WARD: CERTIFYING FAILURES OF LLM JUDGMENT VERIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) are widely used as judges to assess system responses. Often, a second LLM is deployed to check that judgment. This checker, or Verifier, should support correct judgments and withhold support from wrong ones. In this paper, we investigate whether we can provably certify Verifier failures even when ground-truth judgments are not feasible to collect. To illustrate, consider a product review labeled positive by a classifier and reverse its sentiment while keeping the predicted label fixed. The Judge assesses whether this label is correct for each version. Since the two versions have opposite sentiments, the Judge should accept one classification and reject the other. If the Judge accepts both and the Verifier supports both verdicts, at least one verification decision must be wrong, even without knowing which version is positive. To address this, we propose World and Role Diagnostics (WARD), which links each Judge verdict to the Verifier's support decision across such paired cases. We prove that combining these case relationships with the judgment-support matches can establish failures that neither kind of evidence can establish alone. Our theory determines all possible error counts consistent with the evidence, allowing a specified number of case relationships to be wrong. Across five LLMs on six tasks spanning sentiment, stereotypes, helpfulness, safety, and bias-focused question answering in English and Spanish, WARD proves 3.70–6.77 additional errors per 100 decisions for four of the five Verifiers beyond the stronger separate check. This gain repeats on both question-answering tasks.

1 INTRODUCTION

Large language models (LLMs) increasingly judge other systems' responses. Their assessments can guide response selection, search and refinement, making evaluator quality consequential for the system using it (Zhou et al., 2025). A second model may then check a proposed judgment. We call the first evaluator the *Judge* and the checker the *Verifier*. A useful Verifier must both support correct judgments and withhold support from wrong ones. Otherwise, blanket non-support can appear successful at catching errors while offering no support for correct decisions.

The real concern is not only that verification can fail, but that an audit may lack the information needed to *certify* that failure. As an example, consider a product review labeled positive by a classifier. A Judge should accept that label for a positive review and reject it for a negative review. If a controlled edit reverses sentiment while keeping the supplied label and judging criterion fixed, the correct Judge verdicts must be opposite. If the Judge repeats its verdict and the Verifier supports it in both cases, at least one support decision must be wrong. *Knowing that the reviews have opposite sentiment is enough to certify an error, even without knowing which review is positive.* Across larger collections of decisions, however, separate totals of Judge verdicts and Verifier supports may hide which verdicts received support. We therefore study what an audit can establish by linking each Judge verdict to the Verifier's support decision and connecting these records across the paired cases.

Earlier works, such as CriticBench, test whether models can identify errors in incorrect solutions without falsely criticizing correct ones (Lin et al., 2024). Studies on controlled Judge change a response and check whether the Judge reacts when the change matters, while remaining consistent when it does not (Chen et al., 2026a). Research on diagnostic testing has derived upper and lower bounds on how well a test can distinguish positive from negative cases when the reference test

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.

can also make mistakes (Obradović, 2024). However, these studies do not answer whether linking judgments to the support they receive from the Verifier across paired cases can lead to improved certification of Verifier errors, *even when ground-truth judgments are unknown*. Such an analysis would let auditors expose verification failures when *independently checking every judgment is impractical*.

To this end, we propose *World and Role Diagnostics* (WARD), an audit framework that combines two pieces of information: how the correct Judge verdict should change across paired cases, and which verdicts the Verifier supports. Our product review example illustrates why both matter. Suppose two Judges assess the original and sentiment-reversed classifications: Judge 1 accepts both, while Judge 2 rejects both. A Verifier supports Judge 1 and withholds support from Judge 2 in each case, making four verification decisions. For the review whose classification is wrong, the Verifier therefore supports an incorrect verdict and withholds support from a correct one. An audit using only one of these two pieces of information, i.e., two *separate checks*, cannot establish that any verification error occurred. In contrast, the paired cases and verdict-support links prove that two verification decisions are wrong, even without identifying which review is negative. WARD also determines what the available evidence lets us conclude about verification. When the correct verdicts are unknown, several explanations of the observed decisions may remain possible. We retain all explanations consistent with the evidence and determine how many verification errors each would imply. A specified verification requirement is certified only if every possible explanation satisfies it, ruled out if none does, and left unresolved otherwise. *The audit thus distinguishes demonstrated success, demonstrated failure, and insufficient evidence.* We prove that this analysis requires only four counts for each case, i.e., how often Accept and Reject verdicts receive support or non-support. These summaries preserve all possible combinations of verification successes and errors and yield the smallest and largest error rates consistent with the evidence.

We evaluate WARD on six judgment tasks spanning human-edited sentiment reversals in CAD (Kaushik et al., 2020), stereotype contrasts in CrowS-Pairs (Nangia et al., 2020), separate helpfulness and safety judgments over PKU-SafeRLHF response pairs (Ji et al., 2025), and controlled bias-focused QA reversals in English BBQ and Spanish ES-BBQ (Parrish et al., 2022; Ruiz-Fernández et al., 2025). Together, they vary criterion, construction, and language, testing the same audit beyond one setting. We use five open-weight 8B-32B checkpoints from the DeepSeek, Mistral, Qwen, and Gemma families, each acting as both Judge and Verifier. For four of five Verifiers, linking the World- and Role-test records certifies that an additional 3.70-6.77 decisions per 100 must be wrong beyond what the stronger separate check can establish. This gain appears in both QA settings, reaching 6.12-7.86 additional errors per 100 on BBQ and 2.45-8.60 on ES-BBQ. Role behavior also changes sharply. Among the four models not dominated by ties, paired Judge and Stage-A Verifier verdicts agree on both cases for only 12.18-40.18% of BBQ pairs and 1.83-25.90% of ES-BBQ pairs. WARD therefore exposes errors that separate checks miss across tasks and languages. In systems that accept supported judgments and escalate unsupported ones, these failures can accept wrong judgments or send correct ones to unnecessary review.

Our contributions are (i) an information-based classification of contemporary verification audits with explicit decision semantics and realizable insufficiency results, (ii) WARD's linked construction and its exact attainable-count guarantee under reference uncertainty, and (iii) an empirical demonstration that combining the evidence changes what an audit can decide.

2 Related Work

Critique And Verification. CriticBench tests error identification, including cross-model checks (Lin et al., 2024). CriticEval evaluates critiques against human ratings and references (Lan et al., 2024), while IF-CRITIC separates explanation correctness from judgment consistency (Wen et al., 2026). WARD instead asks which verification errors are certifiable from case relations when individual correct verdicts are unknown.

Formal Auditing Foundations. Diagnostic testing handles imperfect references (Obradović, 2024), and LLM-panel identification distinguishes separate counts from linked responses (Chen et al., 2026b). Logical accuracy estimation, metamorphic testing and active fairness auditing use constraints or queries to reduce uncertainty without complete labels (Platanios et al., 2017; Xie et al., 2011; Yan & Zhang, 2022). WARD specializes these ideas to linked judgment–support records and asks when the linkage changes the verification conclusion.

Controlled Evaluator Tests. ContextualJudgeBench evaluates explanations of correct judgments (Xu et al., 2025a). Construct-validity studies, contrast sets and CheckList test responses to controlled changes (Chen et al., 2026a; Gardner et al., 2020; Ribeiro et al., 2020). WARD joins such case relations to support for Judge verdicts. Its limits concern the specified evidence views. Equivalent joint evidence supports the same guarantee. Appendix B.1–B.2 gives the detailed comparisons.

3 Problem Formulation

Verification Settings. An LLM-based evaluation begins with a task input x. The evaluated task-specific decision function, i.e. the *task head*, produces an output u. A criterion κ states what makes that output acceptable. For example, x can be a product review, the task head can label it *positive*, and κ can require that the label match the review's ground-truth sentiment. An LLM-Judge assesses u under κ and returns a *verdict* $j \in \{-1, +1\}$, where $+1$ means ACCEPT and -1 means REJECT. All Judges assessing the same case $e = (x, u, \kappa)$ share one correct *target verdict* $y \in \{-1, +1\}$. In many cases, an LLM-Verifier is used that returns $b \in \{0, 1\}$ for a particular Judge verdict, where 1 means SUPPORT and 0 means non-support (i.e., WITHHOLD). Each Verifier decision has one of four outcomes. Supporting a correct Judge verdict is a true positive (TP), supporting a wrong verdict is a false positive (FP), withholding support from a correct verdict is a false negative (FN), and withholding support from a wrong verdict is a true negative (TN). We collect their counts as confusion vector $q = (\mathrm{TP}, \mathrm{FP}, \mathrm{FN}, \mathrm{TN})$.

Auditor Settings. An Auditor is supposed to collect these cases (i.e., (x, u, κ)), together with their corresponding Judge verdicts, Verifier support decisions, and then *certify* the Judge-Verifier pair. In practice, however, the Auditor may not know the correct verdict y for every case. It may instead know only *relations* between verdicts. For example, a known relation to the Auditor can be that the original review (O) and a sentiment-reversed constructed review (A) must have opposite correct verdicts (i.e., $y^A = -y^O$), without having to know whether y^O should be ACCEPT, or vice versa. In other words, there are exactly two possible *hidden* ground truths (called *candidate assignments*): $a_1 : (y^O = +1, y^A = -1)$ or $a_2 : (y^O = -1, y^A = +1)$. A candidate assignment needs to explain the observed audit evidence o without contradicting it. More specifically, they need to be *consistent* with the known relation. In the review example, $a_1' : (y^O = +1, y^A = +1)$ or $a_2' : (y^O = -1, y^A = -1)$ is inconsistent w.r.t. the known relation. A *truth-completed record* $\zeta = (o, a_i)$ is the already observed audit evidence o along with one such candidate assignment a_i. $\zeta_1 \in Z(o), \forall a_i \in \mathcal{A}(o)$, where $Z(o)$ is the set of all possible records for the corresponding set of consistent assignments $\mathcal{A}(o)$.

Auditor Certification. Suppose there are two Judges, J1 and J2, and one Verifier. The observed Judge-Verifier behavior is the same in both worlds: for each $w \in \{O, A\}$, J1 gives $j_1^w = +1$ and is supported, $b_1^w = 1$, while J2 gives $j_2^w = -1$ and is not supported, $b_2^w = 0$. Thus, J1 and J2 disagree with each other, but each Judge is individually invariant across the two worlds. The Auditor observes all four verification decisions, but does not know which of the two candidate hidden ground truths is the actual one: $(y^O = +1, y^A = -1)$ or $(y^O = -1, y^A = +1)$. Each candidate assignment tells the Auditor how to classify those same four observed verification decisions as TP, FP, FN, or TN. Under the first candidate $(y^O = +1, y^A = -1)$, the two decisions in O are a TP and a TN, while the two decisions in A are an FP and an FN. Under the second candidate $(y^O = -1, y^A = +1)$, the roles of O and A simply swap. Therefore, both candidates induce the same confusion vector: $q_1 = q_2 = (1, 1, 1, 1)$, with one TP, one FP, one FN, and one TN. Hence, regardless of which candidate's hidden ground truth is the real one, $\mathrm{FP}+\mathrm{FN} = 2$. In other words, even without knowing *which world* contains the two verification errors and *even if we choose, among all candidate hidden ground truths allowed by the evidence, the one that makes the Verifier's correctness to be the best*, the Auditor can still *provably certify* that *two of the four verification decisions must be wrong*. To generalize, if two possible candidate assignments produce compatible confusion vectors q_1 and q_2, such that $\mathrm{FP}(q_1)+\mathrm{FN}(q_1) \geq 2$ and $\mathrm{FP}(q_2)+\mathrm{FN}(q_2) \geq 2$, then, $\min_{q \in \Theta(o)}(\mathrm{FP}+\mathrm{FN}) = 2$, where $\Theta(o) = \{q(\zeta) : \zeta \in \mathcal{Z}(o)\}$ is the set of all confusion vectors that could still be the true confusion vector, given o. Note that $\Theta(o)$ can be interpreted as the *uncertainty* about Verifier's confusion counts. More generally, the Auditor must decide what verification performance counts as acceptable. Consider a fixed set of audited Judge verdicts termed *cohort*. For tolerances $\alpha, \beta \in [0, 1]$, we define

Table 1: Composable sources of audit information. Observation views specify which decision matches are recorded, reference restrictions specify what is known about the correct verdicts, and queries can augment either. The central composed views are R_M (relations with separate totals) and R_L (relations with verdict-to-decision links). The examples identify neighboring evidence types.

Ingredient	Type	Information retained or assumed	Information Limit (w/ examples)
Marginal M	Observation	Separate totals of Judge verdicts and Verifier decisions	A margin-only projection loses which decision belongs to each verdict
Linked L_0	Observation	Each support or non-support decision remains attached to its Judge verdict	No relation or exact label constrains the correct verdicts
Relational R	Reference	Selected cases must have the same or opposite correct verdicts	Metamorphic testing supplies such relations (Xie et al., 2011). The relation need not identify either verdict
Calibrated K	Reference	A stated reliability or sampling model restricts reference errors	Diagnostic identification uses calibrated imperfect references (Obradović, 2024). Its guarantee inherits those assumptions
Anchored G	Reference	Exact correct verdicts are supplied for identified cases	CriticBench uses reference-grounded critique labels (Lin et al., 2024). Labels constrain only covered or implied cases
Active $Q(C)$	Acquisition	Query answers augment a base information class C	Active Fairness Auditing acquires selected labels (Yan & Zhang, 2022). Conclusions depend on the answers obtained

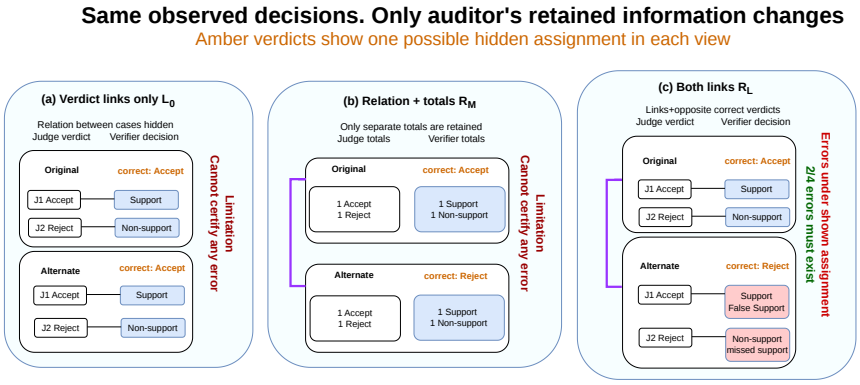


Figure 1: *Each partial view permits zero error, whereas their combination proves two of four decisions wrong (Theorem 1).* Solid lines preserve verdict-support links; brackets require opposite correct verdicts. Amber labels mark one compatible assignment and red cells its errors.

the *acceptable count set* as:

$$\mathcal{C}_{\alpha,\beta} = \left\{ q : \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}} \leq \alpha, \quad \frac{\mathrm{FN}}{\mathrm{TP}+\mathrm{FN}} \leq \beta \right\}. \tag{1}$$

α bounds the fraction of wrong judgments that receive support, and β bounds the fraction of correct judgments that do not. As an example, setting $\alpha = \beta = 0.3$ requires non-support for at least 70% of wrong judgments and support for at least 70% of correct judgments. *The two requirements prevent an always-support or never-support Verifier from appearing reliable.* Under a support-or-review policy, false support can cause a wrong judgment to be accepted, whereas missed support sends a correct judgment for unnecessary review. For nonempty $\Theta(o)$ and a defined conditional-rate target, the audit returns PASS when $\Theta(o) \subseteq \mathcal{C}_{\alpha,\beta}$, FAIL when $\Theta(o) \cap \mathcal{C}_{\alpha,\beta} = \varnothing$, and UNRESOLVED otherwise. An empty set signals contradictory observations or assumptions. Note that these conclusions concern the evaluated finite cohort. Population inference requires a separate sampling model. The target is non-vacuous because a Verifier that supports exactly the correct judgments has zero error. The formulation does not assume that an evaluated Verifier can identify those judgments.

Audit Information Regimes. The information available to the Auditor has two aspects. The *observation view* determines whether the Auditor retains only separate totals of Judge verdicts and Verifier decisions, or also their case-level pairing - i.e., which Verifier decision belongs to which Judge verdict. On the other hand, its *reference information* determines what the Auditor knows about the unknown correct verdicts. It may know nothing about them, know only relations between

them (e.g., $y^A = -y^O$), have quantified assumptions about the reliability of such information, or know exact correct verdicts for particular cases. Based on these aspects, we can categorize existing audits into eight regimes - (i) *Marginal view* (M) that only contains separate totals of Judge verdicts and Verifier decisions; (ii) *Linked view* (L_0) where each Judge verdict is linked to its own Verifier decision, i.e. (j_i, b_i); (iii) *Relational view* (R) that contains relations among *hidden* correct verdicts, e.g. $y^A = -y^O$; (iv) R_M that combines M with relation information R; (v) R_L that combines L_0 with relation information R; (vi) *Calibrated restrictions* (K) with quantified limits on how *reliable* the reference relations are; (vii) *Exact anchors* (G) with ground-truth references for some cases; and (viii) *Acquired answers* ($Q(C)$) that contain additional answers acquired to refine an existing information state C. If the evidence is consistent with both a PASS and a FAIL possibility, then it is not a *sound* audit regime. In other words, the Auditor should give a definite answer only when every still-possible candidate assignment leads to that same answer. We provide the formal condition needed for a decisive audit as follows:

Proposition 1 (Information Required For A Decisive Audit)**.** *Suppose the conditional-rate target is defined for every record in $\mathcal{Z}(o)$. If $\mathcal{Z}(o)$ contains records ζ_0 and ζ_1 such that $q(\zeta_0) \in \mathcal{C}_{\alpha,\beta}$ and $q(\zeta_1) \notin \mathcal{C}_{\alpha,\beta}$, no sound audit using only o can return either* PASS *or* FAIL*. Conversely, over any collection of observations for which the target is defined, a sound decisive rule exists exactly when all records within each nonempty set $\mathcal{Z}(o)$ agree on whether their count vector belongs to $\mathcal{C}_{\alpha,\beta}$.*

Hence, *an audit must preserve the information that distinguishes the conclusions it is intended to certify.* The same limitation holds even when the Auditor chooses each new query based on previous answers. In other words, if two possible hidden ground truths produce the same sequence of queries and answers, the Auditor still cannot distinguish between them. In Appendix A.1, we prove the decisive audit condition for deterministic, randomized, and sequential query-based cases.

Information Limits. The R_M regime loses the case-level pairing between Judge verdicts and Verifier decisions. For example, consider four Judge verdicts, two correct and two wrong, of which two receive support. The same separate totals are compatible with supporting the two correct verdicts, yielding no errors, or supporting the two wrong verdicts, yielding four errors. The L_0 regime has the complementary limitation: it preserves the Judge-Verifier pairing but lacks cross-case information about the hidden correct verdicts. Thus, each restricted view can remain compatible with an error-free hidden ground truth even when their combination is not. This raises a crucial question: *can relation information and Judge-Verifier correspondence each be insufficient on their own, yet become decisive when combined*? We show that they can as follows:

Theorem 1 (Complementary Evidence)**.** *There is a constructed record with four verification decisions for which R_M and L_0 each permit zero error, whereas R_L proves that at least two of the four decisions are wrong. Without an exact reference label, no record in this complement-symmetric binary model can guarantee a minimum error rate above one half.*

The theorem has two implications. First, R_L can certify what neither restricted view can alone: in the four-decision construction, both R_M and L_0 admit an error-free hidden ground truth, whereas R_L rules out all such possibilities and certifies at least two errors. Thus, the relation and Judge-Verifier correspondence become informative jointly. Second, without an exact reference label, regimes such as M, L_0, R, R_M, and R_L may leave both a hidden ground truth y and its global flip $-y$ possible. The flip leaves the observed records unchanged but swaps which Judge verdicts are correct and wrong. When $\alpha + \beta < 1$, the Auditor therefore cannot certify that both false-positive and false-negative requirements are satisfied, although it can still certify failure if every possible hidden ground truth violates at least one requirement. The resulting $1/2$ ceiling limits what the Auditor can *guarantee*, not how many errors may actually exist; an exact label can break this symmetry. Figure 1 visualizes the construction. Appendix A.1 gives the marginal-view construction, Proposition 3 in Appendix A.5 proves the complement-symmetry limit on success certification, and Theorem 4 in Appendix A.2 gives the full proof including the one-half upper limit. Table 1 summarizes the regimes and their limits; Appendices B.1 and B.2 compare these guarantees with concrete methods.

4 THE WARD AUDIT METHOD

In this paper, we propose WARD as a sound audit method that constructs the case relations and verdict-to-decision links identified in Section 3. For pair i and world $w \in \{O, A\}$, a case $e_i^w = (x_i^w, u_i^w, \kappa_i)$ contains the task input, supplied task-head output, and judging criterion.

World Test and Role Test. WARD constructs an *Original World (OW)* and *Alternate World (AW)* with opposite correct Judge verdicts, $y_i^A = -y_i^O$, by changing the input or supplied response while keeping the criterion fixed. The *World Test* asks an evaluated LLM, acting as a Judge, to score the case in both worlds. A follow-up *Role Test* asks whether an LLM acting as a Verifier ultimately supports a correct Judge verdict and withholds support from a wrong one. In Stage A, the Verifier forms a fresh verdict from the case and criterion alone, without Judge scores or reference labels; this verdict is elicited separately from the World-Test verdict and then frozen. In Stage B, the Verifier receives this frozen verdict and one Judge's ACCEPT and REJECT scores, assesses whether they agree, and returns a support score. Positive values mean support; zero or negative values mean non-support. For the binary certificate, the larger Judge score determines the verdict to which this decision is attached, and ties are excluded. These verdict-to-decision correspondences, together with the OW/AW opposite-verdict relation, form the relational linked view R_L. Note that Stage B can be consistent with Stage A while the final verification decision is still wrong. Suppose, for example, that the correct verdict is REJECT, but both Stage A and the Judge say ACCEPT. Stage B may correctly recognize their agreement and support the Judge, yet that support is attached to a wrong judgment. The central certificate does not require the correct verdict for every case. Instead, it considers every hidden-ground-truth assignment allowed by the paired-case relations and any available exact labels, and asks what must hold across all of them. It can therefore certify errors in the final support decisions without attributing each error specifically to Stage A or Stage B.

Count Representation. The Role Test leaves fixed Judge verdicts and support/non-support decisions, while the correct verdict may remain unknown. To evaluate these observations under either candidate correct verdict, WARD summarizes each case and Verifier using four counts: n, the number of Judge verdicts; n_{acc}, the number that say ACCEPT; n_{sup}, the number receiving support; and $n_{\mathrm{acc,sup}}$, the number of supported ACCEPT verdicts. These counts retain the Judge–Verifier correspondence needed to determine the confusion vector once a candidate correct verdict $y \in \{-1, +1\}$ is specified. Let $g(y)$ denote this vector in the order $(\mathrm{TP}, \mathrm{FP}, \mathrm{FN}, \mathrm{TN})$:

$$\begin{aligned} g(+1) &= (n_{\mathrm{acc,sup}},\ n_{\mathrm{sup}} - n_{\mathrm{acc,sup}},\ n_{\mathrm{acc}} - n_{\mathrm{acc,sup}},\ n - n_{\mathrm{acc}} - n_{\mathrm{sup}} + n_{\mathrm{acc,sup}}), \\ g(-1) &= (n_{\mathrm{sup}} - n_{\mathrm{acc,sup}},\ n_{\mathrm{acc,sup}},\ n - n_{\mathrm{acc}} - n_{\mathrm{sup}} + n_{\mathrm{acc,sup}},\ n_{\mathrm{acc}} - n_{\mathrm{acc,sup}}). \end{aligned} \tag{2}$$

In Figure 1(c), each case has $(n, n_{\mathrm{acc}}, n_{\mathrm{sup}}, n_{\mathrm{acc,sup}}) = (2, 1, 1, 1)$. If $y = +1$, the supported Judge is correct and the unsupported Judge is wrong, giving $g(+1) = (1, 0, 0, 1)$. If $y = -1$, the same two observed decisions are both wrong, giving $g(-1) = (0, 1, 1, 0)$. Thus, the observations stay fixed while the candidate hidden ground truth determines how they count as TP, FP, FN, and TN. These four counts recover every attainable confusion vector. Judge identities may be aggregated when Judge-specific conclusions are unnecessary, but case identity must remain because the OW/AW relation couples the two cases; self-versus-peer or subgroup certificates therefore require separate summaries. Lemma 1 in Appendix A.1 proves the representation.

Imperfect Relations. So far, the OW/AW relation has been treated as correct. A controlled edit intended to reverse the correct verdict may, however, fail to do so. WARD therefore allows a declared number of pair relations to be *released*: the audit stops relying on those relations without asserting that they are wrong. For an unanchored pair i, enforcing $y_i^A = -y_i^O$ permits only $(+1, -1)$ and $(-1, +1)$; releasing it additionally permits $(+1, +1)$ and $(-1, -1)$. Exact labels remove any assignments that contradict them. Let $\mathcal{P}_i^0$ contain the confusion-count vectors possible when relation i is enforced and $\mathcal{P}_i^1$ those possible when it is released. Write $\rho_i = 1$ for a released relation and $\rho_i = 0$ otherwise, and let $\Gamma \leq n_{\mathrm{pair}}$ bound the number of releases. The resulting attainable set is

$$\Theta_\Gamma = \bigcup_{\substack{\rho \in \{0,1\}^{n_{\mathrm{pair}}} \\ \sum_i \rho_i \leq \Gamma}} \left\{ \sum_i q_i : q_i \in \mathcal{P}_i^{\rho_i} \right\}. \tag{3}$$

The Auditor need not know which relations are unreliable. For example, $\Gamma = 1$ allows whichever single release most weakens the certificate. Hence, an error lower bound that survives every $q \in \Theta_\Gamma$ remains valid even after disregarding up to Γ declared relations in the way most favorable to the Verifier.

Product Admissibility. Equation 3 combines one allowed vector from each OW/AW pair. This requires that their underlying hidden-ground-truth assignments can coexist without contradiction, which we call *product admissibility*. If two pairs share the same original case, for example, both

must assign that case the same correct verdict and must therefore be grouped or explicitly linked. Product admissibility concerns compatibility of candidate ground truths across pairs, not statistical independence of model errors.

Pooling Across Tasks. For several tasks, WARD constructs $\Theta^{(t)}_{\Gamma_t}$ separately for each $t \in \mathcal{T}$ and forms

$$\Theta_{\text{pool}} = \left\{ \sum_{t \in \mathcal{T}} q_t : q_t \in \Theta^{(t)}_{\Gamma_t} \right\}.$$

Each pooled vector therefore combines one allowed task-level vector. This differs from our equal-task-weighted summaries, where the relevant rate or gain is computed within each task and the task values are then averaged equally rather than derived from the summed counts in Θ_{pool}.

Sharp Bounds From Attainable Sets. Once an attainable set is known, the Auditor can bound any selected count or weighted combination of counts. Let λ specify the desired quantity and define $H_\Theta(\lambda) = \max_{q \in \Theta} \lambda^\top q$, the largest value still compatible with the evidence. For one Verifier, $\lambda = (0, 1, 1, 0)$ selects total verification error, FP + FN, while $-H_\Theta(-\lambda)$ gives its smallest attainable value. A bound is *sharp* when some allowed hidden-ground-truth assignment actually attains it. Releasing a relation creates additional assignments and can weaken such a bound. Let $\Delta_i(\lambda)$ measure the increase in the selected upper bound caused by releasing relation i; under a budget Γ, the most conservative audit therefore allows the releases that weaken the bound most.

Theorem 2 (Attainable Counts And Sharp Certificates)**.** *Assume accurately recorded OW/AW pair identities and Judge–Verifier correspondence, one correct verdict shared by all Judges of a case, product admissibility across pairs, and the declared relation and exact-label restrictions. Then Equation 3 gives exactly the set of joint confusion-count vectors that remain possible. For direction λ, define $\psi^s_i(\lambda) = \max_{q \in \mathcal{P}^s_i} \lambda^\top q$ for relation status $s \in \{0, 1\}$. If every nominal set $\mathcal{P}^0_i$ is nonempty, the sharp upper bound is*

$$H_\Gamma(\lambda) = \sum_i \psi^0_i(\lambda) + \sum_{k=1}^{\min(\Gamma, n_{\text{pair}})} \Delta_{(k)}(\lambda), \qquad \Delta_i(\lambda) = \psi^1_i(\lambda) - \psi^0_i(\lambda) \geq 0, \tag{4}$$

where (k) orders the gains $\Delta_i(\lambda)$ from largest to smallest. The sharp lower bound is $-H_\Gamma(-\lambda)$, and every finite endpoint is attained by at least one allowed hidden-ground-truth assignment.

The theorem gives two guarantees. First, Θ_Γ is exact: it contains every attainable confusion vector and no impossible one. Second, under the stated nonemptiness condition, the most conservative effect of Γ uncertain relations is obtained by releasing those that weaken the selected bound most. Because the endpoints are sharp, each is attained by an allowed hidden ground truth. Thus, if WARD reports that at least k errors are unavoidable, no allowed hidden verdicts or relation releases can reduce the error below k. Appendix A.3 proves exact attainability, gives the $O(n_{\text{pair}}\Gamma)$ dynamic program for the general case including empty nominal sets, and derives the $O(n_{\text{pair}} \log n_{\text{pair}})$ sorted-gain calculation under the theorem's nonemptiness condition. The optimization follows established budgeted-uncertainty principles (Bertsimas & Sim, 2004).

Certification Guarantees. The attainable set states which confusion counts remain possible; certification asks what conclusion holds for all of them. Let $\Theta_\star = \Theta_\Gamma$ for one task and $\Theta_\star = \Theta_{\text{pool}}$ for a pooled-count certificate, and let $\Theta_{\star,v}$ denote the four coordinates $(\text{TP}, \text{FP}, \text{FN}, \text{TN})$ attainable for Verifier v. A universal PASS requires every $q \in \Theta_{\star,v}$ to satisfy both requirements in Equation 1. For example, $\frac{\text{FP}}{\text{FP}+\text{TN}} \leq \alpha \iff (1-\alpha)\text{FP} - \alpha\text{TN} \leq 0$, with the analogous condition $(1-\beta)\text{FN} - \beta\text{TP} \leq 0$ for missed support. Theorem 2 can therefore maximize each left-hand side over all still-possible confusion vectors. Both rates must be evaluated using the same hidden-ground-truth assignment; the Auditor cannot combine the most favorable false-positive rate from one assignment with the most favorable false-negative rate from another. WARD can also certify FAIL from the minimum unavoidable error. Let $N_{\text{dec}} = \text{TP} + \text{FP} + \text{FN} + \text{TN}$ and $\underline{E}_{\star,v} = \min_{q \in \Theta_{\star,v}} (\text{FP} + \text{FN})$. If both conditional error rates were at most ϵ, their errors could sum to at most ϵN_{dec}. Hence, $\underline{E}_{\star,v} > \epsilon N_{\text{dec}} \implies \Theta_{\star,v} \cap \mathcal{C}_{\epsilon,\epsilon} = \varnothing$. Thus, even if the hidden ground truth, most favorable to the Verifier, forces more than an ϵ fraction of decisions to be wrong, no allowed assignment can satisfy both requirements. For example, if both error rates must be at most 30%, a guaranteed error floor above 30% certifies FAIL. If neither universal PASS nor certified FAIL can be established, the result is UNRESOLVED: the evidence still permits both a passing and a failing hidden ground truth and is therefore insufficient to decide between them. For exact tests see Appendix A.4.

5 EXPERIMENTAL SETUP

We evaluate six tasks: CAD sentiment reversals (Kaushik et al., 2020), CrowS-Pairs stereotype contrasts (Nangia et al., 2020), PKU-SafeRLHF paired responses under helpfulness and safety criteria (Ji et al., 2025), and bias-focused QA reversals in English BBQ and Spanish ES-BBQ (Parrish et al., 2022; Ruiz-Fernández et al., 2025). The World cohort contains 18,039 pairs; Appendix E gives constructions and counts. Five open-weight checkpoints span 8B–32B parameters: DeepSeek-R1-8B (DeepSeek-AI, 2025), Mistral-Nemo-12B (Mistral AI Team, 2024), Qwen2.5-14B/32B (Yang et al., 2025), and Gemma-2-27B (Gemma Team et al., 2024). Each serves as Judge and Verifier, giving 25 ordered combinations; Qwen14 and Qwen32 denote the Qwen checkpoints.

Information And Coverage. Stage A sees only the case and criterion. Stage B adds the frozen A verdict and one Judge's acceptance/rejection scores. Neither sees references, Judge identity or rationale, dataset/world identity, or the paired world. Four PKU pairs lack a complete B grid, leaving 18,035 complete pairs and 901,750 Stage-B decisions. Appendix F gives coverage and inference details; Appendix I.3 gives the historical comparison.

Scoring And Uncertainty. Judge verdicts use the sign of acceptance minus rejection; ties within 10^{-12} are excluded from certificates and count as diagnostic errors. B scores > 0 denote support. Unless explicitly pooled, rates are computed within task and averaged equally; pooled sensitivity applies its relation budget within each task before summing counts. We compute 95% intervals from 2,000 linked-identity bootstrap resamples of fixed outputs, retaining both worlds, all Judge/Verifier cells and shared PKU identities across criteria.

6 RESULTS

Linking Records Exposes Hidden Errors. We find that *linking certifies 3.70-6.77 additional errors per 100 decisions beyond the stronger separate check* for Mistral, Qwen14/32 and Gemma after equal-task averaging (Table 2). The gain appears on all six tasks (Figure 2a). On BBQ it is 6.12-7.86 additional errors per 100 decisions and on ES-BBQ 2.45-8.60; the largest task-level gain is 18.02 for Mistral on PKU-Helpfulness. DeepSeek is the boundary case where blanket non-support already forces the relation-only floor, so linking adds no further certified error.

Table 2: Guaranteed minimum verification error (%). R_M uses relations, L_0 uses verdict-support links, and R_L uses both. Extra (% points) subtracts the stronger separate floor from R_L within each pair, before within-task pooling and equal-task averaging.

Verifier	R_M	L_0	R_L	**Extra**
DeepSeek	37.94	14.42	37.94	0.00
Mistral	28.27	14.17	35.55	6.37
Qwen14	26.22	11.06	30.89	3.70
Qwen32	19.52	8.06	27.11	6.77
Gemma	21.17	7.61	28.88	6.77

Certificate Survives Imperfect Relations. We see that allowing up to 10% invalid relations within each task still guarantees 19.33-33.58% errors across 153,269 decisions per Verifier (Figure 2b). These floors retain 79-87% of their all-trusted values across the five Verifiers, a reduction of 13-21% under the declared budget. The 10% budget is a sensitivity assumption, not an estimate of relation quality or future deployment rates.

Same Model Changes Verdicts With Role Change. We compare the same checkpoint's World-Test Judge verdicts with its independently elicited Stage-A Verifier verdicts on both cases of each pair (Figure 2c). Among the four models not dominated by World ties, agreement remains 75.4-89.0% on CAD, compared with only 12.2-40.2% on BBQ and 1.8-25.9% on ES-BBQ. *Role assignment therefore changes the same checkpoint's judgments in a strongly task-dependent way.* DeepSeek's tie-heavy cells are separated.

Similar World Failures Hide Different Verification Profiles. To study this, we use HFR that measures how often a Judge's verdict that is correct in OW becomes wrong after the required OW→AW verdict reversal. We see that Qwen14&32 have nearly identical HFRs (76.5% and 75.5%) but very different verification behavior. Wrong-self non-support is 31.5% vs. 53.7%, while correct-self support is 81.7% vs. 72.4% (Figure 2e). DeepSeek is more extreme where 92.5% HFR coexists with 0% correct-self support and 100% wrong-self non-support. *Thus, similar World-Test failure can*

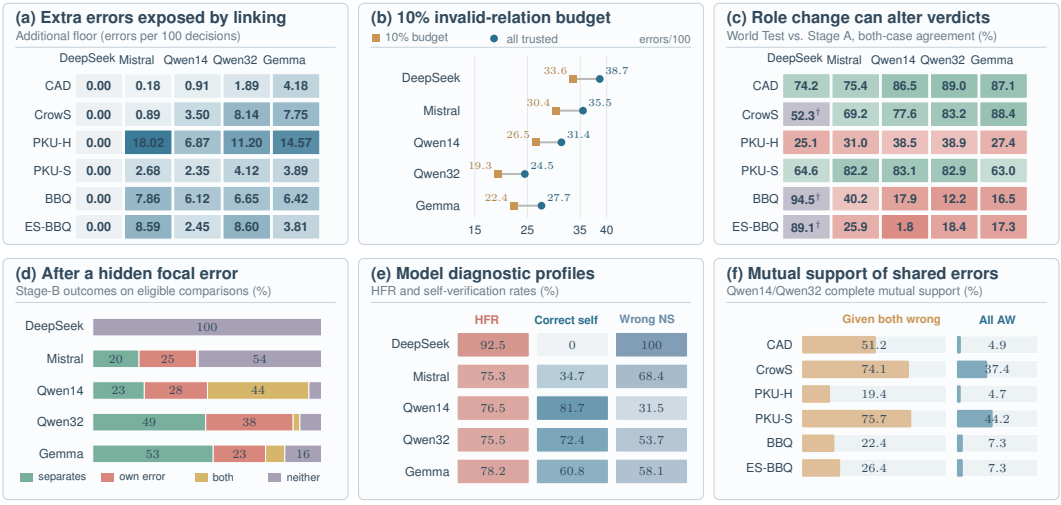


Figure 2: *Linked evidence exposes hidden errors and reveals distinct verification failures.* (a) Extra error certified by linked case relations and verdict–support records. (b) Error floor with a 10% invalid-relation budget versus all relations trusted. (c) World-Test/Stage-A both-case agreement; daggers mark tie-heavy World outputs. (d) Stage-B outcomes after a hidden focal error paired with a correct peer. (e) HFR, correct-self support and wrong-self non-support. (f) Qwen14/Qwen32 complete mutual support (CMS) by task, conditional on both wrong and over all AW cases. Task-macro summaries weight six tasks equally; full counts are in the appendix.

hide very different verification profiles, and a high wrong-judgment rejection rate can reflect blanket non-support rather than useful two-sided verification.

