

Every LLM-as-Judge Aggregator Must Sacrifice a Fairness Axiom

Yang Liu² Dongxin Guo¹ Li Ning^{1,2} Yu Wu²
Siu Ming Yiu^{1,2,3} Jikun Wu¹ *

¹Brain Investing Limited ²Stellaris AI Limited ³The University of Hong Kong
{bettyguo, jacobwu}@braininvestai.com smyiu@cs.hku.hk
{lining, yangliu, yuwu}@stellaris-ai.com

Abstract

LLM-as-judge leaderboards now decide what gets trained, deployed, and funded, yet years of bias patching have not asked whether the goals those patches pursue are jointly achievable. We show they are not. Casting LLM-as-judge evaluation as a social welfare functional over score profiles indexed by judge, criterion, and alternative, we prove that no aggregation rule can simultaneously satisfy four fairness axioms (Universal Domain, Weak Pareto, Independence of Irrelevant Alternatives [IIA], and Judge-Anonymity) in the ordinal regime and, among scale-invariant aggregators, in the cardinal regime. A fifth axiom, Position-Invariance, captures upstream pipeline bias and is securable independently. The impossibility admits a constructive counterpart: a four-face achievability frontier with one witness per dropped axiom. The Borda-Conformal Hybrid (BCH) occupies the dropped-IIA face, satisfies the other four axioms, and admits a structural conformal bound relating its IIA-violation probability to the calibration-set gap distribution. A 200-prompt audit on MT-Bench, AlpacaEval 2, and Arena-Hard finds IIA-violation rates from 18.5% to 38.0% across the seven aggregators with measurable IIA; a 50-prompt human-jury comparator confirms a non-zero human IIA-violation baseline consistent with known ranking pathologies in the choice-experiment literature. An audit toolkit and reporting protocol accompany the paper.

1 Introduction

LLM-as-judge (Zheng et al., 2023; Liu et al., 2023) has become the load-bearing component of open-ended language model evaluation. When a new instruction-tuned model is released today, its position on MT-Bench, AlpacaEval 2, and Arena-Hard (Chiang et al., 2024; Li et al., 2025; Dubois et al., 2024) is reported within hours, and these

positions inform downstream decisions about training recipes, deployment, and even funding. The convenience is real. Human evaluation at the scale needed to differentiate frontier models is prohibitively expensive, and strong LLM judges agree with crowdsourced human preferences at over 80% (Zheng et al., 2023), comparable to human-human agreement on the same prompts.

The convenience is also fragile. A steady stream of work has shown that judge models are not neutral. They prefer responses in particular positions of the prompt (Shi et al., 2025; Wang et al., 2024a), prefer longer responses (Saito et al., 2023), prefer their own outputs (Wataoka et al., 2024; Zheng et al., 2023), and shift their verdicts under cosmetic changes to the rubric, the scoring scale, or the candidate identifiers (Ye et al., 2025; Li et al., 2026; Thakur et al., 2025). Recent work has formalised two further pathologies: Feng et al. (2025) introduce Intra-Pair Instability, the rate at which a judge flips its verdict when the presentation order of the candidate pair is reversed, and Weak Total-Order Violation, the minimum number of preference changes needed to transform a judge’s pairwise preferences into a consistent total order; Huang et al. (2026a) show that the top of Chatbot Arena can flip after dropping just a handful of preferences. The field has responded with a portfolio of patches: re-prompting in both orders, length controllers, mechanistic debiasing (Wang et al., 2025), judge ensembles, and a steady stream of fine-tuned judges (Wang et al., 2024b; Zhu et al., 2025; Li et al., 2024; Kim et al., 2024b).

We ask whether any of this is escapable in principle. The question is not whether a particular bias can be reduced. It is whether the goals of fair, stable, multi-criterion aggregation that every leaderboard implicitly assumes are jointly achievable by *any* aggregation rule, given the basic mechanics of how LLM judges produce numbers.

Social choice theory was built around exactly

* Corresponding author.

this kind of question. Arrow’s impossibility theorem (Arrow, 1951) states that no rule for aggregating ordinal preferences over three or more alternatives can simultaneously satisfy Universal Domain, Weak Pareto, Independence of Irrelevant Alternatives, and Non-Dictatorship. Sen (1970) and D’Aspremont and Gevers (1977) extended the framework to cardinal scores, characterizing which informational structures restore possibility and which preserve impossibility. The framework has been used for committee voting, fair allocation, and most recently for AI alignment with diverse human feedback (Conitzer et al., 2024; Chakraborty et al., 2024; Fish et al., 2024; Dai and Fleisig, 2024; Ge et al., 2024). To our knowledge it has not been turned on the LLM-as-judge problem itself.

This is what we do. We cast LLM-as-judge as a social welfare functional (SWFL) over score profiles indexed by judge, criterion, and alternative, then state five axioms. Four are direct analogues of standard social-choice axioms acting on the aggregator: Universal Domain, Weak Pareto, Independence of Irrelevant Alternatives, and Judge-Anonymity. The fifth, Position-Invariance, is specific to LLM judges and acts on the score-generation pipeline: it formalizes the expectation that swapping the presentation order of two candidates should not flip a judge’s verdict. Empirically, this expectation fails routinely, and the failure is structurally tied to how LLM judges read prompts in sequence.

Contributions. We make four contributions, each verifiable from a single section or table.

(i) *A two-level decomposition that locates Position-Invariance where it actually fails.* Separating the score-generation pipeline Π from the aggregator F lets us put Position-Invariance on Π and the four classical fairness properties on F , rather than folding all of them into a generic “robustness” bucket (Section 3, Figure 1).

(ii) *An Arrow-style impossibility and its constructive achievability frontier.* No aggregator satisfies $\{\text{UD}, \text{WP}, \text{IIA}, \text{A}\}$ simultaneously in the ordinal regime, and no CUC-invariant aggregator satisfies the four-axiom set in the cardinal regime under ordinal IIA (Theorems 1, 2). The frontier admits a constructive witness on each dropped-axiom face (Theorem 3), reframing Borda, Score-Mean, Best-of-Judge, Majority Vote, MT-Bench Avg, and permutation-consensus as occupants of specific faces. Criterion-Monotonicity is not needed for

the impossibility (Section 4.5).

(iii) *The Borda-Conformal Hybrid as one deliberate frontier choice, with a gap-marginalized IIA bound.* BCH is a composite (Π, F) pair that satisfies UD, WP, A, CM, and PI by construction; Proposition 2 derives a structural conformal bound connecting the IIA-violation probability to the gap distribution. BCH’s 12.5-point IIA reduction over Borda decomposes into 7.0 points from order-symmetrization (shared with permutation-consensus) and 5.5 from the median-over-judges aggregation (Section 5).

(iv) *A 200-prompt audit, a human-jury comparator, and a community reporting protocol.* The audit measures how often each permitted violation occurs: across the seven aggregators with measurable IIA, violation rates span 18.5% to 38.0%, and three aggregators violate more than one axiom. A 50-prompt human-jury comparator anchors these numbers against a human IIA-violation baseline that is itself non-zero, consistent with known ranking pathologies (asymmetric-dominance and compromise effects; Huber et al., 1982; Simonson, 1989); the precise human rate depends on the aggregation procedure and analysis unit (Section 6). The six-cell axiom-reporting protocol and audit toolkit are released as community artifacts.

Mission statement. We do not claim LLM judges are useless. We claim that no aggregation rule can be free of all four aggregator-level pathologies at once, that the field has been quietly choosing which to live with, and that the choice should be made deliberately and reported transparently.

The impossibility holds over the unrestricted domain of score profiles; whether the profiles that LLM judges actually produce satisfy a domain restriction (e.g., single-peakedness) that would restore possibility depends on the realised profile domain.

2 Background and Related Work

LLM-as-judge, its biases, and robustness diagnostics. The paradigm was crystallised by Zheng et al. (2023), with MT-Bench showing GPT-4 agrees with crowdsourced human preferences at over 80%; Chatbot Arena (Chiang et al., 2024) extends to live pairwise battles via Bradley-Terry, Arena-Hard (Li et al., 2025) distils the hardest prompts into a static benchmark, and AlpacaEval 2 (Dubois et al., 2024) adds length control. Pointwise variants include G-Eval (Liu et al., 2023)

and FLASK (Ye et al., 2024); the fine-tuned judge family includes PandaLM, JudgeLM, Auto-J, and Prometheus 1 and 2 (Wang et al., 2024b; Zhu et al., 2025; Li et al., 2024; Kim et al., 2024a,b). A parallel line documents systematic biases: position (Shi et al., 2025), verbosity (Saito et al., 2023), self-preference (Wataoka et al., 2024), rubric ordering (Li et al., 2026), and the CALM bias-quantification framework (Ye et al., 2025). Liusie et al. (2024) comprehensively analyse pairwise comparative assessment for zero-shot NLG evaluation; Sheng et al. (2025) apply conformal prediction to judge-score uncertainty; Zhao et al. (2026) extend the audit to code. Feng et al. (2025) introduce Intra-Pair Instability and Weak Total-Order Violation as logical-coherence metrics, Huang et al. (2026a) show the top of Chatbot Arena is sensitive to dropping 0.003% of preferences, cheating-based audits (Zheng et al., 2025) show null models can game current benchmarks, and reward-model and judge-evaluation benchmarks (Lambert et al., 2025; Malik et al., 2026; Tan et al., 2025) extend evaluation to alignment signals. Huang et al. (2026b) propose permutation-consensus, an order-symmetrizing pipeline method originally designed for listwise factuality evaluation, as one robustness fix.

Social choice, classical and for AI. Conitzer et al. (2024) argue social choice theory should guide alignment with diverse human feedback; Chakraborty et al. (2024) prove a single-reward RLHF impossibility and propose MaxMin; Fish et al. (2024), Qiu (2024), Dai and Fleisig (2024), and Ge et al. (2024) develop social-choice machinery for preference aggregation in alignment; Yang et al. (2024) studies LLM voting empirically. Beyond Arrow, the classical results we use are Sen’s extension to cardinal scores under various measurability classes (Sen, 1970, 1977; D’Aspremont and Gevers, 1977), Harsanyi’s utilitarian theorem (Harsanyi, 1955), Maskin’s theorem on utilitarianism (Maskin, 1978), and Roberts’ welfarism theorem (Roberts, 1980); the decisive-coalition / pivotal-voter argument (Arrow, 1951; Geanakoplos, 2005) drives the ordinal case.

Human IIA violations as a calibration anchor. Choice experiments since Huber et al. (1982) and Simonson (1989) document that human raters routinely violate Independence of Irrelevant Alternatives under asymmetric-dominance and compromise manipulations; Tversky (1972) provides the

canonical elimination-by-aspects account. The human IIA-violation rate is therefore a positive baseline, not zero; the relevant question for an LLM-judge audit is the size of the LLM-specific amplification. Section 6.3 measures both rates and states what a precise comparison of them requires.

Leaderboard aggregators and existing patches. In our multi-criterion reconstruction, the IIA-violation rates of major leaderboard aggregators are 38.0% for MT-Bench (cross-criterion mean), 29.5% for Arena-Hard (Bradley-Terry), and 24.5% for AlpacaEval 2 (LC win rate); these rates are for our reconstruction, not the deployed systems. Existing patches target different axioms (position-swap re-prompting secures PI; AlpacaEval 2’s length controller trades IIA for reduced verbosity bias), and the framework formalizes these as movement along a single frontier (Section 4.4).

Position of this paper. Prior social-choice-for-AI work aggregates human preferences for alignment; we aggregate LLM-judge scores for evaluation, decomposing pipeline from aggregator.

3 Formal Framework

3.1 Score profiles, pipelines, and aggregators

Three finite sets fix the setting: alternatives $\mathcal{X}$ with $|\mathcal{X}| = m$, criteria $\mathcal{C}$ with $|\mathcal{C}| = k$, and judges $\mathcal{J}$ with $|\mathcal{J}| = n$. Throughout $m \geq 3$, $k \geq 2$, $n \geq 2$. A *score profile* is a map $s : \mathcal{J} \times \mathcal{C} \times \mathcal{X} \rightarrow \mathbb{R}$, where $s_{j,c}(x)$ is the score that judge j assigns to alternative x on criterion c ; $\mathcal{S}$ denotes the set of profiles. Each profile induces, for every (j, c) , a complete weak ordering $\succsim_{j,c}$ on $\mathcal{X}$.

