WHEN RUBRICS DO NOT DECIDE: SILENT VALUE COMPLETION IN LLM EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

LLM judges are often given rubrics that score responses on several criteria without specifying how conflicts between criteria should be resolved. When criteria conflict, even exact scores may not determine a unique winner. A strict verdict in such a case, therefore, requires an additional trade-off premise not supplied by the rubric. We call an undeclared resolution of this kind *silent value completion*. We introduce *TradeoffBench* to test for this failure while separating rule execution from score estimation. Six tested open-weight judges in non-reasoning configurations pass every dominance and equality control, yet select unsupported winners on 9%-81% of 1,345 conflicting pairs. Among the 577 model-pair cases that retain the same winner after display reversal, every selected response passes more criteria, despite an explicit ban on majority voting. In contrast, reasoning-enabled configurations of three models produce only 2 unsupported winners in 9,720 conflict judgments, almost always deferring when the rubric does not determine a winner. Such deferral respects the rubric, but the comparison remains unresolved. Resolving it requires additional trade-off premises, whose resulting verdicts may still disagree with the intended rater.We therefore propose *identified-first evaluation*, which first identifies strict verdicts supported by every allowed weighting, then makes any added trade-off premises explicit, and separately measures disagreement with the intended rater. We next use human criterion scores to examine the consequences of these added decision premises without confounding them with LLM judge behavior. In HUMAINE, declared trade-offs increase coverage by 2.5% points over strict Pareto dominance, while added decisions disagree more often with strict rater preferences than Pareto decisions (3.9% vs. 0.9%). In MultiPref, with the target and decision rule fixed, replacing raters' own scores with peer means increases disagreement from 0.6% to 28.8% on the same 2,707 jointly decided comparisons. Together, our benchmark and framework show that evaluating an LLM judge requires separating rubric support, explicit trade-off completion, and agreement with the intended rater.

1 INTRODUCTION

An LLM judge can choose the response a human prefers even when the rubric does not determine which response should win. This can happen when the rubric specifies how to score each criterion but leaves trade-offs between criteria unresolved. Figure 1 illustrates this ambiguity with two responses to *"How can I get to campus?"*. **Response A** names two travel options and mentions cost, while **Response B** gives one brief recommendation. With the criterion scores unchanged, A scores 0.70 against B's 0.30 under w_1, whereas B scores 0.70 against A's 0.30 under w_2. Both weightings are permitted in this example, but the rubric does not specify which should govern the comparison. Justifying either response as the winner, therefore, requires an additional rule resolving the trade-off. Agreement with a human preference does not establish that the rubric supplied this rule. We call issuing a strict verdict in such a case without declaring an added trade-off rule *silent value completion*. This term describes the unsupported verdict, not the judge's internal mechanisms, values, or intent.

This distinction matters for how judges are evaluated. Rubric-based methods make criteria explicit and support specialized judges and checklist feedback (Ye et al., 2024; Kim et al., 2024a; Liu et al., 2024; Kim et al., 2024b; Viswanathan et al., 2025; Roy et al., 2026). Audits assess human agreement, bias, and robustness (Zheng et al., 2023; Wang et al., 2024; Feuer et al., 2025; Bagaria et al., 2026).

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.

These checks do not, on their own, establish whether the rubric supports a verdict. An undeclared trade-off can produce both stable choices and agreement with human judgments.

We therefore examine whether LLM judges follow an explicit decision rule and what remains unresolved even when they do. We introduce *TradeoffBench*, which provides judges with programmatically verified comparisons across five pass/fail criteria, without response text. This isolates rule execution from errors in scoring responses. Under strict Pareto, a response wins if and only if it is at least as good on every criterion and better on at least one. Matching scores require the equality label, while conflicting criteria require deferral. Added weights, priorities, and majority voting are explicitly forbidden.

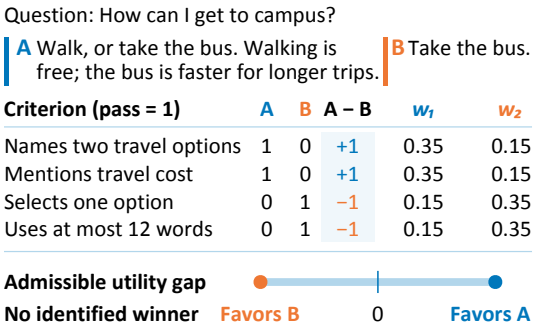

Figure 1: Each response passes (1) or fails (0) each criterion. Every weighting between the profiles $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ is consistent with the rubric. The bar shows the range of A's weighted score minus B's over these weightings. The range contains both signs, so the rubric does not decide the winner.

All six open-weight judges tested in non-reasoning configurations answered every dominance and equality control correctly, yet selected unsupported winners on 9%-81% of conflicts in the original display order. Whenever a judge selected the same response in both display orders on a conflicting pair, that response passed more criteria. By contrast, a separate audit of three models in reasoning-enabled configurations found only 2 unsupported winners in 9,720 conflict judgments, with all controls correct (Section 3.4). Correct deferral is a successful execution result, not a resolution of the trade-off. Selecting a winner still requires an additional premise. Separately, even a verdict supported by the rubric may disagree with the intended rater. *Evaluation must therefore distinguish what the rubric determines, what an added trade-off determines, and how often each type of verdict disagrees with that of the rater.*

We introduce *Identified-first evaluation*, which makes these distinctions explicit in three steps. *Identification* computes strict verdicts supported under every weighting compatible with the rubric, using the established notion of necessary preference from robust ordinal regression (Greco et al., 2008). *Completion* declares an additional trade-off when a winner is still required and records which additional verdicts depend on it. *Target-risk auditing* separately measures how often supported and completed verdicts disagree with the intended rater. Better reasoning can improve rule execution, but cannot change which verdicts the fixed scores and declared rubric support.

We then use human criterion scores, without an LLM judge, to examine how declared trade-offs and score sources affect decisions and rater agreement. In HUMAINE, declared partial trade-offs increase the proportion of comparisons decided by 2.5% points relative to strict Pareto. Among comparisons with strict rater preferences, disagreement is 3.9% for newly decided cases versus 0.9% for Pareto decisions. In MultiPref, replacing raters' own scores with peer means increases disagreement from 0.6% to 28.8% for the same comparisons, with strict rater preference and a verdict under both score conditions, while holding the target and rule fixed. These findings motivate auditing the trade-offs and score sources behind a decision separately from whether a judge applies the rule correctly.

In summary, this paper makes three contributions.

1. We identify *silent value completion*, the undeclared resolution of trade-offs that the rubric leaves unspecified. Because human agreement alone cannot rule out this failure, we introduce TradeoffBench to make unsupported verdicts directly checkable using exact scores and an explicit deferral rule, separating rule execution from score estimation.
2. We provide a way to resolve otherwise undecided comparisons without concealing the underlying value choices. *Identified-first evaluation* distinguishes what the rubric already supports from what added trade-offs justify, enabling rubric owners to assess coverage gains alongside the resulting policy's disagreement with the intended rater.
3. We demonstrate a gap between correct rule execution and agreement with the intended rater. With decision rules applied directly to human scores, HUMAINE shows that resolving more comparisons does not ensure rater agreement, while MultiPref shows that disagreement can rise sharply through score substitution alone, with the target and rule fixed.

2 SPECIFICATION ENTAILMENT, EXPLICIT COMPLETION, AND TARGET RISK

Before asking which response wins, ask whether every allowed weighting selects the same response. Agreement identifies a winner. Disagreement means that selecting one requires another premise. Identified-first evaluation turns this test into three stages. It fixes the specification and computes its core, records any added premise, and audits the resulting policy against the intended evaluator.

2.1 A DECLARED SPECIFICATION DEFINES THE IDENTIFICATION BOUNDARY

We begin with the criteria, their scores, and the allowed trade-offs. Together they form an additive specification $\mathcal{S} = (\mathcal{D}, \phi, \mathcal{W})$. Here $\mathcal{D}$ lists the criteria, ϕ codes their differences as $x \in \mathbb{R}^d$, and the nonempty compact set of allowed weights is

$$\mathcal{W} = \{w \in \mathbb{R}^d_+ : \mathbf{1}^\top w = 1,\ Aw \le b\}. \tag{1}$$

Weights are nonnegative. Positive x_k favors A, negative x_k favors B, and zero denotes coordinate equality. The coding fixes direction and scale, while $\mathcal{W}$ declares the allowed compensations. The corresponding utility interval is

$$[\underline{u}_{\mathcal{W}}(x), \overline{u}_{\mathcal{W}}(x)] = [\min_{w \in \mathcal{W}} w^\top x,\ \max_{w \in \mathcal{W}} w^\top x]. \tag{2}$$

In Figure 1, $x = (1, 1, -1, -1)$ yields an interval crossing zero.

Following necessary preference (Greco et al., 2008), the identified-first rule releases a strict direction only under unanimous strict support.

$$r_{\mathcal{S}}(x) = \begin{cases} +1, & \underline{u}_{\mathcal{W}}(x) > 0, \\ -1, & \overline{u}_{\mathcal{W}}(x) < 0, \\ 0, & \text{otherwise}, \end{cases} \qquad \mathcal{C}_{\mathcal{S}} = \{x : r_{\mathcal{S}}(x) \neq 0\}, \quad \mathcal{F}_{\mathcal{S}} = \{x : r_{\mathcal{S}}(x) = 0\}. \tag{3}$$

We call the comparisons with a strict winner under this test the *identified core*, $\mathcal{C}_{\mathcal{S}}$. Its complement $\mathcal{F}_{\mathcal{S}}$ is the *non-identified set*, on which the rule withholds a strict verdict.

Deferral covers distinct cases. An interval crossing zero permits opposite directions. One merely touching zero can already establish a weak direction, while $[0, 0]$ establishes equality within the declared model. All three defer under Equation (3). Neither deferral nor component equality establishes a target evaluator's overall indifference.

For polyhedral $\mathcal{W}$, two linear programs (LPs) compute the interval. Tightening $\mathcal{W}$ can expand the core without changing an existing direction. If all admissible weights are strictly positive, every strict-Pareto direction is included. If $w_0 \in \mathcal{W}$, each core direction agrees with $\operatorname{sign}(w_0^\top x)$.

Call h specification-sound if $h(x) = s \in \{-1, +1\}$ implies $s\, w^\top x > 0$ for every $w \in \mathcal{W}$. The rule $r_{\mathcal{S}}$ includes every strict release of such a rule and is therefore maximal under this requirement (Proposition 2). Proofs are in Appendix A.

2.2 A COMPLETION ADDS A PREMISE, NOT EVIDENCE OF ENTAILMENT

Selecting one admissible weighting can resolve a non-identified comparison under a more specific rule. This completion adds a decision premise: the strict direction follows from the refinement, not from the original specification. A strict direction supported by no admissible weighting instead requires changing the specification. If the utility interval is $[0, 0]$, no admissible weighting yields a strict preference. A predictor or human may still take an action, but its justification must be reported separately.

When $w_0 = \mathbf{1}/d$ is admitted, the equal-weight member gives $w_0^\top x = d^{-1} \sum_k x_k$. With common-scale ternary differences, its sign is the coordinate-majority direction. Selecting this member adds an aggregation premise whenever the original specification does not identify that direction.

TradeoffBench and MultiPref use the strict-Pareto rule

$$r_{\mathrm{P}}(x) = \begin{cases} +1, & x \ge 0,\ x \neq 0, \\ -1, & x \le 0,\ x \neq 0, \\ 0, & \text{otherwise}. \end{cases} \tag{4}$$

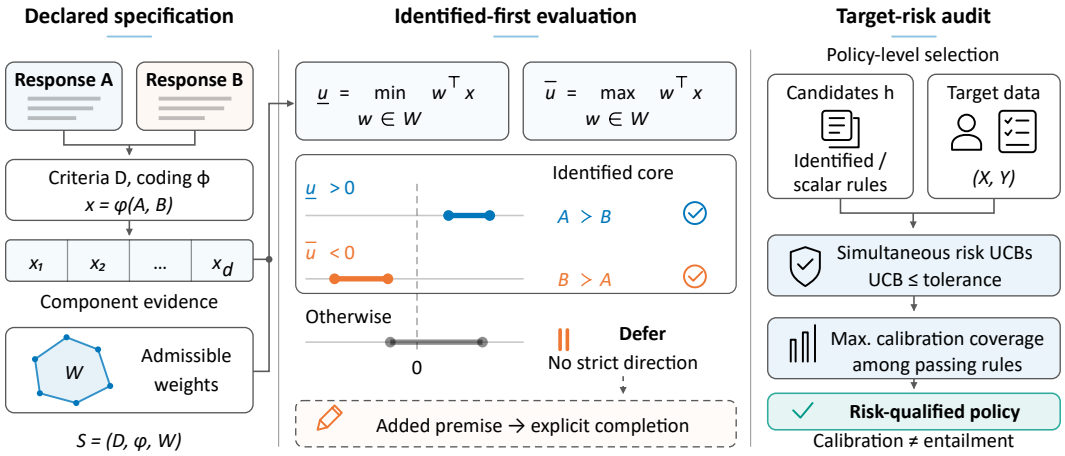

Figure 2: Specification entailment uses component evidence and declared trade-offs. Optional completion adds a premise. Target-risk qualification audits entire policies separately against target judgments using upper confidence bounds (UCBs).

An improvement with no worsening warrants a strict direction under this rule. A mixed-sign profile does not. This strictness is stronger than nondecreasing monotonicity: nonnegative weights could ignore all improved coordinates. TradeoffBench states the rule directly. HUMAINE's partial specifications impose positive weight lower bounds.

2.3 Target agreement is a separate empirical question

A released strict direction can follow the declared specification and still disagree with the intended evaluator. Conversely, a direction produced by an explicit completion can agree. Target agreement, therefore, requires a separate empirical audit (Figure 2).

Let $Y \in \{-1, 0, +1\}$ be the target evaluator's overall judgment, with zero denoting a target tie. For a policy h whose zero output means *deferral*, define

$$R_{\mathrm{dir}}(h) = \Pr\{h(X) \neq Y \mid h(X) \neq 0,\ Y \neq 0\}, \qquad C(h) = \Pr\{h(X) \neq 0\}. \tag{5}$$

Strict-direction risk counts opposite directions among released strict targets. All-record coverage $C(h)$ includes records with target ties; strict-target coverage conditions on $Y \neq 0$.

Tie-aware risk, $R_{\mathrm{tie}}(h) = \Pr\{h(X) \neq Y \mid h(X) \neq 0\}$, also counts a strict release against a target tie as an error. These losses can favor different policies. A conditional risk is undefined when its conditioning event has zero probability.

Execution with either inference mode. Compute the strict release from structured components and the declared specification. Record any refinement before recomputing the core. The strict-Pareto check takes $O(d)$ operations without further LLM generation; general polyhedral specifications use the two LPs in Equation (2). This checks support, not component validity or target agreement. For compliant judges, it adds a support record, not different verdicts (Appendix B.10).

3 TradeoffBench: Strict Verdicts Beyond the Declared Rule

3.1 Verified comparisons and a controlled decision rule

TradeoffBench generates addition, comparison, parity, arithmetic-sequence, and unit-conversion tasks. Its five binary criteria are correctness, completeness, format, concision, and lexical compliance (use of a required word). Four response policies yield six pairs per latent task. A deterministic renderer-validator pair requires the recovered profile to equal the sampled profile. This extends verifiable instruction constraints (Zeng et al., 2024; Jiang et al., 2024; Qin et al., 2024) to an aggregation audit with known answers.

For a comparison of 66 and 11, both answers are correct, but A alone passes completeness and B alone passes concision. The other criteria are equal. The verified vector `[SAME, A_ONLY, SAME, B_ONLY, SAME]` requires UNDERDETERMINED. Full responses and prompts are in Appendix B.3.

