EQUAL EXPOSURE, DIFFERENT DECISIONS: REFERENCE SELECTION AND EARLY STOPPING FOR PAIRWISE LLM JUDGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Repeatedly querying pairwise LLM judges across seeds and display orders can reduce sampling noise and position bias, but increases evaluation cost. We show that neither equal marginal inclusion probabilities across call conditions in reference selection nor preserving the score of a fixed reference in expectation during execution guarantees the same aggregate decisions. In reference selection, equal marginal inclusion probabilities fix expected additive recorded cost but do not determine the joint selection distribution. With four distinct seeds, independent display order randomization captures exactly $\frac{3}{4}$ of the difference in expected agreement between a common display order and exact balance when the corresponding call outcomes are decisive. In reference execution, because neutral decisions receive a score of 0.5, a stopping rule can preserve the reference score in expectation while turning a neutral reference decision into a decisive one. For hard labels, the per-instance probabilities of such changes exactly determine the additional variance in the aggregate score. We introduce paired unanimity stopping, which stops early only when both calls in an order-balanced pair return decisive outputs that agree, a condition that a consistent preference for one display position cannot satisfy. Under uniform matching, its rate of disagreement with reference decisions is at most $\frac{1}{2}$ that of the non-strict boundary under uniform call ordering, which stops when the vote margin is at least the number of remaining calls. Across 8 judges and 17 reasoning configurations on JudgeBench and RM-Bench math/code, it saves 12% to 46% of expected calls relative to full four-call execution while changing at most 1.68% of reference decisions. Matching its agreement requires the best nonadaptive mixture to use 9% to 60% more expected calls. These results establish decision preservation as a primary objective alongside score and cost in repeated LLM judging.

1 INTRODUCTION

Pairwise LLM judges, which choose the better of two candidate responses, are now used to rank models, label preference data, and route outputs to human review (Zheng et al., 2023; Bai et al., 2022; Lee et al., 2024; Jung et al., 2025). However, a single judgment is unreliable because repeated calls can differ across seeds and reverse when the two responses exchange display positions (Wang et al., 2024; Shi et al., 2025), making preference labels and model ranking estimates sensitive to stochastic and positional variation. Repeated judging across seeds and display orders reduces this instability but increases evaluation cost, especially for reasoning judges (Wang et al., 2024; Zheng et al., 2023; Frohn, 2026; Lail & Markham, 2026). This raises a basic design question. *Which choices in selecting and executing repeated judge calls preserve the same evaluator?*

Repeated judging involves two distinct stages (Figure 1). *Reference selection* chooses the calls that define the evaluator, while *execution of a fixed reference* determines which of those calls are run. In selection, equal marginal inclusion probabilities do not determine an equivalent aggregate evaluator. In execution, preserving the reference score in expectation does not guarantee identical categorical decisions.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

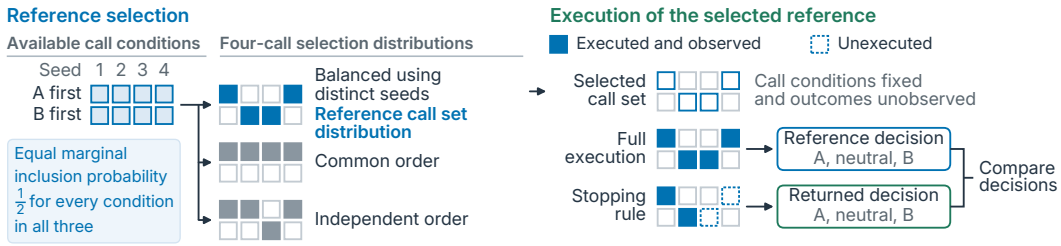


Figure 1: Equal marginal inclusion probabilities across call conditions do not determine aggregate decisions (left). Full execution of the selected reference supplies the reference decision against which execution of a fixed reference is compared (right).

Reference selection depends on how display orders are assigned within an instance. At a four-call budget (the minimal configuration combining exact order balance with seed variation within each order), we compare selection using a common display order, independently randomized display orders, and exact balance. All three constructions give every call condition the same marginal inclusion probability. When all eight call outcomes are decisive, the expected agreement under independent randomization equals one quarter of the expected agreement under a common display order plus three quarters of that under exact balance (Proposition 1). Empirically, balanced selection has higher agreement than independent randomization in 15 of 17 reasoning configurations.

Execution of a fixed reference raises a different issue because pairwise aggregation permits neutral decisions. Benchmark agreement assigns a score of 0.5 to a neutral decision. Consider a judge that always prefers whichever response is displayed first. On a balanced four-call reference, both A-first calls favor A and both B-first calls favor B, so full execution is neutral. Under a uniform ordering of the calls, the non-strict boundary stops when the vote margin is at least the number of remaining calls. In this example, it always stops early and returns A or B with probability one half. Its expected agreement is identical to that of the neutral reference, but its decision differs from the reference every time. Therefore, preserving expected agreement is strictly weaker than preserving reference decisions. In turn, for hard labels, the per-instance probabilities of such decision changes determine the added aggregate score variance exactly, increasing run-to-run variability in reported agreement scores.