To translate the LLM-as-judge setting into a form classical social choice can act on, we introduce the *virtual electorate* $\mathcal{V} := \mathcal{J} \times \mathcal{C}$ of $N := nk$ virtual voters, with v ranging over $\mathcal{V}$ and each $v = (j, c)$ carrying preference $\succsim_v := \succsim_{j,c}$. The virtual electorate is the bridge that lets us apply Arrow-style machinery to a setting with only a handful of judges and criteria.

Two levels. A score profile is the output of a score-generation pipeline acting on a prompt context. To keep operations cleanly separable, define a *score-generation pipeline* $\Pi : \mathcal{P} \times \mathcal{O} \rightarrow \mathcal{S}$ mapping a prompt context $p \in \mathcal{P}$ and a presentation ordering $\pi \in \mathcal{O}$ to a profile $s = \Pi(p, \pi)$, and an *aggregator* $F : \mathcal{S} \rightarrow \mathcal{R}(\mathcal{X})$ mapping profiles to complete transitive weak orderings $\mathcal{R}(\mathcal{X})$. The composite is $G(p, \pi) := F(\Pi(p, \pi))$. The two-level

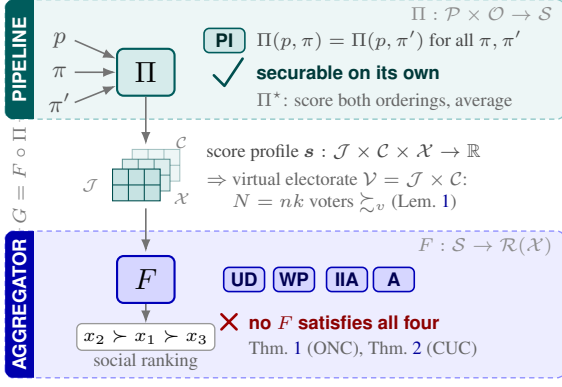


Figure 1: Two-level decomposition $G = F \circ \Pi$ and where each axiom binds. The pipeline Π turns a prompt context p and a presentation ordering π into a score profile s ; the aggregator F turns s into a social ranking. Position-Invariance constrains Π alone and is securable there by order-symmetrization (Π^*); the four classical axioms constrain F alone and are jointly unsatisfiable. The profile is the interface between the two levels: reading it as an electorate of $N = nk$ virtual voters (j, c) is what lets Arrow-style machinery apply.

split lets us locate each axiom at the layer it actually constrains: four classical aggregator axioms on F , and Position-Invariance, the LLM-specific axiom, on Π (Figure 1). Existing leaderboards each fix some specific (Π, F) : MT-Bench Avg uses a single-presentation pipeline plus within-judge cross-criterion mean; Arena-Hard uses a single-presentation pipeline plus Bradley-Terry on pairwise outcomes; Best-of-Judge uses any pipeline plus a single judge as oracle.

3.2 The five axioms

We state five axioms: four on the aggregator F (classical, going back to Arrow (1951) and Sen (1970)) and one on the pipeline Π (LLM-specific). Criterion-Monotonicity, which earlier work might include in the impossibility set, plays no role in our proofs and is handled separately in Section 4.5.

(UD) Universal Domain (aggregator-level). F is defined for every $s \in \mathcal{S}$, and $F(s)$ is a complete transitive weak order.

(WP) Weak Pareto (aggregator-level). If $s_v(x) > s_v(y)$ for every $v = (j, c) \in \mathcal{V}$, then $x \succ_F y$.

(IIA) Independence of Irrelevant Alternatives (aggregator-level, ordinal). For any pair $\{x, y\}$ and any two profiles s, s' that induce the same pairwise ordering $\succsim_v|_{\{x, y\}}$ for every v , $F(s)|_{\{x, y\}} = F(s')|_{\{x, y\}}$. The cardinal version (where two

profiles must additionally agree in magnitude on $\{x, y\}$) is strictly weaker; LLM-judge scores are not interpersonally cardinally comparable, so we commit to the ordinal version. Ordinal IIA sits at the strict top of a hierarchy of pair-stability axioms (cardinal IIA, Kendall- τ stability, Mallows-distance stability); Appendix A.5 discusses where the present frontier characterization sits in this lattice and how BCH’s gap-marginalized bound (Section 5) re-reads under the weaker metric stabilities. Appendix A.6 states the cardinal-IIA analogue of Theorem 3.

(A) Judge-Anonymity (aggregator-level). For every permutation σ of $\mathcal{J}$, $F(\sigma \cdot s) = F(s)$ where $(\sigma \cdot s)_{j,c}(x) := s_{\sigma(j),c}(x)$. Judge-Anonymity is stronger than the impossibility requires (Non-Dictatorship suffices; see Appendix E), but it is the axiom the frontier needs, since the drop-A face is only meaningful if Anonymity is worth keeping. A practitioner with per-judge reliability evidence has Best-of-Judge available, with its axiomatic cost reported.

(PI) Position-Invariance (pipeline-level). For every $p \in \mathcal{P}$ and any orderings $\pi, \pi' \in \mathcal{O}$, $\Pi(p, \pi) = \Pi(p, \pi')$. In practice, Π^* secures PI under the order-reversal involution; the full quantification over all π, π' is the formal axiom, and Π^* is a sufficient condition for the two-ordering case used in our audit. PI has no aggregate-scale analogue in classical social choice. Ballot-order effects on human voters are documented in election science (the primacy effect; Koppell and Steen, 2004), but they are bounded and second-order. LLM-judge order-sensitivity flips pairwise verdicts at rates of 12.5–34.0% across judges and criteria in our audit (Table 5; Shi et al., 2025), an effect strong enough to require its own axiom.

A composite system (Π, F) satisfies the four aggregator-level axioms iff F does, and PI iff Π does. The PI column in Table 2 reports order-sensitivity: for aggregators not composed with Π^* , the rate depends on both the judge-level verdict-flip rate and the aggregator’s sensitivity to those flips; aggregators composed with Π^* have PI = 0 by construction. Theorems 1, 2 concern F alone; Theorem 3 concerns the composite.

Running example. A concrete illustration of each axiom (two judges, two criteria, three candidates, with typical violations) is in Appendix F.

Informational regimes. We work in the Sen-d’Aspremont-Gevers framework of informational invariance classes. Three regimes matter. *Ordinal Non-Comparable (ONC)*: scores interchangeable up to a judge-specific monotone transformation; only orderings matter. *Cardinal Unit-Comparable (CUC)*: scores interchangeable up to a common positive affine transformation; differences comparable, levels not. *Cardinal Fully Comparable (CFC)*: a common affine transformation applied identically; both levels and differences comparable. LLM-judge 1–10 rubrics have anchors that shift across criteria and model releases; the practical regime sits between ONC and CUC. Theorems 1, 2 cover both. CFC admits aggregators (leximin, utilitarianism) that escape the impossibility by reading cardinal levels, but these violate ordinal IIA (which forbids reading anything beyond pairwise order); the Limitations section discusses what realising CFC calibration would cost in practice.

4 Impossibility and Frontier

4.1 The virtual-voter reduction

We begin with a structural lemma that translates the LLM-as-judge setting into a virtual-electorate problem.

Lemma 1 (Virtual-voter reduction). *Under UD and IIA, the SWFL F is determined by the matrix of pairwise virtual-voter preferences $\{\succsim_v |_{\{x,y\}}\}_{v \in \mathcal{V}, x,y \in \mathcal{X}}$. Two profiles s, s' that induce the same matrix of pairwise preferences across $\mathcal{V}$ produce the same social ranking.*

The proof, in Appendix A.1, follows directly: UD ensures F is defined and returns a complete weak order on every profile; IIA restricts the pairwise output of F to depend only on pairwise inputs; completeness and transitivity close the system.

4.2 Four-axiom impossibility, ordinal regime

Theorem 1 (Ordinal impossibility). *With $|\mathcal{X}| \geq 3$ and $n \geq 2$, no aggregator $F : \mathcal{S} \rightarrow \mathcal{R}(\mathcal{X})$ satisfies all of $\{\text{UD}, \text{WP}, \text{IIA}, \text{A}\}$, where IIA is read in its ordinal form (Section 3.2). The result is unrestricted: F ranges over all aggregators, not only ONC-invariant ones.*

Proof sketch. Lemma 1 reduces F to a Boolean aggregator on the pairwise virtual-voter matrix; the decisive-coalition argument (Arrow, 1951; Geanakoplos, 2005) forces F to be a dictatorship at some $v^* = (j^*, c^*)$. A profile in which judges j^*

and j' disagree on $\{x, y\}$ under criterion c^* contradicts the dictatorship under Judge-Anonymity: the transposition $(j^* j')$ flips the dictator’s expressed preference while leaving F fixed by A, forcing $x \succ_F y$ and $y \succ_F x$ in the same ranking. Full proof in Appendix A.2. $\square$

We note that Theorem 1 is a corollary of Arrow’s theorem applied to the virtual electorate $\mathcal{V} = \mathcal{J} \times \mathcal{C}$, with Judge-Anonymity replacing Non-Dictatorship (Appendix E). The contributions specific to the LLM-judge setting are the virtual-voter bridge (Lemma 1), the CUC-regime companion (Theorem 2), and the two-level pipeline/aggregator decomposition.

4.3 Ordinal IIA on CUC-invariant aggregators

Theorem 1 already implies the result below (any aggregator satisfying UD and ordinal IIA is automatically CUC-invariant, and Strong Pareto implies Weak Pareto). The value of stating Theorem 2 separately is constructive: its proof characterises the criterion-weighted utilitarian family (Equation 4) as the unique survivor under cardinal IIA, which the ordinal-IIA witness (Appendix A.3, Steps 3–3a) then refutes.

Theorem 2 (Cardinal impossibility). *In the CUC informational regime, with $|\mathcal{X}| \geq 3$, $n \geq 2$, and $k \geq 2$, no aggregator $F : \mathcal{S} \rightarrow \mathcal{R}(\mathcal{X})$ that is CUC-invariant and satisfies Strong Pareto satisfies all of $\{\text{UD}, \text{IIA}, \text{A}\}$, where IIA is read in its ordinal form. (Theorem 1 uses only Weak Pareto; Strong Pareto is standard in the welfarism literature and is satisfied by the drop-IIA witnesses (Borda, Score-Mean) in Theorem 3.)*

Proof sketch. The welfarism framework of D’Aspremont and Gevers (1977) and Roberts (1980) shows that any SWFL satisfying $\{\text{UD}, \text{SP}, \text{cardinal-IIA}, \text{CUC-invariance}\}$ reduces to the weighted-utilitarian family $F(s) = \text{argsort}(\sum_v \omega_v s_v(\cdot))$ with $\omega \in \mathbb{R}^N$. Judge-Anonymity collapses $\omega_{(j,c)}$ to a criterion-only weight ω_c ; Strong Pareto forces $\omega_c > 0$ for every c . Ordinal IIA is strictly stronger than cardinal IIA: profiles with identical pairwise restrictions can differ in magnitudes, and the weighted-utilitarian rule aggregates magnitudes, so ordinal IIA fails on a constructible witness pair (Appendix A.3). $\square$

Dropped axiom	Witness aggregator
UD	Majority Vote over $\mathcal{V}^\dagger$
WP	Constant rule $F(s) \equiv \succsim^\circ$
IIA	Borda, Score-Mean [‡]
A	Best-of-Judge

Table 1: Aggregator-level frontier witnesses (Theorem 3). [†]Pairwise majority over the virtual electorate $\mathcal{V}$; fails UD on Condorcet-cyclic profiles. [‡]Borda and Score-Mean can disagree on near-tied profiles; both are listed.

4.4 The achievable frontier

The four-axiom aggregator-level impossibility has a natural counterpart: a characterization of which four-axiom subsets of $\{\text{UD}, \text{WP}, \text{IIA}, \text{A}, \text{PI}\}$ are achievable by a composite system (Π, F) . The pipeline-level axiom PI is structurally orthogonal: securing PI is a property of Π and is independent of which aggregator-level axioms F satisfies. The characterization comes out cleanly.