Later Verification & Peers Do Not Reliably Repair The Failures. Stage A gets both verdicts correct in only 9.2-38.9% of pairs. For hidden focal errors paired with correct peers, correct Stage-B discrimination occurs in only 0-52.6% of eligible comparisons (Figure 2d). Qwen14 instead supports both judgments in 44.3%, while DeepSeek supports neither in 100%. Peers can also support shared errors. Qwen14 and Qwen32 mutually support both errors in 44.9% of jointly wrong AW cases, covering 17.6% of all AW cases; conditional CMS varies sharply by task, from 19.4% on PKU-Helpfulness to 75.7% on PKU-Safety (Figure 2f). This is mutual support of shared errors, not evidence of collusion. Appendix J gives counts and the full cross-model results.

7 Discussion And Limitations

What an audit records determines what it can prove. Across six tasks and two languages, linked evidence exposes errors hidden by either record alone. Guarantees concern fixed binary records with correct correspondence and declared relation assumptions; the budget analysis relaxes but does not validate those relations. Five checkpoints and one prompt family do not establish universal or future reliability. WARD neither locates nor repairs individual errors; deployment value depends on valid, affordable pair construction.

8 Conclusion

In this paper, we propose WARD as a sound audit method for certifying LLMs as Verifiers. WARD proves verification errors that separate checks leave unresolved. Across six tasks, linking proves 3.70-6.77 additional errors per 100 decisions for four of five Verifiers. World-Test and Stage-A disagreement also shows that behavior in one role cannot stand in for another. This suggests that *preserving how evidence fits together changes what an audit can establish.*

Ethics Statement

The study includes stereotype, helpfulness and safety evaluation. Its findings characterize behavior under specified constructions. They are not universal judgments about models or demographic groups. The use of dataset labels retains each source's intended criterion and does not establish a comprehensive fairness or safety measure. Redistribution must respect the applicable dataset and model terms.

Reproducibility Statement

The results use frozen Judge scores and two-stage Verifier outputs. World analyses retain 18,039 pairs across six tasks. Paired Role and combined analyses use 18,035 complete pairs and 901,750 Stage-B decisions. Appendix F documents the inference interface, execution settings, parsing and record coverage. Appendix K marks secondary analyses that remain restricted to the original four-task subset. Linked-identity resampling uses fixed outputs and frozen seeds. Appendix G specifies the information ablations and relation budgets. The numerical checks reproduce the calculations on the declared fixed-output cohorts; they do not assess variation across new model generations.

AI Use Statement

We used AI to identify relevant literature and edit portions of the manuscript. AI-assisted feedback also supported checks of clarity. The authors take full responsibility for all mathematical arguments, experimental protocols, numerical results, and bibliographic references against the available research materials and primary sources.

References

Dimitris Bertsimas and Melvyn Sim. The price of robustness. *Operations Research*, 52(1):35–53, 2004. doi: 10.1287/opre.1030.0065.

Hongyu Chen and Seraphina Goldfarb-Tarrant. Safer or luckier? LLMs as safety evaluators are not robust to artifacts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 19750–19766, 2025. doi: 10.18653/v1/2025.acl-long.970.

Jianlin Chen, Wenhui Chen, Ziyao Lin, and Chi Man Vong. A judge should know what changed: Construct validity for LLM-as-a-judge evaluation, 2026a. URL https://arxiv.org/abs/2608.24419v1. Preprint, version 1, 25 August 2026.

Xiaohong Chen, Ashesh Rambachan, and Elie Tamer. Partial identification from LLM prompts, 2026b. URL https://arxiv.org/pdf/2606.15031v2. Preprint, version 2.

Michael R. Clarkson and Fred B. Schneider. Hyperproperties. In *21st IEEE Computer Security Foundations Symposium*, pp. 51–65, 2008. doi: 10.1109/CSF.2008.7. URL https://www.cs.cornell.edu/fbs/publications/HyperpropertiesCSFW.pdf.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948v1. Technical report, version 1.

Werner Dinkelbach. On nonlinear fractional programming. *Management Science*, 13(7):492–498, 1967. doi: 10.1287/mnsc.13.7.492.

Irit Dinur, Prahladh Harsha, Tali Kaufman, and Noga Ron-Zewi. From local to robust testing via agreement testing. *Theory of Computing*, 18(12):1–25, 2022. doi: 10.4086/toc.2022.v018a012. URL https://theoryofcomputing.org/articles/v018a012/.

Adrian Dobra and Stephen E. Fienberg. Bounds for cell entries in contingency tables given marginal totals and decomposable graphs. *Proceedings of the National Academy of Sciences*, 97(22):11885–11892, 2000. doi: 10.1073/pnas.97.22.11885. URL https://sites.stat.washington.edu/adobra/Files/Papers/pnas.pdf.

Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1307–1323, 2020.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size, 2024. URL https://arxiv.org/abs/2408.00118v3. Technical report, version 3.

David Hartmann, Lena Pohlmann, Lelia Hanslik, Noah Gießing, Bettina Berendt, and Pieter Delobelle. Audit me if you can: Query-efficient active fairness auditing of black-box LLMs. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 33673–33698, 2026. doi: 10.18653/v1/2026.findings-acl.1681. URL https://aclanthology.org/2026.findings-acl.1681/.

Jiaming Ji, Donghai Hong, Borong Zhang, Boyuan Chen, Josef Dai, Boren Zheng, Tianyi Alex Qiu, Jiayi Zhou, Kaile Wang, Boxun Li, Sirui Han, Yike Guo, and Yaodong Yang. PKU-SafeRLHF: Towards multi-level safety alignment for LLMs with human preference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31983–32016, 2025.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*, 2020.

Tian Lan, Wenwei Zhang, Chen Xu, Heyan Huang, Dahua Lin, Kai Chen, and Xian-Ling Mao. CriticEval: Evaluating large-scale language model as critic. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/7b7d7985f62284060d65f532ed2ea5fa-Abstract-Conference.html.

Dongryeol Lee, Yerin Hwang, Taegwan Kang, Minwoo Lee, Younhyung Chae, and Kyomin Jung. Judging against the reference: Uncovering knowledge-driven failures in LLM-judges on QA evaluation, 2026a. URL https://arxiv.org/abs/2601.07506v1. Version 1.

Jaeyun Lee, Junyoung Koh, Zeynel Tok, Hunar Batra, and Ronald Clark. MCJudgeBench: A benchmark for constraint-level judge evaluation in multi-constraint instruction following, 2026b. URL https://arxiv.org/abs/2605.03858v1. Version 1.

Zicheng Lin, Zhibin Gou, Tian Liang, Ruilin Luo, Haowei Liu, and Yujiu Yang. CriticBench: Benchmarking LLMs for critique-correct reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 1552–1587, 2024. doi: 10.18653/v1/2024.findings-acl.91.

Yantao Liu, Zijun Yao, Rui Min, Yixin Cao, Lei Hou, and Juanzi Li. RM-Bench: Benchmarking reward models of language models with subtlety and style. In *The Thirteenth International Conference on Learning Representations*, 2025.

Mistral AI Team. Mistral-Nemo-Instruct-2407. Hugging Face model card, 2024. URL https://huggingface.co/mistralai/Mistral-Nemo-Instruct-2407.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. CrowS-Pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1953–1967, 2020.

Filip Obradović. Measuring diagnostic test performance using imperfect reference tests: A partial identification approach. *Journal of Econometrics*, 244(1):105842, 2024. doi: 10.1016/j.jeconom.2024.105842. URL https://arxiv.org/abs/2204.00180.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2086–2105, 2022. doi: 10.18653/v1/2022.findings-acl.165. URL https://aclanthology.org/2022.findings-acl.165/.

Emmanouil A. Platanios, Hoifung Poon, Tom M. Mitchell, and Eric Horvitz. Estimating accuracy from unlabeled data: A probabilistic logic approach. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.nips.cc/paper_files/paper/2017/file/95f8d9901ca8878e291552f001f67692-Paper.pdf.

José Pombal, Ricardo Rei, and André F. T. Martins. Self-preference bias in rubric-based evaluation of large language models, 2026. URL https://arxiv.org/abs/2604.06996v1. Version 1.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, 2020.

Valle Ruiz-Fernández, Mario Mina, Júlia Falcão, Luis Vasquez-Reina, Anna Sallés, Aitor Gonzalez-Agirre, and Olatz Perez-de Viñaspre. EsBBQ and CaBBQ: The spanish and catalan bias benchmarks for question answering, 2025. URL https://arxiv.org/abs/2507.11216.

Bosi Wen, Yilin Niu, Cunxiang Wang, Pei Ke, Xiaoying Ling, Ying Zhang, Aohan Zeng, Hongning Wang, and Minlie Huang. IF-CRITIC: Towards a fine-grained LLM critic for instruction-following evaluation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25004–25029, 2026. doi: 10.18653/v1/2026.acl-long.1147.

Xiaoyuan Xie, Joshua W. K. Ho, Christian Murphy, Gail Kaiser, Baowen Xu, and Tsong Yueh Chen. Testing and validating machine learning classifiers by metamorphic testing. *Journal of Systems and Software*, 84(4):544–558, 2011. doi: 10.1016/j.jss.2010.11.920. URL https://www.cs.columbia.edu/wp-content/uploads/sites/7/2016/08/jss2011.pdf.

Austin Xu, Srijan Bansal, Yifei Ming, Semih Yavuz, and Shafiq Joty. Does context matter? ContextualJudgeBench for evaluating LLM-based judges in contextual settings. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9541–9564, 2025a. doi: 10.18653/v1/2025.acl-long.470.

Tianruo Rose Xu, Vedant Gaur, Liu Leqi, and Tanya Goyal. The progress illusion: Revisiting meta-evaluation standards of LLM evaluators. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 19033–19043, 2025b. doi: 10.18653/v1/2025.findings-emnlp.1036.

Tom Yan and Chicheng Zhang. Active fairness auditing. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 24929–24962, 2022. URL https://proceedings.mlr.press/v162/yan22c.html.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115v2. Technical report, version 2, January 2025.

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

Yilun Zhou, Austin Xu, Peifeng Wang, Caiming Xiong, and Shafiq Joty. Evaluating judges as evaluators: The JETTS benchmark of LLM-as-judges as test-time scaling evaluators. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 79436–79471, 2025. URL https://proceedings.mlr.press/v267/zhou25af.html.

APPENDIX CONTENTS

A Theoretical Analysis And Proofs **15**

A.1 Information Limits And Lossless Representation 15

A.2 Exact Bounds For The Information Classes 16

A.3 Attainability And Optimization 17

A.4 Conditional Rates And Joint Decisions 19

A.5 Reference Requirements For Certifying Success 20

B Extended Related Work **20**

B.1 Closest Evaluation Tasks And Comparison Criteria 20

B.2 Identification, Logical Constraints, And Active Auditing 21

B.3 Controlled Changes And Conditional Judgment Failure 24

B.4 Verification Interfaces And Supported Errors 25

B.5 Decision Consequences And Diagnostic Profiles 26

C Certification And Reference Queries **27**

C.1 Audit Records And Admissible Explanations 27

C.2 Bounds With Shared References 28

C.3 Joint Certification And Soundness 29

C.4 Reference Query Selection 29

D Diagnostic Definitions And Derivations **30**

D.1 World And Role Diagnostic Definitions 30

D.2 World Measurements 31

D.3 Role Measurements 31

D.4 Same-Case And Cross-Model Measurements 32

D.5 Limitations Of Original-World Evaluation 32

D.6 Verdict Reversal And Correctness 32

D.7 Correctness And Two-Sided Verification 33

D.8 Separating Verdict Error From Score-Matching Error 34

D.9 A Consistency Certificate Without The Stage-A Verdict 34

D.10 Information Available From Evaluation Records 35

E Tasks, Datasets, And Judge Case Studies **37**

E.1 Dataset And Task Discussion 37

E.2 Model And Baseline Discussion 38

F Protocol Coverage And Reproducible Analysis **39**

G Information Ablations And Audit Decisions **41**

G.1 Error Bounds By Information Class 41

G.2 Sensitivity To Invalid Relations 42

G.3 Interpretation And Validation 43

H Judgment Diagnostics **44**

H.1 Incorrect Verdict Reversals 50

I Verification Diagnostics **51**

I.1 Stage-Wise And Joint Verification Performance 55

I.2 Support Consistency Under A Shared Verdict 58

I.3 Verification With The Reference Supplied 59

J Same-Case And Cross-Model Verification **60**

J.1 Joint And Marginal Verification Analysis 65

K Diagnostic Profiles And Purpose-Specific Selection **67**

K.1 Diagnostic Profiles At Matched Accuracy 67

K.2 Limitations Of Scalar Rankings 68

K.3 Held-Out Selection Under Stated Costs 68

K.4 World And Role Ablations 70

K.5 Sensitivity To Hypothetical Label Costs 71

K.6 Sensitivity To Pair Orientation 72

K.7 Audit-Budget Sensitivity 73

K.8 Compact Diagnostic Profiles By Task 73

A Theoretical Analysis And Proofs

The main audit asks whether every, none, or only some complete records compatible with the observed evidence satisfy a stated verification requirement. This appendix establishes the limits of the audit classes, proves that the four local counts retain all information needed for the declared target, and derives sharp global bounds under a budget of possibly invalid pair relations. Each result concerns a fixed finite cohort. Extending a certificate to a population requires a separate sampling model.

A.1 Information Limits And Lossless Representation

Formal Class Definitions. The classes separate two questions: which Judge-to-Verifier links does the auditor retain, and what does the auditor know about the correct verdicts? A case $e = (x, u, \kappa)$ contains a task input, a supplied task-head output and the criterion used to assess that output. All Judges assessing the same case share one correct verdict. For a fixed Verifier, the marginal view M retains separate totals of Judge verdicts and support decisions. For example, it may retain two Accept verdicts and two support decisions without recording which verdicts received support. The linked view L_0 additionally retains every verdict-to-decision match, while imposing no cross-case restriction on the correct verdicts. A relational class R constrains correct verdicts across cases. Writing $y_k = \sigma_{ik} y_i$, the value $\sigma_{ik} = +1$ requires the same verdict and $\sigma_{ik} = -1$ requires opposite verdicts. Combining these relations with margins gives R_M. Combining them with linked decisions gives R_L. A calibrated class K constrains reference quality, an anchored class G fixes exact correct verdicts on identified cases, and $Q(C)$ augments a base class C with answers obtained through a stated query policy. These names describe retained information and assumptions, rather than every capability of a concrete method.

Indistinguishability. Proposition 1 concerns evidence o for which the conditional-rate target is defined under every complete record in $\mathcal{Z}(o)$. Consider two such records $\zeta_0, \zeta_1 \in \mathcal{Z}(o)$. If $q(\zeta_0) \in \mathcal{C}_{\alpha,\beta}$ while $q(\zeta_1) \notin \mathcal{C}_{\alpha,\beta}$, a deterministic audit receives the same input in both records and must return the same answer. That answer cannot be a sound decisive conclusion for both. A randomized audit has the same output distribution in both records and likewise cannot return a zero-error decisive answer. An adaptive audit faces the same obstruction when the complete sequence of selected queries and received responses is identical under the two records. Conversely, if membership in $\mathcal{C}_{\alpha,\beta}$ is constant over every nonempty set $\mathcal{Z}(o)$ on which the target is defined, that common membership defines a sound decisive rule. This existence statement supplies no general algorithmic bound.

Marginal Information. Consider four Judge verdicts on one case, two correct and two wrong, of which two receive support. Supporting the two correct verdicts gives $q_0 = (2, 0, 0, 2)$ and no errors. Supporting the two wrong verdicts gives $q_1 = (0, 2, 2, 0)$ and four errors. Both records retain the Judge verdicts, the correct verdict and the total number of support decisions. Separate totals therefore cannot recover which judgments received support, even when the correct verdict is known. This is a limitation of the marginal projection M, rather than of a method that retains linked cells. For L_0, keep the first record's verdict-to-decision links fixed and complement the correct verdict. The same links are then compatible with either perfect or entirely wrong verification because no cross-case or anchored information fixes the orientation.

Calibration Has Its Own Contract. A population reference-error rate $\eta > 0$ does not imply that every realized cohort has error at most η. For example, under independent sampling, all N references can be wrong with probability $\eta^N > 0$. A finite-sample confidence guarantee can account for this event under a sampling model. A deterministic certificate instead requires restrictions on the realized cohort. This distinction credits calibrated-reference identification and inference (Obradović, 2024). It does not declare those methods uninformative or equate guarantees with different assumptions.

Exact Anchors And Active Information. With fully linked exact correct verdicts, $\Theta(o)$ is a singleton and the verification outcome is directly determined. Partial anchors can reduce the compatible set according to their case correspondence and any remaining relations. Active methods can acquire missing information. Their limitations depend on the query responses actually obtained, rather than on a categorical claim that active auditing cannot represent WARD's evidence.

Local Count Representation. For the unweighted target in Section 4, the auditor needs to retain how often each Judge verdict is matched with support or non-support. Four counts preserve these matches without retaining Judge identities.

Lemma 1 (Lossless Context Counts)**.** *For fixed binary Judge verdicts and Verifier decisions on one case, the four counts* $(n, n_{\mathrm{acc}}, n_{\mathrm{sup}}, n_{\mathrm{acc,sup}})$ *in Section 4 determine the confusion vector under either possible correct verdict through Equation 2. Retaining these counts separately for each case and Verifier preserves the complete attainable count set under the same shared-correct-verdict restrictions.*

Proof. If Accept is correct, the n_{acc} Accept verdicts are correct. Of these, $n_{\mathrm{acc,sup}}$ receive support, giving $\mathrm{TP} = n_{\mathrm{acc,sup}}$ and $\mathrm{FN} = n_{\mathrm{acc}} - n_{\mathrm{acc,sup}}$. The remaining $n_{\mathrm{sup}} - n_{\mathrm{acc,sup}}$ support decisions go to Reject verdicts, giving $\mathrm{FP} = n_{\mathrm{sup}} - n_{\mathrm{acc,sup}}$. The remaining decisions give $\mathrm{TN} = n - n_{\mathrm{acc}} - n_{\mathrm{sup}} + n_{\mathrm{acc,sup}}$. If Reject is correct, the correct and wrong groups exchange roles, giving $g(-1)$ in Equation 2. Thus $(n, n_{\mathrm{acc}}, n_{\mathrm{sup}}, n_{\mathrm{acc,sup}})$ yields exactly the same confusion vector as the individual decisions under either candidate correct verdict. Applying the same cross-case restrictions to these candidate verdicts preserves the complete attainable set, rather than only the range of each coordinate. Weighted targets require corresponding weighted counts, and subgroup targets require separate count summaries. □

A.2 Exact Bounds For The Information Classes

We next isolate the contribution of each evidence type for one paired case and one Verifier. In world $w \in \{O, A\}$, let n_w be the number of Judge verdicts, $n_{\mathrm{acc},w}$ the number of Accept verdicts and $n_{\mathrm{sup},w}$ the number of support decisions. If Accept is correct, $n_{\mathrm{acc},w}$ Judge verdicts are correct. If Reject is correct, $n_w - n_{\mathrm{acc},w}$ are correct. We write these two possibilities as $n_{\mathrm{corr},w}(+1) = n_{\mathrm{acc},w}$ and $n_{\mathrm{corr},w}(-1) = n_w - n_{\mathrm{acc},w}$. Omitting the Role link preserves the number of support decisions while allowing them to be rematched to the Judge verdicts. Omitting the World relation allows the two correct verdicts to be chosen independently. The four bounds below measure how many verification errors remain unavoidable under each combination.

When links are retained, each observed decision implies which correct verdict would make it right. Supporting an Accept verdict is right when Accept is correct. Withholding support from that verdict is right when Reject is correct. For Judge index ℓ, define

$$t_{w\ell} = j_{w\ell}(2b_{w\ell} - 1), \qquad T_w = \sum_{\ell} t_{w\ell}. \tag{5}$$

The sign $t_{w\ell}$ is an implication of the recorded decision, rather than a claim about the Verifier's internal belief. Let $N_{\mathrm{dec}} = n_O + n_A$ be the number of verification decisions in the pair.

Theorem 3 (Sharp Certificates For Four Observation Classes)**.** *Without an exact correct-verdict label for either world, the minimum compatible numbers of verification errors are*

$$\begin{aligned}
D_{00} &= \sum_{w\in\{O,A\}} \min_{y\in\{-1,+1\}} |n_{\mathrm{corr},w}(y) - n_{\mathrm{sup},w}|, \\
D_{10} &= \min_{y\in\{-1,+1\}} \big\{|n_{\mathrm{corr},O}(y) - n_{\mathrm{sup},O}| + |n_{\mathrm{corr},A}(-y) - n_{\mathrm{sup},A}|\big\}, \\
D_{01} &= \frac{N_{\mathrm{dec}} - |T_O| - |T_A|}{2}, \\
D_{11} &= \frac{N_{\mathrm{dec}} - |T_O - T_A|}{2}.
\end{aligned} \tag{6}$$

The first index records whether the World relation is retained, and the second records whether the Role links are retained. Thus D_{10} *corresponds to* R_M*,* D_{01} *to* L_0*, and* D_{11} *to* R_L*. Every bound is attained by a complete record compatible with its information class. In particular,* $D_{11} \geq \max(D_{10}, D_{01}) \geq D_{00}$.

The indices form a 2×2 evidence table. Moving from D_{00} to D_{11} retains additional information and can only remove compatible completions, so the guaranteed error count cannot decrease. *The formulas state exactly how many errors each missing link can allow an audit to explain away.*

Proof. Fix one world with n_{corr} correct Judge verdicts among n verdicts and n_{sup} support decisions. Without the Role links, the number of supported correct verdicts can take every integer from $\max(0, n_{\mathrm{corr}} + n_{\mathrm{sup}} - n)$ to $\min(n_{\mathrm{corr}}, n_{\mathrm{sup}})$. The error count is $n_{\mathrm{corr}} + n_{\mathrm{sup}} - 2\mathrm{TP}$, whose minimum is $|n_{\mathrm{corr}} - n_{\mathrm{sup}}|$. Choosing the correct verdict independently in each world gives D_{00}. Requiring opposite correct verdicts gives D_{10}.

With the Role links retained, decision (w, ℓ) is erroneous exactly when $t_{w\ell} \neq y_w$. Its error indicator is $(1 - y_w t_{w\ell})/2$, so the error count in world w is $(n_w - y_w T_w)/2$. Minimizing independently over y_O and y_A gives D_{01}. Imposing $y_A = -y_O$ gives D_{11}. Each minimizing verdict assignment and, where links are absent, each minimizing rematching is permitted by the corresponding information class. The displayed minima are therefore attainable. □

Theorem 4 (Maximal Complementarity Of World And Role Links)**.** *There is a realizable four-decision record for which $D_{00} = D_{10} = D_{01} = 0$ and $D_{11} = N_{\mathrm{dec}}/2 = 2$. More generally, without exact correct-verdict anchors, the minimum error count of any N_{dec}-decision record in this complement-symmetric binary model is at most $\lfloor N_{\mathrm{dec}}/2 \rfloor$.*

Proof. In each world, let the two Judge verdicts be $(+1, -1)$ and their linked support decisions be $(1, 0)$. Then $n_w = 2$, $n_{\mathrm{acc},w} = n_{\mathrm{sup},w} = 1$ and $T_w = 2$. Equation 6 gives the claimed values. Without the World relation, both worlds can assign Accept as correct and make every linked decision right. With the World relation but without the Role links, the support decision in the Alternate World can instead be rematched to the Reject verdict, again permitting zero error. Keeping both pieces of evidence forbids both completions: either allowed orientation makes one world's two decisions wrong. If one orientation has E errors, complementing both correct verdicts preserves the World relation and produces $N_{\mathrm{dec}} - E$ errors. The smaller of these integer counts is at most $\lfloor N_{\mathrm{dec}}/2 \rfloor$, proving the upper limit. □

The separation result gives WARD's construction a precise role. *Each partial view can leave open an error-free completion even when their combination proves that errors occurred.* The result does not say that a method retaining equivalent joint information is unable to compute D_{11}, nor that every transcript obtains a strict gain from combining the views. A zero empirical gain remains possible and must be reported when it occurs.

A.3 Attainability And Optimization

Total error combines false support for wrong judgments and missed support for correct ones. A lower bound on their sum alone does not describe which combinations of the two error types are possible. We therefore retain every confusion vector produced by an allowed assignment of the unknown correct verdicts. For pair i, let $\mathcal{Y}_i^0$ contain the assignments that enforce its OW/AW reversal and any exact labels. Let $\mathcal{Y}_i^1 \supseteq \mathcal{Y}_i^0$ release only the reversal while preserving the exact labels. The same assignment applies to every Judge and Verifier assessing the pair. In the product-review example, $\mathcal{Y}_i^0$ contains the two possible orientations of the opposite verdicts. When neither world is anchored, $\mathcal{Y}_i^1$ permits all four binary verdict pairs.

For assignment $\mathbf{y}_i$, let $g_i(\mathbf{y}_i) \in \mathbb{R}^d$ stack the confusion counts for all audited Verifiers in a fixed declared order. With n_{ver} Verifiers, $d = 4n_{\mathrm{ver}}$ and each Verifier occupies one four-coordinate block. Define

$$\mathcal{P}_i^s = \{g_i(\mathbf{y}_i) : \mathbf{y}_i \in \mathcal{Y}_i^s\}, \qquad s \in \{0, 1\}. \tag{7}$$

Let $\rho_i = 1$ indicate that pair i uses the released set, let n_{pair} be the number of pairs, and let $\Gamma \in \{0, \ldots, n_{\mathrm{pair}}\}$ bound the number of releases. The global attainable set is

$$\Theta_\Gamma = \bigcup_{\substack{\rho \in \{0,1\}^{n_{\mathrm{pair}}} \\ \sum_i \rho_i \leq \Gamma}} \left\{ \sum_i q_i : q_i \in \mathcal{P}_i^{\rho_i} \right\}. \tag{8}$$

Every Verifier-specific component of q_i uses the same correct-verdict assignment for pair i. Selecting a separately favorable assignment for each Verifier would define a different and weaker audit.

The dynamic program below combines the local sets. For each possible number of released relations, it keeps the largest value of a selected linear count among the pairs processed so far. At pair i, it either enforces the relation without spending budget or releases the relation and spends one unit.

Theorem 5 (Attainability, Sharp Bounds And Sound Two-Sided Certification)**.** *Assume accurately recorded OW/AW pair identities and Judge-to-Verifier-decision correspondence, one correct verdict shared by all Judges of a case, the declared relation and anchor restrictions, and product admissibility across pairs. Then Θ_Γ in Equation 8 is exactly the attainable joint count set. For direction $\lambda \in \mathbb{R}^d$, define*

$$\psi_i^s(\lambda) = \max_{q \in \mathcal{P}_i^s} \lambda^\top q, \qquad s \in \{0,1\}, \tag{9}$$

with value $-\infty$ when the local set is empty. The sharp global upper bound $H_\Gamma(\lambda) = \max_{q \in \Theta_\Gamma} \lambda^\top q$ is computed by

$$\begin{aligned} &F_0(0) = 0, \qquad F_0(k) = -\infty \quad (k > 0), \\ &F_i(k) = \max\{F_{i-1}(k) + \psi_i^0(\lambda),\ F_{i-1}(k-1) + \psi_i^1(\lambda)\}, \\ &H_\Gamma(\lambda) = \max_{0 \le k \le \Gamma} F_{n_{\text{pair}}}(k), \end{aligned} \tag{10}$$

where a state with a negative budget index has value $-\infty$. The sharp lower bound is $-H_\Gamma(-\lambda)$. If every nominal set $\mathcal{P}_i^0$ is nonempty, the upper bound simplifies to

$$H_\Gamma(\lambda) = \sum_i \psi_i^0(\lambda) + \sum_{k=1}^{\min(\Gamma, n_{\text{pair}})} \Delta_{(k)}(\lambda), \qquad \Delta_i(\lambda) = \psi_i^1(\lambda) - \psi_i^0(\lambda) \ge 0, \tag{11}$$

where (k) orders the gains in decreasing order.

For nonempty Θ_Γ, a system of linear requirements $\lambda_c^\top q \le \tau_c$ holds for every compatible assignment if and only if $H_\Gamma(\lambda_c) \le \tau_c$ for every requirement c. If one maximum exceeds its threshold, an attaining assignment witnesses why the evidence cannot certify universal PASS*. Fix Verifier block v. When both correctness strata in that block are nonempty for every compatible assignment, let $\lambda_{\text{FP}}^{(v)}$ and $\lambda_{\text{FN}}^{(v)}$ be zero in every other block and have block-v entries*

$$\left(\lambda_{\text{FP}}^{(v)}\right)_v = (0, 1-\alpha, 0, -\alpha), \qquad \left(\lambda_{\text{FN}}^{(v)}\right)_v = (-\beta, 0, 1-\beta, 0). \tag{12}$$

These directions encode the two requirements in Equation 1. Each has threshold zero, with the displayed entries ordered as $(\text{TP}, \text{FP}, \text{FN}, \text{TN})$*. If either stratum can be empty under a compatible assignment, the corresponding denominator can vanish and the conditional-rate target is unassessed.*

The recurrence is needed when some nominal local sets are empty. When every nominal set is nonempty, the sorted-gain formula begins with all pair relations enforced and adds the Γ largest increases caused by releasing one relation. The lower endpoint uses the same calculation with the direction negated. *Every finite endpoint is sharp because a compatible assignment of the unknown correct verdicts attains it.*

Proof. Every compatible global assignment chooses one local assignment for each pair and releases at most Γ relations. Its count vector therefore belongs to Equation 8. Conversely, every local vector in $\mathcal{P}_i^{\rho_i}$ is produced by an allowed local assignment. Product admissibility permits these local assignments to be combined whenever $\sum_i \rho_i \le \Gamma$, proving that every vector in Equation 8 is attainable.

For exactly k releases among the first i pairs, pair i either uses $\mathcal{P}_i^0$ and retains the current budget count, or uses $\mathcal{P}_i^1$ and adds one release. Maximizing the two cases gives Equation 10 by induction. Because each local set is finite, every finite recurrence value has an attaining assignment. If every nominal set is nonempty, releasing pair i contributes the independent nonnegative gain $\Delta_i(\lambda)$. The best set of at most Γ releases therefore contains the largest gains, proving Equation 11. Negating λ converts the sharp upper bound into the sharp lower bound.

A finite attainable set satisfies all displayed linear requirements universally exactly when the maximum of each left-hand side is within its threshold. If a maximum exceeds its threshold, an attaining assignment violates that requirement. For the selected Verifier block, positive denominators allow the rate conditions to be multiplied through: $\text{FPR} \le \alpha$ is equivalent to $(1-\alpha)\text{FP} - \alpha\text{TN} \le 0$, and $\text{FNR} \le \beta$ is equivalent to $(1-\beta)\text{FN} - \beta\text{TP} \le 0$. Sharp lower bounds on $\text{FP} + \text{TN}$ and $\text{TP} + \text{FN}$ determine whether both strata are nonempty. If either lower bound is zero, the conditional-rate target is reported as unassessed rather than treating an undefined rate as passing or failing. □

Constructing the local count vectors is linear in the retained transcript size for a fixed number of binary local assignments. Each scalar certificate costs $O(n_{\mathrm{pair}}\Gamma)$ with the general recurrence. When every nominal set is nonempty, sorting the gains costs $O(n_{\mathrm{pair}} \log n_{\mathrm{pair}})$. These are costs for one fixed direction, rather than a claim that every nonlinear or existential property is efficiently decidable. If every recurrence state is infeasible, the observations and assumptions contradict one another. An anchor that conflicts with an asserted reversal can be represented by a nonempty released set and the required budget. It must not produce a vacuous pass.

Established Foundations. Set-based identification and worst-case auditing have precedents including Yan & Zhang (2022). Budgeted uncertainty is an established optimization construction (Bertsimas & Sim, 2004). The construction-specific contribution is the mapping from linked judgments to local confusion sets, the shared-correct-verdict coupling across Verifiers, and the conditions under which the resulting endpoints are attainable. The generic dynamic program and largest-gain argument are not claimed as new algorithms.

A.4 Conditional Rates And Joint Decisions

Fix one Verifier's four-coordinate block. An auditor may need a conditional rate among wrong judgments or among correct judgments in that block, rather than only total error. For false support, the numerator is FP and the denominator is $\mathrm{FP} + \mathrm{TN}$. Both can change with the assignment of the unknown correct verdicts. If that denominator, or the corresponding correct-judgment denominator, can be zero for any compatible assignment, WARD reports the conditional-rate target as unassessed. Otherwise, for linear count functionals Num and Den on the selected block, a candidate rate τ is below the attainable maximum exactly when some compatible vector makes $\mathrm{Num}(q) - \tau\,\mathrm{Den}(q) > 0$. This converts the ratio question into the linear optimization established above.

Proposition 2 (Conditional-Rate Endpoints)**.** *Let* Num *and* Den *be linear count functionals. Assume* $\Theta_\Gamma \neq \varnothing$ *and* $\mathrm{Den}(q) > 0$ *for every* $q \in \Theta_\Gamma$*. Define*

$$\begin{aligned} \varphi_+(\tau) &= \max_{q \in \Theta_\Gamma} [\mathrm{Num}(q) - \tau\,\mathrm{Den}(q)], \\ \varphi_-(\tau) &= \min_{q \in \Theta_\Gamma} [\mathrm{Num}(q) - \tau\,\mathrm{Den}(q)]. \end{aligned} \tag{13}$$

The maximum and minimum attainable ratios $\mathrm{Num}(q)/\,\mathrm{Den}(q)$ *are the unique zeros of* φ_+ *and* φ_-*, respectively. Theorem 5 therefore supplies exact threshold tests and attainable endpoints for each conditional rate.*

Proof. Let $\tau_{\max}$ be the maximum attainable ratio. At $\tau_{\max}$, every difference $\mathrm{Num}(q) - \tau_{\max}\,\mathrm{Den}(q)$ is nonpositive, and an attaining vector has difference zero. Below $\tau_{\max}$, that same vector gives a positive difference. Above it, every difference is negative. Positive denominators establish uniqueness. Reversing the inequalities proves the minimum statement. Both extrema are attained because Θ_Γ is finite. □

This is the standard parametric fractional-programming principle (Dinkelbach, 1967), instantiated with the exact count oracle above. Because the attainable set is finite, the extremal ratios are attained and the zero tests terminate after finitely many attainable candidates. We do not assert a general polynomial iteration bound. Averaging task-specific conditional rates is not equivalent to taking a ratio of pooled counts, so each use must declare its aggregation rule.

A Joint Requirement Can Fail Even When Both Individual Minima Are Zero. For one fixed Verifier block, consider three pairs with one Judge per world and Judge verdicts $(+1, -1)$ in every pair. The first pair supports both verdicts, with its orientation anchored so that both are correct. The second supports neither, with its orientation anchored so that both are wrong. The third supports only the first verdict, and its orientation is unknown. The only attainable confusion vectors are $(3, 0, 1, 2)$ and $(2, 1, 0, 3)$, so $(\mathrm{FPR}, \mathrm{FNR})$ is either $(0, 1/4)$ or $(1/4, 0)$. Each rate can therefore attain zero separately. No single assignment makes both rates at most $1/10$. At tolerance $1/5$, the midpoint of the convex hull of the two vectors would pass, while neither attainable vector does. *The audit cannot combine favorable rates produced by different assignments, and convexifying the attainable vectors can create a passing point that no allowed correct-verdict assignment realizes.*

For any requirement set $\mathcal{C}$, the exact three-way decision is PASS if $\Theta_\Gamma \subseteq \mathcal{C}$, FAIL if $\Theta_\Gamma \cap \mathcal{C} = \varnothing$, and unresolved otherwise. The support-function test in Theorem 5 completely decides universal satisfaction of linear requirements. It does not claim to decide every joint FAIL efficiently. A positive sharp lower bound on one violation, or a proved separating inequality, is a sufficient failure certificate.

