.753) than both the fine-tune (0.659) and the base judge (0.700), with no weight updates. Since all readers score the same items we test the probe against the base with an exact McNemar test on the 590 paired predictions: $+0.049$

(a) Output-level fixes

Reader	mean	conf.
raw verdict	0.568	ref.
order-average	0.600	+.20
self-consist. k=5	0.581	+.21
calibrated prompt	0.617	+.17
internal probe	**0.682**	*best*

(b) Full fine-tuning

Bench	base	FT	probe
rb2	0.787	0.818	**0.893**
llmbar	0.511	**0.660**	**0.660**
judgebench	0.548	0.495	**0.570**
pku-safe	**0.716**	0.569	0.707
overall	0.700	0.659	**0.753**
excl. pku	0.690	0.715	**0.781**

(c) DPO preference labels

pool·judge	pol.	verd.	probe	gold
llmbar·q7b	1.5b	0.463	**0.787**	0.825
llmbar·q7b	3b	0.625	**0.775**	0.863
rb2·q7b	1.5b	0.906	0.931	0.960
rb2·g9b	1.5b	0.940	0.951	0.963
rb2·q7b	3b	0.923	0.934	0.963

Table 3: Three matched-budget comparisons: (a) readers over 24 cells at 256 labels, with the probe's paired advantage on the **conf**ounded cells (≈ 0 on clean ones); (b) the frozen probe against a *full*-parameter fine-tune of a 24B judge, same 3,349 labels and same held-out 590 items, safety regression in amber; (c) probe- against verdict-labeled DPO, gold-test accuracy, over five pool·judge·policy cells. Per-row n, per-cell Δ, and the label-efficiency sweep: App. Tables 6 and 17.

$[+0.017, +0.083]$, $p = 0.005$.[5] The fine-tune mainly teaches the judge to answer in the expected format (parse rate 1.0 vs. 0.003). Outside the safety split it gains only $+0.025$, and on pku-safe its accuracy falls from 0.716 to 0.569, while the probe leaves the base model unchanged and keeps its safety accuracy (0.707). We used a standard recipe without tuning its hyperparameters, so a tuned fine-tune may avoid this drop, and we do **not** claim that probing beats a well-tuned fine-tune in general. At matched labels the probe recovers about half of the mean LoRA gain ($0.550 \to 0.634$ vs. LoRA 0.724), and most of it on confounded LLMBar ($0.437 \to 0.774$ vs. 0.881), learning 41 parameters where the fine-tune updates 23.6 billion (App. A.8).

Better evaluation matters only if it changes what one can build, so we finally ask whether the recovered signal produces better training data. We label a preference pool three ways (verdict, probe, gold), DPO-tune a small policy on each, and evaluate on a held-out gold split (Table 3c). Probe labels win on four of five cells and are level on the fifth. On adversarial LLMBar the verdict labels are near-random (0.53), so verdict-DPO lands below chance (0.463); on that 1.5B cell **probe labels recover $\approx 90\%$ of the verdict-to-gold gap** ($0.463 \to 0.787$, gold 0.825, non-overlapping CIs, $n = 80$), and the gain survives a 1.5B $\to$ 3B policy swap ($+0.15$). We do not claim probe labels equal human labels: on less-confounded RB2 the gain is small ($+0.016$) but grows as labels get scarce.

A single quantity ties the section together: ranked by cf, the accuracy gain, the abstention advantage, the same-label advantage and the DPO gain all decline together and vanish on length-neutral sets (App. Table 16). In our eight benchmarks, reading internals pays off once $\mathrm{cf} \gtrsim 0.75$, a rule of thumb that new settings should recheck.

8 Conclusion

LLM judges are usually treated as black boxes whose verdict is the evaluation, and our results show that this can leave useful information on the table. Across 64 open-weight evaluators, a judge's final verdict can be wrong while the information needed to agree with humans remains recoverable from its internal activations: a small probe on those frozen activations clearly beats the judge's own verdict on adversarial pairwise evaluation, the gain holds after controlling for surface features and model family, and causal interventions locate the pathways behind the bias.

Reading the internals is not always better: the advantage concentrates where surface features strongly predict the human labels and largely disappears on rubric tasks without those confounds, so it is best used as a targeted diagnostic and repair alongside LLM judging. When a judge's output is unreliable, useful preference information may still be inside the model, and a small probe can recover it without changing the model.

Limitations. We show that preference information is recoverable (L1) and that internal components causally influence the verdict (L2); whether the judge used it (L3) is open (App. A.13).

[5] The paired test uses cached base verdicts scoring 0.703 (two unparsed items resolved differently from the table's 0.700); per-item fine-tune predictions were not retained, so the $+0.094$ gap over it has no paired test.

AI USE STATEMENT

We used general-purpose AI assistants at several points in this work, in every case under author review and revision. In **writing**, they helped polish prose, tighten phrasing, and improve the readability and consistency of the manuscript; the claims, framing, and scientific content are the authors' own. In **argumentation**, we used them as a sounding board to stress-test a few arguments and to surface counter-arguments, which we then checked against our own results. In **implementation**, they produced initial drafts of experiment, analysis, and plotting code, which the authors subsequently reviewed, debugged, and substantially rewrote; the released code reflects that editing rather than any generated draft. They also assisted with **presentation**, such as LaTeX table and figure layout, formatting, and visual design. Every number and claim in the paper was verified by the authors against the released result artifacts. AI assistants were not used to generate any central or core experimental data or labels. LLMs appear in this work as the *research objects* under study (§3, §A.5), which is a separate matter from their assistive use described here.

ETHICS STATEMENT

This work studies existing open-weight models on existing, previously published datasets. It collects no new human data and involves no human subjects. Every model and dataset is used under the terms set by its creators, and we redistribute no model weights and no underlying corpus. One evaluation set is a restricted human-annotated corpus, so we release its schema, prompts, and the results computed from it, but not the corpus itself (§A.13). Human preference labels carry the judgments and the biases of the people who wrote them, so we treat agreement with those labels as a noisy target rather than as ground truth and report inter-annotator agreement where it is available. Finally, the internal directions we use to locate and switch off a judge's position and length bias could in principle be steered the other way to strengthen that bias. We report them so that evaluator failures can be diagnosed and repaired, and we state that risk here rather than leave it implicit.

REPRODUCIBILITY STATEMENT

The deployable read-out recipe is stated as Algorithm 1 in §A.6; the model registry, dataset sheet, and instrument inventory are in §A.5; the fine-tuning and probe configurations, seeds, and determinism settings are in §A.12; and the analytic cost accounting behind the fine-tuning comparison is in §A.8. We release the code, the model registry, per-experiment result artifacts, the evaluation prompts, the 24-cell BUDGET grid, and the multi-node full-parameter fine-tuning configuration at the anonymous link given in §A.12.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations (ICLR) Workshop*, 2017.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations (ICLR)*, 2023.

Lawrence Chan, Adrià Garriga-Alonso, Nicholas Goldowsky-Dill, Ryan Greenblatt, Jenny Nitishinskaya, Ansh Radhakrishnan, Buck Shlegeris, and Nate Thomas. Causal scrubbing: A method for rigorously testing interpretability hypotheses. Alignment Forum, 2022.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled AlpacaEval: A simple way to debias automatic evaluators. In *Conference on Language Modeling (COLM)*, 2024. arXiv:2404.04475.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *Transformer Circuits Thread*, 2022.

Nils Feldhus, Tanja Baeumel, Elena Golimblevskaia, Qianli Wang, Van Bach Nguyen, Aaron Louis Eidt, Selin Kahvecioglu, Christopher Ebert, Wojciech Samek, Jing Yang, Vera Schmitt, Sebastian Möller, and Simon Ostermann. Judge circuits explain format-induced inconsistency in LLM-as-a-judge. *arXiv preprint arXiv:2605.16023*, 2026.