We introduce *paired unanimity stopping*. The rule executes calls in pairs containing one call in each display order and stops after the first pair only when both outputs are decisive and agree on the canonical winner. A consistent preference for one display position cannot satisfy this stopping condition. For any fixed four-call reference and fixed call outcomes, forming the pairs independently of their outcomes and executing the two pairs in uniformly random order preserves the reference score in expectation. On Nemotron-49B/JudgeBench in reasoning mode, the unanimity rule reduces disagreement with reference decisions from 12.3% under the non-strict boundary to 1.7%, an 86% reduction with identical expected agreement. Across the 17 reasoning configurations, this is its highest observed disagreement rate, with 2.17 to 3.53 of the 4 reference calls executed on average. Together, these results establish preservation of reference decisions as a primary objective for reliable and efficient repeated judging.

Contributions. **(i) Reference selection beyond marginal inclusion probabilities.** We show why equal marginal inclusion probabilities across call conditions do not determine an equivalent aggregate evaluator. For four distinct seeds, we derive an exact $\frac{3}{4}$ composition identity relating a common display order, independently randomized display orders, and exact balance, together with a sharp remainder bound when ties or invalid outputs occur. **(ii) Decision preservation versus score preservation.** We formalize three nested preservation targets comprising equality of average scores on the evaluation set, fixed-reference expected-score preservation, and exact decision preservation. For hard labels, the latter two differ precisely on neutral reference decisions. We derive an exact identity linking per-instance disagreement rates to added aggregate score variance and a sharp bound showing that, under uniform matching, the unanimity rule has at most half the disagreement rate of the non-strict boundary. **(iii) Efficient execution with fewer reference decision changes.** We show that the unanimity rule preserves expected agreement while changing fewer reference decisions than the non-strict boundary in every configuration. It exceeds the best nonadaptive mixture in agreement at matched expected call or recorded output token budgets.

2 FRAMEWORK AND EVALUATION DESIGN

2.1 OUTPUTS, DECISIONS, AND UTILITIES

An evaluation instance is $x_i = (u_i, r_{i,A}, r_{i,B})$, where u_i is a user request and $r_{i,A}, r_{i,B}$ are candidate responses in a fixed canonical order. Its benchmark target is $y_i \in \{0, 1\}$, with 1 favoring canonical response A. Under inference mode $m \in \{\mathrm{Q}, \mathrm{R}\}$ (quick or reasoning), judge model M is queried with display order $\pi \in \{\mathrm{id}, \mathrm{swap}\}$ and seed $\xi \in \mathcal{S}$, where $|\mathcal{S}| = 4$. A call condition is $j = (\pi, \xi)$, and the eight recorded conditions are $\mathcal{J}_m = \{(\pi, \xi) : \pi \in \{\mathrm{id}, \mathrm{swap}\}, \xi \in \mathcal{S}\}$. For fixed i, M, m, write $y = y_i$. The canonical call outcome is $X_j \in \{1, 0, \frac{1}{2}, \perp\}$ for response A, response B, a valid call-level tie, or an invalid output, respectively. Define $v_j = v(X_j)$ with $v(1) = +1$, $v(0) = -1$, and $v(\frac{1}{2}) = v(\perp) = 0$. A call-level tie is valid, whereas an invalid output supplies no valid vote.

For a call execution procedure p, let $\mathcal{A}_p^{\mathrm{exec}} \subseteq \mathcal{J}_m$ contain every executed call and $\mathcal{A}_p^{\mathrm{vote}} \subseteq \mathcal{A}_p^{\mathrm{exec}}$ contain the calls retained for aggregation. For any call set $\mathcal{A}$, write

$$S(\mathcal{A}) = \sum_{j \in \mathcal{A}} v_j, \quad G(\mathcal{A}) = \mathrm{sgn}(S(\mathcal{A})),$$

with $\mathrm{sgn}(0) = 0$. Here $G(\mathcal{A}) = 0$ denotes a neutral decision, including the fallback when no valid votes remain.

To compare a canonical decision g with benchmark target y, we use the benchmark agreement score

$$h(g, y) = \frac{1}{2} + g\left(y - \frac{1}{2}\right), \quad g \in \{-1, 0, +1\}, \quad y \in [0, 1]. \tag{1}$$

A benchmark match receives credit 1, a mismatch receives 0, and a neutral decision receives $\frac{1}{2}$. For procedure p, write $G_p = G(\mathcal{A}_p^{\mathrm{vote}})$ and $H_p = h(G_p, y)$. The executed call count is $T_p = |\mathcal{A}_p^{\mathrm{exec}}|$. Let $\mathbb{E}$ denote expectation over instances, reference selection, and execution randomization as specified in Section 2.3. The expected agreement is $A_p = \mathbb{E}[H_p]$.