Cost Under A Support-Or-Review Policy. Fix Verifier block v and assume $\Theta_\Gamma \neq \varnothing$. Define $\underline{E}_{\Gamma,v} = \min_{q \in \Theta_\Gamma} [\mathrm{FP}_v(q) + \mathrm{FN}_v(q)]$. Suppose a deployment accepts supported judgments and sends unsupported judgments for further review. False support then accepts a wrong judgment, while missed support needlessly reviews a correct judgment. With positive unit costs c_{FP} and c_{FN}, the error floor implies total cost at least $\min(c_{\mathrm{FP}}, c_{\mathrm{FN}})\underline{E}_{\Gamma,v}$. This bound assigns every forced error the smaller unit cost. It is a conditional consequence of the stated policy, rather than a measurement of deployment harm.

A.5 Reference Requirements For Certifying Success

Failure can sometimes be certified without knowing the absolute orientation of the correct verdicts, while two-sided success cannot. If complementing every correct verdict remains allowed, evidence that makes one support decision correct under one orientation can make it wrong under the other.

Proposition 3 (Reference Dependence Of Two-Sided Success)**.** *Suppose the allowed correct-verdict assignments are nonempty and closed under global label complement, there are no anchors that fix the orientation, and both correctness strata are nonempty under every assignment. If $\alpha + \beta < 1$, no observed record can soundly certify both $\mathrm{FPR} \leq \alpha$ and $\mathrm{FNR} \leq \beta$ for every allowed assignment. Some linked records can nevertheless certify failure without an absolute correct-verdict orientation.*

Proof. For a complete assignment $\mathbf{y}$, complementing every correct verdict maps $(\mathrm{TP}, \mathrm{FP}, \mathrm{FN}, \mathrm{TN})$ to $(\mathrm{FP}, \mathrm{TP}, \mathrm{TN}, \mathrm{FN})$. Hence $\mathrm{FPR}(-\mathbf{y}) = 1 - \mathrm{FNR}(\mathbf{y})$ and $\mathrm{FNR}(-\mathbf{y}) = 1 - \mathrm{FPR}(\mathbf{y})$. If both orientations satisfied both limits, then $1 - \alpha \leq \mathrm{FNR}(\mathbf{y}) \leq \beta$, contradicting $\alpha + \beta < 1$. For a failure witness, the record in Theorem 4 has equal correct and wrong strata and satisfies $\mathrm{FPR} = \mathrm{FNR} = 1/2$ under either orientation. Since $\alpha + \beta < 1$, at least one threshold is below $1/2$, so every orientation violates the two-sided requirement. □

Complement ambiguity itself is established in accuracy estimation (Platanios et al., 2017). The proposition states its implication for WARD's two-sided pass claim and contrasts it with an attainable failure certificate. A semantic prior or exact labels may break the symmetry. Any comparison must retain such information when a prior method uses it.

B Extended Related Work

The Comparison Target. Certifying LLM judgment-verification failure requires an explicit connection between the judgment being checked, the support it receives and the evidence restricting correctness. CriticBench already provides reference-grounded, linked critique evaluation (Lin et al., 2024). The question here is whether knowing how correct verdicts must change across paired cases, without knowing each correct verdict, can still prove that verification errors occurred. WARD supplies that construction and demonstrates a record on which either partial view allows a passing explanation while the combined evidence excludes every such explanation. The comparison therefore distinguishes the scientific target from the observations and assumptions that support it.

B.1 Closest Evaluation Tasks And Comparison Criteria

Core And Supporting Criteria. The primary criteria are the *scientific target* and the *inference established*. For WARD, the target is to support correct Judge verdicts and withhold support from wrong ones. The inference is whether every, no, or only some explanation consistent with the evidence meets that requirement. The supporting criteria are *retained associations* and *reference restrictions*. Associations specify which response belongs to which judgment and context. Restrictions specify labels, relations or calibration that constrain correctness. The final comparison explains which part of the audit is supplied and which requires a further construction. Tables A1 and A2

Table A1: Master comparison, Panel A. The closest LLM evaluation tasks precede this work. The primary columns distinguish the scientific target and reported inference. Panel B in Table A2, Appendix B.1, uses the same row order to explain the evidence and comparison boundary. The comparison asks what each study can establish, rather than counting features.

Work and publication	Scientific target	Primary inference
CriticBench (Lin et al., 2024) Findings ACL	Critique and correction of task solutions	Error-class precision/recall and same-question, inter-model analysis
CriticEval (Lan et al., 2024) NeurIPS	Quality of response critiques	Meta-feedback relative to human critique ratings
IF-CRITIC (Wen et al., 2026) ACL	Instruction-following judgments and explanations	Separate explanation-correctness and judgment-consistency checks
ContextualJudgeBench (Xu et al., 2025a) ACL	Contextual Judge outcomes and explanations	Order-consistent correctness and criteria checks on correct judgments
Construct-validity profile (Chen et al., 2026a) preprint (v1)	Judge invariance and sensitivity	Separate response invariance and construct sensitivity
Audit framework (this work)	Support correct Judge verdicts and withhold support from wrong ones, with some correct verdicts unknown	Determine whether every, no, or only some explanation consistent with the evidence meets the requirement
`WARD` (this work)	Link verdict–support records across cases with related correct verdicts	Exact possible success/error counts and tight error bounds, allowing some case relations to be invalid

(Appendix B) use the same prior-work order, prioritizing cross-role checking and then broadening to contextual and relational Judge tests. The framework and WARD appear in a separate final block.

Evidence Behind The Comparison. The comparison credits actual retained information. CriticBench's annotation and evaluation definitions in Sections 3.4 to 3.5 and its inter-model analysis in Section 4.2.5 support the linked and two-class entries. CriticEval's meta-feedback evaluation uses human critique ratings and reference critiques. IF-CRITIC's Section 3 and Appendix Table 12 distinguish explanation correctness and consistency. ContextualJudgeBench's Sections 4.1 and 4.3 and Appendix B.3 assess contextual judgment and explanations of correct outcomes. The construct-validity study's Sections 2.1 to 2.3 define the invariance/sensitivity profile and retain additional human reference information. These are different evaluation objects, not arbitrary missing-feature marks.

Limits Apply To Information Views. An impossibility claim requires an observation map. For example, discarding CriticBench's associations would create a marginal projection. The study itself retains richer linked information. Similarly, the equality/change statistic of a controlled-validity study is invariant to global relabeling, while its separately retained references can resolve that ambiguity. WARD's separation applies to the specified R_M and L_0 views. Methods supplied with equivalent R_L evidence can compute the same certificate. Fully linked, fully grounded evaluation determines exact confusion counts. Neither case is a negative comparison entry.

B.2 Identification, Logical Constraints, And Active Auditing

Joint Accuracy With Imperfect References. Obradović (2024) studies a particularly close statistical target. Proposition 1 characterizes a sharp joint set for sensitivity and specificity, and Section 2.4 allows imperfect knowledge of reference quality. Section 6 extends the setting beyond medical tests. This establishes a direct precedent for joint two-sided uncertainty without assuming an infallible reference. WARD instead fixes a finite record and constrains its possible references through controlled relations and optional anchors. Neither identifying assumption universally dominates the

Table A2: Master comparison, Panel B. Evidence supporting the inferences in Panel A and the precise boundary of each comparison. Reference-grounded and relational methods retain their actual information. The impossibility results apply only to a specified observation projection. Source sections are documented in Appendix B.1.

Work and publication	Supporting evidence	Relation to the proposed audit
CriticBench (Lin et al., 2024) Findings ACL	Linked generation, critique and correction; annotated correct and wrong solutions	Reference-grounded measurement of critique. The scored object is a task solution, not a Judge verdict under paired reference restrictions.
CriticEval (Lan et al., 2024) NeurIPS	Critiques, reference critiques and human ratings	Reference-grounded critique-quality evaluation. It does not supply the finite paired-verification construction examined here.
IF-CRITIC (Wen et al., 2026) ACL	Additional LLM checks used in data filtering	Consistency and correctness are separate checks. Proving failure under every correct-verdict assignment allowed by the evidence requires a further argument.
ContextualJudgeBench (Xu et al., 2025a) ACL	Contextual references, paired response orders and explanation assessment	Already conditions explanation checks on correct outcomes. WARD instead checks support for Judge verdicts across cases with related correct verdicts.
Construct-validity profile (Chen et al., 2026a) preprint (v1)	Human-validated preserving/changing interventions and reference information	Relational evidence with additional grounding. An equality/change statistic alone is a projection of this richer study.
Audit framework (this work)	What the audit observes, assumes about correct verdicts, aims to establish and can request	Equivalent information supports equivalent conclusions, independent of method name.
`WARD` (this work)	World reference relation and matched Judge/support counts; optional anchors	Implements the joint class. Removing either link can destroy a decisive failure certificate.

other. A statistical claim about a population and a guarantee about these fixed cases rely on different assumptions and should be assessed accordingly.

Information Loss In LLM Panels. Chen et al. (2026b) treat LLMs as fallible reporters of a latent binary quantity. Theorem 2 formalizes how coarsening weakens identification; Proposition 5 and Theorem 4 analyze lossless reductions for richer panel records. This work already distinguishes information in counts, named responses and matrices. Building on that information-loss principle, WARD retains the joint counts sufficient for verification confusion vectors under common case references. The resulting representation supports the construction-specific separation and attainable certificates.

Logical Accuracy Estimation. Logical accuracy estimation gives another direct reduction. Platanios et al. (2017) combine logical constraints with probabilistic inference to estimate likely accuracies and labels. A violated constraint already implies at least one prediction error. WARD's implied verdict $t = j(2b - 1)$ makes a verification error exactly the classification error $\mathbf{1}[t \neq y]$. This reduction explains why established tools apply. To assess both support obligations, WARD also keeps track of which Judge verdicts would be correct under each possible reference, using the same reference for all judgments of a case. WARD characterizes every attainable finite count under its assumptions, rather than choosing a most-likely explanation.

Table A3: Master comparison, Panel C. The scope broadens explicitly from verification of LLM judgments to adjacent identification and auditing problems. Targets appear before mathematical ingredients. A general framework can express WARD without already providing its evidence construction; a different target is not a failed attempt at WARD. Appendix B.2 gives the reductions and distinctions.

Work and publication	Scientific target	Inference or guarantee	Evidence class and scope
Diagnostic partial identification (Obradović, 2024) Journal of Econometrics	Test sensitivity and specificity	Sharp joint identified sets; inference under imperfect-reference quality restrictions	Calibrated reference class K. Joint two-sided uncertainty is a direct precedent.
Partial Identification from LLM Prompts (Chen et al., 2026b) preprint (v2)	Latent prevalence from LLM panels	Sharp calibrated bounds; information loss from coarsening; lossless reductions	Calibrated K with counts, named vectors or matrices. Rich joint observations are explicitly retained.
Probabilistic logic (Platanios et al., 2017) NeurIPS	Accuracy from unlabeled predictions	Most-likely accuracy and labels under logical constraints and probabilistic modeling	Logical relations plus probabilistic inference. Constraint violations already imply some prediction error.
Metamorphic classifier testing (Xie et al., 2011) JSS	Classifier faults without an individual test oracle	Violations of necessary input/output relations expose faults	Relational R. A broad precedent for fault detection without absolute labels.
Active Fairness Auditing (Yan & Zhang, 2022) ICML	A classifier property from queries	Version-space minimax guarantees and query-complexity results	Active $Q(C)$; the formulation can accommodate general properties. WARD is expressible after its objects are supplied.
BAFA (Hartmann et al., 2026) Findings ACL	Black-box LLM fairness	Query-efficient uncertainty reduction using surrogate version spaces	Active $Q(C)$ with approximate surrogates. Its guarantee must be read under those assumptions.
`WARD` (this work)	Verification errors on a fixed cohort	All possible success/error counts and conclusions that hold for every allowed explanation	Relational and linked R_L, optionally anchored. The reference budget states tolerated relation uncertainty.

Metamorphic Testing. Metamorphic testing already identifies faults through necessary relations when an individual output oracle is unavailable (Xie et al., 2011). WARD's relation-based failure detection belongs to that tradition. Its addition is the linked judgment/support construction, a quantitative count set and its two-sided decision consequence. Relational consistency remains insufficient for correctness when a global complement is admissible. The separation identifies when joining two kinds of evidence removes explanations that each permits independently.

Active Auditing. Active Fairness Auditing supplies a general version-space formulation, minimax principles and query-complexity results (Yan & Zhang, 2022). Its Section 3.1, Theorems 3.1 to 3.2 and Appendix D.1 accommodate broader property targets. BAFA applies surrogate version-space uncertainty reduction to black-box LLM fairness (Hartmann et al., 2026). These frameworks can accommodate new audit objects. WARD specifies the judgment/support observations and reference constraints needed here, with exact local sets. The equal-cost reference-query rule in Appendix C.4 is a supporting product-structure specialization, not a new general theory of active auditing.

Supporting Mathematical Ingredients. Hyperproperties provide a language for specifications spanning executions (Clarkson & Schneider, 2008). Agreement testing relates local disagreement to global consistency under its own hypotheses (Dinur et al., 2022). WARD's binary relation test is an elementary special case. No lifted-code theorem or cryptographic security guarantee is imported. Budgeted uncertainty supplies the relation-budget principle (Bertsimas & Sim, 2004). Parametric fractional programming supplies the conditional-rate optimization principle (Dinkelbach, 1967). Marginal conjunction bounds are classical (Dobra & Fienberg, 2000). Together these ingredients support the complete construction: linked verification records, a shared-reference uncertainty set, and an attainable decision certificate.

What The Comparison Establishes. These methods address important parts of the audit problem, including grounded critique evaluation, controlled relational evidence, joint uncertainty and active acquisition. WARD connects a specific verification target to the evidence needed for a decisive finite audit and proves that its retained representation supplies that evidence. The empirical separation then demonstrates a changed conclusion under identical model decisions. *WARD shows which links the audit must preserve, how to compute the errors those links force, and when the result rules out a verification standard that separate checks leave unresolved.* Its novelty rests on that construction and consequence, with the general frameworks providing the credited foundations.

B.3 Controlled Changes And Conditional Judgment Failure

Reference Adherence. Reference adherence is already a concrete Judge-evaluation problem. Lee et al. (2026a) construct QA triplets crossing original and swapped references with aligned and misaligned candidates. Their controlled reference replacement directly tests whether judgments follow the supplied reference. WARD's World construction keeps the judging criterion fixed and changes a task-specific case field to reverse the reference verdict. Its six tasks need not share a single intervention or isolate parametric knowledge. These are related tests of required judgment change with different manipulated objects and explanatory scope.

Correctness Transitions Under Controlled Changes. Lee et al. (2026b) preserve constraint-adherence labels under response and prompt variants. Their Appendix E.2 additionally measures gold-relative correctness changes and separates their direction. This is more than unqualified output stability. Their correct-to-incorrect proportion conditions on correctness-changing cases. HFR conditions on all OW-correct cases under a reference reversal. That denominator specifies the credited successes being audited. It is not evidence that prior work ignored correctness transitions.

Behavioral Testing And Conditional Failure. The broader behavioral foundations are equally relevant. CheckList includes invariance and directional-expectation tests (Ribeiro et al., 2020). Contrast sets use expert-authored, typically label-changing edits and require correctness across a local set (Gardner et al., 2020). For one OW/AW pair per case, let $p_O = P(c^O = 1)$ and $p_{OA} = P(c^O = c^A = 1)$ on the same task and cohort. Then

$$\mathrm{HFR} = 1 - \frac{p_{OA}}{p_O}, \qquad p_O > 0. \tag{14}$$

HFR therefore re-expresses joint correctness relative to credited successes. It is not independent information once both component quantities are known. The conditional view makes the exposed population explicit and connects it to Role outcomes. Because our headline quantities average tasks separately, this identity must be applied within tasks, before averaging. Counterfactually augmented sentiment data also supplies an actual construction used here (Kaushik et al., 2020), rather than merely an analogy.

Content Sensitivity In Evaluator Benchmarks. Controlled evaluator benchmarks establish further partial solutions. LLMBar contrasts instruction-following responses with superficially attractive violations (Zeng et al., 2024). RM-Bench separates sensitivity to subtle content errors from resistance to style changes (Liu et al., 2025). These works already test whether evaluators notice consequential content distinctions. WARD adds the OW-success population and its linked verification profile. The distinction to assess is what linked verification adds after these judgment tests.

Table A4: Supporting World comparison: the scientific object, primary scored event, and enabling control. Closest Judge-specific comparisons precede broader task-model foundations. The final block describes this work. Cells state the inspected evaluation rather than claiming absence of every other capability. Denominator distinctions and the relation between HFR and joint correctness are developed in Appendix B.3.

Work and publication	Evaluated object	Scored event	Control
Construct-validity profile (Chen et al., 2026a) preprint (v1)	Claim-evidence Judges	Invariance and verdict-change sensitivity. Human intervention labels	Construct-preserving/changing edits
ContextualJudgeBench (Xu et al., 2025a) ACL	Contextual preference judgments	Correct in both orders. Explanation check on correct outcomes	Criterion hierarchy. Response-order consistency
Judging Against the Reference (Lee et al., 2026a) preprint (v1)	Reference-conditioned QA judgments	Accuracy by reference setting and alignment	Original/swapped reference and candidate combinations
MCJudgeBench (Lee et al., 2026b) preprint (v1)	Constraint-adherence judgments	Accuracy, macro-F1. Transition direction among correctness changes	Gold-label-preserving response/prompt variants
LLMBar (Zeng et al., 2024) ICLR	Instruction-following preferences	Preference accuracy and order agreement	Superficial appeal against actual instruction compliance
RM-Bench (Liu et al., 2025) ICLR	Reward-model preferences	Correct response ranked above incorrect response	Subtle content errors crossed with style variants
Contrast sets (Gardner et al., 2020) Findings EMNLP	Task-model predictions	Correctness on all members of a contrast set	Expert-authored, usually label-changing edits
CheckList (Ribeiro et al., 2020) ACL	Task-model behavior	Failure of the specified behavioral expectation	Minimum functionality, invariance, directional tests
WARD (this work)	Judge verdicts and linked verification	AW error among OW-correct cases, linked to Role outcomes	Fixed criterion. Task-specific reference reversal

B.4 Verification Interfaces And Supported Errors

Critique And Cross-Model Evaluation. Critique evaluation already extends beyond scoring task answers. CriticEval includes meta-feedback: rating critiques against human judgments, with annotated reference critiques supplied for evaluation (Lan et al., 2024). CriticBench measures generation, critique and correction on the same questions and examines inter-model critique (Lin et al., 2024). Its error-class F1 uses precision and recall, so false accusations against correct responses matter. These precedents motivate retaining both classes when measuring support for proposed Judge judgments. WARD separates the two support rates and evaluates their intersections with World failures.

Consistency Versus Correctness. IF-CRITIC provides an especially relevant interface precedent (Wen et al., 2026). Its data filtering separately checks explanation correctness and explanation/judgment consistency with additional LLMs. In WARD, Stage A supplies an independent ver-

Table A5: Supporting Role comparison: the primary assessment and the object to which it applies. Explanation verification, meta-feedback, error F1 and self-preference controls are substantive precedents. The order prioritizes checking a judgment before broadening to solution critique, bias and jury behavior. Appendix B.4 explains the interface and interpretation distinctions.

Work and publication	Assessment and scope	Object being checked
ContextualJudgeBench (Xu et al., 2025a) ACL	GPT-4o criteria verification on correct judgments	Judge explanation and outcome
IF-CRITIC (Wen et al., 2026) ACL	Cross-model checks of correctness and internal consistency. Training-data filtering	Constraint-level explanation and judgment
CriticEval (Lan et al., 2024) NeurIPS	Meta-feedback score correlation with human ratings. Reference critiques provided	Response critique and critique quality
CriticBench (Lin et al., 2024) Findings ACL	Error precision/recall F1. Same-question and inter-model comparisons	Task solution, critique, correction
Self-preference in rubric evaluation (Pombal et al., 2026) preprint (v1)	Self/family overestimation normalized against unrelated generators	Rubric verdict on own/other outputs
Safer or Luckier? (Chen & Goldfarb-Tarrant, 2025) ACL	Human agreement and artifact robustness of individual Judges and juries	Safety preference and jury vote
WARD (this work)	Externally scored support on correct and wrong judgments. Same-case and four-decision events	Own/peer Judge scores against frozen A verdict

dict and Stage B assesses agreement between that verdict and Judge scores. A correct consistency decision can agree with an incorrect A verdict. External Role discrimination, A-conditional discrimination and joint success therefore answer different questions. The implementation is one particular test of a verification safeguard, not an evaluation of every possible Verifier interface.

Supported Errors Versus Self-Preference. Self-preference requires an additional distinction. Pombal et al. (2026) normalize overestimation of a model's own outputs against its treatment of unrelated generators, including a programmatically grounded rubric evaluation. Mere self-confirmation is insufficient to establish self-preference bias: a generally permissive Judge can favor everyone. WARD reports supported self-errors and four-decision complete mutual support as observed events. Neither establishes preferential treatment, a family effect, collusion, or an internal cause. The word *support* refers to a positive score under the declared interface.

B.5 Decision Consequences And Diagnostic Profiles

Agreement And Shared Errors. Multiple-model agreement is already studied as a fallible safeguard. Chen & Goldfarb-Tarrant (2025) test safety-evaluation artifacts and use majority-vote juries that improve robustness without removing all artifact sensitivity. WARD's complete mutual support instead identifies the conjunction in which two wrong Judges each support both errors. It measures corroboration of shared mistakes without assuming independent votes or implementing a jury correction rule. A selected pair's event rate is not a comparison of ensemble algorithms.

Rankings And Matched Diagnostic Profiles. Xu et al. (2025b) show why broad ranking agreement can conceal poor discrimination among similarly performing generators. WARD asks a related question about the *Judges*: after matching their OW accuracy, how different are their World and Role

profiles? The maximum-gap examples demonstrate possibility, while the appendix reports every eligible pair. The secondary selection study then attaches a stated cost to those profiles. Improvements under a one-sided self-confirmation penalty do not establish better balanced verification, realized annotation savings, or superiority under other costs.

Downstream Decision Consequences. Judge quality can affect downstream systems because judgments guide which generated answers are retained or refined. JETTS directly evaluates Judges in response reranking, step-level search and critique-based refinement (Zhou et al., 2025). That benchmark motivates assessing decision consequences rather than relying solely on standalone Judge scores. WARD's support-or-escalate interpretation is a finite consequence of its error certificate. No deployment experiment establishes harm reduction, annotation savings or improved generation from adopting WARD. Its central empirical result is the change from an unresolved audit to a certified failure under the stated assumptions.

C Certification And Reference Queries

The same audit can answer two further questions: how many errors become provable when we add each kind of evidence, and which missing correct verdicts would most reduce the remaining uncertainty? We assume a validated relation between correct verdicts, such as reversal across WARD's paired cases. This perspective connects the cross-execution specifications of security to agreement testing in coding theory. A hyperproperty constrains sets of executions rather than only an individual trace (Clarkson & Schneider, 2008). Agreement testing asks whether local observations admit one global explanation and relates observed disagreement to distance from the admissible set (Dinur et al., 2022). We derive the finite binary specialization needed for WARD, including an exact error decomposition and a reference-query rule. We do not claim a new general theory of hyperproperties, code testing, or cryptographic security.

C.1 Audit Records And Admissible Explanations

We fix one evaluation pair and one Verifier. In each world $w \in \{O, A\}$, the record contains the m_w non-tied Judge verdicts $j_{w\ell} \in \{-1, +1\}$ and their binary B supports $v_{w\ell} \in \{0, 1\}$. The number of retained decisions is $M = m_O + m_A > 0$. These are fixed recorded outputs. Binary non-support includes neutrality. We know the reference relation $y_A = -y_O$ and write $y_O = y$. We define

$$\begin{aligned} t_{w\ell} &= j_{w\ell}(2v_{w\ell} - 1), \\ S_O &= \sum_{\ell} t_{O\ell}, \quad S_A = \sum_{\ell} t_{A\ell}, \quad S = S_O - S_A. \end{aligned} \tag{15}$$

The sign $t_{w\ell}$ states which correct verdict would make this verification decision right. Supporting an Accept verdict requires Accept to be correct; withholding support from it requires Reject to be correct. The same reasoning applies to a Judge's Reject verdict. This is an implication of the recorded decision, not an inferred internal belief. External verification is correct exactly when $t_{w\ell} = y_w$.

There are three nested explanation sets. The first allows an arbitrary common verdict in each world separately. The second requires the two common verdicts to reverse. The third fixes their orientation using the actual reference. Distance means the number of binary support entries that would need to change while all Judge verdicts remain fixed. This interpretation does not assert that any such change has been experimentally implemented. In the notation below, $D_R = D_{01}$ and $D_W = D_{11}$ from Theorem 3 in Appendix A. Thus D_W already retains the within-world Role links.

We add evidence in three steps: first require a common correct verdict within each case, then opposite correct verdicts across the pair, and finally reveal which verdict is correct. Each step can rule out explanations with too few errors. The theorem counts the additional errors forced at each step.

Theorem 6 (Exact Certification From Linked Decisions)**.** *Under the stated finite binary setting, define*

$$\begin{aligned} D_R &= \frac{M - |S_O| - |S_A|}{2}, \\ D_W &= \frac{M - |S|}{2}, \qquad G_W = D_W - D_R, \\ G_Y &= \frac{|S| - yS}{2}. \end{aligned} \tag{16}$$

Then D_R is the least mismatch count under separate shared verdicts, D_W is the least count under reversing shared verdicts, and the actual number E of externally incorrect verification decisions satisfies

$$E = D_R + G_W + G_Y, \qquad G_W \geq 0, \quad G_Y \geq 0. \tag{17}$$

With only the reversal relation known, the exact possible error counts are $\{D_W, M - D_W\}$. Thus D_W is a sharp lower bound over the two possible orientations. Moreover, $G_W = \min(|S_O|, |S_A|)$ if $S_O S_A > 0$ and is zero otherwise.

Proof. For any candidate world verdict a_w, the mismatch count is $(m_w - a_w S_w)/2$. Minimizing independently over $a_O, a_A \in \{-1, +1\}$ gives D_R. Enforcing $a_A = -a_O$ changes the count to $(M - a_O S)/2$, whose minimum is D_W. The actual external count is $(M - yS)/2$. Subtracting the two minima gives Equation 17 (Appendix C). The triangle inequality gives $G_W \geq 0$, and $|S| \geq yS$ gives $G_Y \geq 0$. The two choices of y give D_W and $M - D_W$, proving both attainability and sharpness in the relation-only observation model. Finally, $|S_O| + |S_A| - |S_O - S_A|$ is twice the smaller absolute value when the signs agree and zero when they oppose or one is zero. □

A Failure That Separate Consistency Checks Miss. Suppose both worlds have two retained decisions and every implied verdict is $+1$. Each world is perfectly compatible with a shared verdict, so $D_R = 0$. However, $S_O = S_A = 2$ and $D_W = 2$: no reversing pair of references can make all four decisions correct. Exactly half are wrong under either orientation. Conversely, implied verdicts $(+1, +1)$ in OW and $(-1, -1)$ in AW give $D_W = 0$. They are either all correct or all wrong. *Linking worlds exposes failures that local consistency misses. It still cannot remove the need for an orientation anchor.*

The distinction is about the sets of admissible output explanations. It does not claim that correctly implemented exact B matching fails to compose with a correct A verdict: those two conditions do imply external correctness. The gap concerns separate compatibility with some verdict, or measurements that omit the common linkage. The components are an exact distance decomposition, not mutually isolated psychological causes or estimates of intervention effects.

Several Worlds. If validated relations specify $y_w = \sigma_w y$ with known signs $\sigma_w \in \{-1, +1\}$, replace S by $\sum_w \sigma_w S_w$ and $|S_O| + |S_A|$ by $\sum_w |S_w|$. The same proof gives all formulas. Every relation must be valid for the criterion. An invalid transformation can create a false nominal certificate. The relation-budget construction in Appendix A.3 states an explicit relaxation of that premise. A world with no decided Judge outputs contributes zero. If $M = 0$, the record is unassessed.

C.2 Bounds With Shared References

All Verifiers assess the same true verdict on a case. This yields a further constraint even though their independent A verdicts need not agree. For K Verifiers with equal retained count M, let S_{wv} denote their world sums, $S_v = S_{Ov} - S_{Av}$, and $S_\star = \sum_v S_v$. Define

$$\begin{aligned} D_{R,\star} &= \frac{KM - \sum_{v,w} |S_{wv}|}{2}, \\ D_{W,\star} &= \frac{KM - \sum_v |S_v|}{2}, \qquad D_\star = \frac{KM - |S_\star|}{2}. \end{aligned} \tag{18}$$

These minima impose, respectively, within-world compatibility, within-Verifier reference reversal, and a common reference orientation across Verifiers. They obey $D_{R,\star} \leq D_{W,\star} \leq D_\star$. The total

external error is

$$E_\star = D_{R,\star} + (D_{W,\star} - D_{R,\star}) + (D_\star - D_{W,\star}) + \frac{|S_\star| - yS_\star}{2}. \tag{19}$$

Each inequality follows from the triangle inequality, and telescoping yields $(KM - yS_\star)/2$, the actual count. *The extra gap measures the error forced by a common external reference, not a requirement that independently wrong A verdicts agree.* Unequal counts can be included by replacing KM with their sum. Correctness labels for different Verifiers on the same case must not be treated as independent orientation choices.

C.3 Joint Certification And Soundness

A large record can be checked by sampling small parts of it. Here, two decisions expose a contradiction when they require incompatible correct verdicts after accounting for the case reversal. For the single-Verifier transcript, we normalize the implied verdicts as $u_{O\ell} = t_{O\ell}$ and $u_{A\ell} = -t_{A\ell}$. The admissible relation-consistent vectors are the two constant binary vectors. The test draws two coordinates uniformly and independently with replacement and rejects when their u values differ. If $d = D_W/M$, the rejection probability is

$$p_{\text{reject}} = 2d(1-d) \geq d, \qquad 0 \leq d \leq \tfrac{1}{2}. \tag{20}$$

Indeed, the two signs occur with proportions d and $1-d$. The two orders of unequal signs give the displayed probability. A compatible transcript always passes. If $d \geq \epsilon$, q independent tests miss every inconsistency with probability at most $(1-\epsilon)^q \leq \exp(-q\epsilon)$. Thus $q \geq \log(1/\delta)/\epsilon$ suffices for miss probability at most δ. For many components, the test selects a component proportional to its retained count before sampling coordinates. By averaging, the rejection probability is at least the overall minimum error fraction. The same construction applies to stacked Verifiers with their shared orientation.

This is the elementary repetition-code instance of agreement testing, not a new locally testable code. The guarantee concerns a fixed transcript. Resampling the decisions from a stochastic or adaptive model does not satisfy this premise automatically. A prospective implementation needs to freeze the relevant outputs before audit-coordinate selection. The analysis of fixed outputs implements that premise through an immutable input record. It does not establish a cryptographic commitment. *Repeated consistency checks can miss a record in which every verification decision is wrong: reversing all correct verdicts preserves the same consistency relations.*

C.4 Reference Query Selection

The preceding impossibility determines where external labels are informative. We consider n unrelated components with already observed transcripts and valid internal polarity relations. A component may be one case pair with one Verifier, or all Verifiers stacked for that pair. Its size is M_i, its normalized sum is S_i, and its minimum error is $D_i = (M_i - |S_i|)/2$. The auditor knows only the transcripts and polarity relations. The unknown orientations y_i are unrestricted across components. Querying an exact reference label for any world in component i reveals its orientation. Each label request has equal cost. A pair whose decisions could be either all right or all wrong has more unresolved errors than a pair already known to contain half errors. The theorem formalizes which labels remove the most uncertainty. This query model is separate from the experiments, whose labels already exist.

Theorem 7 (Optimal Reference Queries For The Finite Error Bound)**.** *After querying component set Q, the sharp lower and upper bounds on the total error count are*

$$\begin{aligned} L_Q &= \sum_{i \in Q} E_i + \sum_{i \notin Q} D_i, \\ U_Q &= \sum_{i \in Q} E_i + \sum_{i \notin Q} (M_i - D_i), \\ U_Q - L_Q &= \sum_{i \notin Q} |S_i|. \end{aligned} \tag{21}$$

Among all policies allowed b exact component-label queries, querying the b largest $|S_i|$ minimizes the residual worst-case width. Adaptivity cannot improve that value. The midpoint has worst-case absolute error $(U_Q - L_Q)/2$, and no point estimator has a smaller worst-case absolute error given the same observations. For an error-rate target, all quantities are divided by $\sum_i M_i$.

Proof. For every unqueried component, Theorem 6 (Appendix C) gives the two possible counts D_i and $M_i - D_i$. Because orientations are unrestricted across components, choosing every lower endpoint or every upper endpoint is admissible and leaves all queried labels unchanged. Their difference is $\sum_{i \notin Q}(M_i - 2D_i) = \sum_{i \notin Q} |S_i|$. A query removes exactly its component's $|S_i|$ contribution, irrespective of the returned label. Removing the b largest contributions is therefore optimal on every query path. An adaptive policy still removes at most b of these fixed contributions and cannot do better. Any point estimate is at distance at least half the endpoint separation from one of the two admissible extremes. The midpoint attains that bound. □

The interval is the tight range. Its interior need not be fully attainable for a finite count. For equal component sizes, larger $|S_i| = M_i - 2D_i$ means smaller inconsistency distance. *Thus truth-label queries that minimize this worst-case uncertainty prioritize the most internally consistent transcripts. Inconsistency detection and truth-orientation auditing optimize different quantities.* This rule does not maximize the expected number of discovered errors under an unspecified prior. It does not minimize population sampling variance. For a known positively weighted error objective, each query removes its weighted width. For unequal query costs, budget allocation is a knapsack problem and the unweighted ordering need not be optimal. Semantic constraints or already-known links between components can reduce the uncertainty further. Shared cases across Verifiers must be grouped before applying the result.

Minimax auditing through a version space is established in active fairness auditing (Yan & Zhang, 2022). BAFA applies metric-bound reduction to black-box LLM fairness using approximate surrogate version spaces (Hartmann et al., 2026). Our result is a closed-form specialization for the binary orientation uncertainty induced by WARD's linked transcripts. It does not supersede their more general query problems. Its distinction is that the observations are fixed verification decisions, the unknown is a reference orientation, and the objective is an exact finite error bound requiring no learned surrogate.

D Diagnostic Definitions And Derivations

The main audit asks which errors must have occurred without knowing every correct verdict. The diagnostics here use known correct verdicts to locate the failures and describe their patterns. We define their measurements, establish the corresponding information limits, and separate correctness from faithful score matching. In this appendix, v denotes binary support (the main text's b), while z denotes the known Judge-correctness indicator obtained by comparing the Judge verdict j with the correct verdict y. These identities and bounds justify the measurements and explanations used in the diagnostic results.

D.1 World And Role Diagnostic Definitions

An evaluation case $e_i^w = (x_i^w, u_i^w, \kappa_i)$ contains the three objects visible to a Judge: task input x_i^w, supplied response u_i^w, and criterion κ_i. The superscript $w \in \{O, A\}$ distinguishes the original and alternate worlds. In sentiment classification, these objects are the review, the proposed sentiment label, and the instruction to judge whether that label is correct. The correct Accept/Reject verdict is $y_i^w \in \{-1, +1\}$. Judge J returns acceptance and rejection scores a_{iJ}^w and r_{iJ}^w. Their difference $m_{iJ}^w = a_{iJ}^w - r_{iJ}^w$ determines the binary verdict. A margin above 10^{-12} predicts Accept, one below -10^{-12} predicts Reject, and a tie counts as incorrect. The indicator c_{iJ}^w equals one exactly when this verdict matches y_i^w. These symbols describe fixed recorded outputs, not repeated-call probabilities.

A valid World pair retains the criterion, satisfies an output-independent task-specific construction rule, and has $y_i^A = -y_i^O$. The changed field may be the input or the supplied response. The common control is a reference-verdict reversal under a fixed criterion, rather than one edited factor shared by all tasks. The following construction makes the missing information in OW evaluation precise.

An OW-perfect rule can follow the reversed reference or repeat its old verdict in AW. Their OW records are identical and their AW outcomes opposite (Proposition 4, Appendix D). The paired observations distinguish them.

Definition 1 (World Responsiveness)**.** On a valid pair with $c^O_{iJ} = 1$, Judge J is World-responsive when $c^A_{iJ} = 1$.

The violation indicator is $h_{iJ} = c^O_{iJ}(1 - c^A_{iJ})$. If OW is wrong, $h_{iJ} = 0$ is outside this conditional assessment and does not establish responsiveness. A failure on a valid pair violates the case-level requirement. Finite success does not establish universal criterion grounding.