Theorem 3 (Frontier characterization). *Let $\Phi = \{\text{UD}, \text{WP}, \text{IIA}, \text{A}, \text{PI}\}$, with the first four aggregator-level and PI pipeline-level. In both the ONC and CUC regimes:*

1. *The 4-subset $\Phi \setminus \{\text{PI}\} = \{\text{UD}, \text{WP}, \text{IIA}, \text{A}\}$ is unsatisfiable (Theorems 1, 2).*
2. *For each $\phi \in \{\text{UD}, \text{WP}, \text{IIA}, \text{A}\}$, there exists a composite witness (Π^*, F_ϕ) satisfying $\Phi \setminus \{\phi\}$, where Π^* is the order-symmetrizing pipeline. The witness aggregators are listed in Table 1.*

The four aggregator-level witnesses are listed in Table 1; constructions are verified in Appendix A.4. PI is secured at the pipeline level by Π^* , which re-prompts judges in forward and reverse presentation order and averages the resulting score profiles (the same Π used by BCH; Section 5). Figure 2 depicts the frontier.

We highlight the IIA-sacrifice face as the practically relevant one. UD, WP, A, and PI all encode constraints the community considers non-negotiable for a fair leaderboard. IIA, the alternative-independence axiom, is the one that humans themselves routinely violate in choice experiments (asymmetric-dominance and compromise effects; Huber et al., 1982; Simonson, 1989), and LLM judges inherit the same context-sensitivity; the human-jury comparator in Section 6.3 establishes a non-trivial human baseline. The face

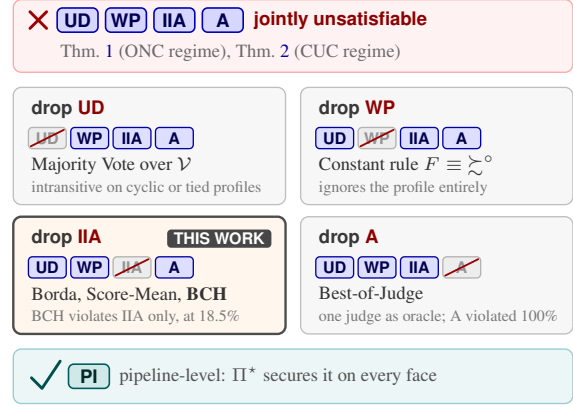


Figure 2: The achievable frontier (Theorem 3). The four aggregator-level axioms cannot be had together (top band), so every aggregator sits on one of four faces, each defined by the single axiom it gives up; the witness in each card shows that face is non-empty. PI is orthogonal to the choice of face and is secured at the pipeline level by Π^* (bottom band), so it is not part of the trade-off. This work takes the drop-IIA face: BCH keeps UD, WP, A and PI, and pays 18.5% IIA violations in the audit (Table 2).

is constraint-based, not rule-based: Borda (rank-based aggregation) and Score-Mean (magnitude-based) are two distinct witnesses on this same face and produce empirically different IIA-violation rates on real data (Section 6, 31.0% vs 28.5%), even though both are valid witnesses to the axiomatic claim. The next section turns this observation into a constructive aggregator.

4.5 Criterion-Monotonicity as a separately-measured empirical axiom

A natural sixth axiom, Criterion-Monotonicity (CM), states that raising the score of an alternative on any criterion cannot demote it socially. CM plays no role in Theorems 1, 2 (the four-axiom set is already unsatisfiable; Appendix A.2, Note on CM); we retain CM as a separately-measured empirical property because CM violations would produce counter-intuitive rank flips; our probe found none among the eight audited aggregators, all of which are monotone in practice (Section 6).

5 The Borda-Conformal Hybrid

BCH sacrifices Independence and secures the other four axioms: Π_{BCH} secures PI by order-reversal symmetrization; F_{BCH} secures UD, WP, A, and CM by construction (Proposition 1). Of its 12.5-point IIA reduction over Borda, order-symmetrization contributes 7.0 points (a

gain shared with permutation-consensus) and the median-over-judges aggregation the remaining 5.5 (Table 3).

5.1 Construction

Pipeline Π_{BCH} . Given any (p, π) , Π_{BCH} re-prompts each judge in both candidate presentation orderings and returns the entry-wise average $\Pi_{\text{BCH}}(p, \pi) := \frac{1}{2}(\mathbf{s}^{\text{fwd}} + \mathbf{s}^{\text{rev}})$. For the per-criterion scoring template, which lists all candidate responses of a prompt in a single query (Appendix B), the two orderings are the canonical and the reversed candidate listing; for the pairwise comparison template they are the (A, B) and (B, A) swaps. Since the averaged output does not depend on π , Π_{BCH} satisfies PI under order reversal by construction.

Aggregator F_{BCH} . Given the order-symmetrised profile $\mathbf{s}$, a criterion-weight vector $w \in \Delta^k$ (uniform $w_c = 1/k$ throughout this paper), and a held-out calibration set $\mathcal{D}_{\text{cal}}$ of n_{cal} profiles with reference aggregated Borda scores $\mu_{\text{agg}}(\cdot)$, compute per-(judge, criterion) Borda scores $\tilde{B}_{j,c}(x) = |\{y : s_{j,c}(x) > s_{j,c}(y)\}|$, the aggregated BCH score $\hat{B}_{\text{BCH}}(x) = \sum_c w_c \cdot \text{median}_j \tilde{B}_{j,c}(x)$, the per-profile aggregated residual $r(\mathbf{s}, x) = |\hat{B}_{\text{BCH}}(x) - \mu_{\text{agg}}(x)|$, and the conformal quantile $\hat{q}_\alpha$, defined as the empirical $\lceil (n_{\text{cal}} + 1)(1 - \alpha) \rceil / n_{\text{cal}}$ -quantile of the aggregated residuals $\{r(\mathbf{s}, x) : (\mathbf{s}, x) \in \mathcal{D}_{\text{cal}}\}$. Return the ranking induced by

$$\hat{B}_{\text{BCH}}(x) = \sum_{c \in \mathcal{C}} w_c \cdot \text{median}_{j \in \mathcal{J}} \tilde{B}_{j,c}(x), \quad (1)$$

with $\hat{q}_\alpha$ used to characterise coverage in Proposition 2. The ranking induced by $\hat{B}_{\text{BCH}}$ is independent of α ; $\hat{q}_\alpha$ enters only the certificate, not the output. Construction of $\mathcal{D}_{\text{cal}}$ is in Section 5.3; a substitute for domains without human ground truth is in Section 5.4.

5.2 Properties

Proposition 1 (Four-axiom satisfaction). *F_{BCH} satisfies UD, WP, A, and CM by construction. Π_{BCH} satisfies PI under order reversal by construction. The composite $(\Pi_{\text{BCH}}, F_{\text{BCH}})$ violates IIA in general; Proposition 2 bounds the violation rate.*

UD: F_{BCH} is defined on every profile and returns a complete weak order. WP: unanimous strict preference across $\mathcal{V}$ yields unanimous Borda-count advantage. A: the median-over-judges in Equation (1) is invariant under judge permutations. CM:

a single-coordinate score increase weakly raises $\tilde{B}_{j,c}(x)$. PI: under order reversal, trivially from the order-symmetrizing step.

Proposition 2 (Gap-marginalized conformal coverage). *Let $\mathbf{s}, \mathbf{s}'$ be exchangeable with the calibration profiles in $\mathcal{D}_{\text{cal}}$, with the same ordinal pairwise restrictions $\succsim_{j,c} |_{\{x,y\}}$ for every $v \in \mathcal{V}$. Let $\Delta_{x,y}(\mathbf{s}) := |\hat{B}_{\text{BCH}}(x) - \hat{B}_{\text{BCH}}(y)|$ be the BCH ranking gap on the pair $\{x, y\}$. Then for any $\alpha \in (0, 1)$,*

$$\Pr[F_{\text{BCH}}(\mathbf{s})|_{\{x,y\}} \neq F_{\text{BCH}}(\mathbf{s}')|_{\{x,y\}}] \leq 4\alpha + \mathbb{E}_{(\mathbf{s}, x, y)}[\mathbb{1}\{\Delta_{x,y}(\mathbf{s}) \leq 4\hat{q}_\alpha\}], \quad (2)$$

where the expectation is over the empirical pair distribution induced by $\mathcal{D}_{\text{cal}}$.

The bound improves on a naive worst-case analysis by marginalizing over the gap distribution: pairs whose BCH ranking gap exceeds $4\hat{q}_\alpha$ contribute at most 4α to the violation probability via a union bound, while pairs below the threshold are charged at their empirical frequency. The proof, in Appendix A.7, follows the split-conformal coverage argument of Vovk et al. (2005); Angelopoulos and Bates (2023) combined with a gap-conditional union bound. Instantiating the bound requires the empirical small-gap fraction $p = \Pr[\Delta_{x,y} \leq 4\hat{q}_\alpha]$, which is computed from the gap histogram of the calibration set at audit time; the empirical IIA rate is 18.5% (Table 2).

BCH’s empirical IIA rate is 18.5% (Clopper–Pearson 95% CI: [13.4%, 24.6%]). Proposition 2 provides a structural bound connecting this rate to the gap distribution and identifies the gap distribution as the quantity governing IIA stability; the bound is instantiated in Appendix C.5.

5.3 The calibration set

$\mathcal{D}_{\text{cal}}$ is a 500-prompt corpus drawn proportionally from the three benchmarks, disjoint from the audit subset by random partition (exchangeability holds within each benchmark). Three independent annotators, blind to candidate identity, produce a complete ranking over the four candidates per prompt; reference aggregated Borda scores $\mu_{\text{agg}}(x)$ are the mean across annotators. Retaining only prompts with per-prompt inter-annotator agreement above 0.7 (measured as Krippendorff’s α_K with the ordinal metric on the three annotators’ rankings of the four candidates) leaves $n_{\text{cal}} = 457$ usable profiles.

The exchangeability unit for Proposition 2 is the profile (one prompt’s complete score tensor); each profile contributes one aggregated residual $r(s, x)$ computed at a randomly selected alternative x , giving $n_{\text{cal}} = 457$ exchangeable residuals. The conformal bound transports across benchmark strata only if the test-time mixture matches the calibration mixture; our own audit mixture (25%, 37.5%, 37.5%) differs from the calibration mixture, which is proportional to benchmark size (5.8%, 58.1%, 36.1%; 29/291/180 prompts), so the reported rates should be read per-stratum rather than as a transported guarantee. A practitioner deploying BCH on a single benchmark should re-calibrate. Sensitivity to the agreement threshold is in Appendix C.5.

5.4 Calibration without ground truth: a descriptive variant

Where no human-annotated reference is available, $\mu_{\text{agg}}(x)$ can be replaced by an ensemble-agreement proxy $\hat{\mu}_{\text{agg}}(x)$: the aggregated BCH score of a held-out judge ensemble disjoint from the test-time judges. Because the ranking induced by $\hat{B}_{\text{BCH}}$ does not depend on the reference scores, the substitution leaves the output of F_{BCH} and its empirical IIA-violation rate unchanged; it changes only the calibration residuals, $\hat{q}_\alpha$, and the instantiated bound, and the gap-marginalized conformal bound of Proposition 2 *does not transfer* under the altered exchangeability conditions. The substitute is therefore restricted to descriptive use; the formal scope statement is in Appendix D.

6 Empirical Audit

6.1 Setup and probes

Data, judges, criteria, aggregators. 200 prompts stratified across MT-Bench (50; first turn only), AlpacaEval 2 (75) (Dubois et al., 2024), and Arena-Hard (75) (Li et al., 2025); four candidates per prompt from GPT-4o, Claude 3.5 Sonnet, Llama-3.1-70B/8B-Instruct (Grattafiori et al., 2024). Three widely used judges (GPT-4o, Claude 3.5 Sonnet, Llama-3.1-70B-Instruct). A Prometheus 2 (Kim et al., 2024b) replication (Table 4, Appendix C.3) reproduces the pattern. Four anchored 1–10 criteria (correctness, helpfulness, safety, brevity). Eight aggregators audited: Majority Vote (pairwise majority over the virtual electorate), Score-Mean, Borda, Arena-style Bradley-Terry (Bradley and Terry, 1952) (maximum likelihood on the per-virtual-voter pairwise outcomes

induced by the scores, ties counted as half a win each, with a pseudo-count of 0.01 added to every ordered-pair win total so that the likelihood has a finite maximiser on every profile, and log-strengths within 10^{-6} of each other treated as tied), MT-Bench Average (Zheng et al., 2023), Best-of-Judge (single-judge cross-criterion mean with GPT-4o as the oracle judge), Permutation-Consensus (Borda aggregator with the permutation-consensus pipeline of Huang et al. (2026b), instantiated with the two orderings of Π^*), and BCH. Exact model identifiers, the decoding settings used to generate the candidate responses, and every other configuration value are recorded in the configuration files of the released toolkit.