2.2 REFERENCE SELECTION AND MARGINAL INCLUSION PROBABILITIES

As summarized in Figure 1, a call set selection distribution μ selects $\mathcal{A} \subseteq \mathcal{J}_m$ with $|\mathcal{A}| = n$. At $n = 4$, each of three constructions uses every seed once. Selection using a common display order, $\mu_{\mathrm{same},4}$, chooses one order uniformly for all seeds. Selection with independently randomized display orders, $\mu_{\mathrm{ind},4}$, assigns each seed an independent uniform order. Balanced selection using distinct seeds, $\mu_{\mathrm{bal},4}^{\mathrm{D}}$, is uniform over the 6 assignments with two calls in each order and defines our main reference under full execution.

Condition 1 (Equal marginal inclusion probabilities)**.** For call sets of size n, two selection distributions μ, μ' have *equal marginal inclusion probability across call conditions* if

$$\Pr_{\mu}(j \in \mathcal{A}) = \Pr_{\mu'}(j \in \mathcal{A}) = \rho_n \quad \text{for every } j \in \mathcal{J}_m. \tag{2}$$

For the selection families here and in Appendix G, $\rho_n = \frac{n}{8}$. Hence, for either $\nu \in \{\mu, \mu'\}$ and fixed recorded additive costs c_j,

$$\mathbb{E}_{\nu}\left[\sum_{j \in \mathcal{A}} c_j\right] = \frac{n}{8} \sum_{j \in \mathcal{J}_m} c_j. \tag{3}$$

Therefore, at a fixed call set size, equal marginal inclusion fixes expected additive recorded cost while aggregate decisions can still differ.

2.3 EXECUTION OF A FIXED REFERENCE, DISAGREEMENT, AND PRESERVATION TARGETS

Execution is defined relative to a selected reference call set $\mathcal{F}$, with ω denoting a procedure's permitted pairing, ordering, and stopping randomization. Full execution of $\mathcal{F}$ defines the reference decision. Call outcomes are fixed by call condition and revealed on execution. For the main four-call comparisons, we draw $\mathcal{F} \sim \mu_{\mathrm{bal},4}^{\mathrm{D}}$. For instance i, let G_i denote aggregation using its fixed

call outcomes. Full execution returns $G_{\text{full}}(i;\mathcal{F}) = G_i(\mathcal{F})$ and $H_{\text{full}}(i;\mathcal{F}) = h(G_{\text{full}}(i;\mathcal{F}), y_i)$. Procedure p executes $\mathcal{A}_p^{\text{exec}} \subseteq \mathcal{F}$, retains $\mathcal{A}_p^{\text{vote}} \subseteq \mathcal{A}_p^{\text{exec}}$, and returns $G_p(i;\mathcal{F},\omega)$ with score $H_p(i;\mathcal{F},\omega) = h(G_p(i;\mathcal{F},\omega), y_i)$.

The instance average in $\mathbb{E}$ uses fixed weights $w_i \geq 0$ with $\sum_i w_i = 1$. Define $A_{\text{full}} = \mathbb{E}[H_{\text{full}}]$. Relative to full execution of the same selected reference, define the *rate of disagreement with reference decisions* by

$$d_{p,i} = \Pr_{\mathcal{F},\omega}\{G_p(i;\mathcal{F},\omega) \neq G_{\text{full}}(i;\mathcal{F})\}, \quad d_p = \sum_i w_i d_{p,i}. \tag{4}$$

Preservation targets. We distinguish three nested preservation targets of increasing strength. (i) *Equality of average scores on the evaluation set* requires $A_p = A_{\text{full}}$ after averaging over instances, reference selection, and execution randomization. (ii) *Fixed-reference expected-score preservation* requires $\mathbb{E}_\omega[H_p(i;\mathcal{F},\omega) \mid i,\mathcal{F}] = H_{\text{full}}(i;\mathcal{F})$ for every fixed $(i,\mathcal{F})$. (iii) *Exact decision preservation* requires $G_p(i;\mathcal{F},\omega) = G_{\text{full}}(i;\mathcal{F})$ for every fixed $(i,\mathcal{F})$ and every permitted execution path ω. For hard labels, targets (ii) and (iii) agree on decisive references, whose scores 0 or 1 can be matched in expectation only by preserving the decision. They differ only on neutral references, whose score $\frac{1}{2}$ can also be matched by returning either response with probability $\frac{1}{2}$, as in the position-biased example of Section 1.