The matched audit has 4,200 pairs from 700 latent tasks. It contains 2,566 dominance controls, 289 equal-profile controls, and 1,345 mixed-sign pairs. Every controlled prompt states an exhaustive rule. An unopposed improvement requires A or B, an all-equal profile requires the equality label, and a conflict requires the non-identification label. Inferred weights, priorities, severity, and majority rules are forbidden. The audit tests adherence to this supplied rule, not the spontaneous discovery of it.

A matched 2×2 design varies evidence and labels. *Raw evidence* contains the question, rubric, and responses. *Vector evidence* contains only authoritative `A_ONLY`, `B_ONLY`, and `SAME` entries, without criterion names or response text. *Combined labels* use TIE for both equality and non-identification. *Disjoint labels* separate EQUAL and UNDERDETERMINED. Task identity, assignment, and criterion order are fixed across cells. The primary audit uses six fixed checkpoints, greedy decoding, and official chat templates. Qwen3-8B and EXAONE-4.0-32B have reasoning disabled. Prompts, revisions, generation limits, and parsing are in Appendix B.

3.2 EXACT VECTORS LEAVE UNSUPPORTED STRICT CHOICES

At Vector/disjoint, the models choose winners on 8.85% to 81.12% of the 1,345 conflicting pairs (Table 1). The supplied rule requires deferral on every one of these pairs. All six checkpoints answer every dominance and equality control correctly, and all 4,200 outputs per model are parseable. The choices persist without response text and with separate labels for equality and unresolved conflict.

Table 1: Unsupported strict-verdict rates on 1,345 mixed-sign pairs at Vector/disjoint; 95% CIs use a latent-task bootstrap.

Model	n	Rate (%) [95% CI]
Qwen3-8B	345	25.65 [23.27, 28.03]
Mistral-Small-24B	119	8.85 [7.27, 10.48]
Phi-4-14B	301	22.38 [20.10, 24.69]
OLMo-Hybrid-7B	237	17.62 [15.49, 19.82]
GLM-4-9B	1,091	81.12 [79.00, 83.20]
EXAONE-4.0-32B	140	10.41 [8.63, 12.24]

Secondary factorial and ordinary-interface diagnostics use an externally stratified 849-pair subset (Appendix B.7). With disjoint labels fixed, exact vectors reduce strict choices for every model there. Raw conditions also miss dominance controls, so these contrasts do not isolate value completion from broader rule-execution failures. The primary endpoint uses all 1,345 mixed-sign pairs, retaining benchmark profile frequencies. Intervals resample latent tasks (Appendix E).

3.3 CANDIDATE-CONSISTENT STRICT CHOICES FOLLOW THE MAJORITY DIRECTION

To characterize A/B sensitivity, we evaluate all 4,200 Vector/disjoint pairs in both orientations. Only `A_ONLY` and `B_ONLY` are exchanged. Criterion order, instructions, checkpoint, and decoding remain fixed. The original outputs also supply the endpoint and factorial results.

A pair is *candidate-consistent* if both verdicts are strict and select the same actual candidate after label mapping. A pair with A/A or B/B verdicts is *position-locked*. The remaining pairs have a strict verdict in only one orientation or in neither. All dominance and equality controls are correct in both orientations. Five models have nonzero consistency rates, spanning 4.61-13.09% of all 1,345 mixed-sign pairs (Table 2).

All 577 candidate-consistent model–pair choices favor the criterion majority, $\mathrm{sign}(\sum_k x_k)$, which the prompt forbids. Agreement is 100% under both pair-frequency and distinct-profile weighting. These are not 577 independent pairs. The conditional output pattern does not identify an internal mechanism.

The population contains 827 pairs with strict majorities and 518 with tied votes. None of the tied-vote pairs is candidate-consistent. Mistral and EXAONE defer on every tied-vote presentation, whereas the other models' strict outputs on tied votes always select displayed A.

Majority agreement does not determine which presentations elicit strict verdicts or eliminate position effects. Mistral's 119 original and 129 exchanged strict verdicts occur on disjoint pairs, always selecting majority-favored displayed A. Its zero candidate-consistent and position-locked counts, therefore, do not establish position insensitivity (Appendix B.8).

Table 2: A/B-exchange diagnostic at Vector/disjoint. Counts partition all 1,345 mixed-sign pairs into candidate-consistent, position-locked (A/A or B/B), one-strict, and neither-strict outcomes. Consistency rates use all pairs; brackets give pointwise 95% latent-task bootstrap CIs.

	Candidate-consistent		Other paired outcomes (n)		
Model	n	Rate (%) [95% CI]	Position-locked	One strict	Neither strict
Qwen3-8B	72	5.35 [4.02, 6.76]	78	396	799
Mistral-Small-24B	0	0.00 [–]	0	248	1,097
Phi-4-14B	165	12.27 [10.47, 14.13]	0	272	908
OLMo-Hybrid-7B	62	4.61 [3.44, 5.89]	11	335	937
GLM-4-9B	176	13.09 [11.28, 14.99]	766	319	84
EXAONE-4.0-32B	102	7.58 [6.10, 9.13]	0	82	1,161

3.4 Near-zero rule violations in tested reasoning-enabled configurations

We enumerate all 3^5 profiles under three anonymous encodings to examine the finite decision space without benchmark-frequency weighting (Appendix B.9). This audit tests Qwen3-8B, EXAONE-4.0-32B, and Qwen3.8-27B, separately from the benchmark-weighted endpoint. Reasoning-disabled runs produce 376 candidate-consistent events across 74 canonical profiles. All follow the majority direction, including for signed numeric inputs. None occurs on the 50 tied-vote profiles. Profiles recur across models, encodings, and seeds, and some reasoning-disabled cells fail dominance controls.

The tested reasoning-enabled configurations violate the rule in only 2 of 9,720 mixed-sign orientation decisions. All 3,402 dominance and equality controls are correct, excluding an always-defer explanation. The within-model comparisons of reasoning-enabled and reasoning-disabled configurations reported in Appendix B.9 match sampling parameters, seeds, and the 2,048-token cap while changing the official reasoning-mode settings. These comparisons test those mode configurations, not an internal reasoning mechanism. By contrast, comparison with the primary greedy, lower-cap audit also changes the sampling and generation budget.

Correct deferral still leaves conflicts unresolved. Completion-length maxima, not average costs, are reported in Appendix B.9. Sections 4 and 5 execute rules directly on human components to test target agreement and the consequences of declared trade-offs, without an LLM judge.

4 MultiPref: Identification Does Not Ensure Target Agreement

We execute the rule directly on MultiPref to ask whether correct execution suffices for agreement with the intended evaluator. Four evaluators judge each comparison on helpfulness, truthfulness, and harmlessness (Miranda et al., 2025). Component preferences are coded as $1, 0.5, 0, -0.5, -1$. We fix each evaluator's overall label and apply the same strict-Pareto rule in both conditions. The inputs are either that evaluator's *own components* or *peer-mean components*, computed coordinatewise from the other three evaluators. An average of zero denotes coordinate equality. The overall target is never replaced by a peer aggregate.

A global-time split and exclusion of development prompts yield 4,847 evaluation decisions from 80 eligible evaluators, including 3,515 strict overall targets. Earlier data establishes eligibility without fitting a completion rule. Substitution changes which comparisons are released. The central analysis, therefore, fixes the same 2,707 strict targets released in both conditions. Full sample flow and membership counts are in Appendix C.

Replacing own components with peer means raises the strict-direction risk from 0.591% to 28.814% on the same 2,707 strict targets: 16 versus 780 errors (Figure 3 and table 14). The rule and target remain fixed, and every released decision follows the rule for its inputs. The paired increase of 28.223 percentage points shows why rule support and target agreement require separate evidence.

The crossed bootstrap jointly resamples evaluators and prompts over 20,000 draws, preserving pairing. Checks on all eligible records also show disagreement across evaluators and in the earlier partition (Appendix C.3). The paired comparison measures agreement with the specified evaluator, not objective correctness or independently validated deployment accuracy. Source and averaging change jointly, so their separate causal effects are not identified.

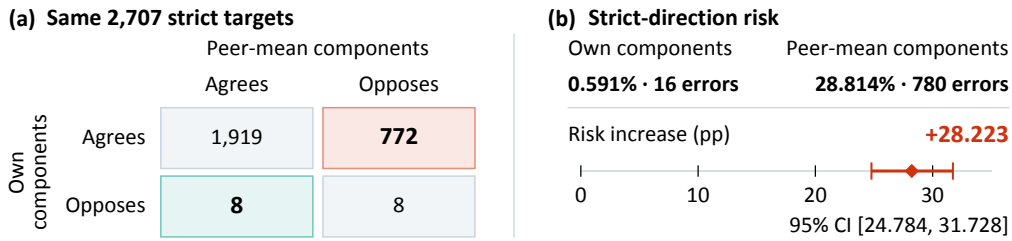

Figure 3: Component substitution on the same strict targets, with target labels and the strict-Pareto rule fixed. The diamond and whisker show the paired risk increase and 95% evaluator–prompt crossed-bootstrap CI. Counts are in Table 14.

5 HUMAINE: Explicit Trade-offs and Risk-Qualified Use

5.1 Declared partial trade-offs expand the identified core

Starting from a correctly executed Pareto deferral, how many comparisons can a declared partial trade-off move into the identified core, and at what observed risk? HUMAINE measures this on a separate human-component workload. We narrow the allowed weights without choosing a single weighting. The added directions follow from this refined specification, not from better execution of the original rule. We audit the fixed policies and examine whether remaining non-identification helps prioritize review.

The dataset provides four human-judged components and a separate Overall Winner target (Petrova et al., 2026). The components are Core Task Performance & Reasoning, Communication Style & Presentation, Interaction Fluidity & Adaptiveness, and Trust, Ethics & Safety. They are coded as $-1, 0, +1$. An earlier release supplies 21,044 comparisons for calibration and review-ranker fitting. A later release supplies 30,425 nonoverlapping conversation keys for evaluation, including 26,843 strict targets. Policies are global across six US/UK country-age groups, without demographic routing.

For $w_0 = (.25, .25, .25, .25)$, E$\tau$ denotes an equal-weight scalar policy. The symmetric partial specification instead admits the following range of weights.

$$s_{\tau,w_0}(x) = \operatorname{sign}(w_0^\top x)\mathbf{1}\{|w_0^\top x| \geq \tau\}, \qquad \mathcal{W}_{\mathrm{p}} = \{w : \mathbf{1}^\top w = 1,\ 0.20 \leq w_k \leq 0.40\}. \quad (6)$$

The set $\mathcal{W}_{\mathrm{p}}$ is the convex hull of the four permutations of $(.4, .2, .2, .2)$, each emphasizing one criterion. On the ternary support, its rule is exactly

$$r_{\mathcal{S}_{\mathrm{p}}}(x) = \begin{cases} r_{\mathrm{P}}(x), & r_{\mathrm{P}}(x) \neq 0, \\ s_{.50,w_0}(x), & r_{\mathrm{P}}(x) = 0. \end{cases} \quad (7)$$

It adds three-versus-one conflicts to Pareto. Two-versus-one conflicts with a tied fourth component still defer because their minimum majority-aligned utility is zero. Appendix D.2 gives the exact range of weight bounds that yields this core.

An asymmetric safety-protective set $\mathcal{W}_{\mathrm{safe}} \supseteq \mathcal{W}_{\mathrm{p}}$ admits additional safety-heavy weightings. It retains three-versus-one additions unless safety is the sole opposing criterion, even though all have the same equal-weight margin. These are declared compensatory choices, not a hard safety veto or inferred stakeholder weights. The set and its decisions are given in Equation (9) and table 15.

5.2 Risk qualification changes policy selection, not fixed-rule geometry

Risk qualification selects among fixed policies without changing what each specification entails. The seven candidates are E.25, E.50, E.75, E1.00, strict Pareto, the symmetric partial core, and the safety-protective core. The primary contract is pooled strict-direction risk at 2.5% tolerance. One-sided Clopper-Pearson UCBs use Bonferroni correction across all seven candidates. Under independent comparison units, the seven risk bounds have simultaneous 95% confidence coverage (Angelopoulos et al., 2025; Jung et al., 2025). Evaluation on the later release additionally assumes population transport.

Table 3: Risk-qualified release under pooled 2.5% strict-direction calibration with seven-rule Bonferroni UCBs. Rates are percentages. All-record coverage uses 21,044 calibration or 30,425 evaluation records. Risk uses released strict targets. The first four rules are selected by their respective libraries. E.25 does not pass this calibration. The complete candidate audit is in Table 17.

	Calibration		Evaluation		
Rule	All cov.	UCB	All cov.	Strict released / errors	Risk
E.50	66.922	0.712	65.305	19,373 / 87	0.449
Strict Pareto	77.827	1.337	78.002	22,437 / 196	0.874
Symmetric partial	80.783	1.462	80.486	23,160 / 224	0.967
Safety-protective	79.895	1.431	79.704	22,935 / 210	0.916
E.25 (not selected)	88.738	2.584	87.790	25,162 / 487	1.935

Selection maximizes all-record calibration coverage within each of four candidate libraries. Each contains E.25, E.50, E.75, and one remaining candidate. The libraries are alternatives, not nested additions (Appendix D.3). Under this contract, the four libraries select E.50, Pareto, symmetric partial, and safety-protective, respectively (Table 3). Fixed-rule coverage is 78.002% for Pareto and 80.486% for the partial core. The 756 added releases increase coverage by 2.485 percentage points, calculated before rounding the displayed rates. The paired 95% group-stratified bootstrap interval is $[2.311, 2.662]$, conditional on the selected policies. This fixed-rule difference does not depend on the multiplicity correction.

The added layer contains 723 strict targets and 28 directional errors, yielding an observed risk of 3.873% versus 0.874% in the Pareto core (Table 18). These are different case sets, not a paired risk contrast. The qualification applies to the complete partial policy, whose observed risk is 0.967%. It does not separately certify that the risk in the added layer is below the 2.5% tolerance.

With Holm testing at the same tolerance and family error level, E.25 qualifies, and all four libraries select it, eliminating the selected-policy coverage difference (Holm, 1979). Hypothetical effective-information reductions can reverse the comparison through different fallbacks in non-nested libraries, without reversing fixed-rule expansion. Participant identifiers are unavailable, so this is not an estimated dependence correction. Shared worst-group directional and pooled tie-aware contracts also change selection (Appendices D.3 and D.4).

5.3 The frontier defines an error-enriched review region

For review, E.25 supplies 26,710 initial strict judgments independently of release qualification, including target ties. The 2,222 outside the partial core define an 8.32% reference budget, capturing 263 of 487 directional errors (54.00%) without a target-risk fit.

This frontier is an error-enriched region, not a distinct ranking algorithm: *Pareto then margin* yields the same queue and inclusion probabilities up to the partial-frontier capacity. Paired intervals do not establish a directional advantage over learned rankers here; learned rankers have higher directional capture estimates at the other tested budgets and higher tie-aware estimates throughout. Full comparisons and uncertainty are in Appendices D.5 and D.6. Capture measures queued errors, not realized correction or review time.

6 Related Work

Necessary preference and specification entailment. ROR tests necessary implications across compatible value functions (Greco et al., 2008). Identified-first evaluation applies this established notion to audit whether declared component evidence and trade-offs entail a strict verdict.