Leander Girrbach, Chi-Ping Su, Tankred Saanum, Richard Socher, Eric Schulz, and Zeynep Akata. Reference-free rating of LLM responses via latent information. *arXiv preprint arXiv:2509.24678*, 2025.

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. Localizing model behavior with path patching. *arXiv preprint arXiv:2304.05969*, 2023.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, 2017.

Srishti Gureja, Lester James V. Miranda, Shayekh Bin Islam, Rishabh Maheshwary, Drishti Sharma, Gusti Winata, Nathan Lambert, Sebastian Ruder, Sara Hooker, and Marzieh Fadaee. M-RewardBench: Evaluating reward models in multilingual settings. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 43–58, 2025.

Hamna, Gayatri Bhat, Sourabrata Mukherjee, Faisal M. Lalani, Evan Hadfield, Divya Siddarth, Kalika Bali, and Sunayana Sitaram. Building benchmarks from the ground up: Community-centered evaluation of LLMs in healthcare chatbot settings. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI)*, 2026. doi: 10.1145/3772318.3791172.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations (ICLR)*, 2024.

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 (NeurIPS)*, 2025. arXiv:2508.03550.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, Lester James V. Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. RewardBench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1755–1797, 2025.

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. In *International Conference on Learning Representations (ICLR)*, 2026.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on Gemma 2. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 278–300, 2024.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2511–2522, 2023.

Sharan Maiya, Yinhong Liu, Ramit Debnath, and Anna Korhonen. Improving preference extraction in LLMs by identifying latent knowledge through classifying probes. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9061–9081, 2025.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Sourabrata Mukherjee, Kalika Bali, and Sunayana Sitaram. Actions speak louder than words: Measuring cross-lingual policy retention in tool-using agents. In *Conference on Language Modeling (COLM)*, 2026a. To appear. arXiv:2608.11110.

Sourabrata Mukherjee, Hamna, Kalika Bali, and Sunayana Sitaram. The geometry of LLM-as-judge: Why inter-LLM consensus is not human alignment. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2026b. To appear. arXiv:2606.03043.

nostalgebraist. Interpreting GPT: The logit lens. LessWrong, 2020.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

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

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, et al. Towards understanding sycophancy in language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

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 *International Conference on Learning Representations (ICLR)*, 2025.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, et al. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10582–10592, 2024a.

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 (ICLR)*, 2023a.

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 (ACL)*, pp. 9440–9450, 2024b.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023b.

Ishaan Watts, Varun Gumma, Aditya Yadavalli, Vivek Seshadri, Manohar Swaminathan, and Sunayana Sitaram. PARIKSHA: A large-scale investigation of human-LLM evaluator agreement on multilingual and multi-cultural data. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7900–7932, 2024.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. Evaluating large language models at evaluating instruction following. In *International Conference on Learning Representations (ICLR)*, 2024.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations (ICLR)*, 2024.

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 (NeurIPS)*, 2023.

A APPENDIX

Table 4 maps this appendix: what each section is for, and what each table and figure in it shows.

Table 4: Guide to the appendix: section purposes, and what each table and figure shows.

Section	What it is for	Tables and figures
§A.1 Additional Definitions	Definitions of dimensional collapse and of the concept directions used in steering and ablation.	–
§A.2 Claim–Evidence Matrix	Every load-bearing claim against the evidence for it.	Tab. 5: claim, evidence, causal or not, scale, and the claims we do *not* make.
§A.3 Main-Paper Figure and Table Guide	Panel and column definitions for Figs. 2–6 and Tabs. 2a, 3a and 3b.	Tab. 6: what every panel and column means.
§A.4 Additional Figures	Companion figures to the main results.	Fig. 4: where the decision forms. Fig. 5: the read-out end to end. Fig. 6: the ten-panel version of Fig. 3.
§A.5 Model Registry and Dataset Sheet	The 64 models and 14 datasets, with roles, sizes, and provenance.	Tab. 7: all 64 models by role. Tab. 8: instruments, controls, compute. Tab. 9: the 8 pairwise benchmarks. Tab. 10: the 6 auxiliary sets.
§A.6 Deployable Read-Out Recipe	The read-out written out as runnable steps, with no weight updates.	Alg. 1: cache the residual stream, order-average, fit the probe.
§A.7 Extended Evaluator Results	Reading against retraining, abstention, cross-judge transfer, and which judge to deploy.	Tab. 11: read vs. train at matched labels. Tab. 12: abstention against the verdict margin. Tab. 13: judge re-ranking.
§A.8 Cost Accounting	Labels, gradient steps, parameters updated, and storage for each arm.	Tab. 14: cost of every arm at the same label budget.
§A.9 Extended Mechanism	Which behavioral failures the mechanism accounts for, and the same story in reward models.	Tab. 15: pathologies and internal repairs. Tab. 16: every measured gain ordered by cf.
§A.10 Consequential Experiments	Full per-cell tables behind §7.	Tab. 17: probe- against verdict-labeled DPO.
§A.11 Extended Related Work	Our precise relation to each closest prior work, one by one.	–
§A.12 Reproducibility Protocol	Seeds, determinism settings, and the numerical guards.	–
§A.13 Limitations and Data Statement	What we do not claim, and where every dataset comes from.	–

A.1 ADDITIONAL DEFINITIONS

Two read-out quantities used in §5 are defined here rather than in the body. **Dimensional collapse** compares the effective rank of the internal score representation with that of the emitted output score; we measure effective rank by the participation ratio, an entropy-based effective-dimension measure of the eigenvalue spectrum, so that a representation whose variance is spread over many directions scores higher than one concentrated on a few. A **concept direction** $\hat{u}$ is the unit weight vector of a linear probe fit on h_ℓ to predict a nuisance or target attribute (position, length, language, quality); it is the direction we add in steering and project out in ablation.

A.2 CLAIM–EVIDENCE MATRIX

Table 5 lists every load-bearing claim in the main paper, the evidence behind it, whether that evidence is causal or correlational, its scale, and the claims we deliberately do *not* make. It is intended to let a reader check the mapping from claim to evidence directly.

Table 5: Claim–evidence matrix for the main paper.

Claim	Evidence	Causal?	Scale	Where
Preference is decodable from frozen activations when the verdict is wrong	CV probe, .456 → .758/.846	no (L1)	50 judges, LLMBar-adv	§4
That signal is not only surface	9-dim residualization, length balancing, shuffle control	no (L1)	50 judges	§4
The gap generalizes	ladder, 8/8 benchmarks, 89% of cells	no	400 cells	§4
… and is not an artifact of treating judges as independent	family-clustered bootstrap; 7/8 CIs exclude 0; positive in 21/21 families	no	21 families	§4
Intervention changes the verdict from mid-stack	residual patching, steering	yes (L2)	41 judges	§5
… before it is lens-decodable	paired per-judge test, 34/39, $p = 2\times10^{-7}$	yes (L2)	39 judges	§5
Attribution is an incomplete causal proxy	DLA vs. causal knockout, $\rho = .31$; 82% indirect	yes (L2)	41 judges	§5
Surface bias is a localizable circuit	head ablation 33–54% vs. ≈3% random; subspace removal; transfer	yes (L2)	41 judges	§5
cf predicts the gain	$\rho = .90$, exact perm. $p = .005$; partial $\rho = .90$; LOO [.86, .96]; held-out seed $\rho = .95$; cell-level $\hat{\beta} = .53$	no	8 benchmarks, 400 cells, 2 seeds	§6
cf predicts *absence*	clean benchmarks $\delta_0 \approx 0$; held-out rubric format $\rho = -.006$	no	2 benchmarks, 94 judge-runs	§6
Read-out beats output-level fixes at matched labels	24-cell BUDGET grid	no	6 judges × 4 benchmarks	§7
Read-out beats one tested full-FT recipe	matched-label comparison, 590 shared items	no	1 judge (24B)	§7
Probe labels train a better DPO policy	5 cells, gold-test accuracy	no	5 cells, 2 policy sizes	§7
*Claims we do **not** make*				
The judge *used* the probe's direction	–	L3, not shown	–	§3
Probing beats fine-tuning in general	one untuned recipe	not shown	–	§7
The read-out transfers across benchmarks without labels	naive transfer fails (−0.29)	refuted	–	§6
cf predicts usefulness in every setting	one source of output-stage error	not shown	–	§6

A.3 GUIDE TO THE MAIN-PAPER FIGURES AND TABLES

Body captions are kept short. Table 6 gives the full reading of every panel and column, with the statistics and controls behind each.