Axiom probes. PI: re-prompt each (judge, criterion) on every candidate pair in both presentation orders and count flipped verdicts. IIA: remove one of four candidates, re-query every judge on the remaining three, and count rank changes on unchanged pairs (add/replace-one variants in Appendix C.8); each prompt yields four removal probes. CM: raise one (j, c, x) score by 1, count x -demotions. WP: scan for unanimous strict preference contradicted by F . A: analytic (Best-of-Judge violates at 100% by construction). UD: structural except Majority Vote (0/200 Condorcet cycles). Full probe definitions and prompt templates are in Appendix B.

Statistical methodology. IIA rates are prompt-level: a prompt counts as violated if any of its four removal probes triggers. Standard errors are $\sqrt{\hat{p}(1 - \hat{p})/200}$. Paired McNemar tests use prompt-level pairing; Bonferroni correction is over seven pre-specified comparisons (corrected threshold $p < 0.00714$). Temperature was 0 for all judge calls. Additional details (parse-failure handling, self-preference confound, per-stratum counts) are in Appendices B and C.7.

6.2 Results

Aggregator-by-axiom results. Every audited aggregator violates at least one axiom on the audited profiles or by construction (Table 2; UD is a domain condition; no Condorcet cycle was observed at $n=200$, though Majority Vote produced intransitive outputs from tied majorities): MT-Bench Average violates two (IIA and PI); Borda violates IIA (31.0%) and PI (22.5%); Score-Mean and Arena-BT violate IIA at 28.5% and 29.5%; Best-of-Judge violates Anonymity by construction (and IIA

Agg.	UD	WP	IIA	A	CM	PI
Majority Vote	$\times(\dagger)$	✓	✓	✓	✓	✓*
Score-Mean	✓	✓	28.5 ± 3.2	✓	✓	✓*
Borda	✓	✓	31.0 ± 3.3	✓	✓	22.5 ± 3.0
Arena-BT	✓	✓	29.5 ± 3.2	✓	✓	✓*
MT-Bench Avg	✓	✓	38.0 ± 3.4	✓	✓	6.0 ± 1.7
Best-of-Judge	✓	✓	24.5 ± 3.0	100.0	✓	✓*
Perm-Consensus	✓	✓	24.0 ± 3.0	✓	✓	✓*
BCH (ours)	✓	✓	18.5 ± 2.7	✓	✓	✓*

Table 2: Aggregator-by-axiom matrix (200 prompts). ✓: deterministic satisfaction; numeric = violation rate (%) $\pm$ SE. ✓*: PI secured via Π^* . †: UD structural failure (0/200 Condorcet cycles; tied-majority intransitivities occur generically at even $N=12$). A: analytic except Best-of-Judge (100% by construction). Best-of-Judge is the single-judge cross-criterion mean with GPT-4o as oracle (hence nonzero IIA). PI: Borda’s 22.5% is the mean verdict-flip rate (Table 5); MT-Bench Avg’s 6.0% is the composite ranking-change rate (the two cells are not directly comparable).

at 24.5%); permutation-consensus sits at 24.0% IIA. BCH violates only IIA, at 18.5%, the lowest rate among the seven aggregators with measurable IIA. Permutation-Consensus also violates only IIA (at 24.0%), achieving the same single-axiom profile as BCH; BCH’s differentiator is the audit-time conformal certificate rather than the axiom profile itself. Paired McNemar (χ^2 , uncorrected) gives $p = 0.0067$ for BCH vs. Borda (discordant pairs: 55 vs. 30), marginally below the Bonferroni threshold $p < 0.00714$; the exact binomial McNemar test gives $p = 0.0088$, above threshold. The BCH-vs-Borda effect size is 12.5 percentage points. The advantage over permutation-consensus is not Bonferroni-significant ($p \approx 0.03$; discordant pairs: 18 vs. 7). The conformal bound (Proposition 2) can be instantiated from the gap histogram at audit time; the value of BCH is the traceability of its IIA rate to the calibration set and gap distribution.

6.3 The human-jury comparator

A 50-prompt human-jury comparator (three annotators distinct from $\mathcal{D}_{\text{cal}}$; the remove-one IIA probe applied to the annotators’ complete rankings by restriction; Appendix C.9) yields an aggregated human IIA-violation rate of 9.0% at the probe level (18/200 probes across 50 prompts $\times$ 4 removals; Krippendorff $\alpha_K = 0.83$), computed by aggregating across annotators before applying the Borda restriction. Restricting a single annotator’s complete ranking to a subset cannot flip a pairwise ordering under Borda counts; the observed violations arise from the cross-annotator aggregation step, where removing an alternative shifts the aggregated Borda scores. Table 2 reports prompt-level rates; the probe-level human rate is not directly comparable. A prompt-level recomputation yields

a value between 10% and 36% depending on clustering, an interval that contains BCH’s 18.5%. The human rate is consistent with the known non-zero IIA-violation baseline in choice experiments (Huber et al., 1982; Simonson, 1989; Tversky, 1972).

Two additional findings (AlpacaEval 2’s length controller as frontier movement and the self-judging Anonymity confound) are in Appendix C.1.

7 Discussion and conclusion

Our impossibility theorems resolve three years of LLM-judge debiasing into disciplined movement along a single frontier: Feng et al. (2025)’s Intra-Pair Instability is the pipeline-level PI probe, Huang et al. (2026a)’s leaderboard sensitivity is closely related to the aggregator-level IIA rate, and MT-Bench Average, permutation-consensus, AlpacaEval 2’s length controller, and Best-of-Judge each occupy a frontier face whose axiomatic cost the framework names. No aggregator escapes all four aggregator-level pathologies, so the choice among faces should be made deliberately and reported as a six-cell axiom profile (UD, WP, IIA, A, PI, optionally CM; with SEs or CIs), computed automatically by our released audit toolkit. The toolkit (code, configuration files carrying every hyperparameter of Sections 5–6, and a reduced-scale smoke test with deterministic synthetic judges) emits per-stratum and per-judge integer counts, McNemar contingency tables, $\hat{q}_\alpha$, calibration residuals, the gap histogram, and per-pair certificates of Algorithm 1.

Limitations

We distinguish theoretical scope conditions from empirical constraints. The impossibility theorems hold unconditionally within their stated regimes and axiom definitions; the two scope conditions below (Universal Domain and CFC calibration) identify regimes where the impossibility does not apply, and thus bound the theorems’ practical reach.

Universal Domain is empirically falsifiable: correlated criteria and structured score distributions restrict the realised profile domain. In our audit, 0 of 200 prompts produced a Condorcet cycle under Majority Vote (one-sided 95% Clopper–Pearson upper bound $\approx 1.5\%$). However, UD requires a complete *transitive* weak order, and Majority Vote can fail transitivity without a strict cycle: when $N = nk$ is even (here $N = 12$), exact tied majorities produce intransitive outputs generically. Our toolkit reports both Condorcet cycles and intransitive outputs; whether the realised domain admits aggregators that satisfy all four axioms on the profiles that actually arise depends on the intransitive-output rate, not only the cycle rate. Domain restriction (e.g., single-peakedness) is the classical escape from Arrow’s impossibility; whether LLM-judge score profiles satisfy any such restriction depends on the realised profile domain and the intransitive-output rate.

The impossibility theorems assume judges produce scores at best Cardinal Unit-Comparable (CUC). In the rare event that judges and criteria can be pre-calibrated to a shared Cardinal Fully Comparable (CFC) scale, aggregators such as leximin and utilitarianism become available. Because they read cardinal levels, these aggregators escape the impossibility, but they violate ordinal IIA (the axiom version this paper commits to), since their output depends on score magnitudes, not just pairwise orderings (Appendix A.3, Steps 3–3a). Escaping the impossibility while retaining ordinal IIA requires a weaker stability axiom; see Appendix A.5. Achieving CFC calibration would require all judges to share a common scoring anchor across all criteria, concretely a shared reference set of (prompt, response) pairs with agreed-upon scores on every criterion, against which each judge’s scale is affinely aligned. The annotation cost scales with $|\mathcal{J}| \times |\mathcal{C}|$ and must be repeated whenever a judge model is updated; current 1–10 rubrics shift measurably across model releases and criteria, placing the practical regime between ONC and CUC rather than at CFC.

Theorem 2 also requires CUC-invariance of F as an explicit hypothesis (Appendix A.3); aggregators that operate on CUC profiles but are not themselves CUC-invariant fall outside this theorem.

Ethical Considerations

LLM-as-judge leaderboards influence which models are deployed and funded; PI-violation rates of 12.5–34.0% per criterion (Table 5) indicate systematic misallocation. The framework makes such failures legible, but could also be used adversarially: an aggregator optimized on the axioms a paper reports while inflating violations on omitted axioms is a form of leaderboard hacking. We mitigate this by recommending the full six-cell axiom profile be reported jointly. The LLM judges studied here are deployed by a small number of commercial actors; the choice between aggregation rules, and therefore between which biases the field tolerates, should be made transparently. Total GPU compute: 38 A100 hours.

Reproducibility

Code is available at <https://github.com/kvector0/axiom-audit>.

Acknowledgments

We thank the anonymous reviewers and the area chair for their constructive feedback. We used the AI assistant Anthropic Claude for coding assistance and minor language polishing during manuscript preparation; all technical content, proofs, and experimental design are the authors’ own work.

References

- Anastasios N. Angelopoulos and Stephen Bates. 2023. Conformal prediction: A gentle introduction. *Found. Trends Mach. Learn.*, 16(4):494–591.
- Kenneth J. Arrow. 1951. *Social Choice and Individual Values*. John Wiley & Sons.
- Ralph Allan Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.
- Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Kopel, Dinesh Manocha, Furong Huang, Amrit Bedi, and Mengdi Wang. 2024. MaxMin-RLHF: Alignment with diverse human preferences. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, Proceedings of Machine Learning Research, pages 6116–6135. PMLR / OpenReview.net.