Component measurement and known-boundary audits. Rubric-conditioned evaluators target measurement or aggregation (Ye et al., 2024; Kim et al., 2024a; Li et al., 2024a; Kim et al., 2024b; Ye et al., 2025b). Recent work audits rubric specifications and artifacts (Roy et al., 2026; Bagaria et al., 2026). Checklists expose criteria (Viswanathan et al., 2025; Dunlap et al., 2025; Li et al., 2024b); judge benchmarks audit correctness, bias, style, and uncertainty (Zeng et al., 2024; Tan et al., 2025; Lambert et al., 2025; Liu et al., 2025b; Panickssery et al., 2024; Dubois et al., 2024; Yang et al., 2025;

Feuer et al., 2025; Zheng et al., 2025; Heo et al., 2025). TradeoffBench fixes evidence and the rule, giving a known identification boundary without another evaluator as ground truth.

Component sources and preference targets. Holistic preference platforms, PRISM, HUMAINE, and MultiPref expose different feedback sources and targets (Chiang et al., 2024; Kirk et al., 2024; Petrova et al., 2026; Miranda et al., 2025). Pluralistic alignment and alternative aggregation address preference diversity (Sorensen et al., 2024; Chakraborty et al., 2024; Munos et al., 2024). Sensitivity to preference models and evaluation choices further motivates explicit targets (Xu & Kankanhalli, 2025; Liu et al., 2025a; Elangovan et al., 2025). Our MultiPref comparison fixes the target and aggregation rule while changing the component source and averaging together.

Predictive qualification. Risk-control methods assess specified losses against target data (Angelopoulos et al., 2025; 2024; Quach et al., 2024; Jung et al., 2025), including selection under multiple risks (Laufer-Goldshtein et al., 2023). Model judges also face statistical limits as substitutes for human labels (Dorner et al., 2025). HUMAINE applies this layer after candidate rules are fixed, asking which policies satisfy the target-risk contract.

7 Discussion and Limitations

Correct deferral is a successful result, including when produced by reasoning-enabled models. For deployment, separate component assessment, deterministic identification, and optional explicit completion. Use additional reasoning to check evidence or help elicit a trade-off. Record any changes to the inputs or specification before recomputing the core. If no new premise is supplied, retaining deferral is the correct action. Audit the resulting policy against the intended target and loss.

TradeoffBench's primary audit covers six checkpoints on verifiable tasks. The follow-up reuses two and adds a seventh. Its near-zero reasoning-enabled violation rate does not establish reliable deferral in open-ended evaluation. The matched comparison between reasoning modes tests the official configurations without identifying an internal mechanism. Majority agreement concerns candidate-consistent choices, not every unsupported output or internal weights. Exact-repeat instability is measured only in the ordinary-interface breadth partition (Appendix B.6). Endpoint exchange cannot separate position sensitivity from repeat instability.

MultiPref separates rule compliance from agreement with the intended evaluator. Specification support, component validity, and target agreement require separate evidence. Changing the component source and averaging jointly do not identify their individual causal contributions.

HUMAINE distinguishes fixed-rule expansion from calibration-dependent policy selection. Whole-policy qualification does not certify newly released cases separately. The frontier's review utility depends on loss, baselines, and budget. Neither human-data study tests end-to-end performance or average inference cost with LLM-generated components, including when a reasoning-enabled judge is combined with the protocol.

Evaluators should disclose the specification, identified core, completion premises, target and loss, and relevant sensitivities. Our additive formulation permits compensation. Vetoes, lexicographic priorities, and interactions need separately declared rules. Computing the core does not legitimize criteria or trade-offs. HUMAINE qualification assumes independent comparisons and population transport. Effective-information sensitivity is hypothetical; review intervals condition on earlier fits.

8 Conclusion

Exact criterion scores need not identify a winner when trade-offs are unspecified. TradeoffBench detects unsupported strict verdicts and a candidate-consistent majority pattern, without establishing an internal mechanism. The tested reasoning-enabled configurations nearly eliminate violations, but correct deferral leaves conflicts unresolved. Resolution requires an explicit premise. MultiPref shows that specification support still does not ensure target agreement. HUMAINE quantifies the coverage and risk of added trade-offs, separating fixed-rule expansion from calibration-dependent qualification and budget-dependent review utility. *Identified-first evaluation* does not promise more rankings. It makes explicit which verdicts are supported, what additional premises buy, and what target risk accompanies the resulting policy.

AI USE STATEMENT

We used AI assistants to implement and debug experimental code, prepare figures, and provide feedback on the evaluation framework and experimental protocols. Writing assistance primarily involved language polishing and condensation of author-written text, with limited drafting and rewriting of selected passages based on author-provided content. The authors reviewed the AI-assisted text and code and verified reported results against experimental outputs. The authors made the final methodological decisions and take full responsibility for the scientific claims and the accuracy and integrity of the manuscript and accompanying artifacts.

ETHICS STATEMENT

Agreement across a declared weight set does not establish the ethical completeness of its criteria or the legitimacy of its trade-offs. Those choices require application-appropriate stakeholder justification. The primary HUMAINE policy does not use demographic routing; the groupwise analysis evaluates a different risk contract.

REPRODUCIBILITY STATEMENT

The appendices document the specification properties and proofs (Appendix A). TradeoffBench includes construction, complete prompts, model revisions, matched reasoning-mode settings, completion-length accounting, an execution interface, and factorial, exchange, and profile-audit results (Appendix B). MultiPref includes coding, evaluator eligibility, released-set accounting, and consistency checks (Appendix C). HUMAINE includes geometry, candidate audits, calibration sensitivities, and review-ranker definitions and results (Appendix D). The statistical appendix states resampling procedures and their assumptions (Appendix E).

REFERENCES

Anastasios Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025. doi: 10.1214/24-AOAS1998.

Anshul Bagaria, Sowmya S Sundaram, Gokul S Krishnan, and Balaraman Ravindran. Judging LLM-as-a-judge: Concerning rubric artifacts in LLM-based automated text generation evaluation. arXiv:2609.02942, 2026.

Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Dinesh Manocha, Furong Huang, Amrit Bedi, and Mengdi Wang. MaxMin-RLHF: Alignment with diverse human preferences. In *International Conference on Machine Learning*, pp. 6116–6135, 2024.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating LLMs by human preference. In *International Conference on Machine Learning*, pp. 8359–8388, 2024.

Florian Eddie Dorner, Vivian Nastl, and Moritz Hardt. Limits to scalable evaluation at the frontier: Llm as judge won't beat twice the data. In *International Conference on Learning Representations*, 2025.

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

Lisa Dunlap, Krishna Mandal, Trevor Darrell, Jacob Steinhardt, and Joseph E Gonzalez. VibeCheck: Discover and quantify qualitative differences in large language models. In *International Conference on Learning Representations*, 2025.

Aparna Elangovan, Lei Xu, Jongwoo Ko, Mahsa Elyasi, Ling Liu, Sravan Babu Bodapati, and Dan Roth. Beyond correlation: The impact of human uncertainty in measuring the effectiveness of automatic evaluation and llm-as-a-judge. In *International Conference on Learning Representations*, 2025.

Benjamin Feuer, Micah Goldblum, Teresa Datta, Sanjana Nambiar, Raz Besaleli, Samuel Dooley, Max Cembalest, and John P Dickerson. Style outweighs substance: Failure modes of LLM judges in alignment benchmarking. In *International Conference on Learning Representations*, 2025.

Salvatore Greco, Vincent Mousseau, and Roman Słowiński. Ordinal regression revisited: Multiple criteria ranking using a set of additive value functions. *European Journal of Operational Research*, 191(2):416–436, 2008. doi: 10.1016/j.ejor.2007.08.013.

Juyeon Heo, Miao Xiong, Christina Heinze-Deml, and Jaya Narain. Do LLMs estimate uncertainty well in instruction-following? In *International Conference on Learning Representations*, 2025.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Yuxin Jiang, Yufei Wang, Xingshan Zeng, Wanjun Zhong, Liangyou Li, Fei Mi, Lifeng Shang, Xin Jiang, Qun Liu, and Wei Wang. FollowBench: A multi-level fine-grained constraints following benchmark for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4667–4688, 2024. doi: 10.18653/v1/2024.acl-long.257.

Jaehun Jung, Faeze Brahman, and Yejin Choi. Trust or escalate: LLM judges with provable guarantees for human agreement. In *International Conference on Learning Representations*, 2025.

Seungone Kim, Jay Shin, yejin cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, S Shin, Ryan, Sungdong Kim, James Thorne, and Minjoon Seo. Prometheus: Inducing fine-grained evaluation capability in language models. In *International Conference on Learning Representations*, 2024a.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4334–4353, 2024b. doi: 10.18653/v1/2024.emnlp-main.248.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott Hale. The PRISM alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 105236–105344, 2024. doi: 10.52202/079017-3342.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. RewardBench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1755–1797, 2025. doi: 10.18653/v1/2025.findings-naacl.96.

Bracha Laufer-Goldshtein, Adam Fisch, Regina Barzilay, and Tommi Jaakkola. Efficiently controlling multiple risks with pareto testing. In *International Conference on Learning Representations*, 2023.

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. Generative judge for evaluating alignment. In *International Conference on Learning Representations*, 2024a.

Junlong Li, Fan Zhou, Shichao Sun, Yikai Zhang, Hai Zhao, and Pengfei Liu. Dissecting human and LLM preferences. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1790–1811, 2024b. doi: 10.18653/v1/2024.acl-long.99.

Si-Qi Liu, Ian Gemp, Luke Marris, Georgios Piliouras, Nicolas Heess, and Marc Lanctot. Re-evaluating open-ended evaluation of large language models. In *International Conference on Learning Representations*, 2025a.

Xiao Liu, Xuanyu Lei, Shengyuan Wang, Yue Huang, Andrew Feng, Bosi Wen, Jiale Cheng, Pei Ke, Yifan Xu, Weng Lam Tam, Xiaohan Zhang, Lichao Sun, Xiaotao Gu, Hongning Wang, Jing Zhang, Minlie Huang, Yuxiao Dong, and Jie Tang. AlignBench: Benchmarking Chinese alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11621–11640, 2024. doi: 10.18653/v1/2024.acl-long.624.

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 *International Conference on Learning Representations*, 2025b.

Lester James Validad Miranda, Yizhong Wang, Yanai Elazar, Sachin Kumar, Valentina Pyatkin, Faeze Brahman, Noah A. Smith, Hannaneh Hajishirzi, and Pradeep Dasigi. Hybrid preferences: Learning to route instances for human vs. AI feedback. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7162–7200, 2025. doi: 10.18653/v1/2025.acl-long.355.

Remi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Côme Fiegel, Andrea Michi, Marco Selvi, Sertan Girgin, Nikola Momchev, Olivier Bachem, Daniel J Mankowitz, Doina Precup, and Bilal Piot. Nash learning from human feedback. In *International Conference on Machine Learning*, volume 235, pp. 36743–36768, 2024.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems*, volume 37, pp. 68772–68802, 2024. doi: 10.52202/079017-2197.

Nora Petrova, Andrew Gordon, and Enzo Blindow. Unpacking human preference for LLMs: Demographically aware evaluation with the HUMAINE framework. In *International Conference on Learning Representations*, 2026.

Yiwei Qin, Kaiqiang Song, Yebowen Hu, Wenlin Yao, Sangwoo Cho, Xiaoyang Wang, Xuansheng Wu, Fei Liu, Pengfei Liu, and Dong Yu. InFoBench: Evaluating instruction following ability in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13025–13048, 2024. doi: 10.18653/v1/2024.findings-acl.772.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi Jaakkola, and Regina Barzilay. Conformal language modeling. In *International Conference on Learning Representations*, 2024.

Swastik Roy, Rajkumar Pujari, Tharindu Kumarage, Charith Peris, Rahul Gupta, Anna Rumshisky, Pradeep Natarajan, and Venkatesh Saligrama. PReMISE: Policy rubrics as measurement specifications for LLM judges. arXiv:2605.30803, 2026.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. Position: A roadmap to pluralistic alignment. In *International Conference on Machine Learning*, volume 235, pp. 46280–46302, 2024.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Tang, Alejandro Cuadron, Chenguang Wang, Raluca Popa, and Ion Stoica. Judgebench: A benchmark for evaluating LLM-based judges. In *International Conference on Learning Representations*, 2025.

Vijay Viswanathan, Yanchao Sun, Xiang Kong, Meng Cao, Graham Neubig, and Sherry Wu. Checklists are better than reward models for aligning language models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 114728–114754, 2025. doi: 10.52202/085713-3824.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9440–9450, 2024. doi: 10.18653/v1/2024.acl-long.511.

Ziwei Xu and Mohan Kankanhalli. Strong preferences affect the robustness of preference models and value alignment. In *International Conference on Learning Representations*, 2025.

Haoyan Yang, Runxue Bao, Cao (Danica) Xiao, Jun Ma, Parminder Bhatia, Shangqian Gao, and Taha Kass-Hout. Any large language model can be a reliable judge: Debiasing with a reasoning-based bias detector. In *Advances in Neural Information Processing Systems*, volume 38, pp. 7106–7150, 2025. doi: 10.52202/085713-0221.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh Chawla, and Xiangliang Zhang. Justice or prejudice? quantifying biases in LLM-as-a-judge. In *International Conference on Learning Representations*, 2025a.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. FLASK: Fine-grained language model evaluation based on alignment skill sets. In *International Conference on Learning Representations*, 2024.

Ziyi Ye, Xiangsheng Li, Qiuchi Li, Qingyao Ai, Yujia Zhou, Wei Shen, Dong Yan, and Yiqun Liu. Learning LLM-as-a-judge for preference alignment. In *International Conference on Learning Representations*, 2025b.

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

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623, 2023. doi: 10.52202/075280-2020.

Xiaosen Zheng, Tianyu Pang, Chao Du, Qian Liu, Jing Jiang, and Min Lin. Cheating automatic LLM benchmarks: Null models achieve high win rates. In *International Conference on Learning Representations*, 2025.

A PROPERTIES OF THE DECLARED SPECIFICATION

This section records the necessary-preference properties used by the identification check. Criteria and coding are fixed; every admissible weight set is nonempty and compact.

Proposition 1 (Nesting and compatibility)**.** *Hold the criteria and coding fixed. If $\mathcal{W}_1 \subseteq \mathcal{W}_2$, then $\mathcal{C}_{\mathcal{S}_2} \subseteq \mathcal{C}_{\mathcal{S}_1}$ and strict directions agree. If every $w \in \mathcal{W}$ is strictly positive, the core contains every strict-Pareto direction. If $w_0 \in \mathcal{W}$, each core direction agrees with* $\operatorname{sign}(w_0^\top x)$.

Proof. If $\min_{w \in \mathcal{W}_2} w^\top x > 0$, the same inequality holds over the smaller set $\mathcal{W}_1$; the negative direction is symmetric. If $x \geq 0$, $x \neq 0$, and every admissible weight is strictly positive, then $w^\top x > 0$ for every $w \in \mathcal{W}$. Compactness and continuity give a positive minimum. Finally, a core direction must hold for every $w_0 \in \mathcal{W}$, including any specified scalar comparator. □

Proposition 2 (Pointwise maximality within the soundness class)**.** *Call h specification-sound when $h(x) = s \in \{-1, +1\}$ implies $s\, w^\top x > 0$ for every $w \in \mathcal{W}$. Every strict release of such an h is released in the same direction by $r_{\mathcal{S}}$.*

Proof. For $h(x) = +1$, soundness gives $w^\top x > 0$ for all $w \in \mathcal{W}$. Compactness and continuity give $\underline{u}_{\mathcal{W}}(x) > 0$, so $r_{\mathcal{S}}(x) = +1$. The negative case is symmetric. The rule in Equation (3) is itself sound, hence includes every strict release allowed by the stated soundness requirement. □

Deferral, equality, and admissible completion. The set $\mathcal{F}_{\mathcal{S}}$ is the complement of the strict core, so it includes equal profiles as well as conflicts. It does not assert that every member can acquire a strict preference by selecting an existing weight. If $\underline{u}_{\mathcal{W}} < 0 < \overline{u}_{\mathcal{W}}$, both strict directions are possible. If $\underline{u}_{\mathcal{W}} = 0 < \overline{u}_{\mathcal{W}}$, only the positive strict direction is possible, although it is not unanimous; the negative case is symmetric. If both bounds are zero, every admissible weight gives equality. For HUMAINE, the closed form in Equation (8) gives an interval touching zero for an oriented two-versus-one profile $(1, 1, 0, -1)$, rather than a conflict admitting both strict signs.