Table 6: Panel and column guide for the main-paper figures and tables.

Panel	What it shows, and how it was computed
Fig. 4: where the decision forms	

Panel	What it shows, and how it was computed
–	A forward-pass residual patch places the causal decision at 0.58 of depth, before the lens first decodes the verdict; per judge this ordering holds on $34/39$ judges (Fig. 2d). On easy items the verdict is linearly readable without extra fitting; on adversarial items the untrained read-out is wrong at every depth, which is why the deployable reader is a *trained* probe.
Fig. 5: the read-out, end to end (50 judges × 8 benchmarks)	
(a)	A small-label lens probe (blue) and a residual-stream probe (green) exceed the stated verdict (rose) on all 8 benchmarks, mean $+0.12$, every one significant under FDR-BH; the margin is largest where the verdict collapses to chance. Training-free order-averaging (sand) beats raw on $7/8$.
(b)	Selective prediction: abstaining on the lowest-confidence items raises accuracy monotonically as coverage falls. At 25% coverage accuracy reaches 0.91–0.98.
(c)	Over all 400 judge×benchmark cells, absolute probe accuracy *rises* with judge quality (slope 0.56) and 89% of cells sit above $y = x$, so the weak-judge gain is range compression rather than headroom.
(d)	The surface-confound diagnostic **cf** orders the eight benchmarks by probe−verdict gain, $\rho = 0.90$ (exact permutation $p = 0.005$), leave-one-benchmark-out $\rho \in [0.86, 0.96]$.
(e)	Given the same 256 human labels the frozen probe recovers about half of the mean LoRA fine-tuning gain with zero gradient steps, and most of it on confounded llmbar (Table 11).
Fig. 2: the gap and its statistics (main body)	
(a)	Reader hierarchy on LLMBar-adversarial over 50 judges with 95% bootstrap CIs, ordered stated verdict → surface-only → internal → surface-residualized internal → shuffled-label control. The hatched bars are the surface-controlled preference signal: the read-out after a 9-dimensional surface vector is regressed out.
(b)	Per-benchmark $\Delta =$ probe $-$ raw, bootstrapped twice: over the 50 judges (sand) and over the 21 model families (blue), $20{,}000$ resamples each. Family clustering is the conservative unit of independence.
(c)	cf vs. mean gain over the eight benchmarks, seed 0 (circles) and the independent replication seed 1 (triangles); the seed-0 diagnostic predicts seed-1 gains at $\rho = 0.95$.
(d)	Per-judge scatter of causal decision depth (residual patch) against lens decoding depth, $n = 39$ judges with both measurements; points above the diagonal decode *after* the intervention-sensitive onset.
Fig. 3: four consequences (main body)	
(a)	The probe's paired advantage over each output-level fix: $+0.17$–$+0.21$ on confounded cells, ≈ 0 or negative on clean ones.
(b)	Label-efficiency sweep on rb2: the downstream DPO gain grows as the labeled pool shrinks, peaking at $+0.029$ at 256–512 labels.
(c)	Ablating the position heads removes 44% of the bias gap and the same circuit transfers to unseen data (51%), while random-head ablation removes $\approx 3\%$.
(d)	Risk–coverage AUC of internal confidence vs. the verdict margin (Table 12).
Fig. 6: consequences, full 10 panels (24 cells, same labels)	
(a)	Mean accuracy of each reader over the 24 cells: probe 0.682 vs. raw 0.568, self-consistency 0.581, calibrated prompt 0.617.
(b)	The probe's paired advantage over each output-level fix is $+0.17$–$+0.21$ on confounded cells and vanishes on clean ones.
(c)	Probe-labeled preference data trains a better DPO policy on four of five cells; on LLMBar verdict-DPO falls below chance (0.463) while probe labels reach 0.787.
(d)	Label-efficiency sweep on rb2: the downstream gain grows as the labeled pool shrinks, peaking at $+0.029$ at 256–512 labels.
(e)	Against a full-parameter fine-tune of a 24B judge on the same labels the frozen probe wins overall (0.753 vs. 0.659) and holds safety where full-FT regresses ($0.716 \to 0.569$ on pku-safe).
(f)	Causal surface control: residualizing length and a 9-dimensional surface vector still leaves human-label-predictive signal well above the shuffled-label control.
(g)	Ablating the position heads removes 44% of the bias gap and the same circuit transfers to unseen data (51%), while random-head ablation removes $\approx 3\%$.
(h)	Stacking the order-averaged final layer with mid-layer content beats either component alone (llmbar 0.738 vs. 0.454 / 0.582).

Panel	What it shows, and how it was computed
(i)	Risk–coverage AUC of internal confidence vs. the verdict margin: the learned internal reader gives a better abstention signal than the judge's own output margin (Table 12).
(j)	The measured gains from internals decline together with the same surface-confound diagnostic **cf** (Table 16).
Table 2a: column definitions	
cf	Surface-confound diagnostic: cross-validated accuracy of a length-only probe on that benchmark, computed with no internal read-out and before any probe is fitted.
raw	The judge's stated verdict.
o-avg	Training-free order-average over both answer orders.
probe	Small-label L_2-logistic probe on the logit-lens read-out.
resid	The same probe on the residual stream (PCA-48).
Δ_{r}	probe − raw; all 8 significant under FDR-BH, both with the judge and with the model family as the resampling unit.
δ_{o}	probe − order-average: the probe's non-redundant value.
mean	Eight-benchmark mean of each column on seed 0 and on the independent replication seed 1; the two seeds agree at cell level with $r \geq 0.96$ for every reader.
fam. CI>0	Number of benchmarks whose family-clustered 95% bootstrap CI for the gain over raw excludes zero, with the 21 model families as the resampling unit.
Table 3a: output-level fixes on the same labels	
cells	24 judge×benchmark cells (6 judges × 4 benchmarks), every arm scored on identical items with the same 256 human labels. The read-out is the 41-parameter logit-lens probe, as everywhere in §7.
self-cons.	$k = 5$ samples at $T = 0.7$, majority-voted.
calibrated	A prompt instructing the judge to ignore length and position.
mean	Accuracy averaged over all 24 cells.
con-founded	The probe's *paired* advantage over that reader on the 12 confounded cells (llmbar, rb2). On the 12 clean cells (judgebench, arena) each of these advantages is ≈ 0 (-0.05 to -0.01), which is why the edge is a confounded-regime phenomenon.
Table 3b: full fine-tuning on the same labels	
setup	A 24B judge (Mistral-Small-24B) is full-parameter fine-tuned on the verdict-format objective over 2×8 B200s with FSDP2; the frozen-judge probe is fit on the *same* 3,349 human labels. Both arms are scored on the same untouched 590-item test set, order-averaged.
base	The judge before any fine-tuning.
full-FT	After full-parameter fine-tuning.
probe	The frozen judge read by the 41-parameter logit-lens probe, with no weight updates.
n per row	Held-out items behind each row: rb2 225, llmbar 47, judgebench 93, pku-safe 225; *overall* 590 and *excl. pku-safe* 365.
pku-safe	The safety third of the pool (amber), where full-FT regresses $0.716 \rightarrow 0.569$ while the probe holds 0.707.
llmbar	The two arms tie at 0.660; $n = 47$ admits ± 0.14, so this cell is not separating.
excl. pku	Overall accuracy with the safety split removed, isolating the non-safety comparison.