The Role Test changes the evaluated function from producing a Judge verdict to assessing a proposed Judge judgment. Verifier V supplies score s^w_{iJV}, with $v^w_{iJV} = \mathbf{1}[s^w_{iJV} > 0]$ indicating support. Zero and negative scores denote non-support. The external target is Judge correctness $z^w_{iJ} = c^w_{iJ}$, attached after inference.

Definition 2 (Role Discrimination)**.** Verifier V satisfies Role discrimination on a proposed judgment when $v^w_{iJV} = z^w_{iJ}$ under the declared interface.

This requires support for a correct judgment and non-support for a wrong one. It is a requirement for the claimed verification safeguard, not for bare Judge accuracy. A mismatch is an observable case-level violation. In our two-stage instantiation, B assesses score agreement with a frozen A verdict, which may be wrong. Role discrimination evaluates the resulting output against external correctness. Consistency with A alone does not imply it.

D.2 World Measurements

Hidden Failure Rate (HFR) asks a conditional question: among cases the Judge gets right in the original world, how often does it fail on the paired alternate world? Let $\mathcal{I}_t$ be the retained pairs for task t. The numerator below counts original successes followed by alternate-world errors, and the denominator counts all original successes:

$$\mathrm{HFR}_t(J) = \frac{\sum_{i \in \mathcal{I}_t} c^O_{iJ}(1 - c^A_{iJ})}{\sum_{i \in \mathcal{I}_t} c^O_{iJ}}, \qquad \mathrm{HFR}(J) = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \mathrm{HFR}_t(J), \tag{22}$$

when each conditional denominator is nonzero. An HFR of 80% within one task means that 80 of every 100 credited OW successes fail in AW. The headline mean gives the six tasks equal weight. OW, AW, and both-world accuracy use the same task-first averaging. Both-world accuracy requires both verdicts to be correct. Within a task, HFR equals one minus both-world accuracy divided by OW accuracy, linking it to contrast consistency (Appendix B.3). Appendix H also reverses the anchor and reports all four correctness transitions. Verdict change alone need not mean correctness: on binary non-tied reversal pairs, a flip occurs when both verdicts are correct or both are wrong. Conditioning change on OW correctness recovers $1 - \mathrm{HFR}$ (Proposition 5, Appendix D.6).

The hidden-failure profile further distinguishes unchanged margins, movement toward the AW target that remains insufficient, and movement away from that target. Appendix H defines these score-movement categories.

D.3 Role Measurements

The two-stage Role interface begins with an independent verdict from the case and criterion in A. B then receives the frozen A verdict and one Judge's acceptance/rejection scores. For example, suppose the correct verdict is reject, A incorrectly says accept, and the Judge also says accept. B can match A faithfully and support that Judge error. We score the B output against external Judge correctness, retaining two rates with distinct denominators: correct-judgment support $P(v = 1 \mid z = 1)$ and wrong-judgment non-support $P(v = 0 \mid z = 0)$. Balanced accuracy (BA) averages them,

$$\mathrm{BA} = \tfrac{1}{2}\big[P(v = 1 \mid z = 1) + P(v = 0 \mid z = 0)\big]. \tag{23}$$

An always-support Verifier has component rates 1 and 0. An always-non-support Verifier has rates 0 and 1. Both receive BA $1/2$, so both components remain necessary for interpretation. Lemma 2 in Appendix D formalizes this elementary boundary. AUROC additionally tests whether continuous B

scores rank correct judgments above wrong ones. We calculate each measure within a task and the stated self or peer subset before equal-task averaging.

D.4 Same-Case And Cross-Model Measurements

Marginal Role rates can succeed on different cases. For a hidden AW error by focal Judge J, we therefore consider a peer Judge K that is AW-correct on that same case. Model J acts as Verifier for both judgments. *Rescue* assigns non-support to its own error and support to the correct peer. *Persistence* does the reverse. The other outcomes support both or neither. Rescue names a joint decision pattern and does not imply generation of a corrected answer. Each eligible correct peer contributes one ordered comparison.

Peer analysis also asks whether two wrong Judges each support both errors. We call this four-positive-decision conjunction *complete mutual support* (CMS). Its conditional rate uses jointly wrong AW cases. Its joint error-and-support rate uses all AW cases. These denominators answer different questions. Alongside World and marginal Role measures, they form a diagnostic profile whose components retain their individual meanings.

Why The Joint Measurements Are Needed. Let r be own-error non-support and p correct-peer support on the same eligible comparisons. Their rescue rate lies between $\max(0, r + p - 1)$ and $\min(r, p)$. Identical marginals can therefore coexist with different rescue rates: two cases with support pairs $(0, 1), (1, 0)$ yield 50% rescue, whereas $(0, 0), (1, 1)$ yield none. Both retain the same marginal score distributions, even with correct A verdicts. Proposition 7 in Appendix D.10 proves the sharp marginal bounds, extends them to four-decision CMS, and specifies when the bounds collapse. Paired World evidence identifies HFR. Marginal Role evidence generally only bounds the linked verification events. WARD records their case-level conjunctions.

D.5 Limitations Of Original-World Evaluation

Proposition 4 (OW Non-Identifiability)**.** *For any finite collection of valid pairs with distinct evaluation cases, two decision rules can agree on every OW verdict while returning y_i^A and y_i^O, respectively, in AW. Identical OW verdicts can therefore coexist with opposite AW correctness.*

Proof of Proposition 4 (Appendix D). We consider n valid pairs whose $2n$ evaluation cases are distinct. We construct two rules that agree on every OW case and disagree in AW. The responsive rule follows the reference verdict in each world. The invariant rule repeats its OW verdict in AW:

$$\begin{aligned} d_{\text{resp}}(e_i^O) &= y_i^O, \qquad d_{\text{resp}}(e_i^A) = y_i^A, \\ d_{\text{inv}}(e_i^O) &= y_i^O, \qquad d_{\text{inv}}(e_i^A) = y_i^O. \end{aligned} \tag{24}$$

The rules may be extended arbitrarily outside the observed cases. Both produce the same perfect OW verdict vector. The responsive rule is correct on every AW case. The invariant rule is wrong on every AW case because $y_i^A = -y_i^O$. No statistic of OW outputs alone therefore identifies AW accuracy, both-world accuracy, or conditional AW failure without assumptions linking the worlds. □

D.6 Verdict Reversal And Correctness

A relational profile records only whether outputs stay equal or change under an intervention. This includes the invariance and construct-sensitivity coordinates of Chen et al. (2026a). The limitation is easiest to see with a binary example. A Judge that always returns the opposite of the correct verdict changes on exactly the same pairs as a perfectly correct Judge. Replacing every verdict J by its complement $-J$ preserves every equality and inequality. The reference rule and its everywhere-wrong complement can therefore have identical relational profiles. *A change pattern describes response structure. Deciding whether the changed verdict is correct additionally requires a label orientation.* This claim concerns those equality/inequality statistics, not a signed transition score that already contains the orientation.

Proposition 5 (Relational Change And Hidden Failure)**.** *On a population with binary, non-tied Judge verdicts in both worlds and reversed references, let $F_J = \mathbf{1}[j^O \neq j^A]$, $p = P(c^O = 1) > 0$,*

and $r = P(F_J = 1)$. Then

$$F_J = \mathbf{1}[c^O = c^A], \qquad \mathrm{HFR} = 1 - P(F_J = 1 \mid c^O = 1). \tag{25}$$

Knowing only p and r gives the sharp unrestricted bounds

$$\max\{0, 1 - r/p\} \le \mathrm{HFR} \le \min\{1, (1 - r)/p\}. \tag{26}$$

At $p = r = 1/2$, identical OW accuracy and verdict-change rate permit every HFR from zero to one. An additional perfect-invariance control does not remove this ambiguity.

Here $F_J = 1$ means that the Judge changes its verdict across the pair, p is original-world accuracy, and r is the overall change rate. On a valid reversal, changing the verdict makes the two correctness indicators equal: the Judge is either right in both worlds or wrong in both. HFR is therefore one minus the change rate among original-world successes. The bounds state how little or how much that conditional change can be when only the separate rates p and r are retained. *Overall sensitivity alone cannot determine whether the Judge responds correctly on the cases for which it was initially credited.*

Proof. Under binary reference reversal, flipping the Judge verdict makes the correctness indicators equal: the pair is either both correct or both wrong. Thus $b = P(c^O = 1, F_J = 1) = P(c^O = c^A = 1)$ and $\mathrm{HFR} = 1 - b/p$. The two-event marginal bounds give $\max(0, p + r - 1) \le b \le \min(p, r)$, yielding Equation 26 (Appendix D). For an explicit two-case example, set both OW references to $+1$, both AW references to -1, and OW predictions to $(+1, -1)$. AW predictions $(-1, -1)$ give HFR zero, whereas $(+1, +1)$ give HFR one. Both records have $p = r = 1/2$. Assigning each extra reference-preserving control the same prediction as its base case makes invariance perfect in both records. Mixtures establish the intermediate probabilities. □

This provides the exact relationship to a relational sensitivity profile. Once change is conditioned on OW correctness, it recovers $1 - \mathrm{HFR}$ on the stated decided population. Once the full paired verdicts and labels are retained, HFR is directly computable. The result justifies preserving that anchor. It does not deny the value of invariance controls, claim a new form of contrast consistency, or evaluate the empirical success of another paper's models. The primary WARD analysis includes Judge ties as errors, so the binary identity is used only in the explicitly restricted sensitivity below. An everywhere-wrong rule has no OW-correct denominator and is not assigned an HFR.

World Responsiveness Violation. Definition 1 (Appendix D) conditions on $c^O_{iJ} = 1$ because conventional evaluation first credits that case as a success. On this population, World responsiveness requires $c^A_{iJ} = 1$. The indicator $h_{iJ} = c^O_{iJ}(1 - c^A_{iJ})$ equals one exactly when an OW success becomes an AW error. When $c^O_{iJ} = 0$, $h_{iJ} = 0$ is outside the conditional assessment and does not establish responsiveness.

D.7 Correctness And Two-Sided Verification

Behavioral Role Target. The external Role target compares the observed binary support decision with true Judge correctness under a declared interface. The implemented protocol additionally records the Verifier's own Stage-A verdict before showing Judge scores. B assesses score agreement with this frozen verdict, which may itself be wrong. Thus $v = z$ is an externally evaluated diagnostic outcome. It is not a theorem that consistency with A establishes truth. Appendix I.1 separates A correctness, B response relative to A, A-conditioned discrimination, and a joint outcome requiring both A correctness and the appropriate B decision.

Two-Sided Verification. The next claim explains why rejecting wrong judgments is only half of verification. Although a Verifier that withholds support from everything rejects every wrong judgment, it also rejects every correct one. Let P denote empirical proportions over a set containing both classes. Two-sided balanced accuracy is

$$\mathrm{BA}(V) = \tfrac{1}{2}\left[P(v = 1 \mid z = 1) + P(v = 0 \mid z = 0)\right]. \tag{27}$$

Lemma 2 (Degeneracy Of One-Sided Verification)**.** *An always-support Verifier and an always-non-support Verifier each attain $\mathrm{BA} = 1/2$. A perfect value for only one component of balanced accuracy therefore does not establish Role discrimination.*

Proof. An always-support Verifier has $P(v = 1 \mid z = 1) = 1$ and $P(v = 0 \mid z = 0) = 0$. An always-non-support Verifier has the reverse values. Substitution into the balanced-accuracy definition gives $\mathrm{BA} = 1/2$ for both constant policies. □

Reporting both terms prevents an always-support or always-withhold rule from looking reliable because it succeeds on only one class.

Finite Testability. Each requirement has an observable witness of failure. For World responsiveness, a valid OW-correct pair followed by an AW error is sufficient. For Role discrimination, a support decision differing from the externally scored Judge correctness is sufficient. The reference verdicts and recorded outputs therefore make violations decidable on the evaluated finite collection. This is a direct consequence of the construction, not a guarantee that a tested model satisfies either requirement. A stronger claim of role-stable evaluative competence requires both correct judgment and the appropriate verification decision under the stated interface.

D.8 Separating Verdict Error From Score-Matching Error

The implemented two-stage interface admits an exact binary decomposition on non-tied Judge decisions. Stage A may choose the wrong verdict, or Stage B may fail to report whether the Judge agrees with that frozen verdict. The symbols below separate these two events.

Lemma 3 (Binary Verification Identity)**.** *Let $j \in \{-1, +1\}$ be the Judge verdict, $a \in \{-1, +1\}$ the frozen A verdict, and y the reference. Define $z = \mathbf{1}[j = y]$, $q = \mathbf{1}[j = a]$, $A = \mathbf{1}[a = y]$, and $F = \mathbf{1}[v = q]$. Here F records fidelity to the binary score-matching rule, not a claim about internal reasoning. Then*

$$\mathbf{1}[v = z] = \mathbf{1}[A = F], \qquad P(v = z) = P(A = 1, F = 1) + P(A = 0, F = 0). \tag{28}$$

The identity says that external verification is correct exactly when A correctness and B fidelity agree. If both are right, B transfers a correct reference. If both are wrong, the two mistakes cancel at the final binary decision. If exactly one is wrong, the external decision is wrong.

Proof. If A is correct, $q = z$, so external correctness requires $v = q$. If A is wrong, binary verdicts imply $q = 1 - z$, so external correctness requires $v \neq q$. These are precisely the two equal-value cases $A = F$. □

Consequently, a faithful B check can propagate an A error, while an unfaithful check can coincide with the externally correct binary decision. Reporting external accuracy alone does not distinguish those routes. This elementary identity is specific to the binary, non-tied setting. Judge ties require separate treatment. It does not apply directly to macro balanced accuracy. The numerical decomposition in Appendix I, Table A37, reports proportions on the stated decided subset.

The same-case contrast gives an additional witness without needing the Stage-A label. If a non-tied wrong Judge and a correct peer produce opposite verdicts on the same case, exact matching to any one frozen binary verdict must support exactly one. Positive support for both cannot be explained by faithful matching to a shared A verdict, whether A is correct or wrong. Non-support for both also fails that binary rule, while potentially representing neutral outputs. This separates the observable decision pattern from an explanation based solely on A correctness.

D.9 A Consistency Certificate Without The Stage-A Verdict

The two-stage interface promises one frozen A verdict for all Judge comparisons on a case. This restricts the binary support patterns that exact matching can produce. For one case, world and Verifier, the record contains $m \geq 1$ non-tied Judge verdicts $j_1, \ldots, j_m \in \{-1, +1\}$ and supports $v_1, \ldots, v_m$. Neutral scores remain non-support. If B supports a Judge, faithful matching requires A to equal that Judge. If B withholds support, it requires the opposite verdict. Each decision therefore implies one possible A verdict, encoded as $t_\ell = j_\ell(2v_\ell - 1)$. Let n_+ and n_- count the two implied signs.

Proposition 6 (Compatibility With One Frozen Verdict)**.** *The least number of binary matching errors over every possible shared verdict is*

$$\min_{a\in\{-1,+1\}} \sum_{\ell=1}^{m} \mathbf{1}\{v_\ell \neq \mathbf{1}[j_\ell = a]\} = \min(n_+, n_-) = \frac{m - |\sum_\ell t_\ell|}{2}. \tag{29}$$

Exact matching to one shared verdict is possible if and only if all t_ℓ have the same sign. Equivalently, equal Judge verdicts must receive equal binary support, and opposite Judge verdicts must receive opposite support. There are exactly n_+n_- incompatible unordered Judge pairs.

The smaller of n_+ and n_- is the fewest B decisions that must violate exact matching under any single frozen A verdict. *The certificate can therefore expose an internally inconsistent Judge-support record even when the actual Stage-A verdict is unavailable.*

Proof. For a proposed verdict a, the matching decision is correct exactly when $j_\ell(2v_\ell - 1) = a$. Choosing $a = +1$ incurs n_- mismatches and choosing $a = -1$ incurs n_+. The minimum and absolute-sum expressions follow. A pair violates the stated equal/opposite rule exactly when its two t signs differ, giving n_+n_- such pairs and the zero-mismatch equivalence. □

This is an elementary distance to the two admissible binary support patterns. It requires neither the external reference nor a Stage-A label. A positive certificate shows that no possible shared A verdict can make all observed binary decisions faithful. It does not locate an internal cause: the common-input contract may have failed, or B may have failed to match the supplied verdict. With verified A linkage, the actual mismatch count is at least this minimum. With missing linkage, the incompatibility statement remains conditional only on the proposed common-A explanation. It does not establish that the common-input rule was followed during inference. A zero certificate establishes compatibility, not correctness: all decisions can faithfully support a wrong verdict. For $m = 1$ it is necessarily zero, and $m = 0$ is unassessed.

Faithfulness Can Preserve Error. On a binary non-tied decision, Equation 28 (Appendix D) implies that perfect B fidelity makes external correctness exactly equal to A correctness. More generally,

$$P(v = z) - P(A = 1) = P(A = 0, F = 0) - P(A = 1, F = 0), \tag{30}$$

so $|P(v = z) - P(A = 1)| \leq P(F = 0)$. Correcting an unfaithful B decision helps external accuracy where A is correct and hurts it where A is wrong. This is a pointwise logical consequence of the binary interface, not a causal estimate of an actual intervention. A reference-informed condition removes one potential information deficit but still requires appropriate use of the supplied information.

D.10 Information Available From Evaluation Records

The related-work comparison concerns information supplied by an evaluation, rather than an exclusive classification of entire papers. A paper may report several diagnostics, retain richer data than its headline score, or couple different objects. We distinguish four records on WARD's declared cases. $\mathcal{O}_0$ retains the OW Judge outputs and reference verdicts. $\mathcal{O}_1$ adds their case-linked AW counterparts. $\mathcal{O}_2$ additionally retains marginal Verifier-score distributions for each task, world, ordered Judge/Verifier pair and Judge-correctness class. It can also retain the full A verdicts and score marginals within A-correctness and eligibility strata. $\mathcal{O}_3$ retains the B scores with their case identities, so the relevant verification decisions are linked. These are observation maps of an evaluation record D, not assumptions that the models have particular abilities. Unlike the main audit classes, these four nested records retain external labels and ask whether a diagnostic is identified exactly.

Consider two complete experiments that become identical after a report discards case identities. If their target values differ, that report cannot determine the target. Identification names the absence of this ambiguity.

Definition 3 (Identification From An Evaluation Record)**.** A target $Q(D)$ is identified by observations $\mathcal{O}(D)$ when $\mathcal{O}(D) = \mathcal{O}(D')$ implies $Q(D) = Q(D')$ for every two admissible records. If two compatible records give different target values, the observations do not identify that target. Bounds may still constrain it.

In other words, an observation identifies Q only when every complete record compatible with it gives the same answer. Otherwise the observation may still bound Q. It cannot recover Q exactly.

World Information. Proposition 4 (Appendix D) supplies two records with identical $\mathcal{O}_0$ and different AW accuracy and HFR. In contrast, $\mathcal{O}_1$ identifies every World statistic reported here by direct evaluation. Indeed, within one task and cohort, OW accuracy and both-world accuracy already identify HFR through Equation 14 (Appendix B). The information claim therefore credits controlled and joint-correctness evaluations. It does not establish HFR as an independent statistic or make WARD the only way to compute it. World records alone leave B outputs unspecified and cannot determine unrestricted verification behavior.

Marginal And Linked Verification. $\mathcal{O}_2$ identifies each individual Role measure, including every score-threshold rate and AUROC when both correctness classes occur. Its same-case conjunctions need not be identified. To see this, take two distinct eligible cases with the same correctness profile: the focal Judge is OW-correct and AW-wrong and a fixed peer is AW-correct. Keep all Judge outputs, references and A verdicts fixed. In the first record, the focal Verifier's own/peer support pairs are $(0, 1)$ and $(1, 0)$, giving rescue $1/2$. In the second they are $(0, 0)$ and $(1, 1)$, giving rescue zero. Each verification stream has one positive and one negative score in either record. Choose scores $+1$ and -1. The own and peer class-conditional score distributions are identical. Both records can have correct A verdicts on both cases, so A-correctness conditioning does not distinguish them. Identical padding cases can supply the opposite correctness classes for AUROC. Distinct score inputs allow the records to be realized by different Verifier rules under the same interface. Thus neither perfect A accuracy nor the complete marginal score summaries identify rescue.

For CMS, take two cases with both Judges wrong and preserve their World outputs and A verdicts. Four support streams with rows $(1, 1, 1, 1)$ and $(0, 0, 0, 0)$ give CMS $1/2$. Rows $(1, 1, 0, 0)$ and $(0, 0, 1, 1)$ give zero. Every stream has the same score multiset in both records. This second construction proves non-identification of the four-way event from its component marginals. Both constructions are logical counterexamples, not experimental results. They establish existence of ambiguity. Some observed marginals determine the joint event exactly.

Knowing how often each action occurs does not tell us how often they occur together. Still, their separate rates limit the overlap. For example, two events each occurring on 90% of cases must overlap on at least 80% and at most 90%. The following result gives the corresponding bounds for any number of actions.

Proposition 7 (Marginal Bounds For Linked Verification)**.** *On a fixed nonempty conditional population, let $X_1, \ldots, X_k$ be binary events with known marginal probabilities $p_1, \ldots, p_k$. With no additional dependence restrictions, their conjunction satisfies the sharp bounds*

$$\max\left\{0, \sum_{j=1}^{k} p_j - (k-1)\right\} \leq P(X_1 = \cdots = X_k = 1) \leq \min_j p_j. \tag{31}$$

Rescue uses $k = 2$, with own-error non-support and correct-peer support on eligible comparisons. Conditional CMS uses $k = 4$, with the four support decisions on jointly wrong AW cases. The conjunction is identified from these marginals alone exactly when the two bounds coincide.

Proof. The intersection is a subset of every event, giving the upper bound. Applying the union bound to the complements gives the lower bound. For the upper endpoint, nest events of measures p_j on $[0, 1]$. For the lower endpoint, place complement intervals of lengths $1 - p_j$ consecutively around the unit circle. Their union has measure $\min\{1, \sum_j (1 - p_j)\}$, giving the stated intersection. Mixing the two constructions preserves the marginals and attains intermediate probabilities. This establishes sharpness over unrestricted probability couplings and the identification criterion. Finite equal-weight tables additionally restrict feasible probabilities to integer multiples of the reciprocal sample size. The endpoint count formulas remain attainable when the specified marginal counts are integral. □

Evaluation reports can use these bounds to distinguish what separate rates already establish from what requires matched decisions. These are classical Fréchet bounds, applied to the specific WARD

events. The bounds themselves are not a new mathematical result (Dobra & Fienberg, 2000). Additional dependence assumptions, repeated-stream structure or retained joint cells can tighten them. In particular, multiplying marginal rates would require an independence assumption. The empirical bounds below retain only the stated within-task conditional marginals. They are valid outer bounds when other record structure is also known and are not confidence intervals. We apply them within each task before averaging. A pooled or macro marginal cannot silently replace those conditional populations. $\mathcal{O}_3$ identifies the events directly by counting their observed conjunctions. This specifies sufficient measurements for WARD's target, without a universal minimality claim for all evaluation tasks.

A Accuracy And B Fidelity. Equation 28 (Appendix D) supplies a second application. On binary non-tied decisions let $a = P(A = 1)$ and $f = P(F = 1)$. Since $P(v = z) = 1 - a - f + 2P(A = 1, F = 1)$, the sharp unrestricted bounds are

$$|a + f - 1| \leq P(v = z) \leq 1 - |a - f|. \tag{32}$$

Marginal A accuracy and B fidelity therefore constrain external accuracy without generally determining it. For example, $a = f = 0.9$ permits external accuracy from 0.8 to 1. If A is perfect, the interval collapses to f. The subsequent check can still be inaccurate when fidelity is low. *Supplying the true reference to a Verifier similarly imposes no nontrivial performance guarantee without a behavioral assumption about using it.* A constant non-support rule receives the same reference and retains BA $1/2$ on a two-class population. This logical boundary motivates the reference-informed condition. Its empirical failure rate is established by the model outputs, not by this construction.

How The Categories Apply To Prior Work. Single-world agreement summaries supply $\mathcal{O}_0$-type evidence. Contrast sets and controlled Judge comparisons supply paired observations on their respective tested objects, potentially identifying their joint correctness. ContextualJudgeBench additionally checks explanations on correct outcomes. CriticEval and IF-CRITIC assess checking or consistency, while CriticBench already couples generation, critique and correction on the same questions. They must not be classified as universally marginal or unlinked. The theorem applies only if the retained observations for the target in question omit the relevant coupling. Appendix B, Table A2, maps the inspected assessments to this distinction. Retaining or collecting WARD's relevant case-linked decisions would close the corresponding information gap, irrespective of the name of the method.

Scope. Every World claim is conditional on the task-specific construction checks. Every Role claim is conditional on the declared Verifier prompt and information set. Observed success on these reference-grounded cases establishes no reliability guarantee outside them. The main audit's universal quantifier instead ranges over admissible references for one fixed observed cohort.

E Tasks, Datasets, And Judge Case Studies

What The Judge Evaluates. The World diagnostic evaluates the Judge's verdict about a supplied task answer. The Role diagnostic then evaluates support for that verdict. Their linkage supplies the evidence used by the central verification audit. Each evaluation case therefore has three parts: a task input, an evaluated answer or response, and a criterion for judging that response. The evaluated response is supplied before the Judge call. The Judge produces a separate verdict about whether that response satisfies the criterion. The World Test changes the case between OW and AW while keeping the criterion fixed, so that the correct Judge verdict reverses. The six tasks differ in which part of the case supplies this controlled contrast.

E.1 Dataset And Task Discussion

CAD Sentiment. Counterfactually Augmented Data (CAD) supplies an original review and a human-edited counterfactual review (Kaushik et al., 2020). The evaluated answer is a proposed sentiment label. OW uses the original review. AW uses its human-edited counterpart while keeping the proposed label and the label-correctness criterion fixed. For example, a proposed label can be correct for the original review, making *accept* the correct Judge verdict, and become wrong after the edit, making *reject* correct. We retain all 2,195 pairs.

CrowS-Pairs Stereotypes. CrowS-Pairs supplies a human-written pair containing a more-stereotypical sentence and a less-stereotypical sentence (Nangia et al., 2020). The evaluated response is one sentence from this pair. The fixed criterion asks whether the sentence is acceptable with respect to the targeted stereotype. Under this criterion, moving from the more-stereotypical sentence in OW to its less-stereotypical partner in AW reverses the reference verdict from *reject* to *accept*. We retain all 1,508 pairs.

PKU Helpfulness And Safety. PKU-SafeRLHF supplies two curated assistant responses for one fixed prompt (Ji et al., 2025). One response appears in OW and its paired response appears in AW. We apply this pair twice: once under a helpfulness criterion and once under a safety criterion. The prompt and the active criterion stay fixed within each task, and the reference verdict reverses between the paired responses. The same 2,500 source records therefore yield 2,500 helpfulness pairs and 2,500 safety pairs. These are two task views of the same 2,500 source records.

BBQ And ES-BBQ Question Answering. BBQ provides English question-answering cases designed to expose social bias, while ES-BBQ adapts this benchmark to Spanish (Parrish et al., 2022; Ruiz-Fernández et al., 2025). We retain only pairs for which the correct Judge verdict strictly reverses under the fixed bias-focused criterion: 6,880 BBQ pairs and 2,456 ES-BBQ pairs. These tasks extend the audit to question answering and a second language. They are separate task cohorts rather than item-matched translations, so their results support cross-setting replication but not a causal comparison between languages.

The construction map fixes the interpretation of the results. The CrowS verdict concerns the targeted stereotype, and the PKU verdict concerns the recorded criterion and paired-response construction. Neither is an unconditional judgment of a sentence or response in every context. In Table A6 (Appendix E), we record the population, and in Table A7 (Appendix E), we identify the changed and fixed fields. Inclusion depends on the source construction rather than any Judge's success. The World claim is conditional on these validated contrasts.

How These Choices Support The Audit. The paired structure supplies the reference relation required by WARD's construction. CAD changes the input while retaining the proposed answer; CrowS, PKU, BBQ and ES-BBQ change the evaluated case under a fixed task criterion. This variation tests the same audit mechanism across distinct contrasts. PKU additionally holds the underlying prompt-response pairs fixed across two criteria, exposing task dependence without treating the two views as independent datasets. BBQ and ES-BBQ test whether the certificate repeats on larger question-answering cohorts and in Spanish. The information ablations can then remove relations or associations from the same model decisions. The resulting comparison tests what linkage establishes, while the construction-specific criteria limit where that conclusion applies.

Guide To The Completed Results. Appendix G reports the primary information ablations and certificate bounds for all six tasks and five Verifiers. Table A17 in Appendix H supplies all thirty World profiles. Tables A30 and A31 in Appendix I supply the thirty self and thirty peer task profiles, while Table A32 retains all ordered model combinations. Secondary selection and uncertainty analyses that were not rerun on BBQ and ES-BBQ are labeled as original-four-task results in their captions. Different cohorts and conditional denominators remain explicit rather than being pooled into one model ranking.

E.2 Model And Baseline Discussion

We use five open-weight checkpoints as diagnostic case studies. Their primary purpose is to test whether the linked audit changes the information and conclusions available from fixed verification records. Grounded World and Role diagnoses explain the observed behavior. We do not use them to create a model leaderboard or to estimate how common the observed behaviors are across all LLMs. The set covers four model families and checkpoint sizes from 8B to 32B.

Every checkpoint fills both roles. In an ordered pair (J, V), model J first produces the Judge judgment and model V later checks it as a Verifier. Five choices for J and five choices for V give $5 \times 5 = 25$ ordered pairs: five self pairs and twenty peer pairs. The complete grid lets us compare self- and peer-verification under the same recorded protocol.

Table A6: Separate World and complete Role populations for six tasks. Counts are task-specific pairs. The two PKU criteria share underlying record identities.

Task	World pairs	Role pairs	A outputs	Observed B
CAD	2,195	2,195	21,950	109,750
CrowS-Pairs	1,508	1,508	15,080	75,400
PKU-Helpfulness	2,500	2,498	25,000	124,950
PKU-Safety	2,500	2,498	25,000	124,950
BBQ	6,880	6,880	68,800	344,000
ES-BBQ	2,456	2,456	24,560	122,800
Total	18,039	18,035	180,390	901,850

The full World cohort contains 36,078 world cases and 180,390 Judge outputs. The complete Role cohort contains 36,070 world cases and 901,750 B decisions over all 25 ordered Judge/Verifier combinations. The remaining 100 observed PKU B decisions belong to four excluded partial task pairs.

Table A7: Task-specific OW/AW construction used in the analysis.

Task	What Changes From OW To AW	What Stays Fixed	Reference Verdict	Pairs
CAD	Original review → human-edited counterfactual	Proposed sentiment label. Label-correctness criterion	accept → reject	2,195
CrowS-Pairs	More-stereotypical → less-stereotypical sentence	Criterion: whether the sentence is acceptable with respect to the targeted stereotype	reject → accept	1,508
PKU-Helpfulness	One curated response ↔ its paired response	Prompt. Helpfulness criterion	helpful ↔ not helpful	2,500
PKU-Safety	One curated response ↔ its paired response	Prompt. Safety criterion	safe ↔ unsafe	2,500

Every dataset-provided controlled pair is retained. Conclusions remain conditional on the listed constructions, and a new source or judging criterion requires fresh quality control.

The two Qwen checkpoints permit a within-family comparison, while the other families broaden the observed decision patterns. These comparisons do not isolate a causal effect of parameter count or training procedure. The certificate need not tighten for every model. For DeepSeek, combining the links leaves the certified error bound unchanged; the other four Verifiers have larger joint bounds (Table 2). The full grid also allows the shared-reference analysis to test whether separate audits rely on mutually incompatible explanations of the same case.

Baselines For The Core Claim. The primary baselines are restricted evidence views of identical decisions, rather than additional competing LLMs. The four-class comparison in Appendix G, Table A10, retains neither link, only the World relation, only the Role association, or both. It therefore isolates the information that makes certification possible. The extra-gain column compares joint evidence with the stronger partial certificate on each pair before aggregation. Reference-grounded diagnostics provide a separate comparison with observed errors when correct verdicts are available. These are comparisons of audit information. They do not claim empirical superiority over every method in the related-work tables. The concrete methods and their retained evidence are assessed separately in Appendix B.

F Protocol Coverage And Reproducible Analysis

World-only diagnostics retain all 18,039 pairs and 180,390 Judge outputs across six tasks. The complete two-stage Role cohort retains 18,035 pairs, 180,350 Stage-A case/model outputs and 901,750 Stage-B decisions. The observed Stage-B archive contains 901,850 outputs; the remaining 100 belong to four partial PKU task pairs. Source identities 31658 and 65455 each lack one Stage-B world under both PKU criteria. World and Role results therefore have different primary denominators. Combined diagnostics use the complete cohort stated in their captions. All 93,360 BBQ and ES-BBQ World score cells exactly match the Judge inputs frozen into Stage B.

Interfaces And Historical Comparison. The two-stage interface supplies the case and criterion to A without Judge scores or references, freezes its verdict, and supplies that verdict with accep-

Table A8: Model checkpoints used as diagnostic case studies. The second column gives the name used in the results. Every checkpoint acts as both a Judge and a Verifier.

Checkpoint / Hub ID	Name In Results	Model Line	Size
deepseek-ai/DeepSeek-R1-Distill-Llama-8B (DeepSeek-AI, 2025)	DeepSeek-R1-8B	DeepSeek-R1 distillation (Llama base)	8B
mistralai/Mistral-Nemo-Instruct-2407 (Mistral AI Team, 2024)	Mistral-Nemo-12B	Mistral NeMo	12B
Qwen/Qwen2.5-14B-Instruct (Yang et al., 2025)	Qwen2.5-14B	Qwen2.5	14B
Qwen/Qwen2.5-32B-Instruct (Yang et al., 2025)	Qwen2.5-32B	Qwen2.5	32B
google/gemma-2-27b-it (Gemma Team et al., 2024)	Gemma-2-27B	Gemma 2	27B

The checkpoints are case studies rather than leaderboard candidates or a sample used to estimate how common these behaviors are across all LLMs. The complete role grid contains five self pairs and twenty peer pairs.

Table A9: Cohort coverage and numerical validation of the fixed model outputs.

Check	Observed Status
Frozen Judge records	180,390 score rows on 18,039 paired cases; World point estimates independently reproduced.
Stage-A records	180,390 supplied verdict outputs, reused across five Judge comparisons.
Observed Stage-B coverage	901,850 outputs; only the original PKU cohort contains four partial task pairs.
Complete paired B grid	901,750 finite B scores on 18,035 pairs; 25 ordered cells in both worlds.
World/Role correspondence	All 93,360 new World score cells exactly match the Judge inputs frozen into Stage B.
Reference alignment	Every retained BBQ and ES-BBQ pair reverses its correct verdict under the frozen curation contract.
Independent numeric checks	Six-task World, Stage-A, self/peer Role, network, and exact certificate quantities reproduced from saved outputs.

tance/rejection scores to B. The same A output is reused for all five Judges. Reference labels are joined offline. By contrast, the historical execution supplied reference-informed original-choice information to the Verifier. It cannot establish blind verification. A single-stage reference-blind interface would assess a case and Judge judgment without reference information. Receiving Judge scores alone would not invalidate that blindness. The study provides no completed evaluation of that interface. Because the reference-informed and two-stage runs differ in several interface features, their difference cannot be attributed to reference-label removal alone.

DeepSeek Final-Output Audit. Across the final B outputs, DeepSeek has 87,009 exact zeros and one score of 0.2 among 87,010 valid outputs. The exception is an OW PKU-helpfulness comparison for source identity 2694 and the Gemma Judge. The generation log contains 87,138 attempts because 128 earlier aborted safety attempts contain 114 empty and 14 truncated outputs. These are not missing final outputs. For all 21,950 CAD outputs, the stronger trace audit checks prepared prompts, raw text, parsed scores, and frozen-A linkage. The inspected generation path passes the returned model text to storage and parsing. No default-zero branch was found in that path, and parser checks preserve positive and negative values while flagging malformed or missing text. This supports an observation of model-returned zero text in the inspected CAD chain, without identifying its cause.

The CAD prompt evidence includes 13,599 score directions agreeing with A, 7,371 opposing A, and 980 ties. Among 803 extreme 1-versus-0 score pairs, 579 agree and 224 disagree with A. All receive zero. Thus, zero output is not restricted to tied Judge scores. The executed prompt contains a zero-valued schema example and an unfilled tie-rule instruction. These are observed prompt properties, not established causes.

Recorded Inference Settings. The audited DeepSeek CAD Stage-B execution uses `llm.chat`, structured JSON decoding, temperature 0, top-p 1, seed 0, and at most 150 generated tokens. It contains 21,950 valid final completions. These settings are documented for that execution only.

Table A10: Equal-task information ablation across six tasks (%). Extra is the pairwise additional floor supplied by retaining both links.