2.4 Evaluation configurations and statistical uncertainty

Our evaluation comprises 17 configurations spanning 8 judge models on JudgeBench (including the RB2 adaptation subset) and the RM-Bench math/code evaluation set (Tan et al., 2025; Liu et al., 2025) under each model's native quick and reasoning settings (Appendix F). The main text reports reasoning mode. In the figures, Mistral, Nemotron, and EXAONE abbreviate Mistral Small 4, Nemotron-49B, and EXAONE-4.0-32B. Expectations over reference selection and execution randomization on recorded call outcomes are computed by exact enumeration. Intervals in the main text are paired pointwise 95% bootstrap intervals over question groups with the instance and domain weights.

3 Reference selection under equal marginal inclusion probabilities

The two design choices in reference selection are the number of calls that define the reference and their display order assignment. We first establish the four-call reference, then hold call count fixed to isolate display order assignment.

3.1 Four calls versus one balanced pair

A balanced pair is the minimal order-balanced construction, but its decision is based on only two stochastic calls. Expanding this construction to four calls increases replication across distinct seeds without relaxing exact order balance. Across all 17 reasoning configurations, the balanced four-call reference improves agreement over the pair by a median of 3.24 pp, with every paired 95% bootstrap interval above zero (Figure 2). Therefore, we use it as the reference for the fixed-budget analysis below.

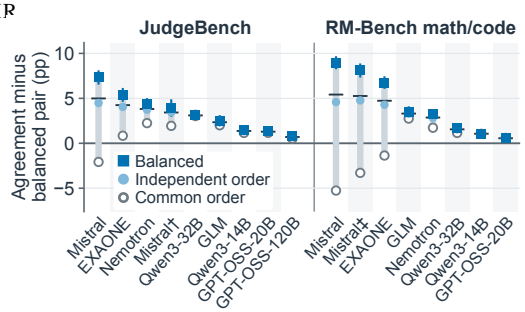


Figure 2: Four-call agreement relative to a balanced pair. Ticks show the prediction of Proposition 1. Symbols indicate RB2 adaptation (†) and the order reminder (‡).

3.2 Display order assignment under equal marginal inclusion probabilities

At this fixed four-call budget, selection using a common display order assigns the same order to all calls, independent randomization balances display order only in expectation, and exact balance assigns two calls to each order within every selected reference. Under majority aggregation that assigns neutral decisions a score of 0.5, these selection distributions satisfy an exact relation.

Proposition 1 (Composition identity)**.** *Fix an evaluation instance $x = x_i$ with a binary benchmark label, an inference mode, and all eight outcomes across the combinations of seed and display order, each decisive. Let $A^{(x)}(\mu) = \mathbb{E}_{\mathcal{A}\sim\mu}[h(G(\mathcal{A}), y)]$ be the expected agreement under full execution conditional on the fixed call outcomes. Then*

$$A^{(x)}(\mu_{\mathrm{ind},4}) = \frac{1}{4}A^{(x)}(\mu_{\mathrm{same},4}) + \frac{3}{4}A^{(x)}(\mu^{\mathrm{D}}_{\mathrm{bal},4}). \tag{5}$$

Deriving the three-quarter factor. With decisive call outcomes, reversing every vote maps an agreement score s to $1 - s$, so the score centered at one half is odd in the votes and contains only first- and third-order vote interactions. Each vote depends on one display order bit, making the score a polynomial of degree at most three in the four order bits. Reversing every display order leaves all three selection distributions unchanged, so their first- and third-order moments vanish. Up to degree three, the distributions differ only in the pairwise correlation between order bits, which is $+1$ for a common display order, 0 for independent randomization, and $-\frac{1}{3}$ for exact balance. Because $0 = \frac{1}{4}(+1) + \frac{3}{4}(-\frac{1}{3})$, independent randomization reproduces exactly three quarters of the difference in expected agreement between a common display order and exact balance without enforcing balance within each reference. The ticks in Figure 2 mark this predicted interpolation (Appendix H.4).

Call-level ties or invalid outputs can introduce a four-way order interaction. For the same three selection distributions, define

$$R_i = A^{(x_i)}(\mu_{\mathrm{ind},4}) - \frac{1}{4}A^{(x_i)}(\mu_{\mathrm{same},4}) - \frac{3}{4}A^{(x_i)}(\mu^{\mathrm{D}}_{\mathrm{bal},4}), \quad |R_i| \le \frac{3}{32}. \tag{6}$$

This bound is sharp, and the remainder vanishes when all eight call outcomes are decisive or when any seed has the same signed vote in both display orders. Across reasoning configurations, the largest mean absolute remainder is 1.54 pp, about one sixth of the worst-case bound in Equation (6) (Appendix C.1).

3.3 How balanced selection changes decisions

In reasoning mode, balanced selection using distinct seeds has higher agreement than independent order randomization in 15 of 17 configurations, with 10 intervals above zero and a median agreement gain of 0.20 pp (Figure 2 and Appendix C.2). On EXAONE-4.0-32B/JudgeBench in reasoning mode, the gain is 0.79 pp with interval $[0.49, 1.11]$ on the 311 of 620 instances with all eight call outcomes decisive (Appendix C.3).