A.4 Additional Figures

This section collects two companion figures to the main results. Figure 4 shows the depth evidence summarized per judge in Table 2b and Fig. 2d; Figure 5 is the end-to-end evaluator panel whose headline rows appear in Table 2.

A.5 Model Registry and Dataset Sheet

Table 7 lists all **64** registry models grouped by the role each plays. Table 9 is the dataset sheet: what each set contains, what the label is, which claim it supports, and how it may be obtained. Table 8 inventories the instruments and controls summarized in §3.

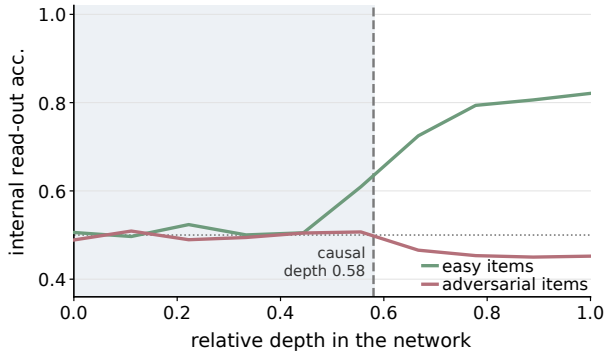

Figure 4: Where the decision forms. A forward-pass residual patch places the causal decision at 0.58 of depth, before the lens decodes the verdict. On easy items the verdict is linearly readable without extra fitting; on adversarial items the untrained read-out is wrong at every depth, which is why the deployable reader is a *trained* probe. Per-judge pairing of the two depths: Fig. 2d.

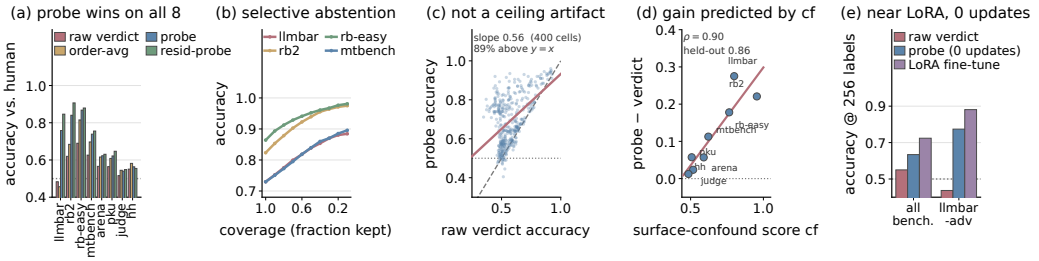

Figure 5: The read-out end to end, over 50 judges and 8 benchmarks. **(a)** Both probes beat the stated verdict everywhere. **(b)** Abstaining raises accuracy monotonically. **(c)** Not a ceiling artifact. **(d)** The gain is predicted by **cf** ($\rho = 0.90$). **(e)** At 256 labels the frozen probe recovers about half of the mean LoRA gain with zero gradient steps. Panel guide: Table 6.

The 21 model families used for clustering. The family-clustered statistics of §4 treat each base-model family as one resampling unit. The 50 panel judges fall into **21** such families: Qwen2.5 (6), Qwen3 (5, incl. QwQ), Gemma (5), Mistral (4, incl. Ministral and Mixtral), Aya/Cohere (4), DeepSeek-R1 (3), Llama-3 (3, incl. Tulu-3), Yi (2), InternLM (2), OLMo-2 (2), SmolLM2 (2), Phi (2), Prometheus (2), and one each of GLM, Falcon, gpt-oss, Sarvam, JudgeLM, OffsetBias, Skywork-Critic and Selene. Purpose-built judges are counted by their *own* identity rather than their base model, which is the conservative choice here: merging them into their bases would reduce the number of clusters and widen the family-clustered intervals only slightly, since each contributes a single judge. Table 7 groups some of these families into shared rows for compactness; the clustering uses the 21 families listed above.

The six auxiliary sets. The evaluator ladder of §4 requires pairwise-preference labels and a length-only confound score, which only the eight ladder benchmarks supply. The remaining six sets carry the mechanism, rubric, and multilingual analyses, and each appears in 20–31 separate experiments; Table 10 states what each contributes and the headline number where one exists. The two Indic rubric benchmarks give the most direct human-grounded check in the paper and the out-of-domain confirmation of §6.

A.6 The Deployable Read-Out Recipe

The recipe below is the frozen-backbone reader referenced in §4: it first decides *whether* to read internals at all, then caches the residual stream over both answer orders, order-averages to cancel position bias, featurizes with the logit lens (or a PCA of the raw residual), and fits a small L_2-

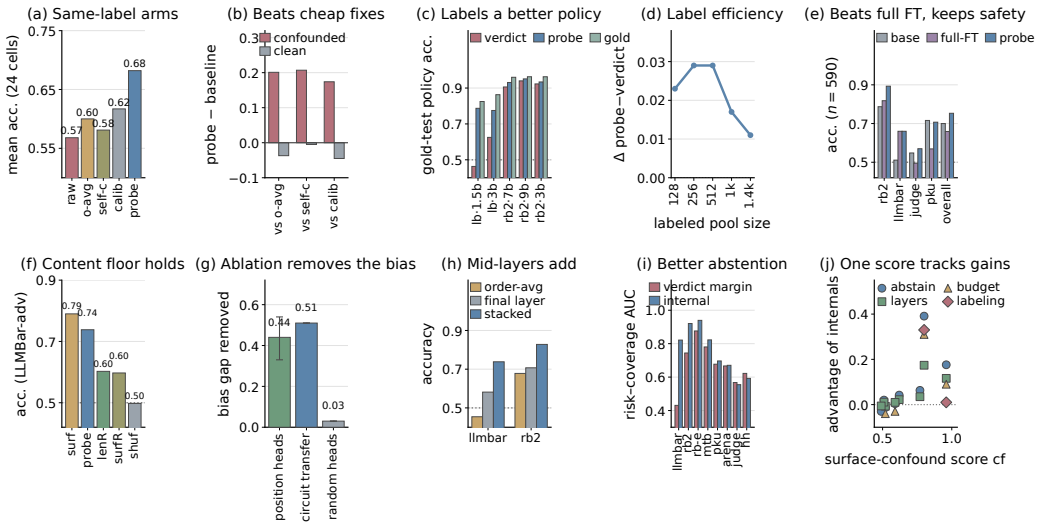

Figure 6: Downstream consequences of the internal read-out, the full ten-panel version of Fig. 3. **(a)** All arms given the same number of labels; probe best. **(b)** Its edge over output-level fixes is confined to confounded cells. **(c)** Probe labels train a better DPO policy on four of five cells. **(d)** That gain grows as labels get scarce. **(e)** Against a *full*-parameter fine-tune of a 24B judge, holding safety where that fine-tune regresses. **(f)** Human-label-predictive signal survives surface residualization (the surface-only bar is high by construction on this adversarial set). **(g)** Ablating the bias heads removes the bias, also on unseen data. **(h)** Stacked mid-layers beat either component. **(i)** Internal confidence supports better abstention than the judge's own verdict margin. **(j)** The measured gains decline with cf. Panel guide: Table 6.

logistic probe. Only the probe is learned; no gradient step touches the judge. Step 0 is the practical contribution of §6: on a low-confound evaluation set the free order-average already suffices and the remaining steps buy nothing. The threshold τ is a heuristic read off the eight benchmarks of §6 rather than a calibrated decision rule, and should be re-estimated on a new evaluation suite.