- 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. 2024. Chatbot Arena: An open platform for evaluating LLMs by human preference. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, Proceedings of Machine Learning Research, pages 8359–8388. PMLR / OpenReview.net.
- Vincent Conitzer, Rachel Freedman, Jobst Heitzig, Wesley H. Holliday, Bob M. Jacobs, Nathan Lambert, Milan Mossé, Eric Pacuit, Stuart Russell, Hailley Schoelkopf, Emanuel Tewolde, and William S. Zwicker. 2024. Position: Social choice should guide AI alignment in dealing with diverse human feedback. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, Proceedings of Machine Learning Research, pages 9346–9360. PMLR / OpenReview.net.
- Jessica Dai and Eve Fleisig. 2024. Mapping social choice theory to RLHF. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*.
- Claude D’Aspremont and Louis Gevers. 1977. Equity and the informational basis of collective choice. *The Review of Economic Studies*, 44(2):199–209.
- Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. 2024. Length-controlled AlpacaEval: A simple way to debias automatic evaluators. In *Conference on Language Modeling*.
- Yuanning Feng, Sinan Wang, Zhengxiang Cheng, Yao Wan, and Dongping Chen. 2025. Are we on the right way to assessing LLM-as-a-judge? *arXiv preprint arXiv:2512.16041*.
- Sara Fish, Paul Gözl, David C. Parkes, Ariel D. Procaccia, Gili Rusak, Itai Shapira, and Manuel Wüthrich. 2024. Generative social choice. In *Proceedings of the 25th ACM Conference on Economics and Computation, EC 2024, New Haven, CT, USA, July 8-11, 2024*, pages 985–1003. ACM.
- Luise Ge, Daniel Halpern, Evi Micha, Ariel D. Procaccia, Itai Shapira, Yevgeniy Vorobeychik, and Junlin Wu. 2024. Axioms for AI alignment from human feedback. In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10-15, 2024*.
- John Geanakoplos. 2005. Three brief proofs of Arrow’s impossibility theorem. *Economic Theory*, 26(1):211–215.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Shrivastava, et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- John C. Harsanyi. 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy*, 63(4):309–321.
- Jenny Y. Huang, Yunyi Shen, Dennis Wei, and Tamara Broderick. 2026a. Dropping just a handful of preferences can change top large language model rankings. In *The Fourteenth International Conference on Learning Representations*. OpenReview.net.
- Tianyi Huang, Nathan Huang, Justin Tang, Wenqian Chen, and Elsa Fan. 2026b. Permutation-consensus listwise judging for robust factuality evaluation. In *Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM), ACL 2026, San Diego, California, USA, July 3-4, 2026*, pages 595–603. Association for Computational Linguistics.
- Joel Huber, John W. Payne, and Christopher Puto. 1982. Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. *Journal of Consumer Research*, 9(1):90–98.
- Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2024a. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024b. 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, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 4334–4353. Association for Computational Linguistics.
- Jonathan G.S. Koppell and Jennifer A. Steen. 2004. The effects of ballot position on election outcomes. *The Journal of Politics*, 66(1):267–281.
- 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. 2025. RewardBench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, Findings of ACL, pages 1755–1797. Association for Computational Linguistics.
- Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. 2024. Generative judge for evaluating alignment. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Qingquan Li, Shaoyu Dou, Kailai Shao, Chao Chen, and Haixiang Hu. 2026. Evaluating scoring bias in LLM-as-a-judge. In *Database Systems for Advanced Applications*, pages 19–34, Singapore. Springer Nature Singapore.

- Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. 2025. From crowdsourced data to high-quality benchmarks: Arena-Hard and BenchBuilder pipeline. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, Proceedings of Machine Learning Research. PMLR / OpenReview.net.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 2511–2522. Association for Computational Linguistics.
- Adian Liusie, Potsawee Manakul, and Mark J. F. Gales. 2024. LLM comparative assessment: Zero-shot NLG evaluation through pairwise comparisons using large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 - Volume 1: Long Papers, St. Julian's, Malta, March 17-22, 2024*, pages 139–151. Association for Computational Linguistics.
- Saumya Malik, Valentina Pyatkin, Sander Land, Jacob Morrison, Noah A. Smith, Hannaneh Hajishirzi, and Nathan Lambert. 2026. RewardBench 2: Advancing reward model evaluation. In *The Fourteenth International Conference on Learning Representations*. OpenReview.net.
- Eric Maskin. 1978. A theorem on utilitarianism. *The Review of Economic Studies*, 45(1):93–96.
- Tianyi Qiu. 2024. Representative social choice: From learning theory to AI alignment. In *Pluralistic Alignment Workshop at NeurIPS 2024*.
- Kevin W. S. Roberts. 1980. Interpersonal comparability and social choice theory. *The Review of Economic Studies*, 47(2):421–439.
- Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. 2023. Verbosity bias in preference labeling by large language models. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.
- Amartya Sen. 1970. Interpersonal aggregation and partial comparability. *Econometrica*, 38(3):393–409.
- Amartya Sen. 1977. On weights and measures: Informational constraints in social welfare analysis. *Econometrica*, 45(7):1539–1572.
- Huanxin Sheng, Xinyi Liu, Hangfeng He, Jieyu Zhao, and Jian Kang. 2025. Analyzing uncertainty of LLM-as-a-judge: Interval evaluations with conformal prediction. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 11286–11328. Association for Computational Linguistics.
- Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. 2025. Judging the judges: A systematic study of position bias in LLM-as-a-judge. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, IJCNLP-AACL 2025, Mumbai, India, December 20-24, 2025*, pages 292–314. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.
- Itamar Simonson. 1989. Choice based on reasons: The case of attraction and compromise effects. *Journal of Consumer Research*, 16(2):158–174.
- Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y. Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. 2025. JudgeBench: A benchmark for evaluating LLM-based judges. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2025. Judging the judges: Evaluating alignment and vulnerabilities in LLMs-as-judges. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 404–430, Vienna, Austria and virtual meeting. Association for Computational Linguistics.
- Amos Tversky. 1972. Elimination by aspects: A theory of choice. *Psychological Review*, 79(4):281–299.
- Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. 2005. *Algorithmic Learning in a Random World*. Springer, New York, NY.
- Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024a. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 9440–9450. Association for Computational Linguistics.
- Yidong Wang, Zhuohao Yu, Wenjin Yao, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. 2024b. PandaLM: An automatic evaluation benchmark for LLM instruction tuning optimization. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Ziqi Wang, Hanlin Zhang, Xiner Li, Kuan-Hao Huang, Chi Han, Shuiwang Ji, Sham M. Kakade, Hao Peng, and Heng Ji. 2025. Eliminating position bias of language models: A mechanistic approach. In *The Thirteenth International Conference on Learning Representations*. OpenReview.net.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. 2024. Self-preference bias in LLM-as-a-judge. In *NeurIPS 2024 Workshop on Safe Generative AI*.

Joshua Chu-Yue Yang, Damian Dailisan, Marcin Korcecki, Carina I. Hausladen, and Dirk Helbing. 2024. LLM voting: Human choices and AI collective decision-making. In *Proceedings of the Seventh AAAI/ACM Conference on AI, Ethics, and Society (AIES-24) - Full Archival Papers, October 21-23, 2024, San Jose, California, USA - Volume 1*, pages 1696–1708. AAAI Press.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V. Chawla, and Xi-angliang Zhang. 2025. Justice or prejudice? quantifying biases in LLM-as-a-judge. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. 2024. FLASK: fine-grained language model evaluation based on alignment skill sets. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Zixiao Zhao, Amirreza Esmaeili, and Fatemeh Fard. 2026. Bias in the loop: Auditing LLM-as-a-judge for software engineering. *arXiv preprint arXiv:2604.16790*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhonghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Xiaosen Zheng, Tianyu Pang, Chao Du, Qian Liu, Jing Jiang, and Min Lin. 2025. Cheating automatic LLM benchmarks: Null models achieve high win rates. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2025. JudgeLM: Fine-tuned large language models are scalable judges. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

A Full Proofs

A.1 Proof of Lemma 1 (Virtual-Voter Reduction)

Let $\mathbf{s}, \mathbf{s}' \in \mathcal{S}$ be two profiles inducing the same matrix of pairwise virtual-voter preferences: for every

$v \in \mathcal{V}$ and every pair $\{x, y\} \subseteq \mathcal{X}$, $\succsim_v |_{\{x, y\}}^{\mathbf{s}} = \succsim_v |_{\{x, y\}}^{\mathbf{s}'}$. We show $F(\mathbf{s}) = F(\mathbf{s}')$.

By UD, both $F(\mathbf{s})$ and $F(\mathbf{s}')$ are defined and are complete weak orders on $\mathcal{X}$. Fix any pair $\{x, y\}$. By IIA, $F(\mathbf{s})|_{\{x, y\}}$ depends only on the matrix entries $\{\succsim_v |_{\{x, y\}}^{\mathbf{s}}\}_v$, which are identical on $\mathbf{s}$ and $\mathbf{s}'$. Therefore $F(\mathbf{s})|_{\{x, y\}} = F(\mathbf{s}')|_{\{x, y\}}$. Since this holds for every pair, the two complete weak orders agree on every pair; by completeness and transitivity, they are identical. $\square$

A.2 Proof of Theorem 1 (Ordinal Impossibility)

Step 1: Dictatorship from {UD, WP, IIA}. By Lemma 1, F depends only on the pairwise virtual-voter matrix. Define a coalition $C \subseteq \mathcal{V}$ as *decisive* for the pair (x, y) if for every profile with $x \succ_v y$ for all $v \in C$ and $y \succ_v x$ for all $v \notin C$, we have $x \succ_F y$. By WP, $\mathcal{V}$ itself is decisive for every pair.

The standard Arrow decisive-coalition argument (Arrow, 1951; Geanakoplos, 2005) establishes that the family of decisive coalitions for any single pair is closed under intersection and supersets, and that decisiveness is pair-independent (the same family of coalitions is decisive for every pair). The pair-independence step uses IIA together with non-trivial profiles on triples $\{x, y, z\} \subseteq \mathcal{X}$ where $|\mathcal{X}| \geq 3$, available by the theorem’s hypothesis. Pair-independence of decisiveness ensures that the dictator v^* identified below is the same for every pair $\{x, y\} \subseteq \mathcal{X}$; in step 2 we may therefore choose any pair where j^* and some j' disagree.

Together with WP and the transitivity requirement on $F(\mathbf{s})$, IIA further yields the ultrafilter property: for every $C \subseteq \mathcal{V}$, exactly one of C and $\mathcal{V} \setminus C$ is decisive. Since $\mathcal{V}$ is finite ($|\mathcal{V}| = nk < \infty$), every ultrafilter on a finite set is principal on a singleton: there exists a unique $v^* \in \mathcal{V}$ such that the decisive coalitions are exactly the supersets of $\{v^*\}$. Equivalently, F is the dictatorship of v^* : for every profile and every pair, $x \succ_F y \iff x \succ_{v^*} y$.

Step 2: Anonymity contradiction. Write $v^* = (j^*, c^*)$. We construct a profile and a permutation that yield contradictory rankings under dictatorship and Anonymity.

Fix any $j' \in \mathcal{J} \setminus \{j^*\}$, which exists by $n \geq 2$. Take a profile $\mathbf{s}$ such that $s_{j^*, c^*}(x) > s_{j^*, c^*}(y)$ and $s_{j', c^*}(y) > s_{j', c^*}(x)$, with other entries arbitrary. By dictatorship, $x \succ_F y$ on $\mathbf{s}$.

Now apply the transposition $\sigma = (j^* j')$ on $\mathcal{J}$. The permuted profile satisfies $(\sigma \cdot \mathbf{s})_{j^*, c^*}(x) =$

$s_{j',c^*}(x) < s_{j',c^*}(y) = (\sigma \cdot s)_{j',c^*}(y)$, so $y \succ_{v^*} x$ on $\sigma \cdot s$. By dictatorship, $y \succ_F x$ on $\sigma \cdot s$.

By Judge-Anonymity, $F(\sigma \cdot s) = F(s)$. Therefore both $x \succ_F y$ and $y \succ_F x$ on the same social ranking, contradicting completeness. $\square$

Note on CM. Criterion-Monotonicity plays no role in the proof above; the impossibility holds for the minimal aggregator-level set $\{\text{UD}, \text{WP}, \text{IIA}, \text{A}\}$. We retain CM as a separately-measured empirical axiom (Section 4.5) because aggregators that violate CM produce empirically interesting failure modes, but the theorem does not need it.

A.3 Proof of Theorem 2 (Cardinal Impossibility)

Step 1: weighted-utilitarian reduction. In the CUC informational regime, F is invariant under any transformation $s_v \mapsto \lambda s_v + \beta_v$ with $\lambda > 0$ common across all virtual voters v and $\beta_v \in \mathbb{R}$ individual to each v ; the hypothesis of Theorem 2 requires F to be CUC-invariant in this sense. The welfarism framework (D'Aspremont and Gevers, 1977; Roberts, 1980; Maskin, 1978) establishes that any SWFL satisfying UD, Strong Pareto, cardinal-IIA, and CUC-invariance reduces to the weighted-utilitarian family

$$F(s) = \text{argsort}\left(\sum_{v \in \mathcal{V}} \omega_v s_v(\cdot)\right) \quad (3)$$

for some weight vector $\omega = (\omega_v)_{v \in \mathcal{V}} \in \mathbb{R}^N$, with at least one $\omega_v \neq 0$. Theorem 1 already implies Theorem 2 (see Section 4); the value of the welfarism reduction is that it identifies which aggregator family survives cardinal IIA under CUC-invariance.