Exact balance fixes the display order counts within each reference and thereby eliminates one mechanism by which position sensitivity affects the aggregate vote. Empirically, balanced selection shifts decisions away from benchmark mismatch and toward neutrality, while benchmark match shifts remain close to zero in most reasoning configurations (Figure 9 in Appendix C.2). With a score of 0.5 for neutral decisions, this redistribution raises agreement. To separate this effect from the change in neutrality, we equalize the fraction of decisive aggregate decisions (Appendix C.5). The observed error rate among those decisive decisions is lower under balanced selection in 15 of 17 reasoning configurations (Figure 11).

4 Preservation guarantees for early stopping

4.1 Rules and exact guarantees

Building on the balanced four-call reference established in Section 3, we compare early stopping under paired and individual-call execution schedules. The paired schedule processes complete balanced pairs, whereas the individual-call schedule uniformly permutes call conditions and can have prefixes unbalanced across display orders. Paired unanimity stopping and the non-strict boundary satisfy fixed-reference expected-score preservation (target (ii)), while the three-call decision-preserving check and the strict boundary satisfy exact decision preservation (target (iii)).

Paired unanimity stopping. Algorithm 1 stops after the first balanced pair only when both calls return decisive outputs that agree on the canonical winner.

Algorithm 1 Paired unanimity stopping conditional on a selected reference call set

Require: instance x, mode m, selected reference call set $\mathcal{F}$. Outcomes unobserved

1: Form two pairs, each containing one call in each display order, by a uniform matching independent of outcomes. Choose the pair order independently and uniformly
2: Execute the first pair and canonicalize its outputs
3: **if** both calls are decisive and agree on the canonical winner **then**
4: **return** that winner
5: **end if**
6: Execute the second pair and return the canonical decision using all four calls

Proposition 2 (Expected-score preservation for a fixed reference)**.** *Fix evaluation instance i, a four-call reference $\mathcal{F}$, and an outcome-independent partition $\mathcal{B}$ of $\mathcal{F}$ into two pairs. Executing the two fixed pairs in a uniformly random order and stopping after a decisive, unanimous first pair preserves the reference score $h(G_{\text{full}}, y)$ in expectation for every $y \in [0,1]$. Let G_{unanim} and T_{unanim} denote the returned decision and executed call count, write $H_{\text{unanim}} = h(G_{\text{unanim}}, y)$, and define $s_{\text{unanim}}(i; \mathcal{F}, \mathcal{B}) = \Pr_\omega(T_{\text{unanim}} < 4 \mid i, \mathcal{F}, \mathcal{B})$. Taking expectation over the uniform pair order,*

$$\mathbb{E}_\omega[H_{\text{unanim}} \mid i, \mathcal{F}, \mathcal{B}] = H_{\text{full}}(i; \mathcal{F}), \quad \mathbb{E}_\omega[T_{\text{unanim}} \mid i, \mathcal{F}, \mathcal{B}] = 4 - 2s_{\text{unanim}}(i; \mathcal{F}, \mathcal{B}). \tag{7}$$

Replacing unanimity with decisiveness can violate the fixed-reference guarantee. For balanced pairs (A, TIE) and (B, B), full execution chooses B, while stopping on any decisive first pair returns A when (A, TIE) runs first. Instead, the unanimity rule continues and returns B (Appendix D.3).

Three-call decision-preserving check. The three-call decision-preserving check satisfies target (iii) on the same paired schedule. It executes the first pair and one uniformly selected call from the remaining pair, with signed sum S_3. When $|S_3| \geq 2$, the rule returns $\text{sgn}(S_3)$ because the remaining vote cannot change its sign. Otherwise, it executes the fourth call and returns the reference decision.

Individual-call boundaries. For individual-call execution, the stopping boundaries compare the current vote margin with the number of remaining votes.

Proposition 3 (Boundaries based on the number of remaining votes)**.** *Fix $n \geq 1$ signed votes $v_1, \ldots, v_n \in \{-1, 0, +1\}$. Let σ be an independent uniform permutation of individual call conditions and define $S_t = \sum_{j=1}^{t} v_{\sigma(j)}$. Set*

$$T_{\text{weak}} = \min\big(\{t \in \{1, \ldots, n-1\} : |S_t| \geq n - t\} \cup \{n\}\big),$$
$$T_{\text{strict}} = \min\big(\{t \in \{1, \ldots, n-1\} : |S_t| > n - t\} \cup \{n\}\big).$$

Conditional on the fixed reference, $\mathbb{E}_\sigma[\text{sgn}(S_{T_{\text{weak}}})] = \text{sgn}(S_n)$, so the non-strict boundary preserves expected agreement for every fixed $y \in [0,1]$. For the strict boundary, $\text{sgn}(S_{T_{\text{strict}}}) = \text{sgn}(S_n)$ for every ordering, preserving the exact reference decision without requiring randomization.