Verifier	Neither	R_M	L_0	R_L	Extra
DeepSeek	14.42	37.94	14.42	37.94	0.00
Mistral	12.73	28.27	14.17	35.55	6.37
Qwen14	9.41	26.22	11.06	30.89	3.70
Qwen32	6.64	19.52	8.06	27.11	6.77
Gemma	6.01	21.17	7.61	28.88	6.77

Other checkpoints and tasks are not assigned these settings. Frozen output files make the diagnostic calculations repeatable without rerunning generation.

Analysis And Reproducibility. We join the fixed Judge matrices with the parsed two-stage outputs. The analysis fixes score signs, the 10^{-12} Judge tie tolerance, the strict $s > 0$ support threshold, task weighting, and cohort membership. Independent checks reproduce the World point estimates, self/peer Role components and AUROCs, the Qwen14/Qwen32 complete-mutual-support counts, and panel totals. The case bootstrap uses 2,000 linked-identity replicates and fixed seed 20260911, while selection uses 500 fixed splits per task and its declared derived seeds. These seeds define the reported resampling analyses. Model outputs remain fixed during these analyses. All calculations use the declared cohorts and scoring rules.

Asset Roles And Usage Terms. The datasets provide the cases and criterion labels: CAD supplies sentiment edits, CrowS-Pairs supplies stereotype contrasts, and PKU-SafeRLHF supplies response comparisons and helpfulness/safety annotations. The CAD repository lists Apache 2. This repository license does not independently establish redistribution rights for the underlying IMDb material. The CrowS-Pairs repository specifies CC BY-SA 4, and the PKU-SafeRLHF dataset card specifies CC BY-NC 4. The five checkpoints are evaluated subjects, with sources cited in Section 5. The Qwen2.5-14B, Qwen2.5-32B and Mistral-Nemo model cards list Apache 2. The Gemma card links its Gemma terms. The DeepSeek card lists MIT and identifies the Llama 3.1 parent license. These source declarations were checked on 13 September 2026. A release must retain applicable attribution and restrictions.

G Information Ablations And Audit Decisions

The same verification decisions can admit an error-free explanation when their records are separated and force errors when those records are linked. We quantify this effect on the complete six-task cohort. One restricted audit retains the opposite-verdict relation and local totals while hiding which verdict received each support. A second audit retains the verdict-support links while omitting the relation between cases. WARD retains both. The model outputs remain fixed throughout.

The complete Role cohort contains 18,035 pairs: 2,195 CAD, 1,508 CrowS-Pairs, 2,498 PKU-Helpfulness, 2,498 PKU-Safety, 6,880 BBQ and 2,456 ES-BBQ pairs. Excluding tied Judge verdicts leaves 153,269 Stage-B decisions per Verifier. Neutral Stage-B scores count as non-support.

G.1 Error Bounds By Information Class

An error floor is the smallest number of wrong verification decisions compatible with the information available to the auditor. A larger floor proves that more errors must exist. The four columns in Table A10 retain neither link, only the case relation, only the verdict-support links, or both. We denote these floors by D_{00}, D_{10}, D_{01} and D_{11}. Thus D_{10} corresponds to R_M, D_{01} to L_0, and D_{11} to WARD's linked view R_L. Hiding the matches allows an explanation to reassign supports while preserving their totals. Hiding the relation allows it to choose the two correct verdicts independently. These are explanations permitted by missing information, not changes to the recorded outputs. Theorem 3 gives the exact local counts.

Table A11: Task-level minimum compatible error (%). Neither retains either link; R_M retains cross-case relations; L_0 retains verdict-support links; R_L retains both. Extra compares the joint certificate with the stronger separate certificate on each pair.

Task	Verifier	Neither	R_M	L_0	R_L	Extra
CAD	DeepSeek	11.19	17.81	11.19	17.81	0.00
CAD	Mistral	15.62	27.15	16.15	27.63	0.18
CAD	Qwen14	5.63	10.33	6.30	11.40	0.91
CAD	Qwen32	3.34	8.59	4.37	10.65	1.89
CAD	Gemma	3.58	9.95	4.47	14.48	4.18
CrowS-Pairs	DeepSeek	8.82	38.90	8.82	38.90	0.00
CrowS-Pairs	Mistral	19.27	34.36	21.00	36.27	0.89
CrowS-Pairs	Qwen14	6.11	33.37	6.73	37.24	3.50
CrowS-Pairs	Qwen32	0.74	29.65	0.87	37.86	8.14
CrowS-Pairs	Gemma	1.77	30.77	2.17	38.76	7.75
PKU-Helpfulness	DeepSeek	20.44	45.16	20.44	45.16	0.00
PKU-Helpfulness	Mistral	3.31	27.31	3.92	45.76	18.02
PKU-Helpfulness	Qwen14	13.86	27.29	18.59	37.44	6.87
PKU-Helpfulness	Qwen32	10.10	24.38	16.01	39.23	11.20
PKU-Helpfulness	Gemma	8.43	22.91	11.65	39.27	14.57
PKU-Safety	DeepSeek	8.39	46.22	8.39	46.22	0.00
PKU-Safety	Mistral	17.72	35.54	20.24	39.62	2.68
PKU-Safety	Qwen14	6.96	41.12	7.90	44.06	2.35
PKU-Safety	Qwen32	22.89	38.93	24.16	44.02	4.12
PKU-Safety	Gemma	8.63	34.35	11.63	40.10	3.89
BBQ	DeepSeek	19.30	40.35	19.30	40.35	0.00
BBQ	Mistral	13.64	24.99	15.08	33.84	7.86
BBQ	Qwen14	9.42	26.08	10.53	32.59	6.12
BBQ	Qwen32	1.88	7.29	2.02	14.00	6.65
BBQ	Gemma	6.92	15.69	7.99	22.82	6.42
ES-BBQ	DeepSeek	18.36	39.18	18.36	39.18	0.00
ES-BBQ	Mistral	6.82	20.26	8.59	30.20	8.59
ES-BBQ	Qwen14	14.49	19.09	16.31	22.64	2.45
ES-BBQ	Qwen32	0.88	8.26	0.93	16.89	8.60
ES-BBQ	Gemma	6.76	13.33	7.73	17.82	3.81

For every case pair, we compare WARD with the stronger of the two restricted audits before aggregation. If i indexes pairs, M_i is their number of retained decisions and $\mathcal{T}$ contains the six tasks, the additional guaranteed error fraction is

$$\frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}}\frac{\sum_{i\in t}\left[D_{11,i}-\max(D_{10,i},D_{01,i})\right]}{\sum_{i\in t}M_i}. \tag{33}$$

Selecting the stronger restricted floor within each pair prevents an apparent gain caused by different summaries working well on different cases. The additional floor is 6.37 points for Mistral, 3.70 for Qwen14, 6.77 for Qwen32 and 6.77 for Gemma. DeepSeek gains zero. *The gain is information recovered by linking unchanged decisions. It is not an increase in errors caused by WARD.*

The task breakdown establishes that the result is not confined to the original four datasets. On BBQ, joint evidence adds 6.12 to 7.86 points for the four nondegenerate Verifiers. On ES-BBQ, it adds 2.45 to 8.60 points. DeepSeek remains the zero-gain boundary on both because its blanket non-support already makes many errors unavoidable under the relation-only view.

G.2 Sensitivity To Invalid Relations

For a trusted pair, the correct verdicts in OW and AW must be opposite. Relaxing the relation also allows them to be equal. We permit up to $\lfloor 0.1n_t \rfloor$ invalid relations separately in task t and select the relations whose relaxation most reduces the error floor. This is the most favorable permitted choice for each Verifier, rather than average random corruption.

With every relation trusted, the pooled floors range from 24.50% to 38.69%. With the 10% task-wise allowance, they remain between 19.33% and 33.58% (Table A12). The corresponding equal-task

Table A12: Exact pooled error floors on 153,269 retained decisions per Verifier across six tasks. The sensitivity permits at most 10% invalid relations within each task.

Verifier	Nominal count	Nominal (%)	Budget count	Budget (%)
DeepSeek	59,300	38.69	51,473	33.58
Mistral	54,432	35.51	46,562	30.38
Qwen14	48,169	31.43	40,671	26.54
Qwen32	37,552	24.50	29,632	19.33
Gemma	42,375	27.65	34,326	22.40

Table A13: CAD sensitivity on 2,195 pairs and 20,970 retained decisions per Verifier. The three rows use relation budgets of 0%, 10%, and 100%.

Verifier	Pair budget	Error floor	FPR range	FNR range
DeepSeek	0	17.81	[0.00, 0.00]	[100.00, 100.00]
DeepSeek	219	13.48	[0.00, 0.00]	[100.00, 100.00]
DeepSeek	2195	11.19	[0.00, 0.00]	[100.00, 100.00]
Mistral	0	27.63	[0.33, 25.86]	[74.14, 99.67]
Mistral	219	23.01	[0.00, 29.94]	[70.06, 100.00]
Mistral	2195	16.15	[0.00, 41.19]	[58.81, 100.00]
Qwen14	0	11.40	[57.39, 99.65]	[0.35, 42.61]
Qwen14	219	6.94	[44.46, 99.94]	[0.06, 55.54]
Qwen14	2195	6.30	[42.05, 99.94]	[0.06, 57.95]
Qwen32	0	10.65	[16.41, 93.68]	[6.32, 83.59]
Qwen32	219	5.77	[0.47, 95.72]	[4.28, 99.53]
Qwen32	2195	4.37	[0.42, 96.11]	[3.89, 99.58]
Gemma	0	14.48	[51.41, 96.44]	[3.56, 48.59]
Gemma	219	9.26	[34.03, 97.70]	[2.30, 65.97]
Gemma	2195	4.47	[15.10, 97.93]	[2.07, 84.90]

ranges are 27.11–37.94% and 21.95–32.91%, respectively. *Uncertainty about a minority of relations weakens the numerical guarantee without removing it.* The allowance is a sensitivity parameter, not an estimate that 10% of the constructions are invalid.

The following four tables retain the original task-specific conditional-rate analysis. BBQ and ES-BBQ were not included in this secondary decomposition, as their corresponding extrema were not rerun. Each table states its own cohort.

A total-error floor does not always reveal which error type dominates. False support accepts a wrong judgment; missed support withholds support from a correct judgment. The tables give the compatible range of each rate under the stated relation budgets. Endpoints from different columns need not occur under the same reference assignment. DeepSeek gives a simple boundary: its zero support avoids false support and misses every correct judgment whenever both strata are nonempty.

G.3 Interpretation And Validation

The certificate uses recorded Judge verdicts and Stage-B support links. It does not require the independently elicited Stage-A verdict or the correct label for each individual case. Exact labels are used only for the grounded diagnostic comparison. The error floor also does not identify which decisions are wrong. It establishes that every explanation consistent with the retained information contains at least that many errors.

Deterministic checks reproduce the six-task World, Stage-A, self and peer Role, network and certificate quantities from the saved outputs. They verify the monotonicity $D_{11} \geq D_{10}, D_{01}$ and the task-wise budget optimization. The complete data contain 901,750 Stage-B decisions, of which 766,345 have non-tied Judge verdicts across all five Verifiers and satisfy the binary Stage-A/Stage-B identity. These checks support the algebra and implementation. They do not validate future model outputs or the semantic quality of a new transformation.

Table A14: CrowS sensitivity on 1,508 pairs and 12,170 retained decisions per Verifier. The three rows use relation budgets of 0%, 10%, and 100%.

Verifier	Pair budget	Error floor	FPR range	FNR range
DeepSeek	0	38.90	[0.00, 0.00]	[100.00, 100.00]
DeepSeek	150	33.83	[0.00, 0.00]	[100.00, 100.00]
DeepSeek	1508	8.82	[0.00, 0.00]	[100.00, 100.00]
Mistral	0	36.27	[12.83, 36.20]	[63.80, 87.17]
Mistral	150	31.31	[6.57, 40.84]	[59.16, 93.43]
Mistral	1508	21.00	[0.14, 55.49]	[44.51, 99.86]
Qwen14	0	37.24	[80.85, 96.45]	[3.55, 19.15]
Qwen14	150	32.17	[76.54, 99.98]	[0.02, 23.46]
Qwen14	1508	6.73	[37.46, 99.99]	[0.01, 62.54]
Qwen32	0	37.86	[75.35, 96.50]	[3.50, 24.65]
Qwen32	150	32.61	[71.12, 99.94]	[0.06, 28.88]
Qwen32	1508	0.87	[5.16, 100.00]	[0.00, 94.84]
Gemma	0	38.76	[76.01, 94.66]	[5.34, 23.99]
Gemma	150	33.59	[71.75, 98.20]	[1.80, 28.25]
Gemma	1508	2.17	[4.52, 99.09]	[0.91, 95.48]

Table A15: PKU-helpfulness sensitivity on 2,498 pairs and 23,830 retained decisions per Verifier. The three rows use relation budgets of 0%, 10%, and 100%.

Verifier	Pair budget	Error floor	FPR range	FNR range
DeepSeek	0	45.16	[0.00, 0.01]	[99.99, 100.00]
DeepSeek	249	40.33	[0.00, 0.01]	[99.99, 100.00]
DeepSeek	2498	20.44	[0.00, 0.02]	[99.98, 100.00]
Mistral	0	45.76	[16.17, 25.98]	[74.02, 83.83]
Mistral	249	40.54	[10.93, 30.79]	[69.21, 89.07]
Mistral	2498	3.92	[0.01, 86.01]	[13.99, 99.99]
Qwen14	0	37.44	[42.03, 67.08]	[32.92, 57.97]
Qwen14	249	32.53	[35.94, 72.07]	[27.93, 64.06]
Qwen14	2498	18.59	[22.34, 90.91]	[9.09, 77.66]
Qwen32	0	39.23	[53.77, 74.77]	[25.23, 46.23]
Qwen32	249	34.00	[48.21, 81.07]	[18.93, 51.79]
Qwen32	2498	16.01	[26.93, 99.29]	[0.71, 73.07]
Gemma	0	39.27	[24.54, 46.79]	[53.21, 75.46]
Gemma	249	34.05	[18.53, 52.16]	[47.84, 81.47]
Gemma	2498	11.65	[1.20, 85.73]	[14.27, 98.80]

H JUDGMENT DIAGNOSTICS

A Judge can perform well on the original case and still fail when the case changes in a way that requires the opposite verdict. We examine where this happens and whether it survives changes in sampling, score boundaries and task weighting. Here OW means the original case and AW its controlled edited counterpart. The World Test checks Judge correctness on both. The main text gives the six tasks equal weight; the breakdown below reveals behavior hidden by that average. Unless a narrower denominator is stated, each row uses its controlled pairs from the 18,039-pair World population.

Task-Level World Profiles. For each model-task row, OW and AW are the percentages judged correctly in the corresponding world, and Both is the percentage judged correctly in both worlds. Higher values are better for these three measures. Hidden Failure Rate (HFR) instead conditions on OW-correct pairs and measures the percentage that become wrong in AW, so lower is better. A Judge output is treated as a tie when the acceptance and rejection scores differ by no more than 10^{-12}, and such ties count as errors in these strict measures.

The task-level values reveal behavior that the model averages conceal. For Qwen2.5-32B, HFR rises from 10.2% on CAD to 99.8% on PKU-Helpfulness, while Both falls from 84.5% to 0.2%. *For the*

Table A16: PKU-safety sensitivity on 2,498 pairs and 24,446 retained decisions per Verifier. The three rows use relation budgets of 0%, 10%, and 100%.

Verifier	Pair budget	Error floor	FPR range	FNR range
DeepSeek	0	46.22	[0.00, 0.00]	[100.00, 100.00]
DeepSeek	249	41.12	[0.00, 0.00]	[100.00, 100.00]
DeepSeek	2498	8.39	[0.00, 0.00]	[100.00, 100.00]
Mistral	0	39.62	[37.06, 58.03]	[41.97, 62.94]
Mistral	249	34.53	[29.93, 63.62]	[36.38, 70.07]
Mistral	2498	20.24	[0.50, 81.60]	[18.40, 99.50]
Qwen14	0	44.06	[87.42, 97.04]	[2.96, 12.58]
Qwen14	249	38.97	[84.58, 99.94]	[0.06, 15.42]
Qwen14	2498	7.90	[50.72, 99.97]	[0.03, 49.28]
Qwen32	0	44.02	[56.69, 69.31]	[30.69, 43.31]
Qwen32	249	38.93	[51.84, 73.63]	[26.37, 48.16]
Qwen32	2498	24.16	[9.04, 89.78]	[10.22, 90.96]
Gemma	0	40.10	[61.88, 82.73]	[17.27, 38.12]
Gemma	249	35.00	[57.19, 88.51]	[11.49, 42.81]
Gemma	2498	11.63	[11.09, 95.35]	[4.65, 88.91]

Table A17: World-Test profile by task and Judge (%). OW, AW and Both use every pair in the task; HFR conditions on OW-correct pairs. Judge ties count as wrong. The six tasks contain 18,039 pairs.

Task	Judge	OW	AW	Both	HFR	OW tie	AW tie
CAD	DeepSeek	84.3	14.5	14.1	83.3	7.3	18.6
CAD	Mistral	85.0	67.4	60.6	28.7	4.1	4.6
CAD	Qwen14	99.3	77.0	76.6	22.9	0.0	1.4
CAD	Qwen32	94.1	89.9	84.5	10.2	1.7	1.3
CAD	Gemma	95.7	70.3	68.3	28.6	1.2	4.4
CrowS-Pairs	DeepSeek	8.8	42.6	1.0	88.7	90.1	56.4
CrowS-Pairs	Mistral	66.2	48.9	17.8	73.2	5.1	2.5
CrowS-Pairs	Qwen14	56.8	49.0	10.6	81.3	2.4	2.5
CrowS-Pairs	Qwen32	71.7	28.0	6.9	90.4	8.7	9.1
CrowS-Pairs	Gemma	68.7	38.5	10.4	84.8	9.0	7.2
PKU-Helpfulness	DeepSeek	63.2	37.3	4.9	92.3	0.0	0.0
PKU-Helpfulness	Mistral	57.0	30.4	3.5	93.8	11.6	11.9
PKU-Helpfulness	Qwen14	32.5	52.0	5.1	84.4	11.1	10.8
PKU-Helpfulness	Qwen32	64.1	35.4	0.2	99.8	0.3	0.4
PKU-Helpfulness	Gemma	64.5	35.9	4.9	92.4	0.0	0.0
PKU-Safety	DeepSeek	59.4	41.7	5.3	91.0	0.0	0.0
PKU-Safety	Mistral	59.4	39.0	3.3	94.4	2.7	2.3
PKU-Safety	Qwen14	56.8	36.7	2.0	96.5	6.8	7.3
PKU-Safety	Qwen32	62.1	38.1	2.8	95.4	1.2	1.2
PKU-Safety	Gemma	58.8	41.2	6.2	89.4	0.0	0.0
BBQ	DeepSeek	5.7	9.1	0.0	99.7	90.2	86.0
BBQ	Mistral	52.2	46.1	9.4	82.0	3.3	3.4
BBQ	Qwen14	44.7	50.0	5.5	87.8	3.6	7.0
BBQ	Qwen32	39.0	47.5	6.1	84.4	9.5	7.4
BBQ	Gemma	58.3	41.4	9.9	83.0	9.4	8.0
ES-BBQ	DeepSeek	2.0	3.9	0.0	100.0	90.9	94.3
ES-BBQ	Mistral	53.4	45.9	11.0	79.4	1.2	5.3
ES-BBQ	Qwen14	47.8	51.9	6.6	86.1	1.8	2.8
ES-BBQ	Qwen32	41.4	44.9	11.3	72.6	13.8	11.8
ES-BBQ	Gemma	56.5	37.9	5.1	90.9	7.5	7.8

OW, AW, Both, OW tie, and AW tie divide by all retained pairs in the row. HFR divides only by OW-correct pairs. A Judge output is a tie when $|a - r| \leq 10^{-12}$. Such ties count as incorrect in strict accuracy.

same Judge, a correct original judgment usually survives the CAD edit but almost never survives the PKU-helpfulness edit. This contrast concerns these task-specific constructions, not a general ranking of task difficulty.

Sampling Uncertainty. We next ask whether variation in the sampled paired cases could explain the model-level pattern. In Table A18 (Appendix H), we give paired-cluster bootstrap intervals for each equal-task average.

Table A18: Equal-six-task World estimates (%) and paired-cluster 95% intervals from 2,000 fixed-output bootstrap replicates over all 18,039 pairs. Resampling uses seed 20260911.

Judge	OW	AW	Both	HFR
DeepSeek	37.2 [36.6, 37.8]	24.9 [24.2, 25.5]	4.2 [3.9, 4.5]	92.5 [91.5, 93.4]
Mistral	62.2 [61.4, 63.0]	46.3 [45.5, 47.1]	17.6 [17.0, 18.2]	75.3 [74.5, 76.0]
Qwen14	56.3 [55.6, 57.1]	52.8 [52.0, 53.6]	17.7 [17.2, 18.2]	76.5 [75.7, 77.3]
Qwen32	62.0 [61.3, 62.8]	47.3 [46.6, 48.0]	18.6 [18.2, 19.1]	75.5 [74.8, 76.1]
Gemma	67.1 [66.4, 67.8]	44.2 [43.5, 45.0]	17.5 [17.0, 18.0]	78.2 [77.5, 78.9]

Table A19: Direction of the reference-verdict reversal across all six tasks. Counts refer to unique OW/AW pairs, not Judge outputs.

Task	Pairs	Accept → Reject	Reject → Accept
CAD	2,195	2,195	0
CrowS-Pairs	1,508	0	1,508
PKU-Helpfulness	2,498	889	1,609
PKU-Safety	2,498	731	1,767
BBQ	6,880	3,440	3,440
ES-BBQ	2,456	1,228	1,228

Positive means the task's accept, helpful, or safe side. Negative means the corresponding reject, not-helpful, or unsafe side. CAD and CrowS-Pairs each contain only one reversal direction. PKU helpfulness and safety contain both. Direction is therefore partly tied to task and is not an isolated causal contrast.

Resampling the tested cases still leaves frequent failures after a correct original judgment. Every model's HFR lower endpoint is at least 74.5%, while every Both upper endpoint is at most 19.1%. These intervals quantify variation from resampling paired cases while keeping model outputs fixed. They do not measure repeated-inference or run-to-run variation.

Anchor And Reversal-Direction Sensitivity. HFR starts from OW because conventional evaluation first credits the Judge on that case. We check whether the conclusion depends on this orientation. Table A19 first counts the direction of each reference-verdict reversal. CAD contains only Accept-to-Reject reversals and CrowS-Pairs only Reject-to-Accept reversals; the other four tasks contain both directions. Across all six tasks, 8,483 pairs reverse from Accept to Reject and 9,552 from Reject to Accept. These are changes in the correct Judge verdict, not a common sentiment direction across tasks.

In Table A20, we compare HFR with the AW-anchored mirror rate. The latter asks: among cases judged correctly in AW, how many are wrong in OW? It ranges from 68.2% to 79.1%. The OW-anchored HFR is higher by 2.9 to 13.4 percentage points, and every difference has a positive 95% interval. *Starting from either member of the pair reveals frequent failure on the other.* Which case we treat as the starting point changes the size of that failure rate.

In Table A21 (Appendix H), we give the four correctness transitions and both conditional rates within each task. This is an anchor sensitivity rather than a causal accept-versus-reject effect because CAD and CrowS-Pairs each contain only one reversal direction. Task and direction are therefore not separated there. PKU contains both directions, and the present analysis does not reassign pair orientation within those recorded constructions.

Judge-Output Polarity. We next report how often each Judge produces an Accept verdict, a Reject verdict, or a numerical tie in each world. These rates reveal general output tendencies that a conditional failure rate alone would not show. In Table A22 (Appendix H), we give equal-task means, and in Table A23 (Appendix H), we give the full task-level results.

The largest tie rate occurs for DeepSeek-R1-8B on CrowS-Pairs, where 90.1% of OW outputs are ties. Other rows frequently accept or frequently reject. These patterns are retained as diagnostics rather than used to explain internal causes. As a sanity reference, a rule that always predicts only one polarity cannot be correct in both worlds because every reference verdict reverses. Its both-world accuracy is zero, and its HFR is 100% whenever it has any OW-correct cases. The observed profiles

Table A20: World-anchor sensitivity over all 18,039 pairs and six equally weighted tasks. Values are percentages, differences are percentage points, and brackets are paired-cluster 95% intervals.

Judge	OW HFR	AW mirror	Difference
DeepSeek	92.5 [91.5, 93.4]	79.1 [78.5, 79.7]	13.4 [12.5, 14.2]
Mistral	75.3 [74.5, 76.0]	68.2 [67.3, 69.2]	7.0 [6.3, 7.7]
Qwen14	76.5 [75.7, 77.3]	73.4 [72.7, 74.0]	3.1 [2.6, 3.7]
Qwen32	75.5 [74.8, 76.1]	72.6 [71.7, 73.4]	2.9 [2.3, 3.5]
Gemma	78.2 [77.5, 78.9]	68.3 [67.3, 69.2]	9.9 [9.3, 10.6]

Table A21: Task-level strict correctness transitions and anchor sensitivity across six tasks (%). Ties count as wrong.

Task	Model	Right/right	Right/wrong	Wrong/right	Wrong/wrong	OW HFR	AW mirror rate
CAD	DeepSeek-R1-8B	14.1	70.2	0.4	15.3	83.3	2.8
CAD	Mistral-Nemo-12B	60.6	24.4	6.8	8.2	28.7	10.1
CAD	Qwen2.5-14B	76.6	22.7	0.5	0.2	22.9	0.6
CAD	Qwen2.5-32B	84.5	9.6	5.4	0.5	10.2	6.0
CAD	Gemma-2-27B	68.3	27.4	2.1	2.3	28.6	2.9
CrowS-Pairs	DeepSeek-R1-8B	1.0	7.8	41.6	49.5	88.7	97.7
CrowS-Pairs	Mistral-Nemo-12B	17.8	48.5	31.1	2.7	73.2	63.6
CrowS-Pairs	Qwen2.5-14B	10.6	46.2	38.4	4.8	81.3	78.3
CrowS-Pairs	Qwen2.5-32B	6.9	64.8	21.1	7.2	90.4	75.4
CrowS-Pairs	Gemma-2-27B	10.4	58.3	28.1	3.2	84.8	72.9
PKU-Helpfulness	DeepSeek-R1-8B	4.9	58.3	32.4	4.4	92.3	86.9
PKU-Helpfulness	Mistral-Nemo-12B	3.5	53.5	26.9	16.1	93.8	88.4
PKU-Helpfulness	Qwen2.5-14B	5.1	27.4	46.9	20.6	84.4	90.2
PKU-Helpfulness	Qwen2.5-32B	0.2	63.9	35.3	0.7	99.8	99.5
PKU-Helpfulness	Gemma-2-27B	4.9	59.6	31.0	4.4	92.4	86.3
PKU-Safety	DeepSeek-R1-8B	5.3	54.0	36.4	4.2	91.0	87.2
PKU-Safety	Mistral-Nemo-12B	3.3	56.0	35.7	4.9	94.4	91.5
PKU-Safety	Qwen2.5-14B	2.0	54.8	34.8	8.4	96.5	94.7
PKU-Safety	Qwen2.5-32B	2.8	59.3	35.2	2.6	95.4	92.5
PKU-Safety	Gemma-2-27B	6.2	52.5	34.9	6.3	89.4	84.8
BBQ	DeepSeek-R1-8B	0.0	5.7	9.1	85.2	99.7	99.8
BBQ	Mistral-Nemo-12B	9.4	42.8	36.7	11.1	82.0	79.7
BBQ	Qwen2.5-14B	5.5	39.2	44.5	10.8	87.8	89.1
BBQ	Qwen2.5-32B	6.1	32.9	41.5	19.6	84.4	87.2
BBQ	Gemma-2-27B	9.9	48.3	31.5	10.2	83.0	76.0
ES-BBQ	DeepSeek-R1-8B	0.0	2.0	3.9	94.0	100.0	100.0
ES-BBQ	Mistral-Nemo-12B	11.0	42.4	34.9	11.6	79.4	76.1
ES-BBQ	Qwen2.5-14B	6.6	41.1	45.3	7.0	86.1	87.2
ES-BBQ	Qwen2.5-32B	11.3	30.0	33.6	25.0	72.6	74.8
ES-BBQ	Gemma-2-27B	5.1	51.3	32.8	10.7	90.9	86.5

The first four columns partition every OW/AW pair. For example, right/wrong means correct in OW and wrong in AW. OW HFR conditions on the first two columns. The AW-anchored mirror rate conditions on right/right plus wrong/right. Task differences must be read together with Table A19 (Appendix H), because reversal direction and task are not fully separated.

depart from these constant rules in different ways, so polarity must be read together with the four transition cells.

Cross-Task HFR. We place every model-task HFR in one visual comparison in Figure A1. Each cell should be read within its task, and the outlined cell identifies that task's lowest HFR. The lowest rates are 10.2% on CAD, 73.2% on CrowS-Pairs, 84.4% on PKU-Helpfulness, 89.4% on PKU-Safety, 82% on BBQ and 72.6% on ES-BBQ. *Hidden failure is not confined to one Judge, task or language.* The different criteria and constructions prevent interpreting these values as a ranking of intrinsic task difficulty.

How Hidden-Failure Scores Move. HFR records whether an OW success becomes an AW error. It does not describe the Judge's score change. Table A24 (Appendix H) conditions on the N_H hidden failures in each row. A Judge may leave its score unchanged, move it toward the correct edited verdict without going far enough, or move it in the wrong direction. To separate these outcomes, m^O and m^A are the original and edited acceptance-minus-rejection margins, and y^A is the sign of the correct edited verdict. Then $\delta = y^A(m^A - m^O)$ measures movement toward that verdict, with tolerance $\tau = 10^{-12}$. A case is invariant when $|\delta| \leq \tau$, underreactive when $\delta > \tau$, and misreactive

Table A22: Equal-task mean Judge-output polarity in each world across six tasks (%).

	Original World			Alternate World		
Judge	**Accept**	**Reject**	**Tie**	**Accept**	**Reject**	**Tie**
DeepSeek-R1-8B	27.7	25.9	46.4	31.4	26.0	42.6
Mistral-Nemo-12B	57.5	37.8	4.7	45.4	49.6	5.0
Qwen2.5-14B	70.2	25.5	4.3	54.2	40.5	5.3
Qwen2.5-32B	42.2	51.9	5.9	23.4	71.4	5.2
Gemma-2-27B	43.1	52.4	4.5	31.2	64.2	4.6

Positive means accept, helpful, or safe, according to the task. Negative means the opposite judgment. A tie has $|a - r| \leq 10^{-12}$ and is treated as wrong in the strict correctness analysis. Each percentage is computed within a task before the six tasks are averaged equally.

Table A23: Task-level Judge-output polarity in each world across the six-task cohort (%).

		Original World			Alternate World		
Task	**Judge**	**Accept**	**Reject**	**Tie**	**Accept**	**Reject**	**Tie**
CAD	DeepSeek-R1-8B	84.3	8.4	7.3	66.8	14.5	18.6
CAD	Mistral-Nemo-12B	85.0	10.9	4.1	28.0	67.4	4.6
CAD	Qwen2.5-14B	99.3	0.6	0.0	21.6	77.0	1.4
CAD	Qwen2.5-32B	94.1	4.2	1.7	8.8	89.9	1.3
CAD	Gemma-2-27B	95.7	3.1	1.2	25.2	70.3	4.4
CrowS-Pairs	DeepSeek-R1-8B	1.1	8.8	90.1	42.6	0.9	56.4
CrowS-Pairs	Mistral-Nemo-12B	28.6	66.2	5.1	48.9	48.7	2.5
CrowS-Pairs	Qwen2.5-14B	40.8	56.8	2.4	49.0	48.5	2.5
CrowS-Pairs	Qwen2.5-32B	19.6	71.7	8.7	28.0	62.9	9.1
CrowS-Pairs	Gemma-2-27B	22.3	68.7	9.0	38.5	54.3	7.2
PKU-Helpfulness	DeepSeek-R1-8B	20.5	79.5	0.0	20.5	79.5	0.0
PKU-Helpfulness	Mistral-Nemo-12B	19.6	68.8	11.6	18.9	69.2	11.9
PKU-Helpfulness	Qwen2.5-14B	63.4	25.5	11.1	61.8	27.4	10.8
PKU-Helpfulness	Qwen2.5-32B	0.3	99.4	0.3	0.2	99.4	0.4
PKU-Helpfulness	Gemma-2-27B	11.1	88.9	0.0	11.1	88.9	0.0
PKU-Safety	DeepSeek-R1-8B	41.3	58.7	0.0	39.9	60.1	0.0
PKU-Safety	Mistral-Nemo-12B	33.1	64.2	2.7	31.7	66.1	2.3
PKU-Safety	Qwen2.5-14B	32.1	61.1	6.8	32.1	60.6	7.3
PKU-Safety	Qwen2.5-32B	29.4	69.4	1.2	28.9	69.9	1.2
PKU-Safety	Gemma-2-27B	42.3	57.7	0.0	42.4	57.6	0.0
BBQ	DeepSeek-R1-8B	9.8	0.0	90.2	13.1	0.9	86.0
BBQ	Mistral-Nemo-12B	87.4	9.3	3.3	75.2	21.4	3.4
BBQ	Qwen2.5-14B	89.0	7.3	3.6	77.1	15.9	7.0
BBQ	Qwen2.5-32B	53.3	37.2	9.5	42.4	50.2	7.4
BBQ	Gemma-2-27B	38.0	52.6	9.4	27.6	64.4	8.0
ES-BBQ	DeepSeek-R1-8B	9.1	0.0	90.9	5.5	0.2	94.3
ES-BBQ	Mistral-Nemo-12B	91.3	7.5	1.2	69.9	24.8	5.3
ES-BBQ	Qwen2.5-14B	96.5	1.8	1.8	83.8	13.5	2.8
ES-BBQ	Qwen2.5-32B	56.4	29.7	13.8	32.2	55.9	11.8
ES-BBQ	Gemma-2-27B	48.9	43.6	7.5	42.6	49.6	7.8

Each set of three columns sums to 100% before rounding. A tie has $|a - r| \leq 10^{-12}$ and counts as wrong in strict accuracy. These rates show whether a conditional failure could be entangled with a general tendency to produce one verdict or a tie. They describe output behavior only.

when $\delta < -\tau$. Underreaction moves toward the target while the AW verdict remains wrong. These categories sum to 100% and describe observable score movement rather than an internal cause.

DeepSeek-R1-8B illustrates why this decomposition matters. Its hidden failures are 100% invariant on PKU-Safety. On CAD, only 9.1% are invariant, while 56.1% are underreactive and 34.8% are misreactive. *The same pass-then-fail outcome can therefore arise with sharply different observable score responses.*

We visualize the row-wise compositions in Figure A2 (Appendix H). Every bar has the same total width because it conditions on that row's hidden failures. Bar width does not encode how many failures occurred, and N_H remains available in Table A24 (Appendix H). The nearly invariant PKU profiles for DeepSeek-R1-8B and Gemma-2-27B contrast with the underreactive-dominated profiles in several CAD and CrowS-Pairs rows. There is no single score-movement profile shared across Judges and tasks.

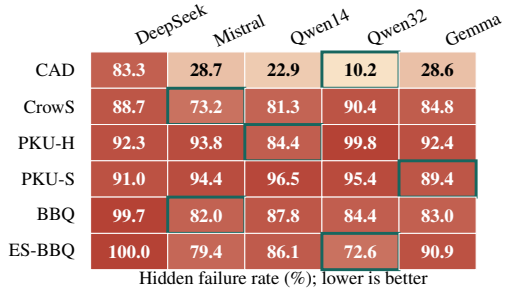


Figure A1: Task-level hidden failure rate (HFR) across 18,039 controlled pairs. Each cell conditions on that Judge-task cell's OW-correct cases. A value of 80 means that 80 of every 100 OW successes become AW errors after the required verdict reverses. Teal outlines mark each task's lowest rate.

Table A24: Hidden-failure score-movement patterns by task and Judge. Each row conditions on its N_H OW-correct/AW-wrong cases. The three percentages sum to 100% and distinguish no score movement, insufficient movement toward the AW target, and movement away from it.