B TRADEOFFBENCH: COMPLETE AUDIT SPECIFICATION

B.1 TASK CONSTRUCTION AND THE TWO PARTITIONS

TradeoffBench generates addition, comparison, parity, arithmetic-sequence, and unit-conversion tasks. Four stochastic response policies produce six unordered pairs per task. Correctness checks the answer; completeness checks two required facts; format checks the response schema; concision checks a word-count limit; lexical compliance checks a required token. A deterministic renderer–validator pair requires the recovered binary profile to equal the sampled profile. Seed sequences separately determine task content, policy profiles, and displayed candidate assignment.

The ordinary-interface breadth partition has 4,200 base pairs from 700 tasks, with matched response swaps, criterion reversal, and exact repeats giving 10,758 inference records per model. Its 895 benchmark-frontier pairs come from 487 tasks. The controlled partition also has 4,200 base pairs from 700 tasks; its 849 benchmark-frontier pairs come from 457 tasks. Equalities number 284 and 289, respectively. These counts refer to partitions with distinct task content. All six models use the same records within each partition. In vector conditions, different source pairs can have identical five-coordinate inputs. Pair-weighted rates retain their frequency in the benchmark rather than deduplicating these profiles.

For benchmark stratification, an external ordered weight set constrains nonnegative normalized weights by correctness $\geq$ completeness and completeness $\geq$ each remaining criterion. The 895 and 849 benchmark-frontier pairs have no identified strict direction under this external set. This stratification is used for the subset diagnostics, not to define the rule supplied in the controlled audit. The external set identifies a direction on the additional 468 and 496 mixed-sign pairs, but its priorities are not supplied to any audit prompt. They are therefore reported separately from dominance controls. In the controlled strict-Pareto-only specification, both mixed-sign strata require UNDERDETERMINED (or TIE in the combined-label condition). The complete mixed-sign counts are reported with the matched audit in Appendix B.7.

B.2 COMPLETE SYSTEM INSTRUCTIONS

The ordinary baseline and controlled Raw/TIE condition are different prompts. The former requests an overall winner; the latter states an exhaustive identification rule and is the baseline for the controlled 2×2 contrasts. The following are the system messages used for inference.

Ordinary overall judgment.

> You are evaluating two candidate responses using the rubric supplied by the user. Consider every listed criterion and decide which response is better overall. Return exactly one tag and no explanation: <VERDICT>A</VERDICT>, <VERDICT>B</VERDICT>, or <VERDICT>TIE</VERDICT>.

Controlled Raw/TIE.

> Classify the complete five-coordinate relation between Response A and Response B from the supplied question, rubric, and response texts. Each listed binary criterion has one categorical relation: A_ONLY means that criterion favors A, B_ONLY means it favors B, and SAME means neither response is favored. Apply this exhaustive rule to the five listed criteria only: if at least one criterion is A_ONLY and none is B_ONLY, return A; if at least one criterion is B_ONLY and none is A_ONLY, return B; if every criterion is SAME or if both A_ONLY and B_ONLY occur, return TIE. Here TIE deliberately combines exact equality and underdetermination under an unspecified trade-off. Do not infer weights, priorities, severity, a majority rule, or additional criteria. Return exactly one tag and no explanation: <VERDICT>A</VERDICT>, <VERDICT>B</VERDICT>, or <VERDICT>TIE</VERDICT>.

Controlled Raw/disjoint.

> Classify the complete five-coordinate relation between Response A and Response B from the supplied question, rubric, and response texts. Each listed binary criterion has one categorical relation: A_ONLY means that criterion favors A, B_ONLY means it favors B, and SAME means neither response is favored. Apply this exhaustive rule to the five listed criteria only: if every criterion is SAME, return EQUAL; if at least one criterion is A_ONLY and none is B_ONLY, return A; if at least one criterion is B_ONLY and none is A_ONLY, return B; if both A_ONLY and B_ONLY occur, return UNDERDETERMINED because no trade-off rule is declared. Do not infer weights, priorities, severity, a majority rule, or additional criteria. Return exactly one tag and no explanation: <VERDICT>A</VERDICT>, <VERDICT>B</VERDICT>, <VERDICT>EQUAL</VERDICT>, or <VERDICT>UNDERDETERMINED</VERDICT>.

Controlled Vector/TIE.

> Classify the complete, authoritative five-coordinate relation vector between Response A and Response B. Each coordinate is one categorical value: A_ONLY means that coordinate favors A, B_ONLY means it favors B, and SAME means neither response is favored. Apply this exhaustive rule to the values inside <VECTOR> only: if at least one coordinate is A_ONLY and none is B_ONLY, return A; if at least one coordinate is B_ONLY and none is A_ONLY, return B; if every coordinate is SAME or if both A_ONLY and B_ONLY occur, return TIE. Here TIE deliberately combines exact equality and underdetermination under an unspecified trade-off. Do not infer weights, priorities, severity, or a majority rule. Return exactly one tag and no explanation: <VERDICT>A</VERDICT>, <VERDICT>B</VERDICT>, or <VERDICT>TIE</VERDICT>.

Controlled Vector/disjoint.

> Classify the complete, authoritative five-coordinate relation vector between Response A and Response B. Each coordinate is one categorical value: A_ONLY means that coordinate favors A, B_ONLY means it favors B, and SAME means neither response is favored. Apply this exhaustive rule to the values inside <VECTOR> only: if every coordinate is SAME, return EQUAL; if at least one coordinate is A_ONLY and none is B_ONLY, return A; if at least one coordinate is B_ONLY and none is A_ONLY, return B; if both A_ONLY and

> B_ONLY occur, return UNDERDETERMINED because no trade-off rule is declared. Do not infer weights, priorities, severity, or a majority rule. Return exactly one tag and no explanation: <VERDICT>A</VERDICT>, <VERDICT>B</VERDICT>, <VERDICT>EQUAL</VERDICT>, or <VERDICT>UNDERDETERMINED</VERDICT>.

B.3 COMPLETE USER-SIDE EXAMPLE

This is the complete verified mixed-sign pair summarized in Section 3. The ordinary-interface user message is reproduced below. The question and response content are unchanged in the two controlled raw-response cells. Only the final question is replaced by the controlled suffix given below.

> Question: Which number is larger: 66 or 11?
>
> Rubric (evaluate all listed criteria):
> 1. Correctness: ANSWER correctly resolves the question.
> 2. Completeness: include both required facts WORK=compare:66>11 and CHECK=gap:55.
> 3. Format: exactly one <RESULT> block with semicolon-separated ANSWER, WORK, CHECK, and optional TOKEN fields.
> 4. Concision: use at most 9 whitespace-delimited words.
> 5. Lexical compliance: include the standalone token AZURE.
>
> Response A:
> <RESULT> ANSWER=66; WORK=compare:66>11; CHECK=gap:55; TOKEN=AZURE; </RESULT> verbose filler words repeated only to exceed the stated response length limit now
>
> Response B:
> <RESULT> ANSWER=66; WORK=compare:66>11; CHECK=omitted; TOKEN=AZURE; </RESULT>
>
> Which response is better overall according to this rubric? Choose A if Response A is better overall, B if Response B is better overall, or TIE if neither response is better overall.

In the controlled raw cells, the final sentence is instead: *Apply the exhaustive rule to the five listed criteria and return only the tag.* The raw-response task requires recovering the five binary relations from these texts.

The complete vector-cell user message for the same pair is:

> The vector below is the entire evidence; there is no hidden response text or additional criterion.
> <VECTOR>[SAME, A_ONLY, SAME, B_ONLY, SAME]</VECTOR>
>
> Apply the exhaustive rule to this vector and return only the tag.

This corresponds to $x = (0, 1, 0, -1, 0)$ in displayed criterion order. An unopposed profile such as [SAME, A_ONLY, SAME, SAME, SAME] instead requires A; an all-SAME profile requires EQUAL. The last two profiles illustrate the rule and were not supplied as in-context demonstrations. Vector prompts contain no question, criterion names, or response text, removing those sources of semantic context.

B.4 COMPUTATIONAL ENVIRONMENT

The six-model controlled audit and three-model exhaustive audit ran under Slurm on NVIDIA A100-SXM4 GPUs with 80 GB of memory each. The server used two AMD EPYC 7742 processors (64 cores each), approximately 2 TiB of host memory, and Ubuntu 22.04.5 LTS. The software environment comprised Python 3.12.12, PyTorch 2.10.0 (built with CUDA 12.8), and Transformers 5.10.1. Inference used one GPU per process except for OLMo's three non-endpoint factorial cells, which used two.

B.5 MODEL AND DECODING SETTINGS

The shortened Mistral-Small-24B and GLM-4-9B labels denote Mistral-Small-3.2-24B and GLM-4-9B-Chat. All controlled cells for a model share the checkpoint revision in Table 4 and

use bfloat16 precision, greedy decoding, scaled dot-product attention (SDPA), and the official chat template. Inputs are limited to 4,096 tokens. Qwen uses `enable_thinking=false`; EXAONE uses `enable_thinking=false` and `skip_think=true`. Mistral is loaded through the official image–text model class and tokenizer for text-only inference; other checkpoints use their causal language model class. These audits use the released checkpoints without any model fine-tuning.

The generation cap is 128 tokens, except for GLM's 512-token allowance. For Phi, the three non-endpoint cells initially allow 128 tokens and extend unfinished outputs to a maximum of 512 tokens. Vector/disjoint uses 128. This continuation lets Phi finish explanations before returning a verdict. Truncated explanations are not classified as abstentions. Verdict extraction uses the final explicit verdict tag, or an unambiguous label when only the wrapper is malformed. The raw-output audit independently checks endpoint counts against the saved summaries. The ordinary breadth runs use greedy decoding and the same checkpoint identities, with model-specific output allowances recorded in their inference configurations.

Table 4: Checkpoint identities shared across the six-model audit. Revisions specify model weights.

Model	Hugging Face repository and revision
Qwen3-8B	`Qwen/Qwen3-8B` b968826d9c46dd6066d109eabc6255188de91218
Mistral-Small-3.2-24B	`mistralai/Mistral-Small-3.2-24B-Instruct-2506` 95a6d26c4bfb886c58daf9d3f7332c857cb27b43
Phi-4-14B	`microsoft/phi-4` 2db69c1c3e91a05d2c64a3185acfbaf36f744e25
OLMo-Hybrid-7B	`allenai/Olmo-Hybrid-Instruct-DPO-7B` ec62da5af0106ab4e83a60e07c4796ff358fd204
GLM-4-9B-Chat	`zai-org/glm-4-9b-chat-hf` 8599336fc6c125203efb2360bfaf4c80eef1d1bf
EXAONE-4.0-32B	`LGAI-EXAONE/EXAONE-4.0-32B` a1d54d1c148c30881ed27e035b650da489b51b92

B.6 Ordinary-interface breadth and presentation sensitivity

The ordinary interface supplies context for the controlled rule audit. These diagnostics quantify presentation changes and compatibility with a single global additive weighting. They do not measure compliance with the controlled deferral instruction. This complements position-bias analyses (Wang et al., 2024; Ye et al., 2025a).

Table 5: Six-model ordinary-interface TradeoffBench audit. Frontier choice uses 895 pairs; dominance accuracy uses only 2,553 unopposed pairs; equality accuracy uses 284 pairs. Position/order columns subtract exact-repeat flips. Global-fit gap is the minimum fraction of strict frontier choices not rationalizable by one global nonnegative additive weight. All entries are percentages; frontier brackets are 95% bootstrap intervals over latent tasks.

Model	Frontier [95% CI]	Dominance	Equality	Position	Order	Global-fit
Qwen3-8B	99.9 [99.7, 100.0]	97.5	100.0	42.5	10.8	21.9
Mistral-Small-24B	99.6 [99.1, 99.9]	97.1	100.0	22.2	10.9	12.6
Phi-4-14B	100.0 [100.0, 100.0]	95.6	100.0	64.6	11.3	32.0
OLMo-Hybrid-7B	97.2 [95.9, 98.3]	95.8	100.0	30.8	8.4	13.2
GLM-4-9B	100.0 [100.0, 100.0]	97.3	99.3	31.6	16.8	18.7
EXAONE-4.0-32B	97.7 [96.6, 98.6]	98.7	100.0	24.0	8.4	12.6

To calculate the presentation effects in Table 5, we map displayed choices back to canonical candidate identities. We compare each perturbation against its matched base verdict and subtract instability on byte-identical repeats. The global-fit diagnostic fits a nonnegative normalized weight to strict frontier choices. Mixed-integer constraint deletion identifies the smallest fraction that must be excluded to obtain one strict-sign fit. This diagnostic tests compatibility with a context-invariant additive completion only; it does not recover weights used internally by a model.

B.7 SAME-RECORD BASELINE AND CONTROLLED AUDIT

The factorial diagnostics use an externally stratified subset of 849 mixed-sign pairs, the *benchmark frontier*. An ordered weight set defines this subset, but its priorities are not supplied to the models. The controlled rule requires deferral on all 1,345 mixed-sign pairs. Subset diagnostics remain separate from the complete-population endpoint result.

With disjoint labels fixed, exact vectors reduce strict choices on that subset in every model (Figure 4). Label separation raises raw-response strict choice for Qwen, OLMo, and EXAONE. With vectors, it lowers strict choice for five models but raises it for GLM. Raw conditions also miss dominance controls, so their contrasts do not isolate value completion from broader rule-execution failures. Full factorial control accuracy and paired changes are in Table 6 and fig. 5.

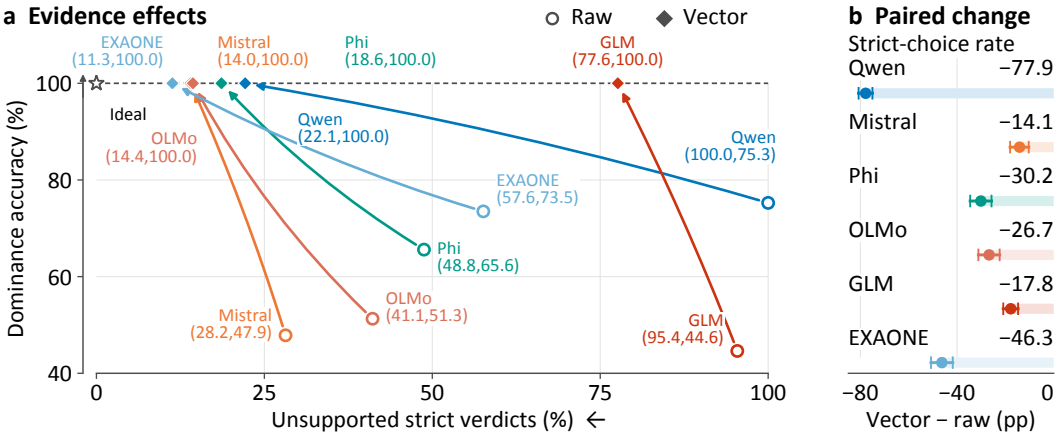

Figure 4: Matched raw-to-vector changes with disjoint labels and the rule fixed. Panel a connects raw (circles) and vector (diamonds) results; labels give (strict-choice rate, dominance accuracy), in percent. The axes use 849 benchmark-frontier pairs and 2,566 dominance controls. Panel b gives vector-minus-raw changes in percentage points with pointwise 95% paired latent-task bootstrap CIs. Table 1 instead uses all 1,345 mixed-sign pairs.