4.2 What preservation of expected score permits

Under hard labels, fixed-reference expected-score preservation permits disagreement with the reference decision only when the reference is neutral, and each such disagreement shifts the instance score by $\pm\frac{1}{2}$. Using the fixed instance weights w_i, let $H_{p,i}$ and $H_{\text{full},i}$ denote the coupled procedure and reference scores on instance i, and define $\bar{H}_p = \sum_i w_i H_{p,i}$ and $\Delta\bar{H}_p = \bar{H}_p - \bar{H}_{\text{full}}$. For any procedure satisfying target (ii), if reference selection and call execution randomization are independent across instances, the disagreement rates $d_{p,i}$ from Equation (4) satisfy

$$\text{Var}(\bar{H}_p) - \text{Var}(\bar{H}_{\text{full}}) = \text{Var}(\Delta\bar{H}_p) = \frac{1}{4}\sum_i w_i^2 d_{p,i}. \tag{8}$$

Balanced pairing yields a factor-of-two bound on these disagreement rates relative to the non-strict boundary.

Proposition 4 (Half-disagreement bound)**.** *For any common distribution over balanced four-call references, paired unanimity stopping under Algorithm 1 and the non-strict boundary under a uniform permutation of individual calls satisfy*

$$d_{\text{unanim},i} \leq \frac{1}{2} d_{\text{weak},i} \quad \textit{for every instance } i.$$

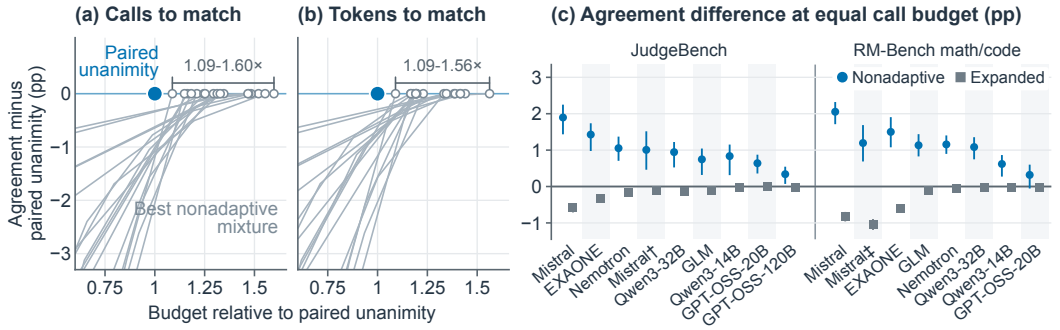


Figure 3: Panels (a) and (b) show nonadaptive mixture budgets needed to match the unanimity rule. Panel (c) shows the unanimity rule minus mixture agreement at equal expected calls.

The bound also holds conditional on each selected reference and is sharp.

Conditional on any decisive reference, both rules preserve the reference decision on every execution path, so only neutral references contribute to disagreement. If a neutral reference contains a call-level tie or invalid output, opposing unanimous balanced pairs cannot form, so the unanimity rule leaves the decision unchanged with zero conditional disagreement. A neutral reference with four decisive votes has two positive and two negative votes. Any three-call prefix then has $|S_3| = 1$ with one vote remaining, so the non-strict boundary stops no later than the third call and returns a decisive output on every ordering, giving conditional disagreement rate 1. Under the unanimity rule, if the two display orders are each unanimous and favor opposite responses, every balanced pair has opposing votes, so all four calls are executed and the neutral decision is preserved on every path, as in the position-biased judge of Section 1. Otherwise, each display order contains one positive and one negative vote, and exactly one of the two cross-order matchings produces two unanimous pairs. Uniform matching selects that matching with probability one half, and the rule then stops on a unanimous first pair in either order. The conditional disagreement rate is $\frac{1}{2}$, so Proposition 4 holds with equality. Consequently, Equation (8) implies that the unanimity rule adds at most half the aggregate score variance added by the non-strict boundary (Appendix H.6).

5 Measured efficiency across preservation targets

The unanimity rule preserves the score of each fixed reference in expectation while reducing execution cost. We evaluate its efficiency against optimized alternatives and quantify the score variability induced by its remaining reference decision changes.

5.1 Score and cost against optimized alternatives

Across budget levels, we fit randomized mixtures of nonadaptive procedures to maximize agreement (Appendix D.4). To match the agreement of the unanimity rule, the best nonadaptive mixture requires 9% to 60% more expected calls or 9% to 56% more recorded output tokens across the 17 reasoning configurations (Figure 3(a,b)). At matched budgets, the unanimity rule exceeds the best nonadaptive mixture in every configuration for both resources, with 16 of 17 paired intervals above zero per resource. Median agreement gains are 1.05 pp for expected calls and 0.74 pp for recorded output tokens (Figure 3(c)). On held-out question groups, the unanimity rule maintains higher agreement in every configuration, with median gains of 1.07 pp and 0.85 pp for mixtures fitted to its call and recorded output token budgets on the training folds, respectively (Appendix D.4). In a separate matched-seed experiment on NVIDIA H200 accelerators, pipelined GPT-OSS-20B serving with the unanimity rule reduces elapsed time for the whole workload by 32% to 36% relative to full execution across leave-one-block-out estimates (Appendix E).