Algorithm 1: Interpretability-native evaluator read-out

input frozen judge J; items $\{(q, a_A, a_B)\}$; labels $\mathcal{L}$; threshold τ
output preference $\hat{y}$ and confidence c per item

0 trigger fit a surface-only probe on $\mathcal{L}$; cf $\leftarrow$ its CV accuracy
  if cf $< \tau$ **then** return the order-averaged verdict and stop
  // $\tau = 0.75$ is a rule of thumb from our eight benchmarks; recheck it on new settings
1 read for $o \in \{AB, BA\}$: run J; cache residual h_ℓ at the decision position $\forall \ell$; $\hat{v}_\ell \leftarrow \mathrm{Unembed}(\mathrm{LN}_f(h_\ell))$
2 debias order-average features over o // cancels position bias
3 featurize $\phi \leftarrow [\hat{v}_\ell]_\ell$ *(lens)* **or** $\phi \leftarrow \mathrm{PCA}_{48}(\Delta h_\ell)$ *(resid.)*
4 fit $g \leftarrow L_2$-logistic probe on (ϕ, y), $\mathcal{L}$ // 5-fold CV + shuffle ctrl
5 predict $\hat{y} \leftarrow g(\phi)$; $c \leftarrow |g(\phi) - \frac{1}{2}|$
6 audit if $c < \tau$ **then** abstain, route-to-human

A.7 Extended Evaluator Results

No weight updates, vs. training (main §7). The main-body comparison gives every arm the **same 256 human labels** (Table 11); the per-benchmark results follow the confound score: on confounded LLMBar the frozen probe and a LoRA/PEFT fine-tune both help substantially (raw $0.437 \rightarrow$ probe $0.774 \rightarrow$ LoRA 0.881), and on length-neutral JudgeBench neither helps. On low-confound PKU-

Family	Models (sizes as loaded)
*Role A, the flagship adversarial panel: **50** generative judges. The forward-pass causal suite intervenes on the **41** of these that expose hookable decoder blocks within our memory budget; all 50 are read out.*	
Qwen2.5	`0.5b, 1.5b, 3b, 7b, 14b, 32b`
Qwen3	`8b, 14b, 30b-a3b, 32b`
Llama-3	`3.2-1b, 3.1-8b, tulu-3-8b`
Gemma-2/3	`2-2b, 2-9b, 2-27b, 3-12b, 3-27b`
Mistral	`7b-v0.3, ministral-8b, small-24b, mixtral-8x7b`
Yi / GLM / InternLM	`yi-1.5-9b, yi-1.5-34b, glm-4-9b, internlm2.5-7b, internlm2.5-20b`
OLMo-2 / SmolLM2	`olmo2-7b, olmo2-13b, smollm2-360m, smollm2-1.7b`
Phi / Falcon / gpt-oss	`phi-3.5-mini, phi-4, falcon3-7b, gpt-oss-20b`
Reasoning	`deepseek-r1-{llama-8b, qwen-7b, qwen-32b}, qwq-32b`
Multilingual / Indic	`aya-23-8b, aya-expanse-{8b,32b}, command-r7b, sarvam-m`
Purpose-built judges	`prometheus-7b, prometheus-8x7b, judgelm-7b, offsetbias-8b, skywork-critic-8b, selene-mini-8b`
*Role B, read-out-only large judges: **5**. Too large to intervene on under our memory budget; used for scale checks only.*	
70B and above	`llama-3.3-70b, qwen2.5-72b, command-r-plus, llama-3.1-405b, qwen3-235b`
*Role C, scalar reward models: **9**. Verify the picture transfers from verdict-emitting judges to scalar heads; **5** are causally intervened.*	
Scalar RMs	`armorm-8b, skywork-rm-llama-8b, skywork-rm-gemma-27b, qrm-llama-8b, urm-llama-8b, internlm2-7b-reward, nemotron-70b-reward, grm-llama-3b, grm-gemma2-2b-rm`

Table 7: The full model registry, all 64 models by role. **50** panel judges + **5** read-out-only large judges + **9** scalar reward models = **64**, spanning 0.36B–405B and 20+ families. The forward-pass causal suite intervenes on **41** panel judges and **5** reward models; every other registry entry is read out but not intervened on. 63 of the 64 are loadable locally.

Safe the confound score predicts no internal advantage and the probe is flat while LoRA helps, so the prediction holds in the negative direction too, which is what makes cf predictive in both directions.

Auditor and selective prediction. Two cross-model auditors must be distinguished, because they support different claims. The **internals-only** auditor, a probe fitted on the other 49 judges' lens features and applied to a held-out judge with *none* of that judge's labels, raises human agreement over the held-out judge's own verdict by $+0.077$ on LLMBar (beating it for 78% of judges) and is flat to slightly negative on JudgeBench (-0.007) and RB2 (-0.031). This is the number we report in the body, and its sign pattern follows cf exactly as §6 predicts. A **combined internal-plus-length** auditor reaches a much larger $+0.322$ (LLMBar) / $+0.273$ (RB2), but on these two sets the length feature is itself a within-benchmark oracle ($\mathrm{cf} = 0.80$ / 0.96), so that figure mostly measures the surface confound rather than transferred internal structure; we therefore do not use it to support any claim about internals. Integrating the internal-confidence risk–coverage curve beats a no-selection baseline on all 8 benchmarks (AUC lift $+0.02$ to $+0.09$; Fig. 3d), and at 25% coverage accuracy reaches 0.91–0.98; a route-to-human policy auto-decides 100% (RB2) / 88% (rb-easy) at a $\geq 90\%$ precision point, and correctly declines to auto-decide any of the hard subjective sets, so it abstains on the sets where automation is least reliable.

What transfers, and what needs labels. *Across judges*, the internals-only zero-label auditor transfers to an unseen judge ($+0.08$ LLMBar, beating its verdict for 78% of judges) and not to the clean benchmarks. *Across benchmarks*, a probe applied naively does not transfer (llmbar→rb2 -0.29; leave-one-benchmark-out beats the verdict on $1/8$); only a probe *pooled* over datasets trans-

Component	What we use
Cached read-out surrogates	
Lenses	logit lens; tuned lens (learned affine per layer)
Attribution	direct logit attribution over heads and MLPs
Patching	activation patching, path patching, causal scrubbing
Features	Gemma-Scope sparse autoencoders
Probes	L_2-logistic, 5-fold CV, shuffled-label control
Hooks	residual / attn-out / mlp-out at the decision position, `float32`
Forward-pass causal suite (the gold standard)	
Steering	dose-response over α, layer sweeps
Ablation	per-head knockout, subspace projection, random-head control
Decomposition	path and mediation decomposition (direct vs. indirect)
Transfer	debias-circuit transfer to unseen data
Statistical controls	
Intervals	bootstrap 95% CIs, 2000 iterations
Tests	paired sign / Wilcoxon, FDR-BH over sweeps and the ladder
Nulls	shuffled labels; random-component null check
Confound controls	length and position residualization; 9-dim surface vector
Seeds	causal suite $0/1/2$; flagship and ladder replicated at seed 1
Compute	
Read-out	cached-activation backend, eight models at a time per node
Fine-tuning	2×8 B200, FSDP2 (full-FT); single-GPU LoRA

Table 8: Instrument, control, and compute inventory for §3. Cached surrogates are validated against the forward-pass suite, which changes the actual generation.

fers leave-one-dataset-out ($+0.03$, CI $[+0.016, +0.076]$). The reader is thus judge-portable and distribution-specific: deployment needs a small label set matched to the evaluation distribution, which is the usual case, since you know your evaluation set in advance.