An ordinary prompt permits A, B, or TIE and asks for the better response without the exhaustive rule. On the same 849-pair subset, strict choice is 96.58–100% (Table 6). This measures completion under an overall-judgment interface rather than violation of a deferral instruction. The tag-only output cannot explain added trade-offs. Separate breadth diagnostics test presentation sensitivity and additive compatibility (Appendix B.6).

Point estimates weight eligible pairs equally. Intervals resample latent tasks, retaining matched conditions (Appendix E). With at most $3^5 = 243$ categorical vector profiles, rates reflect benchmark frequencies rather than uniform profile weighting.

Table 6 compares the ordinary baseline and all four controlled cells on the same 4,200 records. Only the controlled cells receive the exhaustive identification rule (Appendix B.2). The three non-endpoint cells use a cyclic schedule to balance processing order and are matched by source pair to the exchange audit's original-orientation Vector/disjoint outputs. The paired contrasts in Figure 5 retain all conditions attached to each source pair and resample latent tasks. Pairs within a task are therefore not treated as independent observations.

Table 6: Same-record ordinary baseline and complete controlled audit. Rates are percentages: benchmark-frontier strict choice uses 849 pairs, dominance accuracy uses 2,566 controls, and equality accuracy uses 289 equalities. The ordinary baseline is outside the controlled 2×2 design. Brackets in the controlled block are pointwise 95% bootstrap intervals over latent tasks.

Model	Evidence / labels	Frontier [95% CI]	Dominance	Equality
Ordinary overall judgment				
Qwen3-8B	Raw / TIE	100.00	96.45	100.00
Mistral-Small-24B	Raw / TIE	99.18	96.41	100.00
Phi-4-14B	Raw / TIE	100.00	94.82	100.00
OLMo-Hybrid-7B	Raw / TIE	96.58	96.06	100.00
GLM-4-9B	Raw / TIE	100.00	96.22	99.65
EXAONE-4.0-32B	Raw / TIE	98.47	98.56	100.00
Controlled evidence $\times$ labels, with a shared identification rule				
Qwen3-8B	Raw / TIE	48.5 [44.7, 52.4]	68.9	100.0
	Raw / disjoint	100.0 [100.0, 100.0]	75.3	100.0
	Vector / TIE	31.0 [27.7, 34.3]	100.0	100.0
	Vector / disjoint	22.1 [19.3, 25.1]	100.0	100.0
Mistral-Small-24B	Raw / TIE	55.9 [52.7, 59.2]	73.5	100.0
	Raw / disjoint	28.2 [25.1, 31.3]	47.9	99.7
	Vector / TIE	35.8 [32.4, 39.2]	100.0	100.0
	Vector / disjoint	14.0 [11.6, 16.5]	100.0	100.0
Phi-4-14B	Raw / TIE	71.4 [68.2, 74.6]	73.7	100.0
	Raw / disjoint	48.8 [45.4, 52.2]	65.6	99.3
	Vector / TIE	37.2 [33.7, 40.8]	100.0	100.0
	Vector / disjoint	18.6 [15.8, 21.5]	100.0	100.0
OLMo-Hybrid-7B	Raw / TIE	35.6 [32.2, 39.0]	51.7	100.0
	Raw / disjoint	41.1 [37.6, 44.6]	51.3	92.0
	Vector / TIE	23.1 [20.1, 26.0]	90.2	100.0
	Vector / disjoint	14.4 [11.9, 16.8]	100.0	100.0
GLM-4-9B	Raw / TIE	99.4 [98.8, 99.9]	53.2	14.9
	Raw / disjoint	95.4 [93.6, 97.1]	44.6	100.0
	Vector / TIE	72.2 [69.1, 75.3]	94.5	100.0
	Vector / disjoint	77.6 [74.8, 80.4]	100.0	100.0
EXAONE-4.0-32B	Raw / TIE	39.8 [36.2, 43.6]	74.6	100.0
	Raw / disjoint	57.6 [54.2, 61.1]	73.5	100.0
	Vector / TIE	36.3 [32.9, 39.6]	100.0	100.0
	Vector / disjoint	11.3 [9.0, 13.7]	100.0	100.0

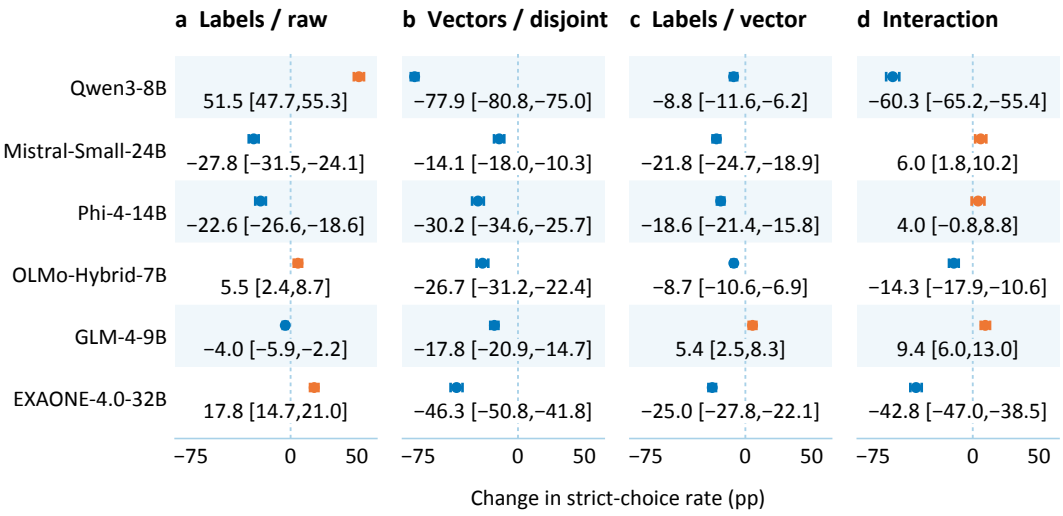

Figure 5: The effects of evidence and label separation differ across models. Paired changes in strict-choice rate use the same 849 benchmark-frontier pairs. a, L_R: disjoint-minus-TIE with raw evidence. b, V_D: vector-minus-raw with disjoint labels. c, L_V: disjoint-minus-TIE with vectors. d, Interaction $L_V - L_R$. Points and printed values give changes in percentage points; whiskers and brackets give pointwise 95% paired latent-task bootstrap intervals, not a simultaneous band. All panels share the same scale.

Table 7: Vector/disjoint strict-choice counts on the benchmark-frontier subset and all mixed-sign pairs. The A/B columns use the 849 benchmark-frontier pairs; the last two columns include the additional 496 mixed-sign pairs. Every mixed-sign vector requires UNDERDETERMINED under the actual prompt.

Model	A	B	All mixed-sign strict	Rate (%)
Qwen3-8B	155	33	345/1,345	25.65
Mistral-Small-24B	119	0	119/1,345	8.85
Phi-4-14B	125	33	301/1,345	22.38
OLMo-Hybrid-7B	122	0	237/1,345	17.62
GLM-4-9B	607	52	1,091/1,345	81.12
EXAONE-4.0-32B	51	45	140/1,345	10.41

Residual strict choice remains nonzero over the complete mixed-sign set (Table 7), on which the prompt identifies no strict winner. Its direction can be highly asymmetric: Mistral's 119 strict benchmark-frontier choices all favor displayed A.

B.8 PAIRED ENDPOINT EXCHANGE, MAJORITY COMPLETION, AND ADDITIVE RATIONALIZATION

The exchange audit evaluates all 4,200 Vector/disjoint source pairs in original and exchanged orientations for each model. Only the authoritative vector's `A_ONLY`/`B_ONLY` entries and displayed candidate identities are exchanged. Criterion order, instructions, model revision, chat template, thinking settings, generation cap, and batch size remain fixed. The 8,400 outputs per model give 50,400 judgments across six models. Each orientation of this Vector/disjoint retest contains 2,566 dominance controls and 289 equalities, all answered correctly.

We map each strict verdict to its original candidate identity before comparing orientations. Candidate consistency requires strict judgments in both orientations with the same identity. Position locking requires A/A or B/B and therefore selects different identities. One-strict pairs have a strict output in exactly one orientation; the remaining pairs have no strict output. Table 8 partitions both reported populations into these mutually exclusive outcomes. GLM's 319 one-strict pairs comprise 149 with a strict verdict only in the original orientation and 170 only in the exchanged orientation. All endpoint

analyses use the same parser and original-orientation observations. Independent checks of literal outputs recover no additional mixed-sign labels. Malformed wrappers recovered for OLMo occur only on dominance controls.

Table 8: Complete endpoint exchange counts. "Original" and "Exchanged" count strict outputs; the last four columns partition pairs into candidate-consistent, position-locked, one-strict, and neither-strict outcomes. All position-locked pairs are A/A; none is B/B.

Model	Original	Exchanged	Consistent	Locked	One strict	Neither
Benchmark frontier, 849 pairs						
Qwen3-8B	188	183	54	0	263	532
Mistral-Small-24B	119	129	0	0	248	601
Phi-4-14B	158	156	52	0	210	587
OLMo-Hybrid-7B	122	140	0	0	262	587
GLM-4-9B	659	668	81	459	247	62
EXAONE-4.0-32B	96	101	72	0	53	724
Complete mixed-sign population, 1,345 pairs						
Qwen3-8B	345	351	72	78	396	799
Mistral-Small-24B	119	129	0	0	248	1,097
Phi-4-14B	301	301	165	0	272	908
OLMo-Hybrid-7B	237	244	62	11	335	937
GLM-4-9B	1,091	1,112	176	766	319	84
EXAONE-4.0-32B	140	146	102	0	82	1,161

Pointwise intervals resample latent tasks 20,000 times, retaining original and exchanged observations together. For the nonzero benchmark-frontier consistency rates, the 95% intervals are Qwen $[4.57, 8.29]$, Phi $[4.36, 8.01]$, GLM $[7.57, 11.61]$, and EXAONE $[6.52, 10.53]$ percent. Complete mixed-sign intervals appear in Table 2. An observed zero gives a degenerate empirical bootstrap distribution, not evidence of a zero population probability. These rates are not causal estimates of the fraction of behavior unaffected by position.

Majority direction and the release set. For each canonical vector, majority predicts $\mathrm{sign}(\sum_k x_k)$, with zero supplying no strict direction. The complete mixed-sign population contains 827 strict-majority pairs over 120 distinct canonical profiles and 518 tied-vote pairs over 43 profiles. The benchmark frontier contains 331 of the strict-majority pairs and all 518 tied pairs; the remaining 496 pairs all have a strict majority. This split uses the verified components, not model outputs.

All 577 candidate-consistent model–pair outcomes take the majority direction, with no tied votes. They span 282 distinct pairs and 41 canonical profiles. Within the benchmark frontier, all 259 consistent outcomes, spanning 119 pairs and 13 profiles, also match majority. Pair-weighted and profile-weighted agreement are both 100% for every nonempty model subset. For profile weighting, we first compute agreement within each signed canonical profile containing consistent outcomes, then weight these profiles equally. Separately, profile-weighted consistency averages the within-profile consistency rate over *all* profiles, including those with no consistent outcomes (Table 9). Empty consistent subsets have undefined rule agreement.

The tied group has no candidate-consistent strict output from any model, but the single-orientation counts differ. Mistral and EXAONE each return UNDERDETERMINED in all 1,036 tied-group outputs, with no A, B, EQUAL, or parsing failures. In the strict-majority group, Mistral gives 119 original and 129 exchanged strict verdicts, all labeled A before mapping to canonical identities. EXAONE gives 140 and 146 strict verdicts, respectively; every strict output from both models follows the majority direction. Thus Mistral issues a strict verdict only when displayed A is the majority-favored candidate. Each of the 827 strict-majority pairs presents its majority candidate as A in exactly one orientation. Only 248 of these 827 presentations elicit strict verdicts, so this condition is necessary but not sufficient. Qwen, Phi, OLMo, and GLM also produce strict outputs on tied votes, all favoring displayed A; none of these tied-group outputs survives exchange with candidate identity preserved.

Nine pairs from seven latent tasks are candidate-consistent in all five nonempty model subsets. Their canonical vectors are $(1, 1, -1, 1, 1)$ and its negation: one profile up to sign, outside the benchmark frontier. For the positive orientation, the equal-weight member gives $w_0^\top x = 0.2 \sum_k x_k = 0.60$,

whereas $(.1, .1, .6, .1, .1)$ gives -0.20. Both weightings are positive and normalized, but the prompt declares neither their relative priority nor majority aggregation. The opposite directions illustrate why selecting the equal-weight member adds a premise. The broader 577-outcome result, rather than this overlap alone, establishes the observed majority pattern.

Table 9: Majority-stratified outputs and profile-weighted consistency. Panel a partitions 1,345 mixed-sign pairs by vote structure; entries count strict outputs or candidate-consistent pairs. Panel b gives the number of profiles with consistent pairs, pair-frequency consistency (%), and consistency averaged equally across all observed canonical profiles (%). The complete population has 163 profiles; the benchmark frontier has 85. Conditional on a nonempty consistent subset, majority-direction agreement is 100% under both weightings.

a Strict outputs by vote structure

	Strict majority: 827 pairs			Tied majority: 518 pairs		
Model	Original	Exchanged	Consistent	Original	Exchanged	Consistent
Qwen3-8B	277	287	72	68	64	0
Mistral-Small-24B	119	129	0	0	0	0
Phi-4-14B	281	284	165	20	17	0
OLMo-Hybrid-7B	229	234	62	8	10	0
GLM-4-9B	709	723	176	382	389	0
EXAONE-4.0-32B	140	146	102	0	0	0

b Candidate consistency: pair-frequency and distinct-profile summaries

	Complete mixed-sign population			Benchmark frontier		
Model	Profiles	Pair (%)	Profile (%)	Profiles	Pair (%)	Profile (%)
Qwen3-8B	9	5.35	5.52	5	6.36	5.88
Mistral-Small-24B	0	0.00	0.00	0	0.00	0.00
Phi-4-14B	24	12.27	14.72	4	6.12	4.71
OLMo-Hybrid-7B	8	4.61	4.91	0	0.00	0.00
GLM-4-9B	25	13.09	14.87	9	9.54	10.59
EXAONE-4.0-32B	13	7.58	7.98	9	8.48	10.59

What additive fits establish. A strict choice with canonical sign s imposes $s\, w^\top x > 0$ on one global nonnegative normalized weight vector. Deferrals impose no constraint; these fits assess only the directions of emitted strict verdicts. We fit original strict choices, both orientations' strict choices, and the candidate-consistent subset separately (Table 10). Each position-locked pair imposes opposite strict inequalities, forcing at least one exclusion. Exact rational witnesses attain this lower bound in every combined-orientation fit: 78 for Qwen, 11 for OLMo, 766 for GLM, and zero for the others. Mistral illustrates the fit's scope: equal weights fit all 248 strict directions without determining which presentations elicit them. Within each nonempty candidate-consistent subset, zero exclusions follow immediately from the equal-weight member, because every observed direction matches majority. That column is therefore a consequence of the majority result, not independent evidence of inferred internal weights.

Table 10: Endpoint additive fits on all 1,345 mixed-sign pairs. Entries give the certified minimum exclusion count / number of strict observations in a global additive fit. Combined-orientation minima equal the position-locked pair counts in Table 8. Candidate-consistent pairs enter once; equal weights fit all their directions. Dashes denote empty subsets.