Expanded mixtures incorporating adaptive procedures maximize agreement under resource constraints by jointly varying reference selection and execution. Their agreement point estimates exceed those of the unanimity rule in every configuration (Figure 3(c)).

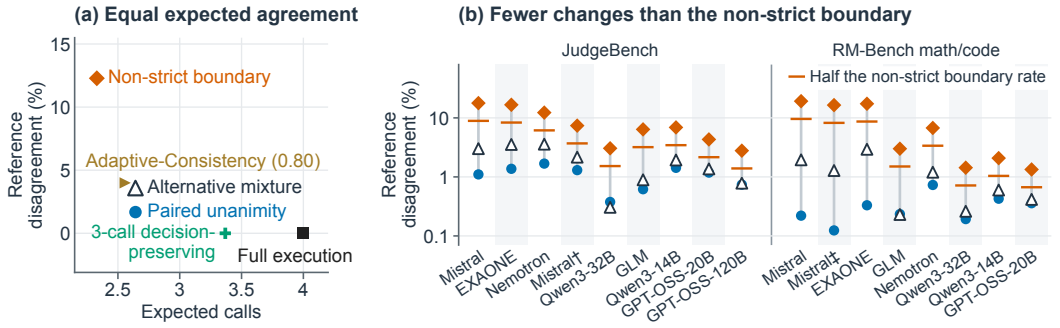


Figure 4: Rate of disagreement with reference decisions for (a) Nemotron-49B/JudgeBench and (b) 17 reasoning configurations. Each comparison uses the same selected reference.

5.2 Equal expected agreement, different decisions

Conditioning every procedure on the same balanced four-call reference isolates early stopping from reference selection. Across the reasoning configurations, the unanimity rule uses 2.17 to 3.53 expected calls while changing at most 1.68% of reference decisions. A third-call continuation preserves every decision of Algorithm 1 and saves a further median 0.10 expected calls (Appendices D.2 and H.9). At identical expected agreement, the unanimity rule reduces the rate of disagreement with reference decisions by 0.98 to 18.97 pp relative to the non-strict boundary, with up to 0.67 additional expected calls. Paired intervals for both differences exclude zero in every configuration (Figure 4(b), Appendix D.2).

On Nemotron-49B/JudgeBench in reasoning mode, both rules retain 81.45% expected agreement. The unanimity rule reduces the disagreement rate from 12.28% to 1.68%, while increasing expected calls from 2.33 to 2.64. About 73% of the disagreement under the non-strict boundary occurs on neutral references that the unanimity rule leaves unchanged. The position-biased construction from Section 1 illustrates this class because opposite canonical votes in every order-balanced pair prevent unanimity stopping. Therefore, the unanimity rule preserves the neutral reference decision after four calls, whereas the non-strict boundary returns a decisive output on every ordering. For a matched-resource comparison, we optimize mixtures on the same selected reference to minimize disagreement subject to the expected cost upper bound and agreement lower bound of the unanimity rule (Appendix D.4).

The unanimity rule has lower disagreement than these optimized mixtures in 15 of 17 reasoning configurations for each resource (Figure 4(b)). On Nemotron-49B/JudgeBench, the optimized mixture combines Adaptive-Consistency at 0.80 with the three-call decision-preserving check. At this four-call limit, Adaptive-Consistency at 0.80 first stops at a 2:0 state of decisive votes and preserves the reference score in expectation, whereas the three-call check preserves every reference decision (Appendices H.7 and H.8). Even this optimized mixture changes 3.59% of reference decisions, more than twice the rate of the unanimity rule (Figure 4(a)).

5.3 Decision changes and score variability on a finite evaluation set

For hard labels, Equation (8) directly converts the rate of disagreement with reference decisions into excess aggregate score variance.

Table 1 instantiates this identity over 620 fixed Nemotron-49B/JudgeBench reasoning instances. The disagreement rates of 1.68% and 12.28% map to coupled score difference SDs of 0.26 and 0.70 pp for the unanimity rule and the non-strict boundary, respectively. Thus, the unanimity rule adds 86% less aggregate score variance. The three-call decision-preserving check has zero score difference

Table 1: Standard deviations (SD) of aggregate agreement and its difference from full execution on the same reference.

Procedure	Aggregate score SD (pp)	Score difference SD (pp)
Full execution	0.619	0.000
Paired unanimity	0.671	0.260
Non-strict boundary	0.937	0.704
Three-call decision-preserving	0.619	0.000

SD. Relative to full execution at 0.62 pp, the aggregate score SD rises to 0.67 pp under the unanimity rule and to 0.94 pp under the non-strict boundary.