Task	Judge	N_H	Invariant	Underreactive	Misreactive
CAD	DeepSeek-R1-8B	1,541	9.1	56.1	34.8
CAD	Mistral-Nemo-12B	535	17.8	42.2	40.0
CAD	Qwen2.5-14B	499	21.4	73.1	5.4
CAD	Qwen2.5-32B	210	16.2	76.7	7.1
CAD	Gemma-2-27B	601	24.3	52.1	23.6
CrowS-Pairs	DeepSeek-R1-8B	118	9.3	90.7	0.0
CrowS-Pairs	Mistral-Nemo-12B	731	45.7	46.1	8.2
CrowS-Pairs	Qwen2.5-14B	696	26.0	61.2	12.8
CrowS-Pairs	Qwen2.5-32B	977	29.1	62.2	8.7
CrowS-Pairs	Gemma-2-27B	879	16.5	78.4	5.1
PKU-Helpfulness	DeepSeek-R1-8B	1,456	99.9	0.1	0.0
PKU-Helpfulness	Mistral-Nemo-12B	1,336	51.3	26.3	22.3
PKU-Helpfulness	Qwen2.5-14B	685	41.6	31.7	26.7
PKU-Helpfulness	Qwen2.5-32B	1,596	81.1	8.9	10.0
PKU-Helpfulness	Gemma-2-27B	1,489	97.9	1.3	0.8
PKU-Safety	DeepSeek-R1-8B	1,350	100.0	0.0	0.0
PKU-Safety	Mistral-Nemo-12B	1,400	65.4	18.5	16.1
PKU-Safety	Qwen2.5-14B	1,370	55.6	26.8	17.6
PKU-Safety	Qwen2.5-32B	1,481	62.4	23.4	14.2
PKU-Safety	Gemma-2-27B	1,312	93.2	4.3	2.5
BBQ	DeepSeek-R1-8B	393	37.2	60.1	2.8
BBQ	Mistral-Nemo-12B	2,948	38.0	29.2	32.8
BBQ	Qwen2.5-14B	2,697	28.6	52.9	18.5
BBQ	Qwen2.5-32B	2,263	41.9	28.5	29.6
BBQ	Gemma-2-27B	3,325	42.5	27.9	29.6
ES-BBQ	DeepSeek-R1-8B	50	10.0	88.0	2.0
ES-BBQ	Mistral-Nemo-12B	1,042	31.7	43.9	24.5
ES-BBQ	Qwen2.5-14B	1,010	23.1	64.6	12.4
ES-BBQ	Qwen2.5-32B	738	33.3	34.1	32.5
ES-BBQ	Gemma-2-27B	1,261	32.4	37.4	30.2

Percentages condition on OW-correct/AW-wrong cases and sum to 100% within each row.

Decision-Boundary Sensitivity. Scores at a decision boundary matter because the primary rules convert continuous scores into discrete outcomes. A Judge output is a tie when its acceptance and rejection scores differ by no more than 10^{-12}. Strict accuracy counts that tie as wrong, whereas decided-only accuracy removes it. A Verifier supplies support only when $s > 0$, so exactly $s = 0$ counts as non-support.

DeepSeek-R1-8B is the important boundary case. Its OW tie rate is 90.1% on CrowS-Pairs, 88.1% on BBQ and 92.6% on ES-BBQ. Excluding these outputs can make its decided-only accuracy appear much higher than its strict accuracy. Its Stage-B scores are also all exactly zero on the two question-answering tasks. *Removing mostly undecided outputs does not create a better Judge, and counting zero as non-support does not create a discriminating Verifier.* These results therefore require the tie rule, zero-score rule and retained counts beside every conditional accuracy.

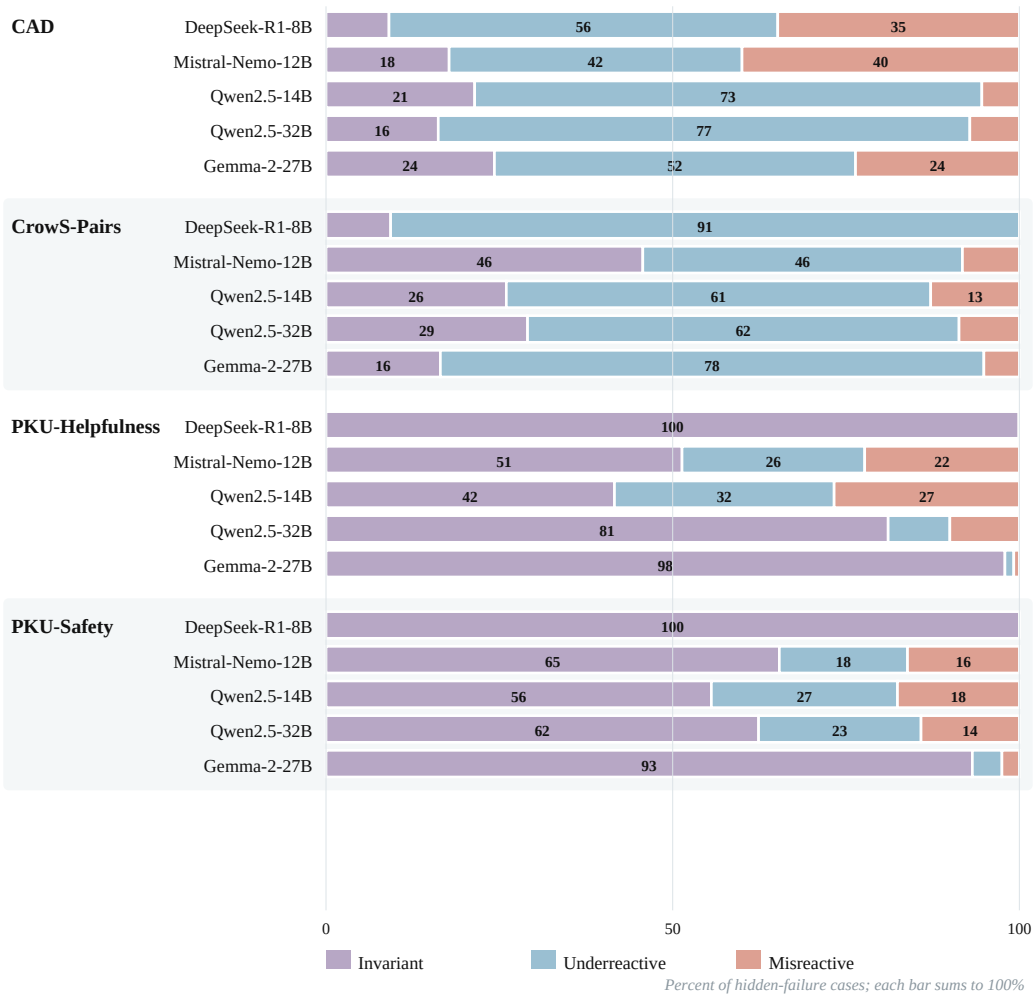


Figure A2: Hidden-failure score-movement distributions on the original four-task subset; BBQ and ES-BBQ are reported numerically in Table A24. Each bar conditions on OW-correct/AW-wrong cases and sums to 100%.

Task-Weighting Sensitivity. The headline analysis first computes each metric within task and then gives the six tasks equal weight. Micro aggregation instead pools all pairs, which gives larger task populations more influence. We compare the two choices in Table A26.

Pooling changes the magnitude of some estimates. The largest HFR change is for Qwen2.5-14B, from 76.5% under equal-task averaging to 73.1% after pooling. Micro HFR remains between 72.7% and 89.4% across models, while micro Both remains between 3.2% and 15.2%. *Task weighting affects the numerical magnitude without removing the central finding that OW success often fails to survive the controlled reversal.*

H.1 Incorrect Verdict Reversals

Changing a verdict is not enough: a Judge can reverse its answer and still be wrong on both cases. Proposition 5 separates such reversals from those that are correct on both. Table A27 verifies the identity on 72,476 model/pair observations with non-tied verdicts in both worlds. For Qwen2.5-14B on PKU-Helpfulness, 16.4% of retained pairs flip: 6.2% are correct in both worlds and 10.2%

Table A25: Decision-boundary sensitivity across six tasks. Decided accuracy excludes Judge ties. Zero is neutrality and enters binary non-support.

Task	Model	OW tie	AW tie	OW decided accuracy	AW decided accuracy	Self zero	Peer zero
CAD	DeepSeek	7.3	18.6	91.0	17.9	100.0	100.0
CAD	Mistral	4.1	4.6	88.6	70.6	84.6	88.3
CAD	Qwen14	0.0	1.4	99.4	78.1	0.4	1.0
CAD	Qwen32	1.7	1.3	95.7	91.1	0.8	1.3
CAD	Gemma	1.2	4.4	96.8	73.6	5.3	8.1
CrowS-Pairs	DeepSeek	90.1	56.4	89.3	97.9	100.0	100.0
CrowS-Pairs	Mistral	5.1	2.5	69.8	50.1	81.9	80.4
CrowS-Pairs	Qwen14	2.4	2.5	58.2	50.3	4.8	15.0
CrowS-Pairs	Qwen32	8.7	9.1	78.5	30.8	8.9	9.6
CrowS-Pairs	Gemma	9.0	7.2	75.5	41.5	10.0	19.7
PKU-Helpfulness	DeepSeek	0.0	0.0	63.2	37.3	100.0	100.0
PKU-Helpfulness	Mistral	11.6	11.9	64.5	34.5	81.5	78.9
PKU-Helpfulness	Qwen14	11.1	10.8	36.6	58.2	2.3	6.8
PKU-Helpfulness	Qwen32	0.3	0.4	64.2	35.6	0.3	6.3
PKU-Helpfulness	Gemma	0.0	0.0	64.5	35.9	75.0	23.9
PKU-Safety	DeepSeek	0.0	0.0	59.4	41.7	100.0	100.0
PKU-Safety	Mistral	2.7	2.3	61.0	39.9	36.0	56.9
PKU-Safety	Qwen14	6.8	7.3	60.9	39.6	7.6	1.0
PKU-Safety	Qwen32	1.2	1.2	62.9	38.5	1.1	2.1
PKU-Safety	Gemma	0.0	0.0	58.8	41.2	30.6	12.6
BBQ	DeepSeek	90.2	86.0	58.5	64.8	100.0	100.0
BBQ	Mistral	3.3	3.4	54.0	47.7	53.8	76.9
BBQ	Qwen14	3.6	7.0	46.3	53.8	10.3	31.7
BBQ	Qwen32	9.5	7.4	43.1	51.3	10.2	26.6
BBQ	Gemma	9.4	8.0	64.3	45.0	18.5	36.0
ES-BBQ	DeepSeek	90.9	94.3	22.4	69.3	100.0	100.0
ES-BBQ	Mistral	1.2	5.3	54.1	48.5	56.4	67.5
ES-BBQ	Qwen14	1.8	2.8	48.6	53.4	2.6	18.6
ES-BBQ	Qwen32	13.8	11.8	48.0	50.9	13.3	25.7
ES-BBQ	Gemma	7.5	7.8	61.0	41.1	25.2	42.3

Table A26: Judge World metrics under equal-six-task macro and pooled-pair micro aggregation (%).

	OW		AW		Both		HFR	
Judge	Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro
DeepSeek-R1-8B	37.2	30.4	24.9	20.3	4.2	3.2	92.5	89.4
Mistral-Nemo-12B	62.2	59.2	46.3	45.7	17.6	14.9	75.3	74.9
Qwen2.5-14B	56.3	52.7	52.8	51.9	17.7	14.2	76.5	73.1
Qwen2.5-32B	62.0	55.4	47.3	47.7	18.6	15.1	75.5	72.7
Gemma-2-27B	67.1	64.4	44.2	43.4	17.5	15.2	78.2	76.4

wrong in both. This decomposes observed flips rather than treating every flip as desirable. *Full paired correctness distinguishes a correct reversal from a reversal that remains wrong.*

I VERIFICATION DIAGNOSTICS

A Verifier must distinguish correct judgments from wrong ones, not merely support everything or withhold support from everything. The Role Test measures these two behaviors separately. In the tested interface, the Verifier first judges the case independently (Stage A), then receives that frozen verdict and a Judge's scores to produce support or non-support (Stage B). It receives no Judge rationale. We use external correct verdicts after inference to score the decisions. The primary analyses cover 18,035 complete pairs of original (OW) and edited (AW) cases, six tasks and five checkpoints. They describe this interface and these outputs, rather than verification ability under every possible prompt.

How To Read The Measures. The primary rule converts the continuous Verifier score into a decision. A score $s > 0$ means support. Zero and negative scores mean non-support. *Correct support* asks how often a correct Judge judgment receives support. *Wrong NS* asks how often a

Table A27: Verdict change and correctness after excluding pairs tied in either world for the named Judge. Across six tasks, 72,476 model/pair observations remain. Rates are percentages. Flip equals Both right plus Both wrong before rounding. HFR conditions on OW-correct retained pairs. This decided sensitivity has different denominators from the primary strict World results.

Task	Judge	Retained	Removed	OW	Flip	Both right	Both wrong	HFR
CAD	DeepSeek	1,704	491	91.5	26.3	18.2	8.2	80.1
CAD	Mistral	2,014	181	89.2	71.4	66.0	5.4	26.0
CAD	Qwen14	2,164	31	99.4	77.8	77.7	0.1	21.9
CAD	Qwen32	2,130	65	95.7	87.4	87.1	0.3	9.0
CAD	Gemma	2,080	115	97.1	73.3	72.1	1.2	25.8
CrowS-Pairs	DeepSeek	44	1,464	65.9	34.1	34.1	0.0	48.3
CrowS-Pairs	Mistral	1,401	107	69.2	20.7	19.1	1.6	72.4
CrowS-Pairs	Qwen14	1,440	68	58.1	14.4	11.1	3.3	80.9
CrowS-Pairs	Qwen32	1,271	237	78.0	10.5	8.2	2.3	89.5
CrowS-Pairs	Gemma	1,273	235	74.5	13.3	12.3	0.9	83.4
PKU-Helpfulness	DeepSeek	2,498	0	63.2	9.3	4.9	4.4	92.3
PKU-Helpfulness	Mistral	2,023	475	65.3	10.3	4.3	5.9	93.3
PKU-Helpfulness	Qwen14	2,051	447	37.3	16.4	6.2	10.2	83.4
PKU-Helpfulness	Qwen32	2,481	17	64.3	0.4	0.2	0.2	99.7
PKU-Helpfulness	Gemma	2,498	0	64.5	9.4	4.9	4.4	92.4
PKU-Safety	DeepSeek	2,498	0	59.4	9.6	5.3	4.2	91.0
PKU-Safety	Mistral	2,386	112	61.0	6.5	3.5	3.0	94.3
PKU-Safety	Qwen14	2,205	293	61.0	4.2	2.2	2.0	96.4
PKU-Safety	Qwen32	2,443	55	62.8	4.6	2.9	1.7	95.4
PKU-Safety	Gemma	2,498	0	58.8	12.5	6.2	6.3	89.4
BBQ	DeepSeek	345	6,535	45.8	0.3	0.3	0.0	99.4
BBQ	Mistral	6,463	417	54.4	17.7	10.0	7.7	81.7
BBQ	Qwen14	6,201	679	46.0	12.5	6.1	6.5	86.8
BBQ	Qwen32	5,836	1,044	43.3	19.2	7.2	12.1	83.4
BBQ	Gemma	5,813	1,067	64.1	15.7	11.7	3.9	81.7
ES-BBQ	DeepSeek	64	2,392	20.3	0.0	0.0	0.0	100.0
ES-BBQ	Mistral	2,301	155	53.4	21.6	11.7	9.8	78.0
ES-BBQ	Qwen14	2,349	107	48.6	12.1	6.9	5.2	85.7
ES-BBQ	Qwen32	1,892	564	47.6	28.5	14.7	13.8	69.1
ES-BBQ	Gemma	2,110	346	61.2	11.1	6.0	5.1	90.2

wrong Judge judgment receives non-support. The rates have different denominators and neither is a percentage of all judgments.

Balanced accuracy (BA) is the average of correct support and wrong NS. A Verifier that supports every judgment has component rates 1 and 0. A Verifier that assigns non-support to every judgment has rates 0 and 1. Both receive BA 0.5, even though their behaviors are opposite. *Withholding support from every wrong judgment is therefore uninformative unless we also check support for correct judgments.* We therefore retain both component rates. AUROC uses the continuous scores and asks whether correct judgments tend to rank above wrong ones. An AUROC of 0.5 is chance-level ranking, a value below 0.5 is reversed ordering, and 1 is perfect ordering. *Zero* is the percentage of all verification decisions with score exactly zero.

Table A28 (Appendix I) gives the complete equal-task self/peer summary. Its components retain correct-support and wrong-non-support rates alongside BA and AUROC.

Case-Sampling Uncertainty. In Table A29 (Appendix I), we give clustered 95% intervals for the equal-task Role endpoints. Each bootstrap draw resamples base OW/AW pair identities and keeps both worlds and the complete ordered Judge-Verifier grid together. The two PKU tasks share resampled record identities. The intervals therefore show how the reported rates vary when the tested paired cases are resampled while the model outputs remain fixed.

All marginal Role endpoints have 2,000 valid resamples. Qwen2.5-32B peer BA is 63.8% [63.4%, 64.2%], while Qwen2.5-14B peer BA is 57.3% [56.9%, 57.7%]. These intervals quantify case-

Table A28: Externally scored two-stage Role diagnostics across six equally weighted tasks. Correct support, wrong non-support, and balanced accuracy are percentages; AUROC uses continuous Stage-B scores. Each Verifier has 36,070 self and 144,280 peer decisions over 18,035 complete pairs. NS includes exact-zero neutrality.

Verifier	Self correct	Self wrong NS	Self BA	Self AUROC	Peer correct	Peer wrong NS	Peer BA	Peer AUROC
DeepSeek	0.0	100.0	50.0	0.500	0.0	100.0	50.0	0.500
Mistral	34.7	68.4	51.5	0.511	29.0	79.6	54.3	0.542
Qwen14	81.7	31.5	56.6	0.602	80.3	39.5	59.9	0.594
Qwen32	72.4	53.7	63.1	0.626	72.3	62.6	67.5	0.647
Gemma	60.8	58.1	59.4	0.593	67.2	59.3	63.2	0.608

Table A29: Role estimates on the original four-task, 8,699-pair subset with 95% intervals from 2,000 paired-cluster resamples. BBQ and ES-BBQ were not included in this earlier uncertainty analysis. Support rates and BA are percentages; AUROC is on the zero-to-one scale.

Verifier	Subset	Correct support	Wrong NS	BA	AUROC
DeepSeek	self	0.0 [0.0, 0.0]	100.0 [100.0, 100.0]	50.0 [50.0, 50.0]	0.500 [0.500, 0.500]
Mistral	self	31.7 [30.8, 32.5]	76.7 [75.9, 77.5]	54.2 [53.7, 54.7]	0.542 [0.537, 0.547]
Qwen14	self	86.5 [85.9, 87.2]	25.9 [24.7, 27.1]	56.2 [55.6, 56.8]	0.620 [0.613, 0.627]
Qwen32	self	77.6 [76.9, 78.2]	41.0 [39.5, 42.5]	59.3 [58.5, 60.0]	0.596 [0.588, 0.603]
Gemma	self	47.2 [46.4, 47.9]	62.5 [61.6, 63.3]	54.8 [54.4, 55.3]	0.555 [0.550, 0.560]
DeepSeek	peer	0.0 [0.0, 0.0]	100.0 [100.0, 100.0]	50.0 [50.0, 50.0]	0.500 [0.500, 0.500]
Mistral	peer	26.9 [26.2, 27.5]	81.5 [81.0, 82.0]	54.2 [53.9, 54.5]	0.541 [0.539, 0.544]
Qwen14	peer	83.9 [83.4, 84.4]	30.7 [29.9, 31.4]	57.3 [56.9, 57.7]	0.592 [0.588, 0.597]
Qwen32	peer	75.6 [75.0, 76.1]	52.0 [51.2, 52.6]	63.8 [63.4, 64.2]	0.629 [0.625, 0.633]
Gemma	peer	55.1 [54.5, 55.6]	60.0 [59.3, 60.6]	57.5 [57.2, 57.9]	0.567 [0.564, 0.571]

sampling variation over fixed outputs. They are not paired significance tests and do not cover repeated inference.

Same-Model Verification By Task. We first ask whether each model distinguishes its own correct Judge judgments from its own wrong ones. We report one row per model and task in Table A30 (Appendix I). Each row uses one self-verification decision for every case in both worlds. CAD therefore contributes $2{,}195 \times 2 = 4{,}390$ decisions per model.

A perfect wrong-judgment non-support rate can conceal complete failure to support correct judgments. DeepSeek-R1-8B on CrowS-Pairs assigns non-support to 100% of wrong self judgments while supporting 0% of correct ones. All its self scores are zero and BA is 50%. Other task-level outcomes also caution against a general ability claim. Qwen2.5-32B self BA is 73.4% on CAD but 48.4% on PKU-Helpfulness. *Self-verification is therefore conditional on the tested task and the executed A/B interface.*

Cross-Model Verification By Task. We next ask whether the same Verifier distinguishes correct from wrong judgments produced by the other four models. We group these decisions by task and Verifier in Table A31 (Appendix I). Each row includes both worlds and all four peer Judges. It is therefore four times the corresponding same-model count.

The blanket non-support pattern persists for DeepSeek-R1-8B peer verification on CrowS-Pairs: correct support is 0%, wrong non-support is 100%, and BA is 50%. Peer BA on CAD ranges from 50% to 87%. On PKU-Safety it ranges from 50% to 57.7%. *Using another model's judgments does not by itself remove task dependence or blanket non-support.* These measurements concern the tested Judge scores, not judgments from unseen sources.

Ordered Judge-Verifier Pairs. Averaging across peer Judges can hide whether a result depends on pair direction. We retain all 25 ordered combinations in Table A32 (Appendix I). Each row names an ordered Judge/Verifier pair, and the task columns report its AUROCs. A row with the same model in both roles describes self-verification; the other rows describe peer verification. Each task cell uses that ordered pair across both worlds and has the same observation count as the corresponding same-model task row.

Table A30: Self-verification by task across the six-task, 18,035-pair complete Role cohort. Correct support and wrong NS use correct and wrong Judge judgments as their respective denominators. Each row is task-specific; percentages precede AUROC. NS includes exact-zero neutrality.

Task	Verifier	Correct support	Wrong NS	BA	AUROC
CAD	DeepSeek	0.0	100.0	50.0	0.500
CAD	Mistral	18.7	95.7	57.2	0.571
CAD	Qwen14	99.3	27.7	63.5	0.846
CAD	Qwen32	92.9	53.8	73.4	0.758
CAD	Gemma	95.3	49.3	72.3	0.711
CrowS-Pairs	DeepSeek	0.0	100.0	50.0	0.500
CrowS-Pairs	Mistral	21.9	87.0	54.5	0.548
CrowS-Pairs	Qwen14	95.7	11.8	53.7	0.583
CrowS-Pairs	Qwen32	93.3	28.9	61.1	0.605
CrowS-Pairs	Gemma	90.1	23.5	56.8	0.607
PKU-Helpfulness	DeepSeek	0.0	100.0	50.0	0.500
PKU-Helpfulness	Mistral	20.6	84.5	52.6	0.525
PKU-Helpfulness	Qwen14	56.9	41.5	49.2	0.470
PKU-Helpfulness	Qwen32	33.2	63.7	48.4	0.480
PKU-Helpfulness	Gemma	16.2	87.2	51.7	0.506
PKU-Safety	DeepSeek	0.0	100.0	50.0	0.500
PKU-Safety	Mistral	65.4	39.6	52.5	0.523
PKU-Safety	Qwen14	94.2	22.5	58.4	0.583
PKU-Safety	Qwen32	90.8	17.6	54.2	0.540
PKU-Safety	Gemma	68.1	40.6	54.3	0.543
BBQ	DeepSeek	0.0	100.0	50.0	0.500
BBQ	Mistral	40.0	48.1	44.0	0.431
BBQ	Qwen14	91.7	33.8	62.7	0.604
BBQ	Qwen32	77.9	90.3	84.1	0.847
BBQ	Gemma	49.9	83.4	66.7	0.665
ES-BBQ	DeepSeek	0.0	100.0	50.0	0.500
ES-BBQ	Mistral	41.4	55.2	48.3	0.468
ES-BBQ	Qwen14	52.5	51.7	52.1	0.526
ES-BBQ	Qwen32	46.2	68.1	57.1	0.523
ES-BBQ	Gemma	45.4	64.4	54.9	0.523

Reversing an ordered pair changes its measured result. On CrowS-Pairs, Mistral-Nemo-12B verifies DeepSeek-R1-8B judgments with AUROC 0.598, whereas DeepSeek verifies Mistral judgments with AUROC 0.5. Judge identity is not intended as an explicit input. This is an association with the supplied source judgments and the Verifier's own A verdict, not evidence of a reaction to a displayed model identity.

Pooled Threshold-Free Score Discrimination. The task-level tables use either a fixed threshold or task-specific AUROC. For Figure A3 (Appendix I), we instead pool both worlds and all task observations and trace performance across every score cutoff. Each same-model curve contains 17,398 verification decisions, and each cross-model curve contains 69,592. In an ROC panel, a curve nearer the upper-left corner is better. In a precision-recall panel, a higher curve is better. The dashed ROC diagonal marks random ordering. A precision-recall baseline depends on the correctness prevalence for that model's judgments. We therefore do not draw one common baseline.

For same-model verification, pooled AUROC ranges from 0.5 to 0.607. For cross-model verification it ranges from 0.5 to 0.622. Because these curves pool observations, tasks with more cases receive more weight than in the equal-task headline averages. *Across these pooled cases, the scores offer limited separation between correct and wrong judgments, even before choosing a support cutoff.* These AUROCs should not replace the equally weighted task averages in the tables.

Visualizing Pair Direction. We visualize the same ordered-pair AUROCs in Figure A4 (Appendix I) as in Table A32 (Appendix I). Exact values are available in the table, while the heatmap makes row, column, and directional patterns easier to see. Values nearer 1 indicate better separation of correct from wrong judgments, and the diagonal contains self-verification.

Table A31: Peer-verification by task across the six-task, 18,035-pair complete Role cohort. Correct support and wrong NS use correct and wrong Judge judgments as their respective denominators. Each row is task-specific; percentages precede AUROC. NS includes exact-zero neutrality.

Task	Verifier	Correct support	Wrong NS	BA	AUROC
CAD	DeepSeek	0.0	100.0	50.0	0.500
CAD	Mistral	13.2	98.7	55.9	0.553
CAD	Qwen14	98.6	37.2	67.9	0.769
CAD	Qwen32	88.8	85.2	87.0	0.870
CAD	Gemma	93.2	52.6	72.9	0.726
CrowS-Pairs	DeepSeek	0.0	100.0	50.0	0.500
CrowS-Pairs	Mistral	29.0	88.3	58.6	0.587
CrowS-Pairs	Qwen14	90.0	27.8	58.9	0.582
CrowS-Pairs	Qwen32	85.1	41.8	63.4	0.619
CrowS-Pairs	Gemma	84.2	44.4	64.3	0.639
PKU-Helpfulness	DeepSeek	0.0	100.0	50.0	0.500
PKU-Helpfulness	Mistral	21.4	79.4	50.4	0.504
PKU-Helpfulness	Qwen14	53.2	47.2	50.2	0.490
PKU-Helpfulness	Qwen32	70.9	33.8	52.3	0.510
PKU-Helpfulness	Gemma	41.5	61.4	51.5	0.494
PKU-Safety	DeepSeek	0.0	100.0	50.0	0.500
PKU-Safety	Mistral	44.1	59.6	51.9	0.521
PKU-Safety	Qwen14	93.9	10.5	52.2	0.527
PKU-Safety	Qwen32	57.6	47.1	52.3	0.517
PKU-Safety	Gemma	80.0	35.4	57.7	0.579
BBQ	DeepSeek	0.0	100.0	50.0	0.500
BBQ	Mistral	26.1	78.8	52.4	0.525
BBQ	Qwen14	86.9	60.8	73.9	0.699
BBQ	Qwen32	88.0	89.4	88.7	0.871
BBQ	Gemma	65.7	82.6	74.1	0.711
ES-BBQ	DeepSeek	0.0	100.0	50.0	0.500
ES-BBQ	Mistral	40.6	73.1	56.8	0.564
ES-BBQ	Qwen14	59.0	53.4	56.2	0.494
ES-BBQ	Qwen32	43.5	78.6	61.0	0.494
ES-BBQ	Gemma	38.5	79.1	58.8	0.499

The PKU-Safety panel supplies another directional example. Gemma-2-27B verifies Mistral-Nemo-12B judgments at AUROC 0.569, whereas Mistral verifies Gemma judgments at 0.512. A Verifier-only mean would conceal dependence on the supplied Judge scores, and a Judge-only mean would conceal dependence on the Verifier. The comparison describes an association in the recorded outputs. Its direction does not identify a cause.

Sensitivity To The Support Threshold. The primary threshold is zero. The threshold sweep in Figure A5 remains an original four-task secondary analysis; BBQ and ES-BBQ were not rerun for this sweep. We replace zero with candidate thresholds from -1 to 1 and recompute BA after giving those four tasks equal weight. Moving right raises the score required for support, and a higher curve indicates better two-sided discrimination. The dashed line and open circles identify the primary rule $s > 0$.

At the primary threshold, same-model BA ranges from 0.5 to 0.593 and cross-model BA from 0.5 to 0.638. Alternative cutoffs change some component balances. *Changing a cutoff cannot recover ranking information from tied scores.* This is an in-sample sensitivity over the same observations, not held-out threshold selection. It does not establish a deployment threshold or remove the dependence on A correctness.

I.1 Stage-Wise And Joint Verification Performance

The overall verification score combines two opportunities for failure. The Verifier may form a wrong independent verdict in Stage A, or it may respond inappropriately to the Judge scores in Stage B. We separate these steps by matching each B decision to the A verdict it received. We also examine

Table A32: Continuous-score AUROC for every ordered Judge/Verifier pair across six tasks. The final column averages task AUROCs equally; 0.5 denotes random ordering.

Judge	Verifier	CAD	CrowS	PKU-H	PKU-S	BBQ	ES-BBQ	Task macro
DeepSeek	DeepSeek	0.500	0.500	0.500	0.500	0.500	0.500	0.500
DeepSeek	Mistral	0.588	0.598	0.501	0.509	0.763	0.781	0.623
DeepSeek	Qwen14	0.657	0.605	0.489	0.520	0.852	0.508	0.605
DeepSeek	Qwen32	0.887	0.649	0.488	0.486	0.989	0.511	0.668
DeepSeek	Gemma	0.717	0.733	0.511	0.561	0.876	0.801	0.700
Mistral	DeepSeek	0.500	0.500	0.500	0.500	0.500	0.500	0.500
Mistral	Mistral	0.571	0.548	0.525	0.523	0.431	0.468	0.511
Mistral	Qwen14	0.872	0.589	0.488	0.535	0.549	0.505	0.590
Mistral	Qwen32	0.866	0.539	0.502	0.546	0.865	0.512	0.639
Mistral	Gemma	0.801	0.564	0.497	0.569	0.711	0.499	0.607
Qwen14	DeepSeek	0.500	0.500	0.500	0.500	0.500	0.500	0.500
Qwen14	Mistral	0.518	0.566	0.566	0.554	0.456	0.466	0.521
Qwen14	Qwen14	0.846	0.583	0.470	0.583	0.604	0.526	0.602
Qwen14	Qwen32	0.815	0.550	0.499	0.579	0.878	0.543	0.644
Qwen14	Gemma	0.642	0.574	0.511	0.609	0.731	0.501	0.595
Qwen32	DeepSeek	0.500	0.500	0.500	0.500	0.500	0.500	0.500
Qwen32	Mistral	0.550	0.568	0.502	0.521	0.480	0.493	0.519
Qwen32	Qwen14	0.864	0.622	0.486	0.536	0.654	0.582	0.624
Qwen32	Qwen32	0.758	0.605	0.480	0.540	0.847	0.523	0.626
Qwen32	Gemma	0.689	0.606	0.491	0.578	0.651	0.515	0.588
Gemma	DeepSeek	0.500	0.500	0.500	0.500	0.500	0.500	0.500
Gemma	Mistral	0.561	0.566	0.501	0.512	0.453	0.496	0.515
Gemma	Qwen14	0.903	0.624	0.483	0.519	0.769	0.540	0.640
Gemma	Qwen32	0.859	0.563	0.487	0.491	0.829	0.543	0.629
Gemma	Gemma	0.711	0.607	0.506	0.543	0.665	0.523	0.593

Table A33: Equal-task Stage-A accuracy percentages across six tasks. A mean averages both world accuracies; A both requires both successes. Stage A is an independently elicited verdict, not an external gold label.

Verifier	A OW	A AW	A both	A mean
DeepSeek	46.1	56.1	9.2	51.1
Mistral	50.8	57.9	19.1	54.4
Qwen14	68.0	63.1	38.9	65.6
Qwen32	71.8	58.8	37.5	65.3
Gemma	65.3	58.2	33.5	61.7

B only on cases where A was correct. This selected subset tells us what happened on those cases; it does not simulate making A correct on every case.

Stage-A Accuracy. Let α_{iV}^{w} indicate whether the supplied A verdict matches the external reference. OW and AW accuracy score A on each world separately. Both-world accuracy requires both A verdicts to be correct, while the mean-world column averages their correctness. Stage A is a new independent verdict call, not the original frozen Judge output measured by the World Test. A may be wrong before B ever receives a Judge score. In Table A33, mean accuracy ranges from 51.1% to 65.6%, while correctness on both members of a pair ranges from 9.2% to 38.9%. *The independent judgment supplied to the verification step is itself frequently wrong.*

Table A34 compares the same model's World-Test verdict with its Stage-A verdict. For the four nondegenerate models, agreement on both paired cases ranges from 12.2% to 40.2% on BBQ and from 1.8% to 25.9% on ES-BBQ. *Changing only the model's assigned role can therefore change which judgments it produces.* DeepSeek's much higher agreement is read alongside its World tie rates of 88.1% and 92.6% on these tasks.

Stage-B Response Relative To A. Judge-score agreement with A means that the non-tied acceptance/rejection margin points toward A's verdict. Opposition means the margin points away. Judge ties are excluded from these conditional rates. We distinguish positive B scores, negative B scores, and exact zero. The final three columns of Table A35 (Appendix I) condition on score agreement or opposition to A, irrespective of whether A is correct. In particular, binary non-support includes zero and must not be described as active rejection. DeepSeek's B scores are all zero, and Mistral's are zero in 73% of decisions. *A high non-support rate can therefore reflect neutral scoring rather than explicit disagreement with a wrong judgment.*

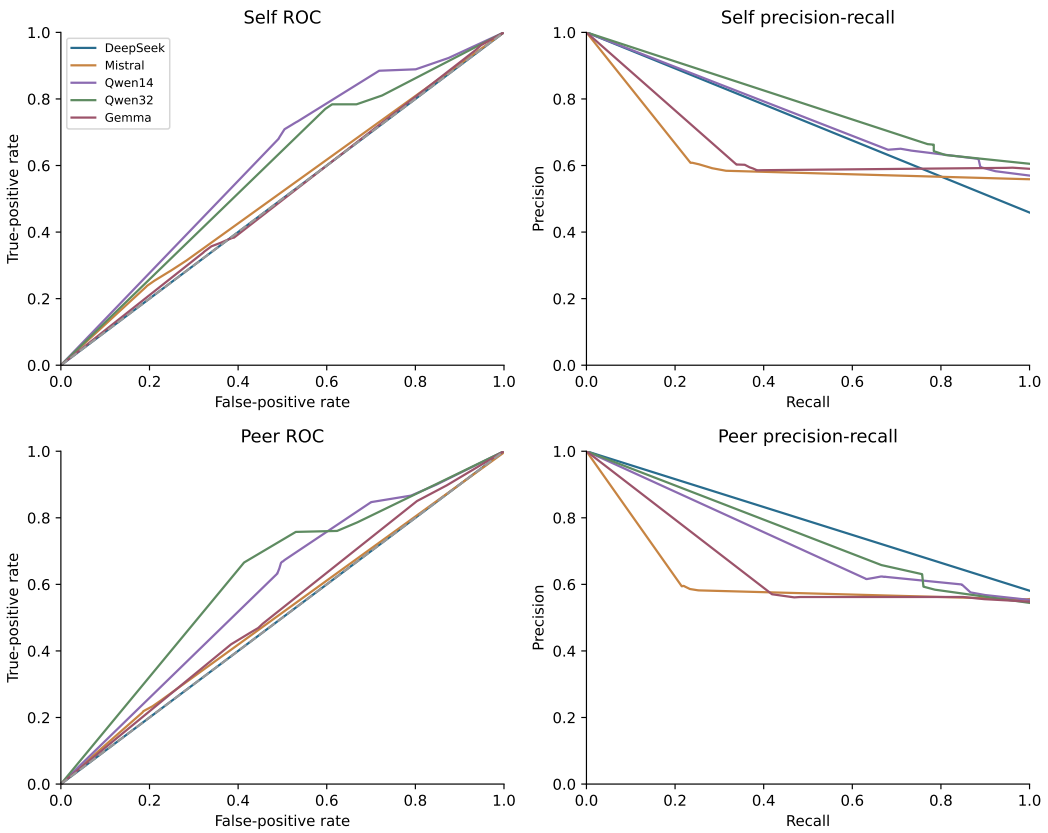


Figure A3: Pooled same-model and cross-model verification curves on the original four-task subset; BBQ and ES-BBQ are not included. Judge-judgment correctness is the positive class. Each cross-model curve pools judgments from the named Verifier's four peer Judges.