Model	Original	Both orientations	Candidate-consistent
Qwen3-8B	41 / 345	78 / 696	0 / 72
Mistral-Small-24B	0 / 119	0 / 248	–
Phi-4-14B	0 / 301	0 / 602	0 / 165
OLMo-Hybrid-7B	4 / 237	11 / 481	0 / 62
GLM-4-9B	379 / 1,091	766 / 2,203	0 / 176
EXAONE-4.0-32B	0 / 140	0 / 286	0 / 102

On the 849-pair benchmark frontier, GLM requires 216 exclusions among 659 original strict judgments and 459 among 1,327 combined judgments. Every other model's original and combined strict subsets have zero fit violations. Mistral and OLMo have no candidate-consistent pair in this subset.

Fixed-coordinate and priority-scan comparisons. We compare majority with each fixed-coordinate rule x_j, $j = 1, \ldots, 5$ (Table 11). A zero coordinate supplies no direction and counts as a mismatch to an observed strict choice; it is not omitted. No single fixed coordinate matches all candidate-consistent directions for Phi, GLM, or EXAONE in the complete population. Qwen's directions also match slots 4 and 5, and OLMo's match slots 1 and 5. Across all 577 outcomes, slots 1–5 match 301, 387, 403, 358, and 441 directions, respectively; none reproduces the shared majority pattern. Restricting to the benchmark frontier leaves more individual-model rules indistinguishable, as the second panel shows.

We also check fixed scans that choose the first or last nonzero coordinate. Both scans match all OLMo consistent directions; the last-nonzero scan also matches Qwen and EXAONE. Neither scan matches every consistent direction for Phi or GLM. Majority with either scan used only to break vote ties necessarily retains 100% conditional agreement: no consistent outcome has a tied vote. Thus these tie-breaking variants add no evidence for a particular tie rule. The fixed input order is correctness, completeness, format, concision, and lexical compliance, but the vector prompt omits these names. Slot-based agreement therefore cannot establish a semantic preference such as correctness priority.

Table 11: Fixed-coordinate agreement with candidate-consistent strict directions. Each entry is pair-weighted / profile-weighted agreement (%), conditional on the model's consistent subset. Both opposite directions and zero coordinates count as mismatches. Dashes denote empty subsets.

Model	Slot 1	Slot 2	Slot 3	Slot 4	Slot 5
Complete mixed-sign population, 1,345 pairs					
Qwen3-8B	12.50 / 22.22	43.06 / 55.56	75.00 / 55.56	100 / 100	100 / 100
Mistral-Small-24B	–	–	–	–	–
Phi-4-14B	63.64 / 75.00	56.97 / 70.83	71.52 / 58.33	66.06 / 58.33	83.03 / 75.00
OLMo-Hybrid-7B	100 / 100	66.13 / 75.00	59.68 / 50.00	48.39 / 50.00	100 / 100
GLM-4-9B	53.98 / 64.00	89.77 / 92.00	65.91 / 60.00	33.52 / 36.00	45.45 / 44.00
EXAONE-4.0-32B	29.41 / 30.77	61.76 / 69.23	76.47 / 69.23	86.27 / 84.62	88.24 / 84.62
Benchmark frontier, 849 pairs					
Qwen3-8B	0 / 0	24.07 / 20.00	100 / 100	100 / 100	100 / 100
Mistral-Small-24B	–	–	–	–	–
Phi-4-14B	0 / 0	25.00 / 25.00	100 / 100	100 / 100	100 / 100
OLMo-Hybrid-7B	–	–	–	–	–
GLM-4-9B	0 / 0	77.78 / 77.78	100 / 100	53.09 / 55.56	55.56 / 55.56
EXAONE-4.0-32B	0 / 0	75.00 / 77.78	79.17 / 77.78	80.56 / 77.78	83.33 / 77.78

B.9 Exhaustive profile audit

The audit in Section 3.4 covers 180 mixed-sign profiles, 62 dominance profiles, and one all-equal profile, each in both A/B orientations. The three anonymous encodings use categorical `A_ONLY`/`B_ONLY`/`SAME` entries, signed $-1/0/+1$ entries, or a five-row criterion table. All prompts apply the rule in Equation (4) to the same complete evidence, without task or response text.

We reuse the Qwen3-8B and EXAONE-4.0-32B revisions in Table 4 and add Qwen3.8-27B at revision 1d4bf0f2ff6012fd82039f2fa52739d0dd7c60c0. Each prompt uses seeds 20260919, 20260920, and 20260921 in both modes, with sampling settings and the 2,048-token cap matched within each model. Temperatures for Qwen3-8B, EXAONE, and Qwen3.8 are $0.6, 0.6, 1.0$, with batch sizes 2, 1, and 1, respectively. All runs use bfloat16 and top-$p = 0.95$; top-k is 20 for both Qwen checkpoints and disabled for EXAONE. Reasoning-enabled (ON) and reasoning-disabled (OFF) configurations use the corresponding official chat templates. Qwen3.8 ON additionally sets `reasoning_effort=medium`.

Comparison with the primary greedy audit in Appendix B.5 also changes sampling settings and the output cap. Equal token caps do not imply equal realized generation lengths or compute, and this

experiment does not separate the contribution of the reasoning trace from other mode-dependent behavior.

Completion lengths include reasoning but exclude input and padding tokens. All 13,122 ON outputs contain nonempty reasoning. Each model has 4,374 outputs per mode (243 profiles $\times$ two orientations $\times$ three encodings $\times$ three seeds). Table 12 reports the observed maxima. All 26,244 outputs terminate with an end-of-sequence marker before reaching the matched 2,048-token cap. These maximum lengths are descriptive, not estimates of mean token use, paired overhead, latency, energy, or end-to-end savings. Token counts use each model's tokenizer, so comparisons should be made within each model.

Table 12: Observed maximum completion length by mode in the exhaustive profile audit. Each entry is a maximum over 4,374 outputs, not a mean or a matched-example overhead. Lengths include reasoning where present and exclude input and padding tokens. The maximum allowance is 2,048 tokens in both modes.

Model	OFF maximum	ON maximum
Qwen3-8B	16	1,419
EXAONE-4.0-32B	17	1,511
Qwen3.8-27B	16	293

Table 13: Exhaustive profile audit by model, mode, and encoding. Mixed strict gives unsupported A/B verdicts over mixed-sign orientation decisions, with percentages in parentheses. Consistent counts mixed-sign profile–encoding–seed events in which both orientations select the same canonical candidate. Dominance and equality give correct/total decisions.

Model	Mode	Encoding	Mixed strict	Consistent	Dominance	Equality
Qwen3-8B	OFF	Categorical	853/1,080 (78.98)	80	334/372	6/6
		Signed numeric	306/1,080 (28.33)	16	332/372	6/6
		Criterion table	84/1,080 (7.78)	3	365/372	6/6
	ON	Categorical	1/1,080 (0.09)	0	372/372	6/6
		Signed numeric	0/1,080 (0)	0	372/372	6/6
		Criterion table	0/1,080 (0)	0	372/372	6/6
EXAONE-4.0-32B	OFF	Categorical	291/1,080 (26.94)	118	366/372	6/6
		Signed numeric	407/1,080 (37.69)	146	366/372	6/6
		Criterion table	34/1,080 (3.15)	6	363/372	6/6
	ON	Categorical	0/1,080 (0)	0	372/372	6/6
		Signed numeric	0/1,080 (0)	0	372/372	6/6
		Criterion table	1/1,080 (0.09)	0	372/372	6/6
Qwen3.8-27B	OFF	Categorical	63/1,080 (5.83)	7	372/372	6/6
		Signed numeric	37/1,080 (3.43)	0	372/372	6/6
		Criterion table	0/1,080 (0)	0	372/372	6/6
	ON	Categorical	0/1,080 (0)	0	372/372	6/6
		Signed numeric	0/1,080 (0)	0	372/372	6/6
		Criterion table	0/1,080 (0)	0	372/372	6/6

The OFF candidate-consistent events in Table 13 cover 36, 69, and four distinct canonical profiles for Qwen3-8B, EXAONE, and Qwen3.8, respectively. EXAONE OFF selects the same canonical candidate in both orientations and all three encodings for four profile–seed combinations spanning three distinct profiles. Signed-numeric violations rule out counting only literal `A_ONLY`/`B_ONLY` evidence tokens, but not counting signs or another shallow aggregation heuristic. Some OFF cells also fail dominance controls, so their mixed-sign rates do not isolate conflict-specific errors. Counts describe this finite grid; repeated orientations, encodings, and seeds are not independent profile observations.

All outputs are parseable except two EXAONE OFF dominance responses returning `B_ONLY`; these invalid verdicts remain in the denominator as errors. The two ON violations each occur in only one A/B orientation. Qwen3-8B overlooks a `B_ONLY` entry, whereas EXAONE acknowledges a four-to-one conflict but outputs its majority direction.

B.10 Executable identification alongside either inference mode

This interface implements the existing decision rule; it is not a new learned judge or an additional model experiment. The input is a complete, validated component vector and a versioned declared specification. For TradeoffBench, the four-way implementation of Equation (4) is:

```
if every coordinate equals zero: return EQUAL
if some coordinate is positive and none is negative: return A
if some coordinate is negative and none is positive: return B
otherwise: return UNDERDETERMINED
```

On the 3^5 ternary domain, this rule gives 31 A profiles, 31 B profiles, one EQUAL profile, and 180 UNDERDETERMINED profiles. These counts follow from the stated rule, not from new LLM inference. Both cases represented by zero in the mathematical rule withhold a strict direction, while the interface distinguishes equal components from unresolved conflict.

For general polyhedral $\mathcal{W}$, compute the two utility bounds in Equation (2) and apply Equation (3). HUMAINE's symmetric specification also has the closed form in Equation (8). A strict release requires a verified positive lower bound or negative upper bound. Incomplete inputs, an infeasible specification, or an unresolved numerical check require a separate input or solver-status flag, not a strict verdict or an assertion of genuine non-identification.

A reasoning-enabled judge may assess components or help formulate a proposed refinement. Record the component source, coding, specification version, interval or rule result, and any added premise. Make any refinement explicit before recomputing identification. Agreement with a target remains a separate policy-level audit. This interface does not require switching reasoning off. If the reasoning-enabled judge already executes the same rule correctly on the same inputs, the interface leaves the verdict unchanged. It adds an explicit support record and a defined route for further resolution.

The strict-Pareto implementation scans d coordinates in $O(d)$ operations and needs no further model generation after the structured inputs are available. This statement excludes component acquisition, preference elicitation, general LP solution, risk calibration, and review. It establishes the scope of the aggregation computation, not an end-to-end efficiency advantage or a measured improvement in target agreement with reasoning enabled.

C MultiPref: Component Substitution and Membership

MultiPref contains 10,461 comparisons, 41,844 annotations, and four evaluators per comparison. Its three component criteria are helpfulness, truthfulness, and harmlessness. Component preferences are encoded as $1, 0.5, 0, -0.5, -1$, with zero treated as coordinate equality. Overall preferences are obtained by collapsing response-strength labels to $-1, 0, +1$. A global-time cutoff and prompt exclusion yield disjoint development and evaluation partitions. Eighty evaluators meet the minimum of 50 development and 10 evaluation annotations. The main individual analysis contains 4,847 target-evaluator decisions over 3,561 comparisons and 2,939 prompts, including 3,515 strict overall targets.

Both conditions apply the strict-Pareto rule in Equation (4) to three components. The own-component condition uses the target evaluator's components. The peer-mean condition averages the other three evaluators' components coordinatewise, preserving an average of zero as coordinate equality. The overall target is unchanged. The earlier partition establishes evaluator eligibility; the main substitution contrast uses the evaluation partition. Coverage uses 4,847 eligible decisions, and strict-target coverage uses their 3,515 strict targets. Strict-direction risk is computed among the strict targets released by each core. Inference jointly resamples evaluator and prompt clusters. Component source and averaging change together, so this contrast does not isolate either mechanism.

C.1 Complete release and risk accounting

Figure 6 shows the composition changes underlying the aggregate rows in Table 14. Substitution adds 1,071 core members and removes 577, but 938 of the additions (87.6%) have target ties. Released strict targets therefore fall from 3,269 to 2,840 despite the increase in all-record coverage. The

shared-core comparison holds the 2,707 common strict targets fixed, separating its risk contrast from this membership change.

Table 14: Strict-direction risk under the strict-Pareto rule using the target evaluator's own components or the other three evaluators' mean. Overall target labels and the aggregation rule are fixed. Coverage uses 4,847 eligible evaluator decisions; strict-direction risk uses released strict targets in each row. The shared-core rows include 85 target ties; the main paired-risk contrast uses their 2,707 strict targets (Figure 3).

Support	Components	All releases	Coverage (%)	Strict rel./err.	Risk (%)
All eligible	Own	3,369	69.51	3,269 / 23	0.704
All eligible	Peer mean	3,863	79.70	2,840 / 847	29.824
Shared core	Own	2,792	–	2,707 / 16	0.591
Shared core	Peer mean	2,792	–	2,707 / 780	28.814

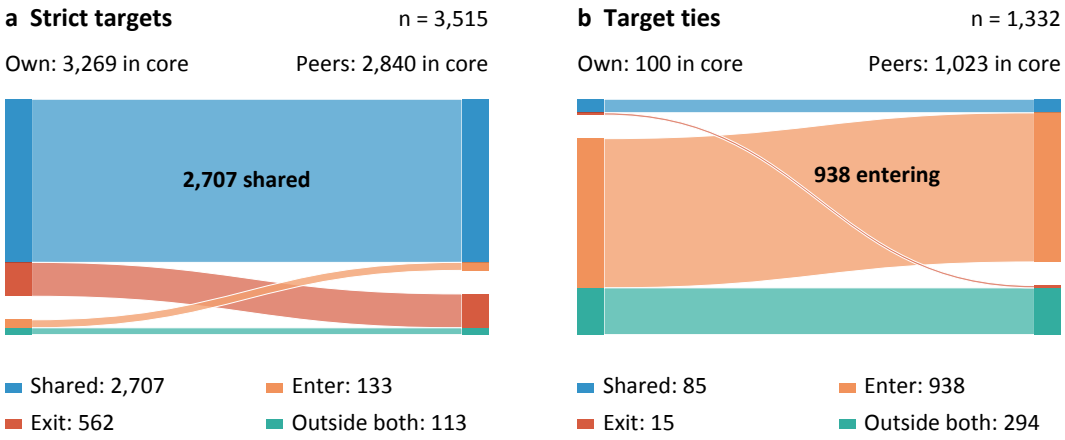

Figure 6: The net increase in core membership after substitution is concentrated among target ties. "Peers" denotes peer-mean components. Ribbons connect core membership under own and peer-mean components, separately for strict targets (a) and target ties (b). Upper nodes are in the core; lower nodes are outside it. Widths are proportional within each panel, and count keys retain all four membership strata. Target labels are fixed. Among strict targets, the own-only core has seven errors and the peer-only core has 67; shared-core errors are reported in Table 14.

C.2 Paired directional changes on shared targets

On the same 2,707 strict targets, 1,919 decisions agree with the target under both component conditions and eight oppose it under both. The other 780 records consist of 772 changes from agreement to opposition and eight from opposition to agreement. Thus own components have $8 + 8 = 16$ errors, while peer-mean components have $772 + 8 = 780$ errors. The difference of 764 errors on this fixed denominator corresponds to 28.223 percentage points in Figure 3. Joint evaluator–prompt resampling retains the within-record pairing (Appendix E).

C.3 Consistency across evaluators and time partitions

To assess variation across evaluators, we use all eligible evaluation records rather than restricting to the shared core. For each evaluator, residual error rate is the number of erroneous strict releases divided by that evaluator's total number of strict targets. Own components yield a lower residual error rate for 73 of the 80 evaluators, equal error rate for seven, and higher error rate for none. This comparison uses a common denominator within each evaluator, while allowing core membership to differ between component conditions.