Abstain: verdict-margin selective baseline (full). Risk–coverage AUC of internal-probe confidence vs. the verdict margin over 50 judges/bench (Table 12): internals beat the trivial verdict-confidence baseline by $+0.39$ (LLMBar) / $+0.18$ (RB2) and converge on length-neutral benches.

Deployment re-ranking (main §6). Table 13 gives the per-benchmark comparison behind the claim that reading internals changes which judge you would deploy.

A.8 Cost Accounting

The body claims the read-out is cheaper than fine-tuning. Table 14 makes that precise for the comparison of §7, in which every arm sees the same 3,349 labels and the same held-out 590 items. We report an analytic accounting (labels, gradient steps, parameters updated, stored artifact, forward passes per item) rather than wall-clock times, which we did not instrument uniformly across arms. The parameter and gradient-step counts are exact and hardware-independent, which is the comparison that transfers across sites; wall-clock would be specific to our cluster.

The headline asymmetry is the parameter count. On Mistral-Small-24B (40 layers, model width 5120) the lens feature vector has one entry per layer, so the learned reader is 40 weights and a bias: 41 **parameters against** 23.6 **billion**, and a 164-byte artifact against a 94 GB checkpoint. The residual-stream reader adds a 48-component PCA basis (≈ 1 MB) and 49 probe parameters. Both readers require two forward passes per item, one per answer order, which is the same cost as the training-free order-average baseline and less than self-consistency at $k = 5$.

Dataset	Label / form	Role in this paper
LLMBar (natural, **adversarial**)	human pairwise, surface decoupled from quality	flagship confounded regime; the verdict falls below chance here
RewardBench 2	human pairwise	highest confound score (cf = .96); largest probe gain
RewardBench (easy)	human pairwise	mid-confound control
JudgeBench	human pairwise, length-neutral	the *null*: no surface cue, no probe gain
MT-Bench (human)	human pairwise	human–human agreement 0.646 (inter-annotator, not a label ceiling)
Chatbot Arena	55k human preferences	scale / in-the-wild preference
HH-RLHF	human helpful/harmless pairs	low-confound control
PKU-SafeRLHF	safety preference pairs	safety regression test for fine-tuning
HelpSteer2	multi-attribute ratings	rubric / attribute decomposition
UltraFeedback	preference pairs	DPO training pool (§7)
BiGGen-Bench	rubric scoring	rubric-format control for the collapse analyses
m-RewardBench	multilingual pairwise (English, Hindi, Bengali, Arabic, Spanish, Chinese)	cross-lingual geometry and transfer
Samiksha (Hamna et al., 2026)	**human** rubric scores, Hindi/Kannada/Malayalam	existing Indic rubric benchmark; supplies the out-of-domain test of §6
Pariksha (Watts et al., 2024)	**human** rubric scores, 8 Indian languages (Bengali, Hindi, Kannada, Malayalam, Marathi, Odia, Tamil, Telugu)	existing Indic rubric benchmark; supplies the out-of-domain test of §6

Table 9: Dataset sheet for all 14 sets. The first eight are the pairwise-preference benchmarks of the evaluator ladder; the last six are the auxiliary sets (Table 10). Amber marks the confounded regime where the internal read-out wins most; green marks the two Indic rubric benchmarks. **All fourteen datasets are existing, previously published resources; none is introduced by this work.** Twelve are public, as is Pariksha (Watts et al., 2024). Samiksha (Hamna et al., 2026) is not openly redistributable; researchers who need it should contact its authors, who can share it under the terms it was collected under. We release the schema, rubric definitions, language coverage, evaluation prompts, and every result computed from it.

Auxiliary set	exps.	Result reported in this paper
Samiksha	31	Out-of-domain confirmation over 50 judges ($n = 80$): emitted score $\rho = -0.006$, held-out best layer -0.005, probe $+0.017$; best layer beats final on 52% of judges, matching the 50% the confound account predicts
Pariksha	31	Out-of-domain confirmation over 44 judges ($n = 80$): $+0.013$ / $+0.003$ / -0.053; best layer beats final on 50% of judges
BiGGen-Bench	27	Rubric-format control: standalone rubric scoring commits at 0.9–1.0 of depth, latest of any format
HelpSteer2	27	Multi-attribute rank-collapse measurements (§5)
m-RewardBench	23	Cross-lingual direction geometry and transfer
UltraFeedback	20	Preference pool for the DPO labeling experiment (§7)

Table 10: The six auxiliary datasets. These do not enter the evaluator ladder, which needs pairwise labels and a length-only confound score, but each is used in 20–31 mechanism experiments. Green marks the two Indic rubric benchmarks, which supply the out-of-domain confirmation of §6.

A.9 Extended Mechanism

Behavioral pathologies the mechanism explains. Table 15 lists the observable failure modes the mechanism accounts for, and two internal *repairs* (rubric decompression; calibration).

Reader (256-label)	Mean	llmbar	judge	pku
Raw verdict	0.550	0.437	0.547	0.667
Probe (0 updates)	0.634	0.774	0.537	0.592
LoRA / PEFT fine-tune	**0.724**	0.881	0.537	0.753

Table 11: Reading the *same* frozen judge vs. training, given the same 256 human labels (4 judges $\times$ 3 benchmarks). The no-update probe (green) recovers about half of the mean LoRA/PEFT gain and most of it on confounded llmbar (amber); against a *full*-parameter fine-tune the probe wins outright (main Table 3b).

Bench	AUC int.	AUC verd.	Δ / % win
llmbar	0.822	0.431	+.391 / 100%
rb2	0.921	0.744	+.176 / 98%
rb-easy	0.939	0.876	+.063 / 74%
mtbench	0.823	0.780	+.042 / 48%
pku-safe	0.697	0.677	+.020 / 56%
arena	0.671	0.667	+.004 / 32%
judgebench	0.555	0.567	−.013 / 30%
hh-rlhf	0.592	0.622	−.030 / 22%

Table 12: Risk–coverage AUC: internal confidence vs. verdict margin. On llmbar the verdict-margin AUC is 0.431 (below chance): the judge's own confidence is *anti*-correlated with correctness.

Reward models and complementarity. Scalar RMs also form the reward mid-stack, where a mid layer reads it +0.2–0.3 better than the scalar head (final 0.055 $\rightarrow$ best 0.355; divergence +0.33). Per-judge signals are complementary: a naive mean fails, but a sign-aligned ensemble beats the best single judge (0.642 vs. 0.438).

Cross-experiment coherence. Table 16 ranks all eight benchmarks by the confound score and shows the selective (ABSTAIN), content (LAYERS), same-label (BUDGET), and downstream (LABEL) advantages declining together; LABEL's cells are reported per bench·judge in Table 17 below.

A.10 CONSEQUENTIAL EXPERIMENTS: FULL TABLES

Labeling downstream RLHF (all cells). Table 17 lists the five headline cells (gold-test accuracy) and the rb2·3B label-efficiency sweep. Probe labels win on four of the five cells; the gain scales with confound and grows as labels get scarce. The rb2·q7b·3B cell is the single null: the row reports the full-pool point of the label-efficiency sweep (+0.011), while a separate run of the same configuration returns 0.000, so we count that cell as level rather than a win.

Layers stacked reader. The full-lens probe is the stack of the order-averaged verdict (final layer) $\oplus$ the mid-layer content. It dominates both components on high-confound benches: llmbar 0.738 vs. order-avg 0.454 / final-only 0.582 (100% of judges); rb2 0.828 vs. 0.678 / 0.707 (96%); neutral on clean benches.