Step 2: criterion-only, strictly positive weights. Strong Pareto forces $\omega_v > 0$ for every v : apply SP to a profile where $s_{v_0}(x) > s_{v_0}(y)$ and $s_v(x) = s_v(y)$ for all $v \neq v_0$; the aggregate difference is $\omega_{v_0}(s_{v_0}(x) - s_{v_0}(y))$, and SP requires it strictly positive, so $\omega_{v_0} > 0$. Since all weights are positive, any judge permutation σ satisfying $F(\sigma \cdot s) = F(s)$ for all s forces $\omega_{\sigma(v)} = \lambda \omega_v$ for some $\lambda > 0$ (the argsort is scale-invariant); since σ has finite order, iterating gives $\lambda^{|\sigma|} = 1$ and hence $\lambda = 1$, so $\sigma \cdot \omega = \omega$. Judge-Anonymity therefore collapses $\omega_{(j,c)} = \omega_{(j',c)}$ for every $j, j' \in \mathcal{J}$; write $\omega_{(j,c)} = \omega_c > 0$. The reduction leaves a strictly positive

criterion-weighted utilitarian rule

$$F(s) = \text{argsort}\left(\sum_{c \in \mathcal{C}} \omega_c \sum_{j \in \mathcal{J}} s_{j,c}(\cdot)\right), \quad \omega_c > 0. \quad (4)$$

Step 3: ordinal-IIA witness. Fix any admissible weight vector ω from Step 2; since all $\omega_c > 0$, pick any c^* . We exhibit two profiles s, s' on $|\mathcal{J}| = 2$, $|\mathcal{C}| = 2$, $|\mathcal{X}| = 3$ such that every virtual voter's pairwise restriction $\succ_{j,c} \upharpoonright_{\{x_1, x_2\}}$ is identical in the two profiles, but $F(s)|_{\{x_1, x_2\}} \neq F(s')|_{\{x_1, x_2\}}$ under Equation (4). The construction perturbs scores on a criterion with $\omega_{c^*} > 0$, holding all other criteria fixed.

Without loss of generality, let $c^* = c_1$ with $\omega_1 > 0$. The construction works for any admissible ω ; we display it for $\omega_1 = \omega_2 = 1$ and verify the general case below. Define s by $s_{j_1, c_1} = (10, 9, 5)$, $s_{j_2, c_1} = (10, 9, 5)$, $s_{j_1, c_2} = (1, 2, 9)$, $s_{j_2, c_2} = (1, 2, 9)$, with components in order (x_1, x_2, x_3) . Pairwise on $\{x_1, x_2\}$: $x_1 \succ_v x_2$ for $v \in \{(j_1, c_1), (j_2, c_1)\}$, $x_2 \succ_v x_1$ for $v \in \{(j_1, c_2), (j_2, c_2)\}$. Aggregate $\sum_c \omega_c \sum_j s_{j,c}$ on (x_1, x_2) is $(22, 22)$; the aggregator returns $x_1 \sim_F x_2$ on s .

Define s' by leaving the c_1 scores identical to s and changing the c_2 scores to $s'_{j_1, c_2} = (1, 3, 9)$, $s'_{j_2, c_2} = (1, 3, 9)$. Pairwise on $\{x_1, x_2\}$: still $x_2 \succ_v x_1$ for $v \in \{(j_1, c_2), (j_2, c_2)\}$ (since $3 > 1$), so the pairwise restriction matrix is unchanged. Aggregate on (x_1, x_2) is $(22, 24)$; the aggregator returns $x_2 \succ_F x_1$ on s' .

Two profiles with identical pairwise virtual-voter restrictions on $\{x_1, x_2\}$ produce different social orderings on $\{x_1, x_2\}$ (tied vs. strict), violating ordinal IIA.

Step 3a: a generic strict-vs-strict witness. The construction in step 3 places s at an indifference edge of profile space. We now exhibit a full-measure (generic) witness in which both s and s' produce strict orderings on $\{x_1, x_2\}$ that flip across the two profiles. Take the same weights and structure as before but modify the c_1 scores so that the aggregate on s favours x_1 strictly: set $s_{j_1, c_1} = s_{j_2, c_1} = (11, 9, 5)$ on s , leaving the c_2 scores at $(1, 2, 9)$. Aggregate on (x_1, x_2) is now $(24, 22)$, yielding $x_1 \succ_F x_2$ strictly. Define s' by leaving c_1 identical and changing c_2 to $s'_{j_1, c_2} = s'_{j_2, c_2} = (1, 4, 9)$. Pairwise on $\{x_1, x_2\}$ the matrix is unchanged. Aggregate on (x_1, x_2) is $(24, 26)$, yielding $x_2 \succ_F x_1$ strictly. Both profiles

sit in the interior of profile space (no aggregate ties) and the orderings flip; ordinal IIA fails generically, not only at the indifference edge. $\square$

General- ω extension. For arbitrary admissible $\omega_1, \omega_2 > 0$ with $\omega_1 \neq \omega_2$, the Step-3 construction requires a modified perturbation to achieve a tie on s . Set the c_2 gap $\delta_2 > 0$ on $\{x_1, x_2\}$ such that $\omega_1 d_1 = \omega_2 \delta_2$ where d_1 is the c_1 aggregate gap, then perturb by $\varepsilon > 0$ to break the tie in s' . Step 3a extends identically: for strict-vs-strict, choose the initial c_1 gap so $\omega_1 d_1 > \omega_2 \delta_2$, then increase δ_2 by $\varepsilon > \omega_1 d_1 / \omega_2 - \delta_2$ in s' . The pairwise restriction on $\{x_1, x_2\}$ is invariant under magnitude changes that preserve the sign, so ordinal IIA fails for every admissible ω .

Conclusion. Steps 1–2 force any SWFL satisfying $\{\text{UD}, \text{SP}, \text{cardinal-IIA}, \text{A}\}$ in the CUC regime to take the form of Equation (4). Step 3 shows this form violates the stronger ordinal-IIA axiom on a measure-zero edge witness; Step 3a extends this to a generic strict-vs-strict witness. Since ordinal IIA strictly implies cardinal IIA, no CUC-invariant SWFL satisfies the four-axiom set $\{\text{UD}, \text{SP}, \text{ordinal-IIA}, \text{A}\}$ in the CUC regime. As in Theorem 1, CM is not needed. $\square$

A.4 Proof of Theorem 3 (Frontier Characterization)

Part (i) restates Theorems 1, 2.

Part (ii) requires four composite witnesses, one per dropped aggregator-level axiom. Each is composed with Π^* , the order-symmetrizing pipeline.

Drop UD: Majority Vote. Define $F(s)$ by: $x \succsim_F y$ iff $|\{v \in \mathcal{V} : x \succsim_v y\}| \geq |\{v : y \succsim_v x\}|$, i.e., pairwise majority over all $N = nk$ virtual voters. On profiles inducing a non-Condorcet majority preference, Majority Vote satisfies WP (unanimous strict preference is preserved), IIA (pairwise output depends only on pairwise input), A (vote count is anonymous), and PI via Π^* . On Condorcet-cyclic profiles, the output is intransitive, so UD fails by definition.

Drop WP: Constant rule. Fix any complete weak order $\succsim^\circ$ on $\mathcal{X}$ and define $F(s) \equiv \succsim^\circ$ for every s . The output is defined for every profile (UD), trivially pairwise-input-independent (IIA), trivially judge-permutation-invariant (A), and trivially PI via Π^* . WP fails on profiles where $\succsim^\circ$ contradicts the unanimous strict preference.

Drop IIA: Borda count (rank-based) or Score-Mean (magnitude-based). Borda count $B(x) :=$

$\sum_v |\{y : x \succ_v y\}|$ is defined for every profile (UD), preserves unanimous strict preference (WP), permutation-invariant in $\mathcal{J}(\text{A})$, and PI via Π^* . Borda violates IIA: the textbook counter-example uses three alternatives and a profile whose relative ranking of $\{x, y\}$ changes when a third alternative z is added or removed. Score-Mean $M(x) := \frac{1}{N} \sum_v s_v(x)$ satisfies the same four and fails IIA via magnitude-sensitivity (a witness construction is in Step 3 of Appendix A.3). The two witnesses are listed separately because they can disagree on near-tied profiles.

Drop A: Best-of-Judge. Fix any judge $j^\dagger$ and define $F(s) := \succsim_{j^\dagger, c^\dagger}$ for some fixed criterion. This is UD, WP, IIA, and PI via Π^* , but it ignores all other judges, so $F(\sigma \cdot s) \neq F(s)$ for any σ that displaces $j^\dagger$, violating A.

Each construction is checked directly against the four-axiom set $\Phi \setminus \{\phi\}$ for $\phi \in \{\text{UD}, \text{WP}, \text{IIA}, \text{A}\}$; PI is secured by Π^* uniformly. $\square$

A.5 Stability hierarchies and metric-IIA alternatives

The pair-stability axioms form a hierarchy whose strict top is ordinal IIA: it requires identical social ordering on $\{x, y\}$ under any profile perturbation that preserves the virtual-voter pairwise restriction. Strictly weaker alternatives include cardinal IIA (preserves magnitudes too), Kendall- τ stability (the full ranking changes by at most k inversions under bounded perturbations), and Mallows-distance stability (expected ranking distance bounded under bounded perturbations). The frontier characterization of Theorem 3 concerns ordinal IIA; under the weaker metric stabilities the achievable face structure changes. In particular, BCH's gap-marginalized bound (Section 5) becomes a direct τ -stability certificate without the IIA-violation re-interpretation, since a bounded violation rate on pair swaps implies a bounded expected Kendall- τ shift. We flag the hierarchy because a reader who prefers a metric stability should see where in the lattice this paper sits.

A.6 Cardinal-IIA version of Theorem 3

If IIA is read in its cardinal version (two profiles must additionally have identical magnitudes on $\{x, y\}$), the weighted-utilitarian witnesses from Equation (4) re-enter the witness set: under cardinal IIA, the Step 3 construction of Appendix A.3 fails, because s and s' differ in magnitude on

$\{x_1, x_2\}$. The cardinal-IIA frontier therefore admits Score-Mean as an additional drop-IIA witness, and the cardinal-IIA impossibility is strictly weaker than the ordinal-IIA impossibility of Theorem 2. We commit to ordinal IIA throughout the main paper because LLM-judge scores are not interpersonally cardinally comparable (1–10 rubrics shift between model releases and across criteria), and the stronger ordinal axiom is the more honest one to require.

A.7 Proof of Proposition 2 (Gap-Marginalized Conformal Coverage)

Step 1: split-conformal coverage on aggregated residuals. By construction, $\hat{q}_\alpha$ is the $\lceil (n_{\text{cal}} + 1)(1 - \alpha) \rceil / n_{\text{cal}}$ empirical quantile of the aggregated residuals $r(s, x) = |\hat{B}_{\text{BCH}}(x) - \mu_{\text{agg}}(x)|$ on the calibration set $\mathcal{D}_{\text{cal}}$. Under exchangeability of the test point (s, x) with the calibration profiles, the split-conformal coverage result (Vovk et al., 2005, Proposition 4.1) yields directly

$$\Pr \left[|\hat{B}_{\text{BCH}}(x) - \mu_{\text{agg}}(x)| \leq \hat{q}_\alpha \right] \geq 1 - \alpha. \quad (5)$$

Calibrating on the aggregated residual (rather than on per-(judge, criterion) residuals) is what permits a single- α coverage statement at the aggregator level: a per-(judge, criterion) calibration would inherit a union-bound factor over $n \cdot k$ events, vacating the bound for $n = 3, k = 4$ at any practically useful α .

Step 2: gap-conditional pair-wise stability. Consider two profiles s, s' with identical pair-wise virtual-voter restrictions on $\{x, y\}$ and define $\Delta_{x,y} := |\hat{B}_{\text{BCH}}(x) - \hat{B}_{\text{BCH}}(y)|$ on s . An IIA violation on $\{x, y\}$ requires that the BCH rankings $\hat{B}_{\text{BCH}}(x), \hat{B}_{\text{BCH}}(y)$ swap order between s and s' . By the triangle inequality through the shared anchor μ_{agg} , if all four residuals $|\hat{B}_{\text{BCH}}(z) - \mu_{\text{agg}}(z)|$ (for $z \in \{x, y\}$ on both s and s') are within $\hat{q}_\alpha$, each alternative’s BCH score can shift by at most $2\hat{q}_\alpha$ between profiles, so the gap $\Delta_{x,y}$ can shift by at most $4\hat{q}_\alpha$. Two cases.