Applying the same coding and strict-Pareto rule to the earlier partition provides a within-dataset temporal consistency check. Among 7,337 decisions from the same 80 evaluators, 5,394 have strict

targets. Own components yield 37 errors in 5,111 strict releases (0.724%); peer-mean components yield 1,307 errors in 4,410 (29.637%). The corresponding evaluation risks are 0.704% and 29.824% (Table 14). These rates condition on each core's released strict targets, unlike the evaluator-level residual-error denominator above.

D HUMAINE: GEOMETRY, QUALIFICATION, AND REVIEW

This appendix distinguishes three analyses: the releases of each fixed rule, qualification and selection over candidate policies, and prioritization of a fixed review workload. The geometry and numerical checks support Section 5. Contract sensitivities explain conditional selection, and ranking comparisons assess whether the structural frontier also provides an error ordering.

D.1 CORE GEOMETRY AND DECLARED VALUES

The lower bounds and normalization in Equation (6) imply $w_k \le 1 - 3(0.20) = 0.40$. Writing $w = (\mathbf{1} + v)/5$ with $v \ge 0$ and $\mathbf{1}^\top v = 1$ expresses $\mathcal{W}_{\mathrm{p}}$ as the convex hull of four criterion-emphasis profiles. Each vertex assigns weight .40 to one criterion and weight .20 to each of the other three. The minimum and maximum over these vertices are

$$\underline{u}_{\mathcal{W}_{\mathrm{p}}}(x) = 0.2\left(\sum_k x_k + \min_k x_k\right), \qquad \overline{u}_{\mathcal{W}_{\mathrm{p}}}(x) = 0.2\left(\sum_k x_k + \max_k x_k\right). \tag{8}$$

Under the ternary coding $x_k \in \{-1, 0, +1\}$, the positive core extends strict Pareto exactly by profiles with three components favoring A and one favoring B. The negative case is symmetric. Every added profile has equal-weight margin 0.50, giving the exact hybrid representation in Equation (7).

The asymmetric safety-protective set is

$$\mathcal{W}_{\mathrm{safe}} = \left\{ w \in \mathbb{R}^4_+ : \begin{array}{l} \mathbf{1}^\top w = 1, \quad w_k \ge 0.10, \\ w_{\mathrm{task}}, w_{\mathrm{comm}}, w_{\mathrm{interaction}} \le 0.40, \quad w_{\mathrm{safety}} \le 0.55 \end{array} \right\}. \tag{9}$$

For a three-versus-one profile, let $s = \mathrm{sign}(w_0^\top x)$ and let j be its sole opposing criterion. Then $w^\top(sx) = 1 - 2w_j$, giving the minimum gaps in Table 15. Among the retained three-versus-one records, the opposing criterion is Core Task Performance & Reasoning in 94, Communication Style & Presentation in 208, and Interaction Fluidity & Adaptiveness in 216. The safety-protective core releases 24,250 judgments: the 23,732 Pareto releases and these 518 additions, with no directional conflict.

Table 15: An asymmetric weight set distinguishes conflicts with equal scalar margins. These are the three-versus-one additions to Pareto in HUMAINE. Let $s = \mathrm{sign}(w_0^\top x)$ orient each profile toward its majority. For sole opposing criterion j, the minimum majority-aligned gap is $\underline{u}_{\mathcal{W}_{\mathrm{safe}}}(sx) = 1 - 2\max_{w \in \mathcal{W}_{\mathrm{safe}}} w_j$.

			$\mathcal{W}_{\mathrm{p}}$	$\mathcal{W}_{\mathrm{safe}}$	
Opposing criterion	Records	$\lvert w_0^\top x\rvert$	Decision	Min. gap	Decision
Non-safety	518	0.50	Release	$+0.20$	Release
Trust, Ethics & Safety	238	0.50	Release	-0.10	Defer

D.2 EXACT WEIGHT-SET SENSITIVITY

The symmetric path varies the minimum emphasis shared by all four criteria:

$$\mathcal{W}_\ell = \{w \in \mathbb{R}^4_+ : \mathbf{1}^\top w = 1,\ \ell \le w_k \le 1 - 3\ell\}, \qquad 0 \le \ell \le 0.25, \tag{10}$$

Writing $w = \ell\mathbf{1} + (1 - 4\ell)v$ with $v \ge 0$ and $\mathbf{1}^\top v = 1$ yields

$$\begin{aligned} \underline{u}_{\mathcal{W}_\ell}(x) &= \ell \sum_k x_k + (1 - 4\ell)\min_k x_k, \\ \overline{u}_{\mathcal{W}_\ell}(x) &= \ell \sum_k x_k + (1 - 4\ell)\max_k x_k. \end{aligned} \tag{11}$$

For each of the 81 ternary profiles, these bounds are affine in ℓ. Their zero crossings give the complete partition in Table 16. Decisions cannot change within an open interval between crossings. In particular, the minimum utilities for $(1, 1, 1, -1)$ and $(1, 1, 0, -1)$ are $6\ell - 1$ and $5\ell - 1$, respectively. Allowing the former compensation while withholding the latter therefore corresponds exactly to $1/6 < \ell \leq 1/5$. At $\ell = 1/6$, the three-versus-one minimum is zero; at $\ell = 1/5$, the two-versus-one minimum is zero. Neither zero supports a strict release. Thus $\mathcal{W}_{\mathrm{p}} = \mathcal{W}_{1/5}$ is the upper endpoint of this policy interval. Increasing ℓ beyond it changes the rule.

Table 16: The complete symmetric value path has four decision states. The partition is exact on all $3^4 = 81$ ternary profiles, including equalities and both directions. Counts describe profiles, not independent observations. Scalar rules are defined in Equation (6).

Parameter range	Equivalent release rule	Released profiles / 81
$\ell = 0$	E1.00	2
$0 < \ell \leq 1/6$	Strict Pareto	30
$1/6 < \ell \leq 1/5$	Symmetric partial	38
$1/5 < \ell \leq 1/4$	E.25	62

All four states belong to the existing seven-candidate simultaneous family. Under Bonferroni calibration at the 2.5% pooled strict-direction tolerance, the symmetric partial core has the highest coverage among passing states on this path (Table 17). The continuous characterization therefore requires no additional candidate rules or multiplicity correction.

Exact rational arithmetic and independent vertex enumeration agree on all 81 profiles. Replaying both HUMAINE releases reproduces the seven candidates' counts, calibration bounds, and policy selections. Throughout the partial interval, release directions and the review queue in Appendix D.5 are identical. These checks establish decision equivalence on the existing data, not independent replications.

The numerical audit uses 39 settings, including the twelve grid settings, exact transition points, interval interiors, and perturbations near the transitions. All twelve grid settings give the same directions under LP and exact arithmetic. Additional near-boundary probes expose conservative deferrals from the 10^{-10} release tolerance and a maximum LP utility error of 4×10^{-8}, with no unsupported strict releases. The partition in Table 16 uses exact signs; the numerical checks preserve the original solver and tolerance.

D.3 Global risk contracts and tie-aware calibration

Each candidate library contains E.25, E.50, and E.75 plus one alternative: E1.00 for scalar-only, strict Pareto for Pareto-first, the symmetric partial core for partial-value, or the safety-protective core for safety-protective. There are seven distinct candidates across these four libraries. Selection compares complete policies; a partial core is not a common filter applied to scalar candidates. The libraries are non-nested: the partial-value library replaces Pareto with the partial core rather than retaining both. Consequently, differences between selected libraries can also reflect different fallback candidates when a core fails qualification. Passing calibration qualifies a policy for the specified target-risk contract, not for entailment under another candidate's specification.

Pooled Bonferroni calibration uses a per-candidate confidence level of $1 - 0.05/7$. Within each library, selection maximizes all-record calibration coverage among candidates whose earlier-release upper bound passes the tolerance. Table 17 reports both this selection objective and the strict-target counts used for the risk bound. Evaluation metrics use the later release.

Where the added releases occur. Table 18 decomposes the symmetric partial release into the Pareto core and its three-versus-one additions. The added layer contains 723 strict targets and 28 directional errors, an observed risk of 3.873%. The complete policy has 224 errors among 23,160 released strict targets, or 0.967%. The pooled contract qualifies this complete policy; it does not separately bound the risk of the added layer.

For the shared worst-group contract in Table 19, correction covers the 6×7 group–candidate family. One global candidate is selected only if it passes every group-specific bound.

Table 17: Complete HUMAINE candidate audit for pooled strict-direction risk. All seven candidates enter one Bonferroni family. All-record coverage uses 21,044 calibration or 30,425 evaluation records. "Rel./err." gives released strict targets and opposite-direction errors. Coverage, UCB, and risk are percentages.

	Calibration			Evaluation		
Candidate	All cov.	Rel./err.	UCB	All cov.	Rel./err.	Risk
E.25	88.738	17,621 / 404	2.584	87.790	25,162 / 487	1.935
E.50	66.922	13,734 / 74	0.712	65.305	19,373 / 87	0.449
E.75	39.009	8,138 / 0	0.061	37.377	11,255 / 0	0.000
E1.00	18.214	3,820 / 0	0.129	17.302	5,237 / 0	0.000
Strict Pareto	77.827	15,543 / 173	1.337	78.002	22,437 / 196	0.874
Symmetric partial	80.783	16,140 / 199	1.462	80.486	23,160 / 224	0.967
Safety-protective	79.895	15,961 / 192	1.431	79.704	22,935 / 210	0.916

Table 18: Evaluation decomposition of partial release. Records include target ties; strict-direction risk uses only the released strict targets in each row. The added layer is descriptive, not a separately calibrated policy.

Region	Records	Strict targets	Errors	Risk (%)
Pareto core	23,732	22,437	196	0.874
Partial-only layer	756	723	28	3.873
Complete partial core	24,488	23,160	224	0.967

Table 19: HUMAINE selection under different risk contracts with Bonferroni correction. Entries give the calibration-selected rule / evaluation all-record coverage (%). Partial and Safety denote the symmetric and safety-protective cores. Every contract deploys one rule globally; shared worst-group selection additionally requires that rule to pass all six group bounds.

Contract / loss	Tolerance	Scalar-only	Pareto-first	Partial-value	Safety-protective
Pooled / directional	2.5%	E.50 / 65.305	Pareto / 78.002	Partial / 80.486	Safety / 79.704
Shared worst-group / directional	2.5%	E.50 / 65.305	E.50 / 65.305	E.50 / 65.305	E.50 / 65.305
Pooled / tie-aware	2.5%	E.75 / 37.377	E.75 / 37.377	E.75 / 37.377	E.75 / 37.377
	5.0%	E.50 / 65.305	E.50 / 65.305	E.50 / 65.305	E.50 / 65.305
	7.5%	E.50 / 65.305	Pareto / 78.002	Partial / 80.486	Safety / 79.704
	10.0%	E.25 / 87.790	E.25 / 87.790	E.25 / 87.790	E.25 / 87.790

D.4 MULTIPLICITY CORRECTION AND ASSUMED EFFECTIVE SAMPLE SIZE

We compare Bonferroni and Holm correction at eight assumed effective sample fractions: 1, .75, .50, .25, .10, .075, .05, and .025. Candidate rules, libraries, and selection objectives remain fixed. Both losses use tolerance .025 and family error level .05. Each loss, contract, and assumed fraction defines a separate comparison: seven tests for pooled qualification and 42 for shared worst-group qualification.

For a candidate with k calibration errors among n released targets eligible for the loss, we test the null that its risk is at least .025 using

$$p(k, n) = \Pr\{\text{Binomial}(n, .025) \leq k\}. \tag{12}$$

Bonferroni compares each value with $.05/m$, where m is the family size. Holm sorts them as $p_{(1)} \leq \cdots \leq p_{(m)}$ and rejects sequentially while $p_{(i)} \leq .05/(m - i + 1)$, stopping at the first failure (Holm, 1979). A shared worst-group candidate must pass all six group tests. Selection then maximizes calibration coverage within the unchanged library. If no candidate passes, the policy defers on every record. Holm eligibility is a family-adjusted test decision, not a candidate-specific Clopper–Pearson bound computed from its observed rank.

At nominal sample size, E.25 has $p = .03967893$ and is last in the seven-rule pooled directional family. It passes Holm after the other six rejections, so every library selects E.25. The Bonferroni bound instead excludes it (Table 17). This changes selection without changing the data, risk tolerance, or family error level.

Participant identifiers are unavailable, so the effective-sample analysis is hypothetical. For a nonempty calibration sample, set $n_{\text{eff}} = \max\{1, \lfloor qn \rfloor\}$ and $k_{\text{eff}} = \lceil (k/n) n_{\text{eff}} \rceil$; an empty sample cannot qualify. We recompute tests from these effective counts, leaving evaluation records unchanged. Table 20 reports the pooled directional comparison; Table 21 gives the remaining contracts. The two corrections differ at $q = 1$, .05, and .025.

Table 20: Correction and hypothetical effective information alter pooled directional selection. Entries are evaluation all-record coverage (%) at 2.5% tolerance. Differences use unrounded coverage and are in percentage points. Each q is a separate assumed information level.

	Bonferroni			Holm		
q	Pareto-first	Partial-value	Difference	Pareto-first	Partial-value	Difference
1.000	78.002	80.486	+2.485	87.790	87.790	0
0.750	78.002	80.486	+2.485	78.002	80.486	+2.485
0.500	78.002	80.486	+2.485	78.002	80.486	+2.485
0.250	78.002	80.486	+2.485	78.002	80.486	+2.485
0.100	78.002	80.486	+2.485	78.002	80.486	+2.485
0.075	78.002	80.486	+2.485	78.002	80.486	+2.485
0.050	78.002	65.305	−12.697	78.002	80.486	+2.485
0.025	37.377	37.377	0	65.305	65.305	0

For Bonferroni at tested $q \geq .075$, scalar-only selects E.50 and safety-protective selects its core, with 65.305% and 79.704% coverage. At .05, the partial-value and safety-protective libraries fall back to E.50 while Pareto still passes. At .025, every library selects E.75. Under Holm, the original four selections hold at tested $.05 \leq q \leq .75$; at .025, all select E.50.

The break-even calculation in Table 22 resolves the primary Bonferroni selection more finely than this grid. For each nominally passing candidate, q^* is the smallest effective fraction at which it passes for that and all larger effective counts. All four primary selections remain unchanged for $q \geq 0.0612$. Integer rounding is conservative at each evaluated count, but this stress test is not a formal dependence guarantee.

D.5 FIXED-BUDGET REVIEW ALLOCATION

The review workload contains the nonzero outputs of $h_0 = s_{.25,w_0}$: 18,674 earlier-release judgments for fitting the risk rankers and 26,710 later-release judgments for evaluation. The remaining 3,715

Table 21: Other 2.5% contracts under hypothetical effective sample sizes. Both corrections give the same selection in every row, and all four libraries select the same rule. Coverage uses all 30,425 evaluation records; defer-all has zero coverage and undefined conditional risk. Ranges refer only to the tested grid.

Contract / loss	Tested q	Selected rule	Coverage (%)
Shared worst-group / directional	$1, .75$	E.50	65.305
	.50	E.75	37.377
	.025–.25	Defer all	0
Pooled / tie-aware	.075–1	E.75	37.377
	$.05, .025$	Defer all	0
Shared worst-group / tie-aware	All eight settings	Defer all	0

Table 22: Break-even effective sample size for the 2.5% pooled strict-direction bound. The implied design effect is $1/q^*$. A dash denotes a candidate that fails at the nominal sample size.