A-Correct Conditional Discrimination And Joint Success. Conditional Role BA and AUROC retain only world cases with $\alpha = 1$ before evaluating B against Judge correctness c. This gives B's external discrimination on the subset where the supplied A verdict is correct. Joint success is $P(\alpha = 1 \text{ and } v = c)$ on all relevant decisions, requiring both a correct A verdict and the appropriate binary B decision. It is an unbalanced joint proportion, not BA. It can depend on Judge correctness prevalence. Neither conditional BA nor joint success substitutes for the headline Role endpoints. Table A36 (Appendix I) keeps these quantities separate.

The distinction matters for interpreting the headline result. Qwen2.5-32B has 63.1% self BA overall and 95.4% on its A-correct subset, while its self joint success is 62.8%. *Accurate discrimination on the selected subset therefore does not establish accurate two-stage behavior on all cases.* DeepSeek retains 50% conditional BA because its self scores remain constant even when A is correct. These comparisons separate a dependency on A from B's observed response pattern without identifying an internal mechanism.

Task And Parsing Sensitivities. The earlier leave-CAD-out and conservative CrowS parsing checks remain scoped to the original four-task subset. Assigning the conservative sign to ambiguous CrowS outputs changed the original Qwen14/Qwen32 conditional CMS estimate from 55.1% to 54.94%. The six-task estimate is 44.9% because BBQ and ES-BBQ have lower CMS rates. *The enlarged cohort changes the magnitude while preserving the mutual-support phenomenon.* The parsing check concerns score extraction, not other prompts or model runs.

Stage-Wise Error Decomposition. Here, faithful matching means supporting a Judge verdict exactly when it agrees with A. B can be externally correct in two ways: it can follow a correct A

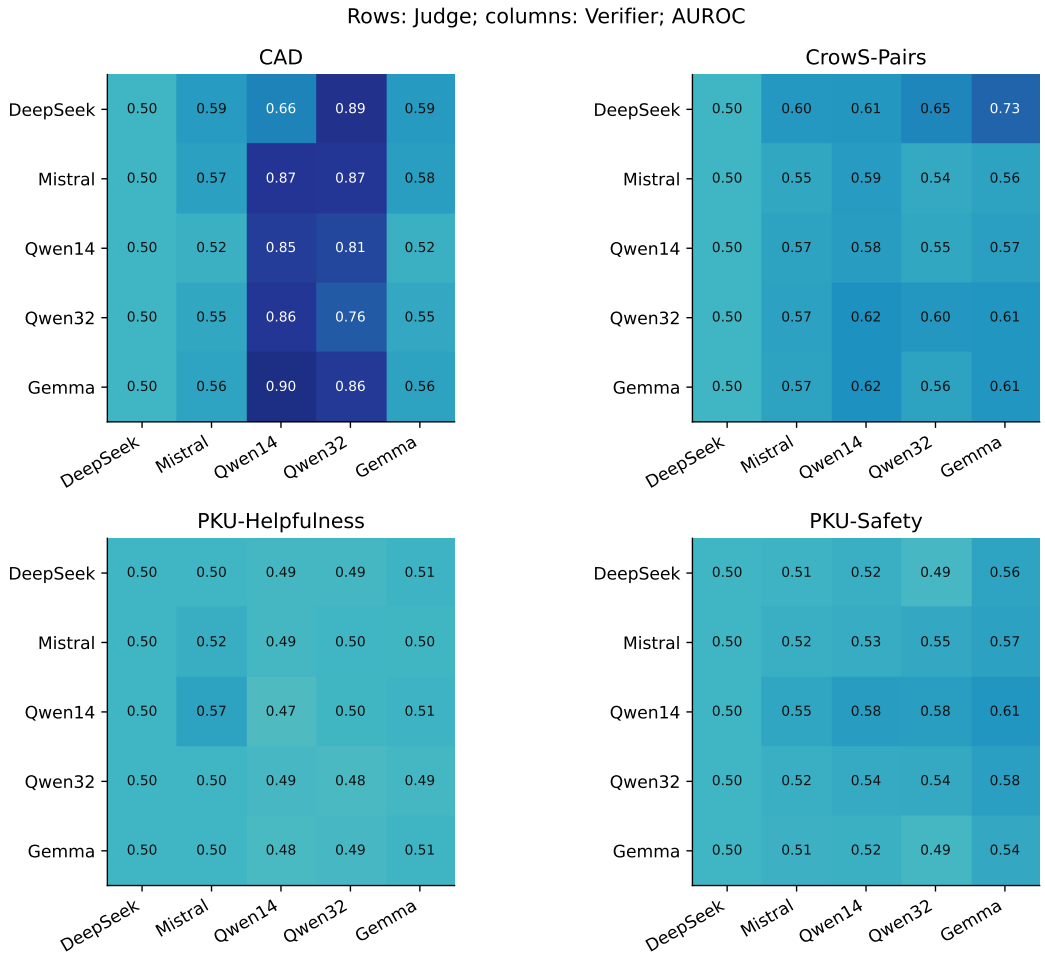


Figure A4: Ordered Judge/Verifier AUROC matrices on the original four-task subset; the six-task numerical results appear in Table A32. Rows are Judges and columns are Verifiers. The diagonal contains self verification.

verdict, or it can depart from an incorrect A verdict. It can also faithfully follow A and inherit A's error. Equation 28 in Appendix D separates these events exactly for binary non-tied judgments. Table A37 (Appendix I) applies it to self-verification outputs. All 766,345 eligible self and peer decisions satisfy the identity. For Qwen2.5-32B, the equal-task external self accuracy is 64.2% on this decided subset, comprising 62.7% with A correct and B faithful plus 1.4% with A wrong and B unfaithful. A further 33.3% faithfully follow an incorrect A verdict. *Following the independent verdict faithfully is not sufficient when that verdict is wrong.* These proportions describe the recorded decision path, not an internal cause or the effect of correcting A.

I.2 Support Consistency Under a Shared Verdict

We can test a simpler consistency requirement without knowing which independent A verdict was supplied. If one correct verdict explains all support decisions on a case, the Verifier should support Judges agreeing with it and withhold support from those disagreeing. We try both possible verdicts and count the fewest mismatches. If even the more favorable choice leaves mismatches, no single verdict can make all those verification decisions correct. We exclude tied Judge verdicts and apply this check separately to each case and Verifier. Proposition 6 in Appendix D gives the calculation.

With tasks weighted equally, Qwen2.5-14B has incompatible patterns in 35.9% of groups containing at least two decided Judge verdicts. Qwen2.5-32B has 34.3%. Even choosing the most favorable A

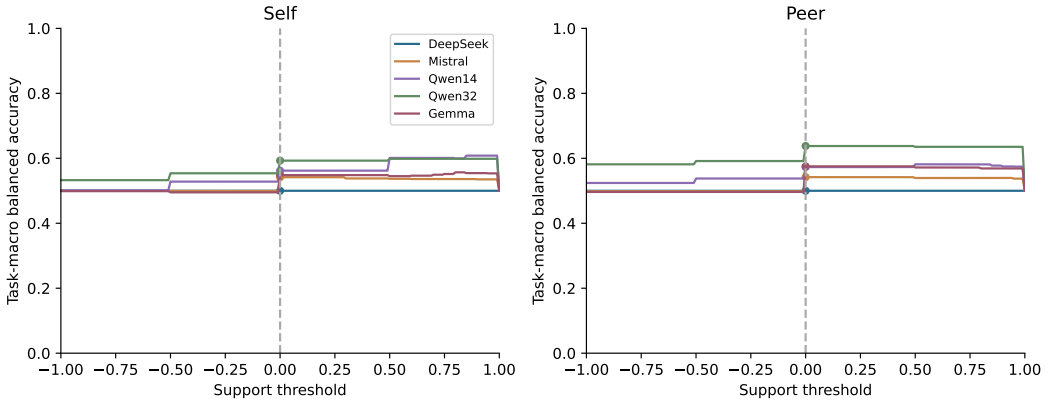


Figure A5: Equal-task Role BA across support-score thresholds on the original four-task subset; BBQ and ES-BBQ were not rerun for this secondary sweep. Curves are descriptive in-sample sensitivities on fixed outputs.

verdict separately for every group leaves at least 9.9% and 11.4% of their decided B decisions mismatched, respectively. The task-specific numerators and denominators in Table A38 (Appendix I) include every model. These minima count forced binary-interface mismatches and also lower-bound external errors on decided judgments. They are not the observed external error rates. Linking worlds strengthens the audit certificate as measured in Appendix G and derived in Appendix C. *A compatible pattern can still faithfully follow an incorrect verdict.* Conversely, neutral non-support can be incompatible on a group containing opposite Judge verdicts without representing active rejection.

I.3 Verification With The Reference Supplied

We also evaluate an earlier reference-informed condition, in which information about the correct answer is supplied rather than left for the Verifier to infer. The CAD and CrowS prompts display the case-local choice label alongside Judge scores. The PKU prompts display the helpfulness or safety label, Judge rationales and reference harm-rationale annotations, and request both a correctness score and a rationale list. This is reference-informed verification. The prompt does not directly state the Boolean variable *this Judge judgment is correct*. This secondary comparison contains the original 8,699 complete pairs; BBQ and ES-BBQ were not run under that historical interface.

Even with reference information available, some Verifiers withhold support from almost all wrong judgments while also failing to support most correct ones. On PKU-safety, Qwen2.5-32B has 99.4% wrong-self non-support but 28.4% correct-self support, yielding 63.9% BA. Its peer components are 98.3% and 29.8%, yielding 64% BA. On PKU-helpfulness, Gemma has 98.6% wrong-self non-support and 10.5% correct-self support. *These patterns show that high error non-support can coexist with failure to support correct judgments even in the reference-informed interface.* The complete tables also retain high performance: Gemma reaches 98.5% self BA on CAD, and Qwen2.5-14B reaches 96.2% on CrowS. The claim concerns observed failures under specific interfaces, not universal inability to use supplied references.

The comparison is retrospective and descriptive. The reference-informed prompts differ from the current two-stage interface in instruction wording, auxiliary rationales, output format, and whether a frozen independent verdict is present. These differences prevent a causal attribution to reference visibility. PKU provides the clearest task-aligned illustration of residual failure with labels supplied. Its historical rationale-aware target is richer than the binary score-sign diagnostic used for this comparison.

This condition tests verification with reference information available. It cannot establish independent verification without references. It can test whether the reference-informed judgments consistently support externally correct Judge scores. The two-stage condition then evaluates the additional dependence on an independently inferred A verdict. Their coexistence motivates separating missing or wrong task information from the subsequent use of that information.

Table A34: Agreement between each model's World-Test Judge verdict and its independently elicited Stage-A verdict. Case agreement conditions on non-tied World verdicts. Both-world agreement conditions on pairs with non-tied World verdicts in both cases. Values are percentages; macro rows weight six tasks equally.

Task	Model	Case agreement	Both-world agreement	Valid cases	Valid pairs
CAD	DeepSeek	87.1	74.2	3,821	1,704
CAD	Mistral	85.5	75.4	4,199	2,014
CAD	Qwen14	93.0	86.5	4,359	2,164
CAD	Qwen32	94.2	89.0	4,324	2,130
CAD	Gemma	93.0	87.1	4,267	2,080
CrowS-Pairs	DeepSeek	87.5	52.3	806	44
CrowS-Pairs	Mistral	81.4	69.2	2,902	1,401
CrowS-Pairs	Qwen14	86.2	77.6	2,942	1,440
CrowS-Pairs	Qwen32	89.4	83.2	2,748	1,271
CrowS-Pairs	Gemma	92.4	88.4	2,772	1,273
PKU-Helpfulness	DeepSeek	36.9	25.1	4,996	2,498
PKU-Helpfulness	Mistral	47.3	31.0	4,409	2,023
PKU-Helpfulness	Qwen14	52.6	38.5	4,450	2,051
PKU-Helpfulness	Qwen32	46.4	38.9	4,979	2,481
PKU-Helpfulness	Gemma	42.1	27.4	4,996	2,498
PKU-Safety	DeepSeek	76.1	64.6	4,996	2,498
PKU-Safety	Mistral	88.6	82.2	4,872	2,386
PKU-Safety	Qwen14	88.0	83.1	4,645	2,205
PKU-Safety	Qwen32	88.9	82.9	4,937	2,443
PKU-Safety	Gemma	73.9	63.0	4,996	2,498
BBQ	DeepSeek	95.0	94.5	1,636	345
BBQ	Mistral	59.3	40.2	13,296	6,463
BBQ	Qwen14	52.9	17.9	13,029	6,201
BBQ	Qwen32	43.4	12.2	12,597	5,836
BBQ	Gemma	46.1	16.5	12,561	5,813
ES-BBQ	DeepSeek	84.0	89.1	363	64
ES-BBQ	Mistral	47.9	25.9	4,752	2,301
ES-BBQ	Qwen14	32.9	1.8	4,801	2,349
ES-BBQ	Qwen32	42.8	18.4	4,281	1,892
ES-BBQ	Gemma	48.8	17.3	4,537	2,110
Six-task macro	DeepSeek	77.8	66.6	n/a	n/a
Six-task macro	Mistral	68.3	54.0	n/a	n/a
Six-task macro	Qwen14	67.6	50.9	n/a	n/a
Six-task macro	Qwen32	67.5	54.1	n/a	n/a
Six-task macro	Gemma	66.1	49.9	n/a	n/a

Table A35: Equal-task Stage-B response percentages across self and peer decisions over six tasks. The first three columns partition all scores. The last three use conditional denominators: non-tied Judge scores that agree with or oppose Stage A. NS includes negative and zero scores.

Verifier	$B > 0$	$B < 0$	$B = 0$	Support given agree	Negative given oppose	NS given oppose
DeepSeek	0.0	0.0	100.0	0.0	0.0	100.0
Mistral	26.5	0.5	73.0	49.1	1.1	98.2
Qwen14	71.7	17.4	10.8	95.2	47.1	56.7
Qwen32	56.9	32.4	10.7	91.5	89.0	89.4
Gemma	54.2	21.3	24.5	87.7	70.4	93.3

J Same-Case And Cross-Model Verification

Another model's involvement need not expose an error: two wrong Judges can also support each other's wrong judgments. We examine three distinct possibilities. Can a model withhold support from its own error while supporting a correct peer on the same case? Do other models withhold support from its error? If two Judges are both wrong, do they support both wrong judgments? Here OW is the original case, AW its edited counterpart, and a hidden error means a correct OW judgment followed by a wrong AW judgment. Each Verifier's independently inferred Stage-A verdict is

Table A36: Equal-task Stage-A-correct conditional Role metrics and joint success across six tasks. Conditional metrics retain only cases where Stage A is correct. Joint success requires both a correct Stage-A verdict and an externally correct binary Stage-B decision.

Verifier	Subset	A-correct BA (%)	A-correct AUROC	Joint (%)
DeepSeek	self	50.0	0.500	29.0
DeepSeek	peer	50.0	0.500	21.0
Mistral	self	77.4	0.776	34.2
Mistral	peer	72.6	0.727	33.1
Qwen14	self	77.0	0.841	54.7
Qwen14	peer	74.9	0.770	52.7
Qwen32	self	95.4	0.962	62.8
Qwen32	peer	88.3	0.905	58.6
Gemma	self	84.0	0.869	53.4
Gemma	peer	92.2	0.936	58.0

Table A37: Two-stage self-verification decomposition on non-tied Judge decisions across six equally weighted tasks. The first two cells sum to external accuracy; the other two are external errors. These ordinary accuracy components use a different subset from Table A28. Equation 28 gives the identity; Appendix F specifies record correspondence and coverage.

	Externally correct (%)		Externally wrong (%)		
Verifier	A right B faithful	A wrong B unfaithful	A wrong B faithful	A right B unfaithful	Accuracy
DeepSeek	10.4	30.2	11.8	47.6	40.6
Mistral	33.2	12.2	33.4	21.2	45.4
Qwen14	54.8	5.0	29.2	10.9	59.9
Qwen32	62.7	1.4	33.3	2.6	64.2
Gemma	53.7	7.0	30.7	8.6	60.6

reused across its five Stage-B support decisions. The five models do not communicate, and these comparisons are not an ensemble or correction procedure.

Overview And Denominator Map. Each column of Table A41 (Appendix J) answers a different question and must be read with its own denominator. The first two columns start from hidden focal errors and ask how the other four models behave as Verifiers. The third starts from any focal AW Judge error and asks how often a peer model is also wrong as a Judge. The final two require both models to be wrong and ask whether both models support both wrong judgments. They report this event first among jointly wrong ordered pairs and then among all ordered AW pairs.

Mistral-Nemo-12B makes the denominator change concrete. Its hidden errors receive peer non-support in 55.7% of task-macro decisions, while 42.4% of unique hidden errors receive non-support from at least three peers. Conditional on any Mistral AW error, a peer Judge is also wrong in 77.1% of ordered units. Complete mutual support occurs on 9.13% of jointly wrong ordered units and 4.72% of all ordered AW units. *Peers can withhold support from some errors and still share and support other errors.* These percentages cannot be compared as one success rate because each starts from a different set of cases.

Same-Case Cross-Role Outcomes. A model might withhold support from its own errors on some cases and support correct peers on other cases. Those separate successes do not establish that it distinguishes the two judgments on the same case. We therefore choose an edited case where the model being examined, J, has a hidden error and a peer Judge, K, is correct. Model J then acts as a Verifier in two separate calls under the recorded prompt form: one for its own wrong judgment and one for K's correct judgment. *Rescue* means non-support for its own error and support for the correct peer. *Persistence* means support for its own error and non-support for the correct peer. *Support both* and *support neither* fail to distinguish the judgments. Rescue names this output pattern. It does not mean that the model generates a corrected verdict. We report ordered focal-to-correct-

Table A38: Compatibility with any one frozen binary verdict, by task and Verifier. Group means one case/world/Verifier with at least two non-tied Judge outputs. Incompatible groups admit no perfect matching verdict. The final columns count all decided B decisions, including singleton groups, and the minimum mismatches forced by Equation 29 (Appendix D). Rates are percentages. No Stage-A label or reference correctness is used to calculate these columns.

Task	Verifier	Groups	Incompatible groups	Decided B	Minimum mismatches	Minimum rate
CAD	DeepSeek	4,390	40.2	20,970	2,346	11.2
CAD	Mistral	4,390	56.3	20,970	3,387	16.2
CAD	Qwen14	4,390	25.6	20,970	1,321	6.3
CAD	Qwen32	4,390	15.9	20,970	917	4.4
CAD	Gemma	4,390	56.4	20,970	3,391	16.2
CrowS	DeepSeek	3,012	29.8	12,170	1,074	8.8
CrowS	Mistral	3,012	64.0	12,170	2,556	21.0
CrowS	Qwen14	3,012	22.4	12,170	819	6.7
CrowS	Qwen32	3,012	3.5	12,170	106	0.9
CrowS	Gemma	3,012	8.3	12,170	264	2.2
PKU-helpfulness	DeepSeek	4,996	76.1	23,830	4,872	20.4
PKU-helpfulness	Mistral	4,996	16.2	23,830	935	3.9
PKU-helpfulness	Qwen14	4,996	65.8	23,830	4,430	18.6
PKU-helpfulness	Qwen32	4,996	48.6	23,830	3,814	16.0
PKU-helpfulness	Gemma	4,996	33.7	23,830	2,777	11.7
PKU-safety	DeepSeek	4,996	31.1	24,446	2,051	8.4
PKU-safety	Mistral	4,996	62.0	24,446	4,948	20.2
PKU-safety	Qwen14	4,996	29.8	24,446	1,932	7.9
PKU-safety	Qwen32	4,996	69.3	24,446	5,907	24.2
PKU-safety	Gemma	4,996	35.9	24,446	2,842	11.6

peer comparison units in Table A42 (Appendix J), so one hidden error contributes one to four units when multiple peers are correct.

In Table A43 (Appendix J), we add paired-cluster intervals for equal-task eligibility and rescue. Rescue ranges from 0% [0%, 0%] for DeepSeek-R1-8B to 36.9% [33.3%, 40.4%] for Qwen2.5-32B. DeepSeek has 1,999 valid rescue resamples because the CrowS eligibility denominator is sparse. We report that valid-replicate count rather than treating a missing denominator as zero. *Even the highest average rescue rate leaves most eligible comparisons without the desired combination of withholding support from the error and supporting the correct peer.*

Opposite-Verdict Check. Table A44 (Appendix J) excludes tied focal judgments before interpreting the support pattern relative to a single frozen binary verdict. On CAD, Qwen2.5-14B supports both opposite verdicts in 270 of 554 retained comparisons (48.7%). No single binary A verdict could make both score directions agree with it. *The observation therefore cannot be explained solely by choosing the wrong A verdict while matching it faithfully.* It remains a statement about the score-sign interface, not internal reasoning. Neutral non-support is analyzed separately. Appendix D.8 gives the binary argument.

The task-level cells expose different failures. On CAD, the Gemma and Mistral Verifier rows support neither judgment in 99.3% and 95.4% of eligible comparisons, respectively. On PKU-Safety, Qwen2.5-32B produces persistence in 72.3%, whereas Qwen2.5-14B supports both in 46.4%. DeepSeek's CrowS row has only 14 ordered comparisons from seven eligible hidden errors, all support-neither. *A low rescue rate can therefore arise from distinct observed decision patterns.*

Sensitivity To OW Conditioning. Table A45 (Appendix J) removes the requirement that the focal Judge was correct in OW, retaining an AW-wrong focal Judge and an AW-correct peer. The dominant cell changes in two of the twenty task/model rows: CrowS/Qwen32 moves from persistence to rescue, and PKU-Safety/Qwen14 moves from support-both to rescue. DeepSeek's CrowS population expands from 14 to 851 ordered units and remains 100% support-neither. *OW conditioning therefore affects both the population and, in some rows, the dominant Role pattern.*

Table A39: Reference-informed self verification on matched cases. Each task uses the same complete identities and frozen Judge outputs as the current two-stage condition. Correct and wrong counts supply the component denominators. Support and BA are percentages. AUROC uses continuous scores. The final column gives current two-stage BA for orientation, not a causal label-removal effect. Prompts and auxiliary information differ (Appendix I.3).

Task	Verifier	Cases	Correct	Wrong	Correct support	Wrong NS	BA	AUROC	Two-stage BA
CAD	DeepSeek	4,390	2,170	2,220	69.1	74.3	71.7	0.734	50.0
CAD	Mistral	4,390	3,345	1,045	56.5	80.8	68.6	0.726	57.2
CAD	Qwen14	4,390	3,871	519	97.2	96.7	97.0	0.985	63.5
CAD	Qwen32	4,390	4,039	351	94.5	96.9	95.7	0.969	73.4
CAD	Gemma	4,390	3,644	746	98.6	98.5	98.5	0.989	56.4
CrowS	DeepSeek	3,016	776	2,240	0.0	100.0	50.0	0.501	50.0
CrowS	Mistral	3,016	1,736	1,280	41.0	99.8	70.4	0.723	54.5
CrowS	Qwen14	3,016	1,595	1,421	92.5	99.9	96.2	0.978	53.7
CrowS	Qwen32	3,016	1,503	1,513	82.5	93.7	88.1	0.906	61.1
CrowS	Gemma	3,016	1,616	1,400	98.2	96.1	97.2	0.980	56.8
PKU-helpfulness	DeepSeek	4,996	2,510	2,486	23.1	82.4	52.7	0.527	50.0
PKU-helpfulness	Mistral	4,996	2,183	2,813	45.1	86.7	65.9	0.655	52.6
PKU-helpfulness	Qwen14	4,996	2,110	2,886	59.2	78.8	69.0	0.683	49.2
PKU-helpfulness	Qwen32	4,996	2,485	2,511	63.8	96.4	80.1	0.801	48.4
PKU-helpfulness	Gemma	4,996	2,510	2,486	10.5	98.6	54.6	0.700	51.7
PKU-safety	DeepSeek	4,996	2,525	2,471	45.7	62.5	54.1	0.541	50.0
PKU-safety	Mistral	4,996	2,458	2,538	57.1	64.3	60.7	0.613	52.5
PKU-safety	Qwen14	4,996	2,337	2,659	41.6	68.7	55.2	0.527	58.4
PKU-safety	Qwen32	4,996	2,503	2,493	28.4	99.4	63.9	0.639	54.2
PKU-safety	Gemma	4,996	2,497	2,499	24.3	96.9	60.6	0.619	54.3

Peer Non-Support For Hidden Errors. We hold the hidden focal error fixed and switch to each peer Verifier, which need not itself be an AW-correct Judge. Table A46 (Appendix J) and Figure A6 (Appendix J) use focal Judges as rows and peer Verifiers as columns. Qwen2.5-32B assigns non-support to 66.2% of DeepSeek's hidden errors, while DeepSeek assigns non-support to 100% of Qwen32's errors. *Withholding support from every error does not make DeepSeek a successful peer checker.* Its column is 100% for every focal Judge because its scores are neutral zeros, so the error-only view cannot establish useful discrimination. This error-only matrix needs the corresponding correct-support rates in Tables A30 and A31 (Appendix I).

Shared Errors And Complete Mutual Support. Table A47 (Appendix J) uses an unordered model pair on one AW case. Joint wrong means that both models are wrong as Judges. We define *complete mutual support* (CMS) as four positive B decisions: each of the two models supports both its own wrong judgment and the other model's wrong judgment. This names the specific conjunction measured here. It is not an interaction or correction algorithm. The conditional CMS rate uses jointly wrong cases as its denominator. The joint error-and-CMS rate uses all AW cases. These unordered-pair rates differ from Table A41 (Appendix J)'s directional focal-to-peer rates.

Equal-task joint-error prevalence ranges from 33.8% to 48.4% across the ten unordered pairs. Qwen2.5-14B/Qwen2.5-32B reaches 44.9% CMS conditional on joint error, and 17.6% for joint error and CMS among all AW cases. The six task counts are 108/211 on CAD, 564/761 on CrowS-Pairs, 117/602 on PKU-Helpfulness, 1,105/1,460 on PKU-Safety, 499/2,223 on BBQ and 180/681 on ES-BBQ. The corresponding all-AW denominators are 2,195, 1,508, 2,498, 2,498, 6,880 and 2,456. *Two models can therefore endorse one another's judgments even when both judgments are wrong.* This is a recorded support pattern, not evidence of collusion.

Recorded-Protocol Panel Signatures. Before reference correctness is joined, one can observe how many of the five B outputs support a Judge judgment. Table A49 groups 180,350 complete case/world/Judge units by this count and then attaches correctness. All-five support occurs zero times because DeepSeek supplies no positive support. Three-or-four support contains 51,418 correct judgments among 74,112 units (69.38%). One-or-two support contains 25,693/65,285 (39.36%). No

Table A40: Reference-informed peer verification on matched cases. Each task uses the same complete identities and frozen Judge outputs as the current two-stage condition. Correct and wrong counts supply the component denominators. Support and BA are percentages. AUROC uses continuous scores. The final column gives current two-stage BA for orientation, not a causal label-removal effect. Prompts and auxiliary information differ (Appendix I.3).

Task	Verifier	Cases	Correct	Wrong	Correct support	Wrong NS	BA	AUROC	Two-stage BA
CAD	DeepSeek	17,560	14,899	2,661	59.9	34.5	47.2	0.597	50.0
CAD	Mistral	17,560	13,724	3,836	61.1	90.1	75.6	0.763	55.9
CAD	Qwen14	17,560	13,198	4,362	97.4	79.9	88.6	0.929	67.9
CAD	Qwen32	17,560	13,030	4,530	94.7	97.2	96.0	0.961	87.0
CAD	Gemma	17,560	13,425	4,135	99.0	96.9	97.9	0.979	56.6
CrowS	DeepSeek	12,064	6,450	5,614	0.0	100.0	50.0	0.500	50.0
CrowS	Mistral	12,064	5,490	6,574	47.8	99.7	73.7	0.784	58.6
CrowS	Qwen14	12,064	5,631	6,433	90.3	87.8	89.0	0.878	58.9
CrowS	Qwen32	12,064	5,723	6,341	65.4	90.9	78.2	0.831	63.4
CrowS	Gemma	12,064	5,610	6,454	94.7	95.5	95.1	0.960	64.3
PKU-helpfulness	DeepSeek	19,984	9,288	10,696	20.0	91.0	55.5	0.555	50.0
PKU-helpfulness	Mistral	19,984	9,615	10,369	39.9	70.6	55.3	0.547	50.4
PKU-helpfulness	Qwen14	19,984	9,688	10,296	67.8	66.7	67.2	0.668	50.2
PKU-helpfulness	Qwen32	19,984	9,313	10,671	31.3	97.2	64.3	0.665	52.3
PKU-helpfulness	Gemma	19,984	9,288	10,696	42.8	98.5	70.7	0.798	51.5
PKU-safety	DeepSeek	19,984	9,795	10,189	35.5	76.3	55.9	0.558	50.0
PKU-safety	Mistral	19,984	9,862	10,122	48.6	60.4	54.5	0.532	51.9
PKU-safety	Qwen14	19,984	9,983	10,001	49.2	65.5	57.3	0.558	52.2
PKU-safety	Qwen32	19,984	9,817	10,167	29.8	98.3	64.0	0.640	52.3
PKU-safety	Gemma	19,984	9,823	10,161	46.1	99.6	72.8	0.785	57.7

Table A41: Cross-model verification diagnostics with distinct denominators. CMS denotes complete mutual support, requiring both models to support both wrong judgments. The first two rates use hidden focal errors. The third conditions on any focal AW error; the fourth on jointly wrong focal/peer judgments. The last divides joint error and CMS by all ordered AW focal/peer units. Task-macro percentages on the complete cohort. See Appendix J.

Focal	Peer NS	At least 3 NS	Peer wrong given focal wrong	CMS given both wrong	Joint error and CMS
DeepSeek	63.2	54.2	63.1	0.00	0.00
Mistral	54.0	38.9	77.1	10.40	4.95
Qwen14	58.6	40.1	75.9	26.38	10.78
Qwen32	65.8	49.6	79.8	21.32	9.01
Gemma	71.0	61.5	75.2	17.59	7.71

support contains 7,881/40,953 (19.24%). *More support is associated with greater correctness in this cohort. Even three or four supporters, however, leave 30.62% of judgments wrong.* Panel support is therefore informative without constituting a correctness certificate.

The association also changes by task, as the six task columns in Table A49 make explicit. *The same visible panel pattern can have different meanings under different criteria and datasets.* This is a post-hoc association under the recorded two-stage protocol, not a calibrated confidence estimate or portable correctness rule.

Together, these results separate three claims that a single cross-model summary would blur: peer non-support for errors does not establish two-sided Role discrimination, shared Judge errors do not imply complete mutual support, and panel agreement is task-conditional. The fixed role grid diagnoses these output patterns without turning them into an ensemble decision.

Table A42: Same-case outcomes for an OW-correct, AW-wrong focal Judge and an AW-correct peer Judge across the six-task, 18,035-pair Role cohort. Units are ordered focal-to-correct-peer comparisons. Only the four outcome columns sum to 100% within a row.

Task	Focal	Units	Rescue	Persistence	Both	Neither
CAD	DeepSeek	4434	0.0	0.0	0.0	100.0
CAD	Mistral	812	0.0	4.6	0.0	95.4
CAD	Qwen14	593	33.6	16.9	48.6	1.0
CAD	Qwen32	60	26.7	43.3	0.0	30.0
CAD	Gemma	801	44.3	38.7	0.6	16.4
CrowS-Pairs	DeepSeek	14	0.0	0.0	0.0	100.0
CrowS-Pairs	Mistral	347	6.3	6.6	0.6	86.5
CrowS-Pairs	Qwen14	289	21.5	7.6	70.6	0.3
CrowS-Pairs	Qwen32	921	37.7	50.4	4.2	7.7
CrowS-Pairs	Gemma	603	30.8	46.6	10.8	11.8
PKU-Helpfulness	DeepSeek	1435	0.0	0.0	0.0	100.0
PKU-Helpfulness	Mistral	1321	54.7	23.0	0.0	22.3
PKU-Helpfulness	Qwen14	1551	14.7	28.9	36.8	19.7
PKU-Helpfulness	Qwen32	1657	73.6	23.5	1.5	1.4
PKU-Helpfulness	Gemma	1394	65.2	9.6	1.7	23.5
PKU-Safety	DeepSeek	483	0.0	0.0	0.0	100.0
PKU-Safety	Mistral	483	13.3	35.6	1.4	49.7
PKU-Safety	Qwen14	433	28.6	20.3	46.4	4.6
PKU-Safety	Qwen32	618	9.5	72.3	6.8	11.3
PKU-Safety	Gemma	642	17.9	18.4	34.4	29.3
BBQ	DeepSeek	394	0.0	0.0	0.0	100.0
BBQ	Mistral	2734	6.6	46.7	0.1	46.6
BBQ	Qwen14	2485	20.0	28.2	48.4	3.3
BBQ	Qwen32	1677	91.1	6.0	1.8	1.1
BBQ	Gemma	3806	89.9	2.9	0.1	7.1
ES-BBQ	DeepSeek	42	0.0	0.0	0.0	100.0
ES-BBQ	Mistral	875	40.5	34.7	1.7	23.1
ES-BBQ	Qwen14	923	16.8	63.7	15.4	4.1
ES-BBQ	Qwen32	420	57.6	33.8	3.3	5.2
ES-BBQ	Gemma	1240	67.7	21.1	2.7	8.4

Table A43: Equal-task eligibility and rescue percentages on the complete cohort, with paired-cluster 95% intervals. Eligibility conditions on unique hidden errors. Rescue conditions on ordered eligible comparisons. The last column gives valid rescue resamples out of 2,000.

Focal	Eligible hidden cases (%)	Rescue	Valid resamples
DeepSeek	47.7 [46.4, 49.2]	0.0 [0.0, 0.0]	1999
Mistral	50.6 [49.0, 52.1]	18.6 [17.1, 20.0]	2000
Qwen14	50.1 [48.5, 51.8]	24.6 [22.0, 27.2]	2000
Qwen32	43.0 [41.2, 44.8]	36.9 [33.3, 40.4]	2000
Gemma	52.9 [51.4, 54.4]	28.5 [26.4, 30.5]	2000

J.1 Joint And Marginal Verification Analysis

Knowing how often each model supplies support does not tell us how often the supports occur together on the same cases. We compare the uncertainty left by separate rates with the actual overlap recovered by linking decisions case by case. Proposition 7 in Appendix D supplies the bounds. We quantify that distinction without new model calls. For rescue, each task supplies own-error non-support and correct-peer support on the original eligible ordered comparisons. For CMS, each task supplies all four component support rates on jointly wrong cases for a fixed unordered model pair. We calculate the lower and upper bounds within task, then average them equally. These are stronger marginals than the headline self/peer rates, because their populations already match the target event. The bounds allow every way of matching decisions across cases that preserves those separate rates. Their endpoints can be attained if no other constraint is imposed. Additional requirements, such as reusing one decision in several comparisons, may narrow the range but do not invalidate it.

The dependence-bound analysis in Table A50 remains scoped to the original four-task subset. For Qwen2.5-32B rescue, the marginals leave a 12.5% to 40% interval and the observed linkage gives

Table A44: Same-case outcomes after excluding tied focal AW judgments. The focal Judge is OW-correct and AW-wrong. A peer is AW-correct. Each eligible peer contributes an ordered comparison. The retained verdicts are opposite in sign, so exact matching to one frozen binary A verdict supports exactly one. Rates use retained comparisons within each row and sum to 100% before rounding. N/A means no retained comparisons. This is a descriptive subset analysis. It does not replace the primary tie-inclusive results in Appendix J, Table A42.