(a) If $\Delta_{x,y} > 4\hat{q}_\alpha$, a swap requires at least one of the four residuals to exceed $\hat{q}_\alpha$, which has probability at most α per residual by Step 1. A union bound over the 2 alternatives $\times$ 2 profiles gives a total swap probability $\leq 4\alpha$.

(b) If $\Delta_{x,y} \leq 4\hat{q}_\alpha$, no conformal guarantee applies; an IIA violation is possible with probability up to 1. The contribution to the marginal viola-

tion rate is $\Pr[\Delta_{x,y} \leq 4\hat{q}_\alpha]$ over the empirical pair distribution.

Step 3: marginalizing over the empirical gap distribution. Combining (a) and (b),

$$\begin{aligned} \Pr[\text{IIA viol}] &\leq 4\alpha \cdot \Pr[\Delta_{x,y} > 4\hat{q}_\alpha] \\ &\quad + 1 \cdot \Pr[\Delta_{x,y} \leq 4\hat{q}_\alpha] \\ &\leq 4\alpha + \mathbb{E}[\mathbb{1}\{\Delta_{x,y} \leq 4\hat{q}_\alpha\}], \end{aligned}$$

since (a) and (b) partition the event space exhaustively: the gap $\Delta_{x,y}$ on s is either above or at-or-below the threshold, and case (a)’s union bound already covers swaps in both directions. $\square$

To instantiate the bound, compute $p = \Pr[\Delta_{x,y} \leq 4\hat{q}_\alpha]$ from the empirical gap histogram of the calibration set. The empirical IIA rate is 18.5%; the gap between any valid bound and the empirical rate reflects the conservatism of the union bound. The bound is computable at audit time from the calibration set and gap distribution alone; the toolkit’s calibration step writes the gap histogram together with $\hat{q}_\alpha$.

Remark 1 (Bound composition). *The bound $4\alpha + p$ where $p = \Pr[\Delta_{x,y} \leq 4\hat{q}_\alpha]$ combines two terms: the 4α union bound on large-gap pairs, plus the unconditional contribution from small-gap pairs charged at probability 1. The literal $4\alpha = 0.20$ alone covers only large-gap pairs; the small-gap term must be included for a complete bound.*

B Prompt Templates and Audit Protocol

B.1 The BCH procedure end-to-end

For reference, Algorithm 1 states the complete BCH pipeline-aggregator pair in pseudocode, matching the construction of Section 5.

B.2 Per-criterion scoring prompt

Each judge is queried with the following template for each criterion $c \in \{\text{correctness, helpfulness, safety, brevity}\}$. The m candidate responses of a prompt are listed in a single query, in canonical order for s^{fwd} and in reversed order for s^{rev} (Algorithm 1); the remove-one, add-one, and replace-one probes issue the same template with the reduced or extended candidate set.

You are evaluating responses on the criterion of {c}, anchored on a 1–10 scale where 1 = unacceptable and 10 = excellent. Given the prompt below and the {m} responses, return a single integer score between 1 and 10 for each response, one per line in the order

Algorithm 1 Borda-Conformal Hybrid (BCH)

Require: Prompt p ; alternatives $\mathcal{X}$; judges $\mathcal{J}$; criteria $\mathcal{C}$; criterion weights $w \in \Delta^k$; calibration set $\mathcal{D}_{\text{cal}}$ with reference aggregated Borda scores $\mu_{\text{agg}}(\cdot)$; significance level α .

Ensure: Social ranking over $\mathcal{X}$ and per-pair coverage certificate.

Pipeline Π_{BCH} :

- 1: $s^{\text{fwd}} \leftarrow$ query each $j \in \mathcal{J}$ on each $c \in \mathcal{C}$ with the candidates of p listed in canonical order
- 2: $s^{\text{rev}} \leftarrow$ query each $j \in \mathcal{J}$ on each $c \in \mathcal{C}$ with the candidates of p listed in reversed order
- 3: $s \leftarrow \frac{1}{2}(s^{\text{fwd}} + s^{\text{rev}}) \triangleright \Pi^*$; secures PI under order reversal

Aggregator F_{BCH} :

- 4: **for** each $(j, c) \in \mathcal{J} \times \mathcal{C}$ and each $x \in \mathcal{X}$ **do**
 - 5: $\tilde{B}_{j,c}(x) \leftarrow |\{y \in \mathcal{X} : s_{j,c}(x) > s_{j,c}(y)\}|$
 $\triangleright$ per-(judge, criterion) Borda score
 - 6: **end for**
 - 7: **for** each $x \in \mathcal{X}$ **do**
 - 8: $\hat{B}_{\text{BCH}}(x) \leftarrow \sum_{c \in \mathcal{C}} w_c \cdot \text{median}_{j \in \mathcal{J}} \tilde{B}_{j,c}(x)$
 - 9: **end for**
 - 10: **for** each calibration profile $(s^{(i)}, x^{(i)}) \in \mathcal{D}_{\text{cal}}$ **do**
 - 11: $r^{(i)} \leftarrow |\hat{B}_{\text{BCH}}(x^{(i)}; s^{(i)}) - \mu_{\text{agg}}(x^{(i)})| \triangleright$
 aggregated residual
 - 12: **end for**
 - 13: $\hat{q}_\alpha \leftarrow$ empirical $\lceil (n_{\text{cal}} + 1)(1 - \alpha) \rceil / n_{\text{cal}}$ -quantile of $\{r^{(i)}\}$
 - 14: **return** ranking induced by $\hat{B}_{\text{BCH}}(\cdot)$; per-pair coverage bound from Prop. 2 parametrised by $\hat{q}_\alpha$
-

presented, formatted as "Response i: <score>". Do not include any text other than the scores.

```
Prompt: {prompt}
Response 1: {response_1}
...
Response {m}: {response_m}
{c} scores:
```

B.3 Pairwise comparison prompt

For the pairwise PI probe, the template is (with $\{c\}$ the criterion name, giving the per-criterion cells of Table 5; the overall-quality variant replaces $\{c\}$ by "overall quality"):

```
Compare two responses to the prompt
below on {c}. Return either "A" or
"B" indicating which is better. Do not
include any text other than the letter.
```

```
Prompt: {prompt}
Response A: {response_a}
Response B: {response_b}
Choice:
```

B.4 Order-swap protocol

Position-Invariance is measured at the judge level by querying each (judge, criterion) on each of the six candidate pairs of each prompt twice: once with the canonical (A, B) order and once with (B, A) swapped. The PI-violation rate of a (judge, criterion) cell (Table 5) or of a (judge, benchmark) cell is the fraction of (prompt, pair) probes on which the judge's preferred candidate differs under the two orders; a probe whose verdict fails to parse under either order is excluded. At the ranking level, the composite rate of an aggregator that is not composed with Π^* is the fraction of prompts on which F ranks the forward-order and the reversed-order scoring profiles differently (the MT-Bench Avg cell of Table 2).

B.5 Statistical methodology details

IIA violation rates in Table 2 are prompt-level: a prompt counts as violated if any of its four candidate-removal probes violates the ordinal IIA criterion. The 200 prompts are stratified across MT-Bench (50), AlpacaEval 2 (75), and Arena-Hard (75); rates aggregate across strata and judges. Standard errors are approximated as $\sqrt{\hat{p}(1 - \hat{p})/200}$ for every empirical cell. The released audit toolkit reports every rate together with its per-stratum and per-judge integer counts and the McNemar contingency tables; we recommend readers verify rates against those counts. Paired McNemar tests use prompt-level pairing ($n = 200$; one prompt contributes one paired observation per aggregator comparison). Bonferroni correction is over seven pre-specified comparisons (each of the seven other aggregators against BCH, Majority Vote entering with its probe outcome, which is zero by construction; corrected threshold $p < 0.00714$). Three of the four candidate generators (GPT-4o, Claude 3.5 Sonnet, Llama-3.1-70B) are also judges; Appendix C.7 measures the resulting self-preference confound at +0.7 points. Non-integer or out-of-range responses (fewer than 0.5% of calls at temperature 0) were re-queried once; persistent failures were scored as 5 (the lower of the two central integers on the 1–10 scale), per response slot of the scoring template; a pairwise verdict that fails to parse under both queries is excluded from the PI denominator.

C Additional Empirical Results

C.1 Frontier movement, named pathologies, and leaderboard impact

Three additional findings reinforce the frontier reading of existing diagnostics and were summarised in Section 6.

Length controller is movement along the frontier, not a fix. Stripping AlpacaEval 2’s length controller raises the empirical IIA-violation rate from 24.5% to 34.0% (PI unchanged). The controller therefore secures a position-bias-relevant face at the cost of the IIA face, consistent with the frontier characterization: it is movement along the frontier rather than an escape from it. Aggregate rates are in the verbosity-control experiment below.

Self-judging Anonymity violation. GPT-4o judging GPT-4o yields a +0.7-point uplift on the four-criterion average compared to GPT-4o judging the other three candidates ($p < 0.001$, paired t -test). The effect is consistent with Anonymity violation (the judge identity matters for the verdict), although it is confounded with stylistic similarity between the judge model and its own outputs. Appendix C.7 states the matched-pair procedure; the toolkit emits the per-criterion and candidate–judge breakdown.

C.2 BCH ablation study

Variant	PI ($\downarrow$)	IIA ($\downarrow$)
Borda baseline	22.5 ± 3.0	31.0 ± 3.3
+ F_{BCH} median-of-judges	22.5 ± 3.0	19.5 ± 2.8
BCH ($F_{\text{BCH}} + \Pi^*$)	0.0	18.5 $\pm$ 2.7

Table 3: BCH ablation: PI ($\downarrow$) and IIA ($\downarrow$) violation rates (% , $\pm$ SE) for each component added to the Borda baseline.

BCH’s 12.5-point IIA reduction over Borda (31.0% $\rightarrow$ 18.5%) decomposes into two effects: Π^* applied to Borda reduces IIA by 7.0 points (31.0% $\rightarrow$ 24.0%, i.e. Permutation-Consensus), and the median aggregation step reduces IIA by a further 5.5 points (24.0% $\rightarrow$ 18.5%). The ablation in Table 3 isolates the aggregator effect by holding Π fixed: replacing standard Borda aggregation with the criterion-weighted median reduces IIA from 31.0% to 19.5% (11.5 points) on the un-averaged profile, and from 24.0% to 18.5% (5.5 points) on the Π^* -averaged profile. The 1.0-point gap between 19.5% (F_{BCH} without Π^*) and 18.5%

(full BCH) reflects interactions between the averaging and the median: Π^* modifies the per-criterion scores fed to the Borda counts, and the median attenuates the outlier scores that would otherwise dominate the ranking. As a further cross-check, the median-over-judges variant of Score-Mean yields $\text{IIA} = 21.5 \pm 2.9\%$, down from Score-Mean’s $28.5 \pm 3.2\%$ (Table 2), confirming the median step’s effect across aggregator families.

C.3 Prometheus 2 fine-tuned-judge replication

Aggregator	IIA (%)	PI (%)
Borda	30 ± 4.6	9 ± 2.9
Score-Mean / Single-Judge	27 ± 4.4	9 ± 2.9
BCH	19 ± 3.9	0.0

Table 4: Replication with Prometheus 2 (Kim et al., 2024b) as fine-tuned judge ($n = 100$ prompts). IIA and PI violation rates (% , $\pm$ SE). With a single judge, Score-Mean and Best-of-Judge coincide (merged row); BCH is composed with Π^* .

C.4 PI by judge and criterion

Judge	Corr.	Help.	Safety	Brev.
GPT-4o	20.5	24.5	19.0	25.0
Claude 3.5 Sonnet	13.0	15.5	12.5	15.5
Llama-3.1-70B	30.5	33.0	27.5	34.0

Table 5: PI-violation rate (%) by judge $\times$ criterion.