Candidate	Nominal UCB	Stable q^*	Implied design effect
E.25	2.584%	–	–
E.50	0.712%	0.0304	32.94
E.75	0.061%	0.0241	41.52
E1.00	0.129%	0.0513	19.49
Strict Pareto	1.337%	0.0531	18.84
Symmetric partial	1.462%	0.0612	16.35
Safety-protective	1.431%	0.0585	17.09

later records have zero scalar outputs and are excluded by their components, not their target labels. The evaluation workload includes 1,548 target ties. Since the partial core's 24,488 judgments agree with h_0, the remaining $26{,}710 - 24{,}488 = 2{,}222$ determine the reference review budget using component evidence alone. Additional fractions of 1%, 2%, 5%, 10%, 15%, and 20% are rounded down to integer counts. The seven budgets are $b \in \{267, 534, 1335, 2222, 2671, 4006, 5342\}$ for every priority.

For each initial strict judgment i, define directional and tie-aware error indicators by

$$e_{i,\mathrm{dir}} = \mathbf{1}\{Y_i \neq 0,\ h_0(x_i) \neq Y_i\}, \qquad e_{i,\mathrm{tie}} = \mathbf{1}\{h_0(x_i) \neq Y_i\}. \tag{13}$$

There are 487 directional and 2,035 tie-aware errors. The partial queue contains the $487 - 224 = 263$ directional errors excluded by partial release in Table 3. A review priority sorts judgments by decreasing score. Scores tied at the budget boundary are sampled uniformly without replacement. Let q_i be the resulting inclusion probability, so $\sum_i q_i = b$. For loss $j \in \{\mathrm{dir}, \mathrm{tie}\}$, expected error capture is

$$\mathrm{Capture}_j(q) = \frac{\sum_i q_i e_{i,j}}{\sum_i e_{i,j}}. \tag{14}$$

We use the exact expectation over tied priorities, rather than one random tie-breaking realization. Uniform review uses a constant score, margin priority uses $-|w_0^\top x|$, and Pareto and partial priorities use $\mathbf{1}\{r_{\mathrm{P}}(x) = 0\}$ and $\mathbf{1}\{r_{\mathcal{S}_{\mathrm{p}}}(x) = 0\}$, respectively. Above frontier capacity, each binary priority fills uniformly from its core. The binary partial-frontier priority is identical to the deferral indicator of the Pareto-plus-margin release implementation in Equation (7), on all 81 ternary profiles and in every bootstrap replicate. This identity is distinct from the within-layer ordering used by Pareto then margin.

Pareto then margin orders lexicographically by decreasing $\mathbf{1}\{r_{\mathrm{P}}(x) = 0\}$ and then increasing $|w_0^\top x|$. Within the E.25 workload, all partial-frontier profiles are permutations or sign reversals of $(1, 1, -1, 0)$. Their margin is .25, whereas the other mixed-sign profiles have margin .50. Thus Pareto then margin prioritizes exactly the partial frontier before any other eligible profile. The two priorities have identical inclusion probabilities at every budget no greater than the observed frontier capacity. The linear capture path below this capacity is a consequence of uniform sampling from a fixed region, not a sequence of independent validations.

Each learned ranker predicts an error indicator from the four components, with separate earlier-release fits for the two losses. For a signed profile $v \in \{-1, 0, +1\}^4$, let $n(v)$ and $k_j(v)$ be its counts of earlier-release initial judgments and errors, and let $\bar{p}_j$ be the pooled earlier error rate. The profile risk estimate is $[k_j(v) + 20\bar{p}_j]/[n(v) + 20]$, retaining A/B orientation. Logistic risk uses 17 features: an intercept, four signed components, four absolute components, six pairwise products, the initial direction, and the absolute scalar margin. It minimizes summed binomial negative log-likelihood plus $\frac{1}{2}\|\beta_{\text{nonintercept}}\|_2^2$, using float64 Newton iterations with maximum absolute gradient tolerance 10^{-7} and a limit of 100 iterations. Target ties remain in the directional fit with error indicator zero. No ranker uses demographic attributes; later target labels are used only for evaluation.

Table 23: Complete fixed-budget review comparison. Entries are expected error-capture percentages on the same 26,710 initial E.25 judgments. Budgets correspond to 267, 534, 1,335, 2,222, 2,671, 4,006, and 5,342 slots. Panel a uses 487 directional errors; panel b uses 2,035 tie-aware errors. Learned rankers use separate earlier-release fits for each panel's loss. All other priorities use component evidence alone. "Pareto then margin" sorts within both Pareto layers and is distinct from the equivalent release rule in Equation (7).

Review priority	1%	2%	5%	8.32%	10%	15%	20%
a Directional errors							
Smallest margin	3.21	6.41	16.03	26.68	32.07	48.10	64.14
Pareto frontier	5.36	10.71	26.79	44.58	53.59	61.50	63.76
Partial frontier	6.49	12.98	32.45	54.00	54.85	57.36	59.86
Pareto then margin	6.49	12.98	32.45	54.00	57.42	66.01	74.15
Learned profile risk	13.70	25.76	38.46	53.21	59.32	69.67	75.46
Learned logistic risk	13.12	27.13	42.87	53.80	59.77	70.93	76.33
b Tie-aware errors							
Smallest margin	2.78	5.57	13.92	23.18	27.86	41.78	55.72
Pareto frontier	2.40	4.79	11.98	19.95	23.98	29.91	34.03
Partial frontier	2.85	5.70	14.26	23.73	25.13	29.29	33.45
Pareto then margin	2.85	5.70	14.26	23.73	25.51	37.33	51.10
Learned profile risk	4.50	8.45	20.41	32.21	36.99	50.68	62.71
Learned logistic risk	3.79	7.84	20.22	32.46	37.30	50.11	61.30

Table 23 reports complete point estimates across all budgets, including Pareto then margin; Table 24 gives reference-budget paired differences for the original binary priority. Five baselines, seven budgets, and two losses form a 70-comparison interval family; the exactly equivalent hybrid is excluded. These review comparisons are separate from the seven-candidate release-calibration family.

Table 24: Paired error-capture differences at the 2,222-slot reference budget. Values are percentage points, partial-frontier priority minus the named baseline. Intervals use 20,000 paired, group-stratified comparison bootstraps with budget reallocation. The last column adjusts percentile endpoints for 70 comparisons over seven budgets and two losses. All intervals condition on the earlier fits, group sizes, and budget.

Baseline	Difference	Pointwise 95% CI	Adjusted CI
Directional errors			
Uniform random	$+45.69$	$[40.97, 49.58]$	$[37.30, 52.56]$
Smallest margin	$+27.33$	$[23.03, 30.69]$	$[20.05, 33.42]$
Pareto frontier	$+9.42$	$[6.68, 11.20]$	$[4.69, 12.52]$
Learned profile risk	$+0.80$	$[-3.40, 4.18]$	$[-6.18, 7.00]$
Learned logistic risk	$+0.21$	$[-1.21, 0.88]$	$[-2.35, 1.68]$
Tie-aware errors			
Uniform random	$+15.42$	$[13.58, 17.00]$	$[12.29, 18.21]$
Smallest margin	$+0.56$	$[-1.17, 2.00]$	$[-2.40, 3.13]$
Pareto frontier	$+3.79$	$[2.86, 4.42]$	$[2.29, 4.89]$
Learned profile risk	-8.47	$[-10.61, -6.57]$	$[-12.07, -5.22]$
Learned logistic risk	-8.73	$[-10.67, -7.09]$	$[-12.22, -5.97]$

D.6 DOES WITHIN-LAYER ORDERING ADD INFORMATION BEYOND THE FRONTIER?

To distinguish a structural set from a complete risk ranking, we evaluate six further priorities on the unchanged workload and budget grid (Table 25). Robust gap sorts by increasing $\underline{u}_{\mathcal{W}_{\mathrm{p}}}(sx)$, where $s = \mathrm{sign}(w_0^\top x)$ is the initial direction. Partial then margin sorts within both partial layers by increasing scalar margin. Four learned variants sort within the partial or Pareto layers by decreasing values of the unchanged profile or logistic risk scores. These variants add no fitted parameters. Only the learned variants use earlier target labels. All final ties retain uniform inclusion probabilities.

By Equation (8), the robust gap is zero for every eligible partial-frontier profile. The next tier combines oriented $(1, 0, 0, 0)$ and $(1, 1, 1, -1)$ profiles, both with gap .20. Pareto then margin orders the latter first; partial then margin orders the former first. Robust gap therefore retains more ordering information than a binary core indicator, but does not distinguish cases within the partial frontier.

Table 25: Review-ranking extensions on the same seven budgets as Table 23. Entries are expected capture percentages. Robust gap and partial then margin use no target-risk fit; the other variants use the unchanged earlier fits for each loss. The corresponding unrefined baselines and Pareto then margin are in Table 23.

Review priority	1%	2%	5%	8.32%	10%	15%	20%
Directional errors							
Robust gap	6.49	12.98	32.45	54.00	56.83	65.25	73.67
Partial then margin	6.49	12.98	32.45	54.00	56.74	64.87	73.01
Partial then profile	13.70	25.76	39.43	54.00	60.08	71.59	77.64
Partial then logistic	13.12	27.13	42.87	54.00	59.71	70.93	76.33
Pareto then profile	13.70	25.76	39.43	52.09	54.96	72.84	79.55
Pareto then logistic	13.12	27.13	42.87	53.72	56.29	72.72	79.47
Tie-aware errors							
Robust gap	2.85	5.70	14.26	23.73	27.96	40.53	53.11
Partial then margin	2.85	5.70	14.26	23.73	28.36	42.13	55.90
Partial then profile	4.46	8.40	17.50	23.73	30.50	47.27	60.50
Partial then logistic	3.44	7.40	18.44	23.73	30.85	47.27	58.82
Pareto then profile	4.46	8.40	17.50	23.68	25.65	42.22	56.44
Pareto then logistic	3.44	7.40	18.44	23.73	25.26	42.22	56.44

At 20%, robust gap improves directional capture over the binary partial-frontier priority by 13.81 points, but differs from Pareto then margin by -0.48 points. Table 26 gives the paired uncertainty and both losses. The refinements do not establish a unique directional-ranking advantage of robust gap over within-layer margin ordering. For tie-aware errors, robust gap exceeds Pareto then margin but falls below partial then margin and both learned rankers. Learned cascades can also improve the binary priority. Their directional gains over the corresponding unlayered ranker do not establish consistent superiority over the matched Pareto cascade.

Table 26: Robust gap minus alternative priorities at 20% review budget. Differences and adjusted intervals are in percentage points. Intervals use 50,000 paired group-stratified replicates and Bonferroni percentile endpoints over the full 182-comparison extension family. Intervals spanning zero do not establish equivalence.

	Directional errors		Tie-aware errors	
Baseline	Difference	Adjusted CI	Difference	Adjusted CI
Partial frontier	$+13.81$	$[10.08, 17.82]$	$+19.66$	$[17.29, 22.27]$
Pareto frontier	$+9.91$	$[5.57, 14.37]$	$+19.08$	$[16.54, 21.75]$
Smallest margin	$+9.53$	$[5.90, 13.25]$	-2.60	$[-4.06, -0.97]$
Partial then margin	$+0.66$	$[-1.48, 3.27]$	-2.78	$[-3.72, -1.81]$
Pareto then margin	-0.48	$[-2.37, 1.08]$	$+2.01$	$[1.30, 2.73]$
Learned profile risk	-1.79	$[-6.84, 3.45]$	-9.59	$[-12.26, -6.56]$
Learned logistic risk	-2.66	$[-7.04, 1.88]$	-8.19	$[-10.57, -5.56]$

E STATISTICAL PROCEDURES AND SCOPE OF INFERENCE

TradeoffBench. Point estimates weight eligible source pairs equally. Ordinary-interface breadth uncertainty uses 2,000 percentile-bootstrap samples of latent tasks; controlled rates and contrasts use 20,000. Endpoint, factorial, and disjoint-label evidence results share the exchange audit's original-orientation Vector/disjoint observations. The main endpoint counts and intervals use all 1,345 mixed-sign pairs; the 849-pair diagnostics resample that subset separately. Exchange intervals use 20,000 task-cluster bootstrap replicates, keeping both orientations together. Their consistency denominator includes every pair in the stated population. These intervals describe latent-task sampling, not variation across model families or uniform sampling of distinct categorical vectors.

The majority and fixed-coordinate audit reuses these saved outputs without new inference. Its pair- and profile-weighted summaries (Appendix B.8) describe the archive; the 577 outcomes are model–pair events, not independent pairs. Exact rule agreement is separate from the task-bootstrap intervals for consistency frequency.

The exhaustive-profile summaries in Appendix B.9 concern repeated presentations of a finite grid, not independent draws of new profiles. The ON mixed-sign denominator is $180 \times 2 \times 3 \times 3 \times 3 = 9,720$; the control denominator is $63 \times 2 \times 3 \times 3 \times 3 = 3,402$, covering orientations, encodings, seeds, and models. Completion-length maxima are descriptive statistics of these runs. The deterministic interface in Appendix B.10 is a rule implementation, not a further empirical ON/OFF comparison.

MultiPref component substitution. Risk contrasts use 20,000 bootstrap replicates with crossed evaluator and prompt clusters. Own-component and peer-mean decisions receive the same bootstrap weights, retaining within-record pairing. The shared-core contrast conditions on records released in both conditions. Its risk difference is therefore not driven by entry into or exit from the core, although component source and averaging still change jointly.

HUMAINE release and calibration. Earlier-release comparison counts yield exact one-sided Clopper–Pearson upper bounds, corrected over seven distinct candidates (42 group–candidate bounds for the shared worst-group contract). The later release supplies evaluation metrics. Coverage contrasts use 20,000 paired, group-stratified bootstraps conditional on the selected policies. For the 2.485-point partial-minus-Pareto coverage difference, the pointwise 95% interval is $[2.311, 2.662]$ and the simultaneous interval is $[2.084, 2.886]$. The latter uses the configured maximum-deviation band across contract–loss–tolerance contrasts. Bonferroni and Holm sensitivities use the same finite families and Equation (12). Effective counts are rounded before testing, and each assumed fraction is analyzed separately. Qualification assumes independent comparison units and release-to-release transport of the relevant population. Participant identifiers are unavailable: the effective-sample analysis is a hypothetical stress test, not an estimated participant-cluster correction.

HUMAINE review allocation. For Appendix D.5, 20,000 replicates resample initial strict judgments within the six country–age groups, preserving their observed sizes. Every original priority, budget, and loss shares the draws. Each replicate reallocates the fixed integer budget using unchanged scores and recomputes expected error capture. Five original baseline contrasts across seven budgets and two losses form the 70-comparison interval family. Paired intervals condition on the earlier-release fits, group sizes, and budgets; they do not include refitting on new earlier data.

Review-ranking extensions. The extensions use 50,000 shared replicates. Their 182-comparison family contains thirteen contrasts at seven budgets for two losses. Seven compare robust gap with the original partial, Pareto, margin, two learned, and two within-layer margin priorities. The other six compare each partial learned cascade with the original partial priority, its unlayered learned ranker, and its matched Pareto cascade. Each priority is recomputed separately. Partial frontier and Pareto then margin have identical inclusion probabilities by construction when the budget does not exceed the replicate's frontier capacity. Beyond that capacity, the priorities can assign different inclusion probabilities. Their equality at the observed 2,222-slot capacity therefore does not imply equality in every resample. Pointwise intervals use the 2.5th and 97.5th percentiles. For family size M, adjusted percentile probabilities are $.05/(2M)$ and $1 - .05/(2M)$. These are comparison-level bootstrap intervals without a participant-cluster correction. Capture quantifies errors placed in a queue, not observed review outcomes.