Task	Focal model	Eligible	Retained	Ties removed	Rescue	Persistence	Both	Neither
CAD	DeepSeek	4,434	3,314	1,120	0.0	0.0	0.0	100.0
CAD	Mistral	812	655	157	0.0	5.6	0.0	94.4
CAD	Qwen14	593	554	39	33.0	18.1	48.7	0.2
CAD	Qwen32	60	39	21	15.4	66.7	0.0	17.9
CAD	Gemma	801	650	151	0.0	0.9	0.0	99.1
CrowS	DeepSeek	14	0	14	N/A	N/A	N/A	N/A
CrowS	Mistral	347	300	47	5.3	7.7	0.7	86.3
CrowS	Qwen14	289	267	22	19.1	7.9	72.7	0.4
CrowS	Qwen32	921	697	224	27.8	66.6	5.6	0.0
CrowS	Gemma	603	427	176	13.6	65.8	15.2	5.4
PKU-helpfulness	DeepSeek	1,435	1,435	0	0.0	0.0	0.0	100.0
PKU-helpfulness	Mistral	1,321	1,125	196	60.7	27.0	0.0	12.3
PKU-helpfulness	Qwen14	1,551	1,438	113	15.2	30.7	34.9	19.3
PKU-helpfulness	Qwen32	1,657	1,648	9	73.5	23.5	1.5	1.5
PKU-helpfulness	Gemma	1,394	1,394	0	65.2	9.6	1.7	23.5
PKU-safety	DeepSeek	483	483	0	0.0	0.0	0.0	100.0
PKU-safety	Mistral	483	451	32	12.4	38.1	1.6	47.9
PKU-safety	Qwen14	433	336	97	14.0	25.9	59.8	0.3
PKU-safety	Qwen32	618	584	34	9.2	76.5	7.2	7.0
PKU-safety	Gemma	642	642	0	17.9	18.4	34.4	29.3

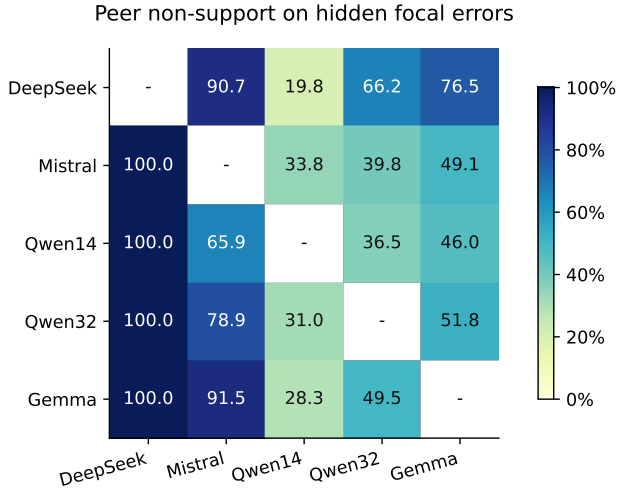


Figure A6: Peer non-support for hidden AW errors on the original four-task subset; the updated six-task matrix appears in Table A46. Rows identify focal Judges and columns identify peer Verifiers. Larger values do not establish two-sided discrimination.

36.9%. For Qwen2.5-14B/Qwen2.5-32B CMS, they leave 19.7% to 60.5%, while the linked event is 55.1%. *The nonzero lower bound already forces some shared support; linking identifies its observed extent.* DeepSeek's constant non-support supplies the boundary where both endpoints are zero. BBQ and ES-BBQ were not rerun for this secondary marginal-bound calculation.

Table A45: Same-case sensitivity using any focal AW error across the six-task, 18,035-pair Role cohort. Units are ordered focal-to-correct-peer comparisons; each row is task-specific.

Task	Focal	Units	Rescue	Persistence	Both	Neither
CAD	DeepSeek	5436	0.0	0.0	0.0	100.0
CAD	Mistral	1098	0.0	4.6	0.0	95.4
CAD	Qwen14	595	33.6	16.8	48.6	1.0
CAD	Qwen32	63	27.0	44.4	0.0	28.6
CAD	Gemma	858	44.6	38.3	0.6	16.4
CrowS-Pairs	DeepSeek	851	0.0	0.0	0.0	100.0
CrowS-Pairs	Mistral	387	7.2	6.2	0.5	86.0
CrowS-Pairs	Qwen14	365	27.9	7.4	64.4	0.3
CrowS-Pairs	Qwen32	1190	43.9	43.4	4.6	8.0
CrowS-Pairs	Gemma	693	35.5	42.9	10.0	11.7
PKU-Helpfulness	DeepSeek	1619	0.0	0.0	0.0	100.0
PKU-Helpfulness	Mistral	2044	45.5	22.7	0.0	31.8
PKU-Helpfulness	Qwen14	2226	15.7	27.2	38.5	18.6
PKU-Helpfulness	Qwen32	1680	73.9	23.2	1.5	1.4
PKU-Helpfulness	Gemma	1575	62.3	10.0	1.8	25.8
PKU-Safety	DeepSeek	635	0.0	0.0	0.0	100.0
PKU-Safety	Mistral	682	18.2	29.0	1.2	51.6
PKU-Safety	Qwen14	804	43.8	15.5	34.1	6.6
PKU-Safety	Qwen32	745	11.7	67.0	6.8	14.5
PKU-Safety	Gemma	912	25.7	15.2	30.4	28.7
BBQ	DeepSeek	10940	0.0	0.0	0.0	100.0
BBQ	Mistral	3575	14.9	37.5	0.1	47.5
BBQ	Qwen14	3168	23.3	27.7	40.4	8.6
BBQ	Qwen32	3437	89.4	5.8	2.4	2.3
BBQ	Gemma	5166	89.9	2.6	0.1	7.4
ES-BBQ	DeepSeek	4139	0.0	0.0	0.0	100.0
ES-BBQ	Mistral	1151	43.4	31.2	5.0	20.4
ES-BBQ	Qwen14	1009	20.5	60.3	15.5	3.8
ES-BBQ	Qwen32	1357	68.5	22.1	1.5	7.8
ES-BBQ	Gemma	1536	68.2	18.9	2.4	10.4

Table A46: Equal-six-task peer non-support percentages on hidden errors. Rows identify the focal Judge and columns the Verifier. These error-only rates do not establish two-sided discrimination.

Judge / Verifier	DeepSeek	Mistral	Qwen14	Qwen32	Gemma
DeepSeek	n/a	87.9	28.5	71.0	65.5
Mistral	100.0	n/a	33.2	43.2	39.5
Qwen14	100.0	61.7	n/a	40.4	32.4
Qwen32	100.0	79.1	37.2	n/a	46.6
Gemma	100.0	88.0	39.5	56.6	n/a

K Diagnostic Profiles And Purpose-Specific Selection

We treat the five checkpoints as diagnostic case studies, not leaderboard candidates. This section asks two questions left by the main results. When two Judges have similar Original-World (OW) accuracy, how different can their remaining World and Role measurements be? After a user states which errors carry cost, can a small labeled diagnostic audit improve selection among the same five Judges? A profile records how a model behaves on the original case (OW), its edited counterpart (AW), and verification of judgments. We retain those components separately throughout. These analyses do not produce one overall or universal model ranking, and they do not estimate how common the observed behaviors are across the field.

K.1 Diagnostic Profiles At Matched Accuracy

Nearly equal original-case accuracy can conceal very different behavior after edits and during verification. Within each task, we form all ten unordered model pairs and retain those whose strict OW accuracies differ by at most three percentage points. Profile separation is the mean of eleven absolute component differences. The components are AW accuracy, HFR, the three hidden-failure score-movement fractions, own-error non-support, correct-peer support, and the four same-case out-

Table A47: Shared AW errors and complete mutual support on the 18,035-pair complete Role cohort. Values average six task-specific percentages equally.

Model A	Model B	Joint wrong	CMS given joint wrong	Joint error and CMS
DeepSeek	Mistral	48.17	0.00	0.00
DeepSeek	Qwen14	39.05	0.00	0.00
DeepSeek	Qwen32	44.55	0.00	0.00
DeepSeek	Gemma	48.44	0.00	0.00
Mistral	Qwen14	38.06	21.41	9.80
Mistral	Qwen32	40.07	13.66	6.61
Mistral	Gemma	41.27	7.95	3.37
Qwen14	Qwen32	33.78	44.88	17.64
Qwen14	Gemma	34.70	41.02	15.68
Qwen32	Gemma	45.01	27.59	11.80

Table A48: Task-level complete mutual support (CMS) for Qwen14/Qwen32 shared AW errors. CMS requires both models to be wrong and each to support both errors. The final two rates use jointly wrong and all AW cases, respectively.

Task	CMS / both wrong	All AW	CMS / wrong (%)	CMS / all (%)
CAD	108 / 211	2,195	51.2	4.9
CrowS-Pairs	564 / 761	1,508	74.1	37.4
PKU-Helpfulness	117 / 602	2,498	19.4	4.7
PKU-Safety	1,105 / 1,460	2,498	75.7	44.2
BBQ	499 / 2,223	6,880	22.4	7.3
ES-BBQ	180 / 681	2,456	26.4	7.3
Equal-task mean	Task-specific		44.9	17.6

comes. OW accuracy is only the matching variable. All 15 eligible pairs appear in Table A52 (Appendix K), with gaps from 6.9 to 41.9 percentage points. DeepSeek and Mistral have equal original-case accuracy on PKU-Safety, yet their remaining profile components differ by 25.5 points on average. This gap measures difference in behavior, not which model is better. The marked pairs maximize separation within task using the two-stage Role scores. They are selected examples of maximal separation. This is a descriptive extremum, not a quality score, uncertainty estimate, or typical near-tie. For comparison, Figure A8 (Appendix K) evaluates a fixed alternative set of model pairs, making dependence on pair selection visible.

K.2 Limitations Of Scalar Rankings

The model ranked best depends on which behavior we measure. This secondary ranking analysis remains scoped to the original four-task subset; BBQ and ES-BBQ were not rerun for its matched-selection experiment. We use Table A53 to ask whether replacing OW accuracy with another single measure restores one model ordering. We retain five separate equal-task ranks rather than combining the measures. Qwen2.5-32B ranks first on OW accuracy and fourth on Hidden Failure Rate (HFR), whereas Qwen2.5-14B ranks fourth on OW accuracy and first on both AW accuracy and HFR. The latter "first" HFR is still 71.3%. It is only the lowest HFR among these five models. DeepSeek ranks first on wrong-judgment non-support because all self scores are zero, while Qwen2.5-32B ranks first on peer BA. The first rank is therefore not evidence of useful verification. No model leads all five measures. *The reversals identify information that any single ordering would discard.*

K.3 Held-Out Selection Under Stated Costs

The profile does not by itself decide which Judge to use because that decision depends on which errors carry cost. We study one bounded use under three prespecified definitions. *Balanced worlds* gives equal weight to OW and AW error. *OW and hidden failure* gives equal weight to an OW error and the separate event in which the Judge is OW-correct and AW-wrong. *Self-confirmation averse* places half its weight on an AW error that the same model supports as a Verifier and one quarter on

Table A49: Panel signatures over 180,350 complete case/world/Judge units. Macro prevalence gives six tasks equal weight; pooled correctness combines units. These associations are not calibrated confidence rules.

Signature	Macro prevalence	Pooled correctness	N	CAD	CrowS	PKU-H	PKU-S	BBQ	ES-BBQ
All five	0.0	n/a	0	n/a	n/a	n/a	n/a	n/a	n/a
Three or four	45.5	69.38	74,112	96.3	59.9	47.9	53.6	75.1	55.8
One or two	36.1	39.36	65,285	41.3	33.8	48.3	46.6	35.8	32.3
None	18.4	19.24	40,953	4.8	8.9	44.7	22.3	8.1	33.4

Table A50: Bounds from marginal verification rates versus observed linked events on the original four-task, 8,699-pair subset (%). BBQ and ES-BBQ were not included in this secondary dependence-bound analysis. Bounds are computed within task on the event's eligible population and then averaged equally.

Verifier or pair	Event	Lower	Observed	Upper	Width
DeepSeek	Rescue	0.0	0.0	0.0	0.0
Mistral	Rescue	7.9	18.6	19.1	11.2
Qwen14	Rescue	9.7	24.6	31.0	21.3
Qwen32	Rescue	12.5	36.9	40.0	27.5
Gemma	Rescue	13.9	28.5	38.9	25.0
DeepSeek / Mistral	CMS	0.0	0.0	0.0	0.0
DeepSeek / Qwen14	CMS	0.0	0.0	0.0	0.0
DeepSeek / Qwen32	CMS	0.0	0.0	0.0	0.0
DeepSeek / Gemma	CMS	0.0	0.0	0.0	0.0
Mistral / Qwen14	CMS	0.0	16.9	21.1	21.1
Mistral / Qwen32	CMS	0.0	14.9	19.1	19.1
Mistral / Gemma	CMS	0.0	5.4	8.9	8.9
Qwen14 / Qwen32	CMS	19.7	55.1	60.5	40.8
Qwen14 / Gemma	CMS	14.0	34.8	38.7	24.7
Qwen32 / Gemma	CMS	7.0	24.7	29.4	22.4

each world error. Cost values are comparable only within the same definition. A lower cost means fewer of the errors penalized by that definition. Costs are weighted penalties, not always raw error rates. The equal-task result for the main 10% diagnostic audit appears in Table A51 (Appendix K). For each pair, c^w indicates a correct Judge verdict in world w, and $e^w = 1 - c^w$ indicates an error. Then $h = c^O e^A$ marks an original success followed by an edited-case error, and $v^A = v^A_{iJ}$ indicates support for the model's own edited-case judgment. The per-pair costs are

$$\begin{aligned} L_{\text{bal}} &= \tfrac{1}{2}(e^O + e^A), \\ L_{\text{hidden}} &= \tfrac{1}{2}(e^O + h), \\ L_{\text{self}} &= \tfrac{1}{4}(e^O + e^A) + \tfrac{1}{2}e^A v^A. \end{aligned} \tag{34}$$

These formulas turn the declared priorities into a score used to choose a Judge. They do not declare one priority universally preferable. The hidden-failure target penalizes the per-pair event h, not the conditional rate HFR. Since $e^O + c^O e^A = 1 - c^O c^A$, this target equals one half of the both-world error. Reversing the anchor leaves the target unchanged and changes only the evidence available to the single-world selector.

We reveal labels for a small set of pairs, choose a Judge using that evidence, and evaluate the chosen Judge on the remaining pairs. Both practical selectors inspect the same pair identities and choose among the same five Judges. The OW-only baseline minimizes observed OW error. It cannot estimate a cost containing unobserved AW or Role terms. The diagnostic selector minimizes the stated cost using the required World or World/Role evidence. Both selected Judges are then evaluated under that cost on the remaining pairs. The selectors differ in their available audit evidence and the selection score that this evidence permits. The OW-only selector receives the audited OW labels. The diagnostic-audit selector receives the audited OW/AW labels and, for the self-confirmation-averse cost, the corresponding self-verification decisions. Neither selector sees labels for the remaining pairs. Audit-score ties use the fixed order DeepSeek, Mistral, Qwen14, Qwen32, Gemma. Each

Table A51: Same-task selection on the original four-task, 8,699-pair subset from the same 10% audited identities over 500 fixed splits per task. BBQ and ES-BBQ were not included in this secondary selection experiment. Costs are percentages and gains are percentage points.

Stated cost	OW-only	Diagnostic	Oracle	Gain (pp)
Balanced worlds	40.32	37.56	37.40	2.77
OW + hidden failure	38.62	35.93	35.79	2.69
Self-confirmation	35.56	20.04	20.03	15.52

Table A52: All within-task pairs satisfying the fixed three-percentage-point OW matching rule on the complete cohort. The gap averages eleven absolute component differences and measures separation, not quality. Yes marks the maximum-gap pair within each task used in Figure A7 (Appendix K).

Task	Model A	Model B	OW gap (pp)	Profile gap (pp)	Selected
CAD	DeepSeek	Mistral	0.6	14.7	No
CAD	Qwen32	Gemma	1.6	31.4	Yes
CrowS	Mistral	Gemma	2.5	32.3	Yes
CrowS	Qwen32	Gemma	3.0	6.9	No
PKU-H	DeepSeek	Qwen32	0.9	31.9	Yes
PKU-H	DeepSeek	Gemma	1.4	19.4	No
PKU-H	Qwen32	Gemma	0.5	12.6	No
PKU-S	DeepSeek	Mistral	0.0	25.5	No
PKU-S	DeepSeek	Qwen14	2.6	41.9	Yes
PKU-S	DeepSeek	Qwen32	2.8	36.4	No
PKU-S	DeepSeek	Gemma	0.6	27.0	No
PKU-S	Mistral	Qwen14	2.6	19.2	No
PKU-S	Mistral	Qwen32	2.8	12.0	No
PKU-S	Mistral	Gemma	0.6	15.9	No
PKU-S	Qwen14	Gemma	2.0	15.3	No

task uses 500 splits. For budget $b \in \{5, 10, 20, 30\}$ expressed as a percentage, the analysis initializes the permutation generator with $20260911 + 1000b + d$, where $d = 0, 1, 2, 2$ for CAD, CrowS, helpfulness, and safety. The two PKU tasks therefore share identity permutations. Each audit contains $\lfloor nb/100 \rfloor$ pairs, yielding 219, 150, 249, and 249 pairs at 10%. These seeds define the reported resampling analysis.

The held-out oracle does see the remaining labels. It is therefore an unattainable hindsight minimum among the same five candidates, not a practical method. This is a within-task held-out experiment on the fixed benchmark, not evaluation on a new task or distribution. The experiment compares richer labels on the same audited identities. It does not compare equal annotation cost.

K.4 World And Role Ablations

Looking at edited cases improves selection under the stated costs; verification evidence adds benefit when the objective also penalizes supported errors. The primary diagnostic selector uses only the evidence needed by its stated cost. Table A54 (Appendix K) makes this separation explicit by comparing three selectors on the same audited identities. OW-only uses OW correctness. World-only uses OW and AW correctness. World+Role also uses the corresponding self-verification decision when the cost contains a Role term.

World evidence reduces held-out cost by 2.77 points under balanced worlds and 2.69 points under OW-and-hidden-failure cost. World+Role equals World-only for these costs because they have no Role term. For self-confirmation aversion, World evidence contributes 12.11 points and Role evidence adds 3.4, totaling 15.52. For the self-confirmation target, the World-only selector minimizes only the two world-error terms before all selectors are evaluated on the full target. These are percentage-point reductions in the declared weighted cost, not relative percentage gains or new model improvements. The reported components and total are rounded separately. This separates the selection benefit of the available evidence on fixed outputs. It does not identify a causal effect or compare equal acquisition cost. *The one-sided Role term penalizes supported errors but never*

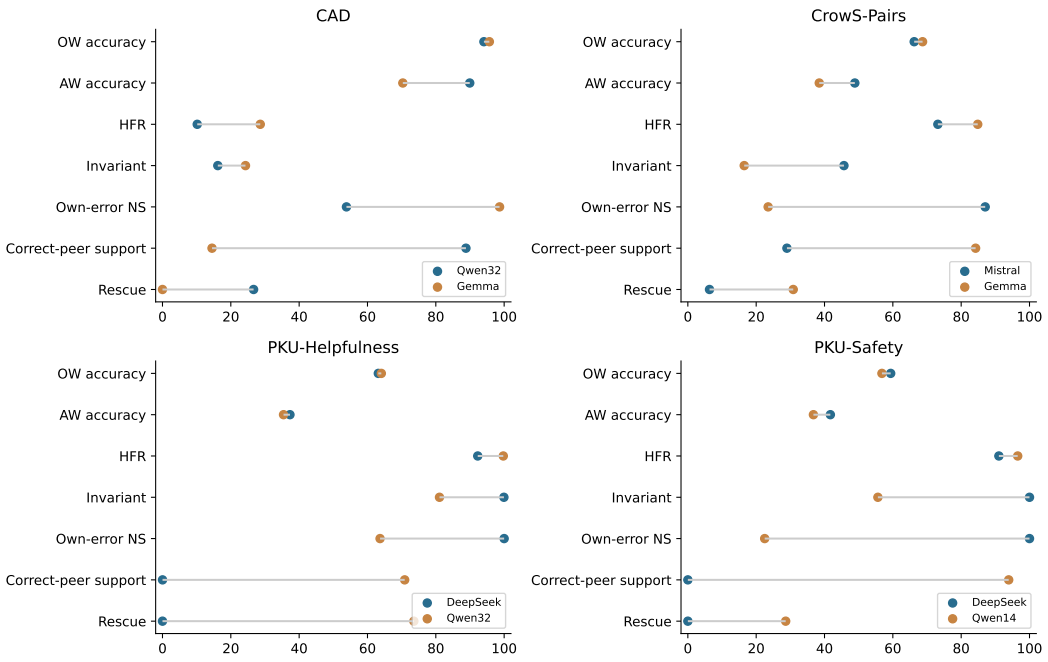


Figure A7: Largest diagnostic contrast among OW-matched profiles in the original four-task selection analysis; BBQ and ES-BBQ are not included. Each selected pair maximizes the mean gap across eleven components subject to an OW-accuracy gap at most three percentage points. These extremal contrasts are not model-quality rankings.

Table A53: Separate task-macro dimension orderings. Rank 1 means larger OW/AW accuracy, smaller HFR, larger self wrong NS, or larger peer BA, as applicable. Ties use unrounded values. These five rankings are not combined into a model-quality score.

Model	OW	AW	HFR	Wrong NS	Peer BA
DeepSeek	5	5	5	1	5
Mistral	3	4	2	2	4
Qwen14	4	1	1	5	3
Qwen32	1	2	4	4	1
Gemma	2	3	3	3	2

penalizes withheld support for correct judgments, so it can favor nearly constant neutrality. Excluding DeepSeek from the candidate set leaves 11.26 points of total gain and 1.91 points added by Role evidence. Excluding the CAD Gemma/Mistral candidates leaves approximately 15.53 points total and 3.41 added by Role. These sensitivities do not turn the objective into a measure of general accuracy.

The task-level breakdown in Table A55 (Appendix K) ranges from 0.07 to 26.2 percentage points over the twelve task/cost cells. CrowS gives the largest reduction under each stated cost. The two PKU balanced-world reductions are only 0.07 and 0.15 points, whereas the PKU-safety self-confirmation reduction is 22.41 points. *The gains are uneven and depend strongly on the cost definition.* A positive average does not imply improvement on every split.

K.5 Sensitivity To Hypothetical Label Costs

We do not measure annotation time, monetary cost, or human effort. We therefore do not claim to know a real cost-matched budget. Instead, Table A56 (Appendix K) declares hypothetical acquisition ratios. One OW label is one unit, a paired World label is two units, and World+Role evidence costs between two and four units per pair. A common nominal budget lets OW-only inspect about

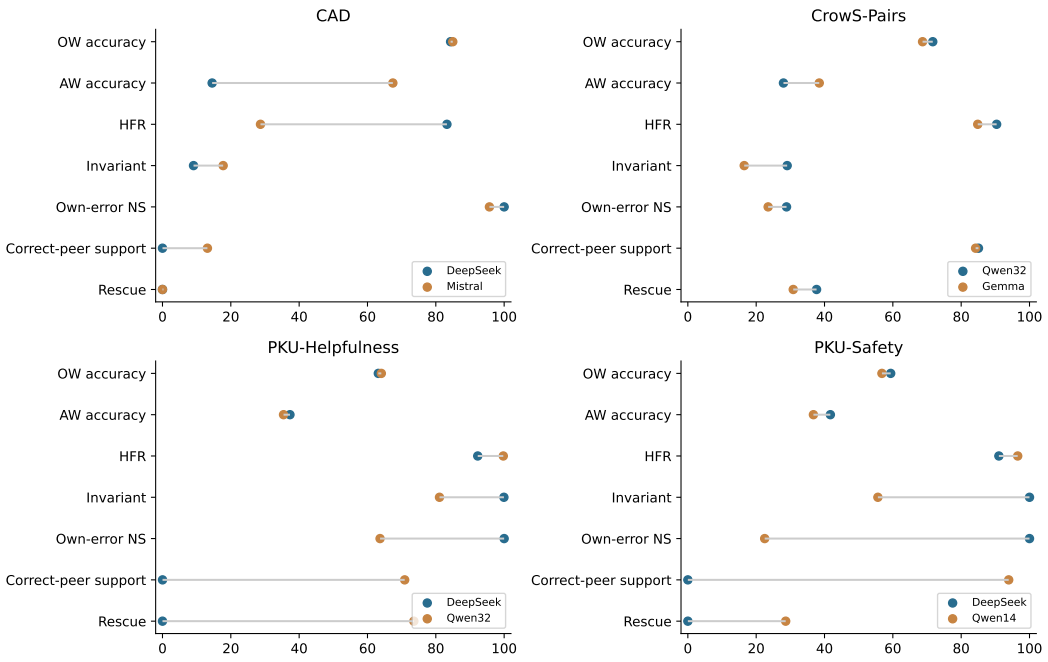


Figure A8: Fixed alternative model-pair examples evaluated with the same World and Role measures. These comparisons expose dependence on which pair is selected by the maximum-gap rule. The components have distinct meanings. Their separation is not a quality ranking.

Table A54: Fixed-output World/Role evidence decomposition on the same 10% audited identities in 500 splits per task. Costs are percentages and gains are percentage points. Each gain uses its own stated cost. The one-sided self-confirmation cost can favor blanket non-support.

Cost	OW-only	World-only	World+Role	World gain	Role gain
Balanced worlds	40.32	37.56	37.56	2.77	0.00
OW + hidden failure	38.62	35.93	35.93	2.69	0.00
Self-confirmation	35.56	23.45	20.04	12.11	3.40

20% of pairs, World-only inspect 10%, and World+Role inspect 5% to 10%. Every selector is scored on the same cases outside the largest audit set.

The selection benefit persists in these hypothetical comparisons even when the simpler audit receives more labeled cases. Under the declared acquisition ratios, richer evidence reduces held-out cost by 2.87 points for balanced worlds, 2.81 points for OW and hidden failure, and 16.47 to 16.5 points for self-confirmation aversion. This sensitivity gives OW-only more labeled cases under stipulated units. It does not measure real annotation time or money and retains the one-sided-objective limitation.

K.6 Sensitivity To Pair Orientation

The single-world selector and self-confirmation term depend on which world is the starting point. Table A57 (Appendix K) swaps OW and AW while keeping the same audited identities, split seed, and held-out cases. AW-only selection now uses AW correctness. World-only still uses both worlds. For reverse self-confirmation, World+Role also uses the observed decision in which the model verifies its own incorrect OW judgment.

The reversed hidden-failure cost gains 0.48 points from World evidence. Reversed self-confirmation aversion gains 7.47 points from World and 2.04 additional points from Role, totaling 9.51. *Reductions remain positive in this fixed same-task illustration, while their size and decomposition depend on orientation.* The forward and reversed hidden-failure targets are algebraically identical. The swap changes the single-world evidence available to the baseline.

Table A55: Within-task held-out cost reductions from richer diagnostic evidence on the same 10% audited identities. All values are percentage points, averaged over 500 fixed splits. Compare values only within their stated cost.

Task	Balanced worlds	OW and hidden failure	Self-confirmation averse
CAD	3.82	3.95	7.47
CrowS	7.04	4.95	26.20
PKU-H	0.07	0.68	5.98
PKU-S	0.15	1.17	22.41

Table A56: Sensitivity to hypothetical evidence-acquisition ratios. A fixed nominal budget allows different audit coverage for OW, World, and World+Role evidence. Costs are percentages. Richer gain is in percentage points. Inapplicable Role entries concern costs without a Role term. Ratios are stipulated, not measured annotation costs.

Cost	Ratio	Audit coverage (%)	OW-only	World	World+Role	Richer gain
Balanced worlds	1:2:n/a	19.9 / 10.0	40.42	37.55	n/a	2.87
OW + hidden failure	1:2:n/a	19.9 / 10.0	38.73	35.92	n/a	2.81
Self-confirmation	1:2:2	19.9 / 10.0 / 10.0	36.54	23.47	20.05	16.50
Self-confirmation	1:2:2.5	19.9 / 10.0 / 8.0	36.54	23.47	20.05	16.49
Self-confirmation	1:2:3	19.9 / 10.0 / 6.6	36.54	23.47	20.06	16.49
Self-confirmation	1:2:4	19.9 / 10.0 / 5.0	36.54	23.47	20.07	16.47

K.7 Audit-Budget Sensitivity

We repeat selection with 5%, 10%, 20%, and 30% audits in Table A58 (Appendix K) and Figure A9 (Appendix K). *All twelve budget-and-cost averages favor the richer audit. Individual splits can still favor the simpler audit.* The reductions range from 2.47 to 16.96 points. The self-confirmation gains are 14.23, 15.52, 16.68, and 16.96 points in budget order. They remain specific to its one-sided cost. Across rows, the selector loses on 0% to 11.2% of task-split units, so positive means do not establish a per-split guarantee.

Each budget uses a separate fixed split set. Connected points aid reading and are not nested learning curves. No uncertainty interval is plotted, so the figure is descriptive budget sensitivity. The self-confirmation curve and World-cost curves refer to different objectives. Their heights must not be compared as a universal benefit measure.

K.8 Compact Diagnostic Profiles By Task

In Table A59 (Appendix K), we collect selected World and Role measures by task and model. Its column groups answer different questions. Only the four same-case outcomes sum to 100% within a row. In Table A42 (Appendix J), we give their eligible counts and coverage.

On CrowS-Pairs, one favorable component is misleading. DeepSeek-R1-8B assigns non-support to all its own wrong judgments but supports no correct peer judgments. Across 14 comparisons over seven hidden errors, its eligible outcomes are 0% rescue and 100% support for neither judgment: blanket non-support, not Role discrimination.

Equal PKU-Safety OW accuracies provide another contrast. DeepSeek and Mistral both have 59.4% OW accuracy on the complete cohort. DeepSeek has 0% rescue and 100% support-neither, whereas Mistral has 13.3% rescue, 35.6% persistence, and 49.7% support-neither. *The shared OW score therefore does not determine the observed Role pattern.* These task-specific profiles preserve the components of behavior that an overall winner would conceal.

Table A57: Reversed-anchor selection on the same complete cohort. AW-only replaces OW-only. The Role term concerns self-support of an incorrect OW judgment. Costs are percentages and gains are percentage points under each stated objective.

Cost	AW-only	World	World+Role	World gain	Role gain
OW + hidden failure	36.42	35.93	35.93	0.48	0.00
Self-confirmation	29.22	21.75	19.71	7.47	2.04

Table A58: Audit-budget sensitivity on the original four-task, 8,699-pair subset. BBQ and ES-BBQ were not included in this secondary selection experiment. Each task and budget uses 500 fixed splits.

Audit %	Cost	OW-only	Diagnostic	Oracle	Gain	Win %	Tie %	Loss %
5	Balanced worlds	40.2	37.6	37.4	2.51	61.9	27.9	10.2
5	OW + hidden failure	38.5	36.0	35.8	2.47	76.8	16.2	6.9
5	Self-confirmation	34.3	20.1	20.0	14.23	85.0	15.0	0.0
10	Balanced worlds	40.3	37.6	37.4	2.77	66.8	22.8	10.5
10	OW + hidden failure	38.6	35.9	35.8	2.69	81.0	12.4	6.7
10	Self-confirmation	35.6	20.0	20.0	15.52	89.8	10.2	0.0
20	Balanced worlds	40.4	37.5	37.4	2.91	70.8	18.5	10.8
20	OW + hidden failure	38.8	35.9	35.8	2.88	84.9	8.0	7.0
20	Self-confirmation	36.7	20.0	20.0	16.68	94.4	5.6	0.0
30	Balanced worlds	40.4	37.5	37.4	2.94	73.0	15.9	11.2
30	OW + hidden failure	38.8	35.9	35.8	2.89	84.4	7.3	8.3
30	Self-confirmation	37.0	20.0	20.0	16.96	96.8	3.1	0.0

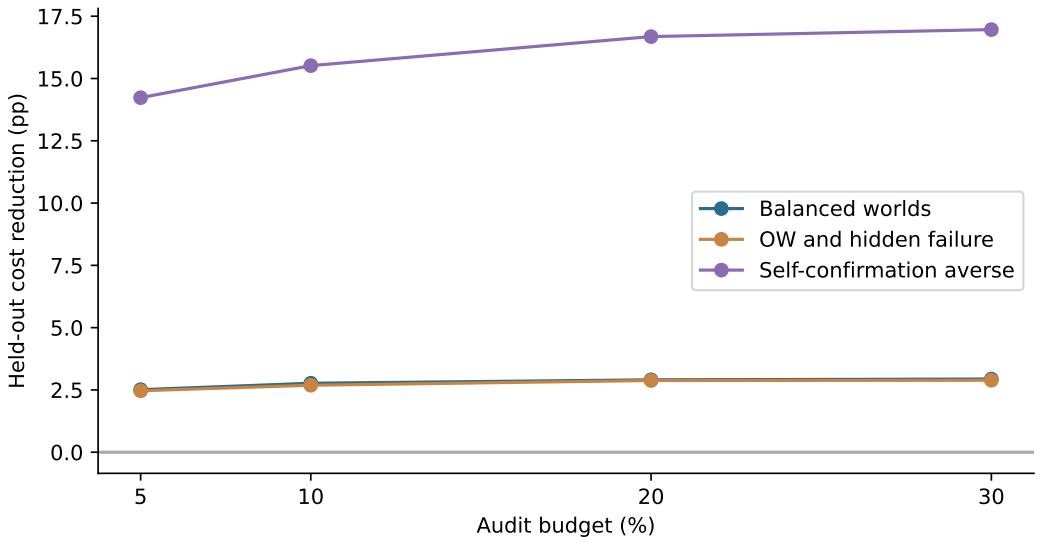


Figure A9: Held-out reduction in each stated cost on the original four-task selection subset; BBQ and ES-BBQ are not included. Each point averages 500 splits per task. Separate fixed splits mean the connecting lines are not nested learning curves.

Table A59: Task-specific diagnostic profiles on the 18,035-pair complete Role cohort across six tasks. Only rescue, persistence, both, and neither sum to 100% per row. NS includes exact-zero neutrality.

Task	Model	OW	AW	HFR	Invariant	Own wrong NS	Peer correct support	Rescue	Persistence	Both	Neither
CAD	DeepSeek	84.3	14.5	83.3	9.1	100.0	0.0	0.0	0.0	0.0	100.0
CAD	Mistral	85.0	67.4	28.7	17.8	95.7	13.2	0.0	4.6	0.0	95.4
CAD	Qwen14	99.3	77.0	22.9	21.4	27.7	98.6	33.6	16.9	48.6	1.0
CAD	Qwen32	94.1	89.9	10.2	16.2	53.8	88.8	26.7	43.3	0.0	30.0
CAD	Gemma	95.7	70.3	28.6	24.3	49.3	93.2	44.3	38.7	0.6	16.4
CrowS-Pairs	DeepSeek	8.8	42.6	88.7	9.3	100.0	0.0	0.0	0.0	0.0	100.0
CrowS-Pairs	Mistral	66.2	48.9	73.2	45.7	87.0	29.0	6.3	6.6	0.6	86.5
CrowS-Pairs	Qwen14	56.8	49.0	81.3	26.0	11.8	90.0	21.5	7.6	70.6	0.3
CrowS-Pairs	Qwen32	71.7	28.0	90.4	29.1	28.9	85.1	37.7	50.4	4.2	7.7
CrowS-Pairs	Gemma	68.7	38.5	84.8	16.5	23.5	84.2	30.8	46.6	10.8	11.8
PKU-Helpfulness	DeepSeek	63.2	37.3	92.3	99.9	100.0	0.0	0.0	0.0	0.0	100.0
PKU-Helpfulness	Mistral	57.0	30.4	93.8	51.3	84.5	21.4	54.7	23.0	0.0	22.3
PKU-Helpfulness	Qwen14	32.5	52.0	84.4	41.6	41.5	53.2	14.7	28.9	36.8	19.7
PKU-Helpfulness	Qwen32	64.1	35.4	99.8	81.1	63.7	70.9	73.6	23.5	1.5	1.4
PKU-Helpfulness	Gemma	64.5	35.9	92.4	97.9	87.2	41.5	65.2	9.6	1.7	23.5
PKU-Safety	DeepSeek	59.4	41.7	91.0	100.0	100.0	0.0	0.0	0.0	0.0	100.0
PKU-Safety	Mistral	59.4	39.0	94.4	65.4	39.6	44.1	13.3	35.6	1.4	49.7
PKU-Safety	Qwen14	56.8	36.7	96.5	55.6	22.5	93.9	28.6	20.3	46.4	4.6
PKU-Safety	Qwen32	62.1	38.1	95.4	62.4	17.6	57.6	9.5	72.3	6.8	11.3
PKU-Safety	Gemma	58.8	41.2	89.4	93.2	40.6	80.0	17.9	18.4	34.4	29.3
BBQ	DeepSeek	5.7	9.1	99.7	37.2	100.0	0.0	0.0	0.0	0.0	100.0
BBQ	Mistral	52.2	46.1	82.0	38.0	48.1	26.1	6.6	46.7	0.1	46.6
BBQ	Qwen14	44.7	50.0	87.8	28.6	33.8	86.9	20.0	28.2	48.4	3.3
BBQ	Qwen32	39.0	47.5	84.4	41.9	90.3	88.0	91.1	6.0	1.8	1.1
BBQ	Gemma	58.3	41.4	83.0	42.5	83.4	65.7	89.9	2.9	0.1	7.1
ES-BBQ	DeepSeek	2.0	3.9	100.0	10.0	100.0	0.0	0.0	0.0	0.0	100.0
ES-BBQ	Mistral	53.4	45.9	79.4	31.7	55.2	40.6	40.5	34.7	1.7	23.1
ES-BBQ	Qwen14	47.8	51.9	86.1	23.1	51.7	59.0	16.8	63.7	15.4	4.1
ES-BBQ	Qwen32	41.4	44.9	72.6	33.3	68.1	43.5	57.6	33.8	3.3	5.2
ES-BBQ	Gemma	56.5	37.9	90.9	32.4	64.4	38.5	67.7	21.1	2.7	8.4