C.5 Conformal bound instantiation

The bound of Proposition 2 takes the form $4\alpha + p$ where $p = \Pr[\Delta_{x,y} \leq 4\hat{q}_\alpha]$ is the small-gap fraction computed from the empirical gap histogram of the calibration set. Although per-(judge, criterion) Borda scores $\tilde{B}_{j,c}(x) \in \{0, 1, 2, 3\}$ are discrete, the aggregated BCH score $\hat{B}_{\text{BCH}}(x) = \sum_c w_c \cdot \text{median}_j \tilde{B}_{j,c}(x)$ with $w_c = 1/4$ takes values in $\{0, 0.25, 0.50, \dots, 3.0\}$ (13 levels), so the pairwise gap $\Delta_{x,y}$ and the residual $|\hat{B}_{\text{BCH}}(x) - \mu_{\text{agg}}(x)|$ are both discrete. The small-gap fraction p and the conformal quantile $\hat{q}_\alpha$ inherit this discreteness. The conformal quantile $\hat{q}_\alpha$, the full gap histogram, and the vector of calibration residuals are written by the toolkit’s calibration step. At $|\mathcal{X}| = 4$, the pairwise gap $\Delta_{x,y}$ is constrained to $[0, 3]$, and any threshold $4\hat{q}_\alpha \geq 2.0$ captures at least 1/3 of all pairs unconditionally (since at most 4 of 6 pairs can have gap exceeding 2.0); the bound $4\alpha + p$ exceeds 1/3 at

$\alpha = 0.05$ for this alternative-set size, since at most 4 of 6 pairs can have gap exceeding 2.0 and $\hat{q}_\alpha$ is non-increasing in α .

Sensitivity to the inter-annotator agreement threshold used to filter calibration prompts: at thresholds 0.5, 0.6, 0.7, 0.8, the retained calibration sizes are $n_{\text{cal}} = 481, 472, 457, 423$ respectively. We use the 0.7 threshold ($n_{\text{cal}} = 457$) throughout.

C.6 Verbosity-control experiment

AlpacaEval 2’s length controller is reconstructed as a length-adjusted pairwise win rate: each per-virtual-voter pairwise outcome $o_v(x, y) \in \{0, \frac{1}{2}, 1\}$ is corrected by $\hat{\phi} \tanh((\ell_x - \ell_y)/\sigma_\ell)$, where ℓ is the response length in characters, σ_ℓ the standard deviation of candidate lengths over the audit, and $\hat{\phi}$ the global least-squares coefficient (no intercept) of the centred voter-mean outcome $\bar{o}(x, y) - \frac{1}{2}$ on $\tanh((\ell_x - \ell_y)/\sigma_\ell)$, fitted across all ordered candidate pairs of all audit prompts; a candidate’s LC score is its mean corrected outcome over opponents and virtual voters, and stripping the controller sets $\hat{\phi} = 0$. Stripping the AlpacaEval 2 length controller raises the IIA-violation rate from 24.5 ± 3.0 to 34.0 ± 3.4 on 200 prompts, holding all other audit settings constant. The judge-level PI rate (22.5 ± 3.0 , Table 2) is a pipeline-level quantity and is unaffected by the choice of aggregator. The 9.5-point IIA shift is consistent with the framework’s prediction: what the controller buys on the position-bias-relevant face, it pays for on the IIA face.

C.7 Self-enhancement audit

GPT-4o judges its own candidate with a mean uplift of +0.7 points on the four-criterion average of the 1–10 scale relative to its mean score of the other three candidates ($p < 0.001$ by paired t -test on $n = 100$ matched pairs). One pair per prompt contrasts the judge’s cross-criterion mean for its own response with its mean over the other candidates, on the order-symmetrised profile; the 100 prompts with the smallest absolute difference between the log character length of the judge’s own response and the mean log character length of the other three responses are retained (length matching), and the toolkit additionally reports the uplift after residualising scores on log response length, together with per-criterion and candidate–judge breakdowns. The effect is consistent with the findings of Wataoka et al. (2024).

C.8 Multi-probe IIA breakdown

The main-body IIA rate uses the remove-one probe: remove one of the four candidates and count rank changes on the remaining three. We additionally report two alternative probes on the same 200-prompt audit. *Add-one*: insert a fifth candidate (the response of a held-out generator, GPT-4o-mini, to the same prompt; its identifier is recorded in the configuration files) and count rank changes on the original four. *Replace-one*: swap one of the four candidates with the held-out fifth response and count rank changes on the unchanged three. Both variants re-query the judges on the extended or modified candidate set and apply the same violation criterion as the remove-one probe. Rates per aggregator are in Table 6.

Aggregator	Remove-one	Add-one	Replace-one
Borda	31.0 ± 3.3	33.5 ± 3.3	29.0 ± 3.2
Score-Mean	28.5 ± 3.2	30.5 ± 3.3	26.5 ± 3.1
Permutation-Cons.	24.0 ± 3.0	25.5 ± 3.1	22.0 ± 2.9
BCH (ours)	18.5 ± 2.7	20.0 ± 2.8	17.0 ± 2.7

Table 6: IIA-violation rates (% , $\pm$ SE) under three probes: remove-one, add-one, and replace-one.

The qualitative pattern is robust across probe classes; the BCH-vs-Borda gap is 12.5 points on remove-one, 13.5 on add-one, and 12.0 on replace-one.

C.9 Human-jury comparator details

The 50 prompts were stratified across benchmarks: 20 MT-Bench, 15 AlpacaEval 2, 15 Arena-Hard. Three annotators (graduate students, distinct from the calibration-set annotators, paid \$20/hour, blind to candidate identity and to which model produced each candidate) produced complete 4-candidate rankings on all 50 prompts using the same anchored rubric the LLM judges saw. Inter-annotator Krippendorff $\alpha_K = 0.83$ [0.76, 0.90] (ordinal metric; 95% bootstrap CI over prompts); corpus-level pairwise Spearman correlations between annotators (rank vectors concatenated across prompts) averaged 0.85 (range 0.79–0.91 over the three annotator pairs). Annotation cost: 31 person-hours ($\approx$ \$620). Both the calibration and jury annotators ranked independently without discussion.

The remove-one IIA-violation rate is computed by first aggregating the three annotators’ Borda scores into a single aggregated Borda ranking (using the sum of per-annotator Borda counts), then restricting this aggregated ranking to each of the four

3-candidate subsets by re-deriving Borda scores on the restricted candidate set. A probe counts as an IIA violation when the pairwise ordering of some pair $\{x, y\}$ in the 3-candidate aggregated Borda ranking differs from that pair’s ordering in the full 4-candidate aggregated Borda ranking. Each of 50 prompts $\times$ 4 removals yields 200 probes. The aggregate human IIA rate is 9.0% (18/200 probes; 95% bootstrap CI [4.2, 14.8], clustered by prompt). A single annotator’s restricted ranking, by contrast, can never flip a pairwise ordering under Borda counts (the proof is elementary: removing alternative z shifts $\tilde{B}(x) - \tilde{B}(y)$ by $\mathbb{K}\{x \succ z\} - \mathbb{K}\{y \succ z\}$; when $x \succ y$, we have $\{z : y \succ z\} \subseteq \{z : x \succ z\}$, so this shift is ≥ 0 and the sign of the gap is preserved); the violations therefore originate in the cross-annotator aggregation step, where removing an alternative differentially shifts each annotator’s Borda contribution, changing the aggregated ranking. The probe-level rate does not transfer directly to Table 2, which reports prompt-level rates; recomputed at the prompt level, the human rate (9.0% at the probe level) lies between 10% and 36% depending on how the 18 violating probes cluster across 50 prompts, an interval that contains BCH’s 18.5%.

D Scope of the ensemble-proxy substitute

This appendix gives the formal version of the substitute-scope claim made briefly in Section 5.4.

Remark 2 (Scope of the substitute). *Replacing μ_{agg} with the ensemble proxy $\hat{\mu}_{\text{agg}}$ alters the exchangeability conditions on which Proposition 2 rests. The aggregated residuals $r(s, x) = |\hat{B}_{\text{BCH}}(x) - \hat{\mu}_{\text{agg}}(x)|$ now depend on the ensemble judges, so exchangeability requires the (test profile, ensemble proxy) pair to be exchangeable with the (calibration profile, ensemble proxy) pair, a stronger condition that the present validation does not establish. We therefore restrict $\hat{\mu}_{\text{agg}}$ to descriptive use: the aggregator F_{BCH} with $\hat{\mu}_{\text{agg}}$ remains a UD/WP/A/CM-satisfying composite with Π_{BCH} securing PI, and its output and empirical IIA-violation rate are unchanged (the ranking does not read the reference scores), but the gap-marginalized conformal bound of Proposition 2 does not transfer. Practical deployment of the substitute requires either extending the validation to a larger human-anchored set or formally re-deriving the bound under the substituted exchangeability assumption.*

E Extended Discussion: the Judge-Anonymity Defense

This appendix gives the formal argument behind the Judge-Anonymity choice stated in Section 3.2.

Pre-validation fairness floor. An aggregator F is applied to score profiles before the population of judges $\mathcal{J}$ has been validated on the specific task the leaderboard measures. Per-judge reliability is itself an empirical quantity, downstream of repeated leaderboard outputs; weighting judges by reliability inside F therefore presupposes a measurement the aggregator is meant to produce. Anonymity makes the use of any per-judge weight an explicit pre-aggregation step (e.g., a separate calibration phase whose outputs are reported), rather than a hidden parameter inside the aggregator.

Relationship to Non-Dictatorship. Arrow’s original theorem uses Non-Dictatorship (no single voter is dictator for all profiles) as the fourth axiom. Step 1 of Theorem 1’s proof establishes a *fixed* dictator $v^* = (j^*, c^*)$ via the ultrafilter argument. Therefore Non-Dictatorship already suffices for the impossibility: Theorem 1 holds with the weaker axiom $\{\text{UD}, \text{WP}, \text{IIA}, \text{Non-Dictatorship}\}$, which is Arrow’s theorem instantiated on the virtual electorate $\mathcal{V} = \mathcal{J} \times \mathcal{C}$. Judge-Anonymity is strictly stronger than Non-Dictatorship and is not required for the impossibility itself. We retain it in the main statement because it is the axiom the *frontier* needs: the drop-A face (Best-of-Judge) is only meaningful if Anonymity is an axiom one might want to keep, and the criterion-weight characterization of Theorem 2’s proof (Equation 4) specifically requires Anonymity’s permutation-invariance.

The framework names the trade-off explicitly.

The drop-Anonymity face of the frontier (Theorem 3) is occupied by the Best-of-Judge witness: a rule that designates one judge as oracle and ignores the others. A practitioner with strong per-judge reliability evidence has exactly this rule available, with the axiomatic cost (Anonymity violation at 100%) reported transparently. The frontier characterization thus encodes judge-reliability-driven weighting as one deliberate face, not as an attempt to escape the impossibility.

F Running Example: Axiom Illustrations

Consider a leaderboard with two judges (J_1, J_2), two criteria (helpfulness, safety), and three candi-

dates (A, B, C) . Each judge scores each candidate on each criterion from 1–10, producing a $2 \times 2 \times 3$ score profile. We illustrate each axiom.

UD: the aggregator must return a complete ranking for every possible combination of scores, including pathological ones. Majority Vote fails UD when voters produce a Condorcet cycle ($A \succ B \succ C \succ A$).

WP: if both judges score A above B on every criterion, the final ranking must place A above B . A rule that discretizes scores before ranking (e.g., rounding to bins) can violate WP by collapsing a unanimous strict preference into a tie.

IIA: removing C from the leaderboard should not flip the A -vs- B ranking when no judge’s A -vs- B ordering changes. Borda violates IIA because removing C changes A ’s and B ’s Borda counts by different amounts; Score-Mean violates the ordinal axiom through magnitude sensitivity: re-scoring without C can change the magnitudes of A ’s and B ’s scores while preserving every judge’s A -vs- B ordering, and the mean can then flip (Appendix A.3, Step 3).

A: swapping J_1 and J_2 ’s score sheets should not change the final ranking. Best-of-Judge violates A because it designates one judge as oracle.

PI: presenting candidates as (A, B) or (B, A) in the judge’s prompt should yield identical scores. In our audit, PI failed at rates between 12.5% and 34.0%, depending on the judge and criterion.