Surface full 9-dim surface control. On the four aligned benches the preference signal survives residualizing the full 9-dim surface vector; *surface-resid* $\approx$ *length-resid* on all four (within 0.005), so length is the dominant surface confound and the extra eight features (markdown, lists, code, headers, bold, sentence length, digits) remove nothing more. On rb2 the surface vector alone predicts the label at 0.980 yet residual content still survives (+0.086 over shuffle).

A.11 EXTENDED RELATED WORK

Our work sits at the intersection of three threads; Table 1 summarizes the positioning.

Bench.	ρ	by verdict	by internals
rewardbench2	.58	skywork-crit-8b	*unchanged*
llmbar-adv	.20	skywork-crit-8b	mistral-sm-24b
rb-easy	.71	qwen2.5-14b	qwen2.5-32b
mtbench	.54	phi-4	ds-r1-qwen-32b
arena	.71	mistral-sm-24b	gemma-3-27b
judgebench	.32	qwen2.5-32b	ds-r1-qwen-32b
pku-safe	.74	gemma-2-27b	qwen2.5-32b
hh-rlhf	.59	qwen2.5-32b	offsetbias-8b
mean	**.55**		**changed on 7 of 8**

Table 13: Deployment re-ranking: Spearman ρ between ranking the 50 judges by emitted verdict and by internal read-out, with the top-ranked judge under each. Rows follow the confound ordering of Table 2. The best judge to deploy changes on 7 of 8 benchmarks and the two rankings agree only moderately (mean $\rho = 0.55$). Agreement is lowest on LLMBar-adversarial (0.20), where the emitted verdict is least trustworthy, and also low on JudgeBench (0.32), where near-ties among judges make the ranking unstable under either view (amber). RewardBench-2 is the one benchmark whose top judge is stable. Names abbreviated (ds-r1 = deepseek-r1, sm = small, crit = critic).

Arm	fwd/item	grad. steps	params upd.	stored
raw verdict	1	0	0	–
order-average	2	0	0	–
self-consistency (k=5)	5	0	0	–
calibrated prompt	1	0	0	–
lens probe	2	0	41	164 B
residual-stream probe	2	0	49	$\approx$1 MB
LoRA (r=16, α=32)	2	yes	adapter	adapter ckpt
full-parameter FT (24B)	2	3 epochs	23.6 B	94 GB

Table 14: Cost of each arm when all arms are given the same $3{,}349$ human labels in §7. Parameter counts are for Mistral-Small-24B (40 layers, width 5120); the lens probe learns one weight per layer plus a bias. Green marks the arms that touch no model weights; amber the arm they are compared against. This is an analytic accounting, not a wall-clock benchmark.

(i) Evaluation. LLM judges (Zheng et al., 2023; Liu et al., 2023) and reward models (Christiano et al., 2017; Ouyang et al., 2022) are audited *behaviorally* through benchmarks (Lambert et al., 2025; Tan et al., 2025; Gureja et al., 2025) that report agreement with human preference but say nothing about the computation producing it. Known failure modes include position bias (Wang et al., 2024b), length/verbosity bias (Dubois et al., 2024; Saito et al., 2023), self-preference (Panickssery et al., 2024), sycophancy (Sharma et al., 2024), and adversarial-surface failure (Zeng et al., 2024). Such confounds also distort measurement beyond judging: in multilingual agent evaluation, trace length and chance agreement can each flip a conclusion, and a single trace-extraction step, with the model unchanged, manufactured an apparent failure (Mukherjee et al., 2026a). Remedies act at the *output*: preference optimization (Rafailov et al., 2023), multi-objective reward heads (Wang et al., 2024a), and cheap debiases such as swap-and-average or self-consistency (Wang et al., 2023b). Reward hacking (Skalse et al., 2022) frames the risk of optimizing such a proxy. None of these opens the evaluator.

(ii) Mechanistic interpretability. We use the logit (nostalgebraist, 2020) and tuned (Belrose et al., 2023) lenses, direct logit attribution (Elhage et al., 2021), activation/path patching and causal tracing (Meng et al., 2022; Wang et al., 2023a; Goldowsky-Dill et al., 2023), causal scrubbing (Chan et al., 2022) and mediation analysis (Vig et al., 2020), sparse autoencoders (Bricken et al., 2023; Huben et al., 2024; Templeton et al., 2024; Lieberum et al., 2024), and linear probing (Alain & Bengio, 2017; Belinkov, 2022) under superposition (Elhage et al., 2022). We confirm at roster scale the caution that attribution need not predict causal effect (Zhang & Nanda, 2024), and quantify *why*: the effect is mostly MLP-mediated, which a direct-path surrogate cannot see.

Property (behavioral probe)	Measure	Value
Bias & safety pathologies		
Position-bias flip rate	swap-flip frac.	0.95–1.0
Swap non-invariance	1−agreement	up to 1.0
Eval-awareness shift	lens σ	2.0–5.6
Self-preference (cross-gen)	win rate	0.553
scaling with size	Spearman	+0.63
Contamination inflation	lens σ	+0.41
CoT pre-commit	frac. before	$\approx$0.80
Reassuring separability		
Safety $\perp$ quality	1−overlap	0.98
Reward-hacking (spurious)	R^2	≈ -0.18
Interpretability-native repairs		
Rubric decompression	eff. rank	$2.95 \rightarrow 3.5$
Calibration (confidence)	Spearman	$-0.108 \rightarrow +0.072$
Calibration error	ECE	$0.405 \rightarrow 0.262$

Table 15: Behavioral pathologies the mechanism explains, with operational units. The last block are internal repairs.

Bench.	**cf**	MEASURE	ABSTAIN	LAYERS	BUDGET	LABEL
	diagnostic	*acc.*	*AUC*	*content*	*budget*	*DPO*
rewardbench2	.96	+.22	+.18	+.12	+.09	+.01
llmbar-adv	.80	+.28	+.39	+.17	+.31	+.33
rb-easy	.77	+.18	+.06	+.04	–	–
mtbench	.62	+.11	+.04	+.02	–	–
arena	.59	+.06	+.00	+.01	−.03	–
judgebench	.52	+.02	−.01	−.01	−.04	–
pku-safe	.51	+.06	+.02	+.01	–	–
hh-rlhf	.49	+.01	−.03	−.01	–	–

Table 16: One score tracks the measured gains. The measurement gain (MEASURE, probe−verdict accuracy), selective-prediction advantage (ABSTAIN, risk–coverage AUC Δ), mid-layer content (LAYERS), same-label advantage over cheap fixes (BUDGET), and downstream DPO gain (LABEL) all track the surface-confound diagnostic **cf** and shrink to ≈ 0 on length-neutral sets. Amber rows are the confounded regime where every column peaks (green); the three full columns span all 8 benchmarks, while BUDGET/LABEL were run on 4- and 2-benchmark grids (– = not in that grid).

(iii) Reading internals for evaluation. Probes can detect latent knowledge that the output does not state (Burns et al., 2023), and selective prediction supplies principled abstention (El-Yaniv & Wiener, 2010; Guo et al., 2017). Four recent lines apply this to evaluators directly, and we state our relation to each precisely.

Maiya et al. (2025) train supervised and unsupervised linear probes on contrastive prompt pairs and show they extract preferences more accurately than generation-based judgment across four model families and six datasets, generalize under domain shift, and can match fine-tuned evaluators at equal data. This establishes the descriptive claim (L1) that our §4 also observes, at larger roster scale and under a harder adversarial split. *We do not claim that observation as a contribution.* What we add is orthogonal: a causal account of *where* the evaluator's decision and biases are established (§5) and a pre-read-out diagnostic for *when* the read-out pays off (§6), including a case where it does not.

Lai et al. (2025) (LAGER) improve point-wise judge–human correlation on Flask, HelpSteer, and BIGGen by aggregating score-token logits across layers with the backbone frozen, reporting that mid-to-upper layers align better with human scores than the final layer, independent support for the mid-stack picture of §5. We differ in target and method: they aggregate logits into a single point-wise score with no training, whereas we intervene causally and work mainly in the pairwise setting where surface cues separate candidates. §6 discusses the apparent sign disagreement between their

bench·judge	pol.	verd.	probe	gold	Δ
llmbar·q7b	1.5b	0.463	0.787	0.825	+.325
llmbar·q7b	3b	0.625	0.775	0.863	+.150
rb2·q7b	1.5b	0.906	0.931	0.960	+.026
rb2·g9b	1.5b	0.940	0.951	0.963	+.011
rb2·q7b	3b	0.923	0.934	0.963	+.011
rb2·3b label-efficiency (pool → Δ probe−verd.)					
128		0.880	0.903		+.023
256		0.900	0.929		+.029
512		0.923	0.951		+.029
1024		0.923	0.940		+.017
1387		0.923	0.934		+.011

Table 17: Probe- vs. verdict-labeled DPO, gold-test accuracy. The gain is largest on confounded llmbar (verdict-DPO below chance) and, on rb2, grows as the labeled pool shrinks.

point-wise gains and our absolute-rubric null, and proposes the reconciliation that internal read-out helps in proportion to the error the output stage introduces, of which surface confounding is one source and score-token miscalibration another.

Girrbach et al. (2025) derive scalar ratings from three latent signals (probability-weighted scores over integer ratings, verifier-style "yes" probabilities, and probes on activations at the rating position) and show they improve reference-free evaluation, Best-of-N selection, and routing, with multi-teacher distillation proposed as a further use. Their finding that probes recover signal specifically when output logits are miscalibrated is the closest prior statement of a *conditional* account, and is consistent with ours; we make the condition measurable before the read-out is fitted for pairwise preference and validate it on a held-out replication seed. Li et al. (2026) frame evaluation as representation probing in small models (INSPECTOR), predicting aspect-level scores from intermediate features; their aim is efficiency, replacing a large judge, rather than diagnosing a large judge's failures.

Finally, concurrent with this work, Feldhus et al. (2026) apply position-aware edge attribution patching to judges in three model families and identify a sparse "latent evaluator" sub-graph in mid-to-late MLPs that is shared across judgment tasks and feeds fragile, format-specific output branches. That is the closest mechanistic result to ours and it agrees with two of our findings on independent evidence: the evaluation computation is established before the output stage, and the MLP pathway carries much of it. Their question is why scores change *across output formats*; ours is whether the human-labeled preference survives the output stage at all, measured with forward-pass interventions in 41 judges and connected to evaluation and preference-learning outcomes.

Unlike prior interpretability work, which targets *capabilities*, we target the *evaluator* itself, at roster scale, and convert the findings into a deployed reader whose labels demonstrably improve downstream training.

Operationalizing causal decision depth. For each judge we sweep layers on a fixed grid of 10 normalized depths. At each depth we patch the residual stream at the decision position from a counterfactual run (the same item with the two answers swapped) and re-run the forward pass to completion, recording whether the *emitted* verdict changes. A depth counts as flipping the verdict if it does so on a majority of the item pairs for that judge ($> 50\%$ of 8–12 paired items, depending on the judge's memory budget), and **causal decision depth** is the shallowest such depth on the grid. We take the shallowest rather than the peak because the quantity of interest is the *onset* of causal sufficiency; because it is an onset on a coarse grid, it is stable to seed (reported across seeds $0/1/2$) but should not be read to a precision finer than the grid spacing. Judges for which no depth reaches the majority criterion are excluded from the depth statistics, which is why the per-judge depth comparison of §5 uses 39 of the 41 causally intervened judges.

A.12 Reproducibility Protocol

All runs are deterministic (`PYTHONHASHSEED`=0, offline caches). Model seeds are $0/1/2$: the causal suite runs at all three and is seed-stable, the flagship adversarial study and the evaluator ladder are reported at seed 0 and replicate at seed 1 within ± 0.035. The fixed seed 42 is used only for dataset splitting and stratified subsampling, never for model execution. Numerical pipelines are guarded by unit checks against a random-component null.

The full-parameter fine-tune of §7 uses Mistral-Small-24B-Instruct trained on the verdict-format objective with AdamW at learning rate 6×10^{-6}, cosine schedule, warmup ratio 0.03, weight decay 0, gradient clipping 1.0, 3 epochs, micro-batch 2 with gradient accumulation 2 (effective batch 4), sequence length 2048 with sample packing, bf16, gradient checkpointing and flash attention, sharded with FSDP2 over 2×8 B200s. These are the framework defaults; we ran no hyperparameter search, and the shipped configuration is released with the code. The LoRA baseline uses $r = 16$, $\alpha = 32$, dropout 0.05. Probes are L_2-regularized logistic regressions under 5-fold cross-validation with shuffled-label controls.

We release the code, the model registry, per-experiment result artifacts, the evaluation prompts, the 24-cell BUDGET grid, and the multi-node full-FT configuration at `https://anonymous.4open.science/r/llm-judge-internals-B775`. Twelve of the fourteen datasets are public, as is Pariksha (Watts et al., 2024). Samiksha (Hamna et al., 2026) is a restricted human-annotated corpus that we cannot redistribute: its schema, rubric definitions, language coverage, evaluation prompts, and all results computed from it are released, and the corpus itself is available from its custodians under the data-use terms under which it was collected. Every model and dataset is used under the license and terms of use set by its original creators; we redistribute no model weights and no underlying corpus.

A.13 Extended Limitations and Data Statement

Decodability is not causal use. Our probes establish that human-labeled preference is recoverable from frozen activations (L1) and our interventions establish that specific components causally change the emitted verdict (L2); neither establishes that the judge used the probe's direction to form its own verdict (L3, §3). We report a result consistent with that separation, since causal decision depth does not predict probe gain, and avoid "the judge knows" phrasings throughout. Relatedly, our evidence combines forward-pass interventions, run on the registry subset exposing hookable decoder blocks within our compute budget, with cached read-out surrogates run across the full registry and validated against those interventions (median $\rho = 0.78$, positive for $40/41$ judges); we report the agreement between the two so readers can weigh each.

Scope of the read-out and the diagnostic. The read-out is fitted per evaluation distribution and does not transfer naively across benchmarks; cf requires a small labeled sample of that distribution, so it is pre-read-out, not label-free. It captures one concrete source of output-stage error, namely surface confounding, and is not a general predictor of every setting in which internals might help; §6 discusses one such setting reported in prior work. Cost is parameter- and training-efficient rather than uniformly cheap: both readers need two forward passes per item, and residual-stream features must be cached.

Baselines and labels. Fine-tuning baselines use standard published configurations rather than a per-benchmark hyperparameter search, so they are representative reference points rather than upper bounds, and we explicitly do not claim that probing outperforms a well-tuned fine-tuning procedure in general. Human labels are noisy annotator-derived targets, not ground truth; where we report inter-annotator agreement it characterizes label consistency and does not bound the accuracy attainable against one annotator-derived label set.

Data statement. All fourteen datasets are existing, previously published resources and none is introduced by this work. Thirteen are publicly obtainable. Samiksha (Hamna et al., 2026) is not openly redistributable; we release its schema, rubric definitions, language coverage, evaluation prompts, and every result computed from it, and direct interested readers to its authors, who can share it under the terms it was collected under (App. A.5).