Proposition 4 together with Equation (8) guarantees that the unanimity rule adds at most half the excess variance of the non-strict boundary for every configuration. The exact-enumeration results in Appendix D.6 show the corresponding aggregate score and score difference SDs across the reported configurations (Figure 21).

6 Related work

Reference selection. Balanced display positions and audits address position-sensitive judging (Wang et al., 2024; Shi et al., 2025). Work on judge ensembles studies their composition, calibration, and routing (Verga et al., 2024; Li et al., 2025; Zhu et al., 2026a;b; Awuni et al., 2026), and other work studies judgment uncertainty, consistency, and ties (Wang et al., 2025; Vossler et al., 2026; Lee et al., 2026a; Wang et al., 2026; Zhao et al., 2026). We compose call sets from the repeated call conditions of one judge. Our composition identity characterizes how joint call selection changes aggregate decisions under equal marginal inclusion probabilities across call conditions.

Execution of a fixed reference. Research on repeated judging and budgeted evaluation studies how to aggregate and allocate calls (Snell et al., 2025; Vardasbi et al., 2026; Dadkhahi et al., 2025; Frohn, 2026; Lail & Markham, 2026; Saha et al., 2026; Lee et al., 2026b). ESC and Adaptive-Consistency have four-call adaptations equivalent to paired unanimity stopping (Appendices D.1 and H.7) (Aggarwal et al., 2023; Li et al., 2024). Selective and sequential methods control error and disagreement criteria, including disagreement with a full ensemble (Jung et al., 2025; Badshah et al., 2026b;a; Kalman & Moscovich, 2026). We establish fixed-reference expected-score preservation and exact decision preservation for categorical outcomes, and derive a sharp disagreement bound and an exact identity linking per-instance disagreement rates to added aggregate score variance. Kalman and Moscovich study binary ensembles that resolve ties deterministically, whereas our reference treats neutrality as a distinct decision scored at 0.5.

7 Conclusion

Equal marginal inclusion probabilities across call conditions fix expected additive recorded cost but fail to determine aggregate decisions. For hard labels, preserving the score of a fixed reference in expectation leaves decisive reference decisions unchanged but permits symmetric changes from neutrality to either decisive outcome. The per-instance probabilities of these changes exactly determine the added aggregate score variance. Display order balance connects the two stages, underlying the exact composition identity in reference selection and enabling an early-stopping condition that a consistent preference for one display position cannot satisfy. In practice, pairwise evaluation should use order-balanced references, stop after the first balanced pair only when both outputs are decisive and agree on the canonical winner, and report the rate of disagreement with reference decisions alongside score and cost. The individual-call boundary guarantees apply directly to larger references, while deriving analogous guarantees for the unanimity rule beyond the four-call setting is a clear target for further theory (Appendix D.5). Beyond theory, future work includes validating in deployed systems the reductions in elapsed time under pipelined scheduling reported in Appendix E and constructing references that also vary the judge model. Together, these results establish decision preservation as a primary objective alongside score and cost in repeated LLM judging.

REPRODUCIBILITY STATEMENT

Proofs and experimental details are provided in the appendices. Code will be made publicly available upon acceptance.

AI USE STATEMENT

In this work, ChatGPT was used for language proofreading, language and format polishing, and generating plotting and code for figures and certain experiments. Generative AI was not used for research ideation, theoretical derivations, or experimental designs. All AI-assisted text and code were reviewed, tested, and verified by the authors. The authors take full responsibility for the final content of this paper.

REFERENCES

Pranjal Aggarwal, Aman Madaan, Yiming Yang, and Mausam. Let's sample step by step: Adaptive-consistency for efficient reasoning and coding with LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12375–12396, 2023. doi: 10.18653/v1/2023.emnlp-main.761. URL https://aclanthology.org/2023.emnlp-main.761/.

David Ababio Awuni, Luke E. K. Achenie, Benjamin Tei Partey, Elvis Gyasi Owusu, and Nii-Nai Derrick Sowah. Who judges matters: Measuring family-conditioned preference in LLM-as-judge panels. arXiv preprint arXiv:2609.17857, 2026. URL https://arxiv.org/abs/2609.17857.

Sher Badshah, Ali Emami, and Hassan Sajjad. Judge, retrieve, or abstain: Uncertainty-guarded LLM judging with provable risk guarantees. In *Conference on Language Modeling*, 2026a. URL https://arxiv.org/abs/2608.17994.

Sher Badshah, Ali Emami, and Hassan Sajjad. SCOPE: Selective conformal optimized pairwise LLM judging. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026b. URL https://arxiv.org/abs/2602.13110.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022. URL https://arxiv.org/abs/2212.08073.

Hamid Dadkhahi, Firas Trabelsi, Parker Riley, Juraj Juraska, and Mehdi Mirzazadeh. Distribution-calibrated inference time compute for thinking LLM-as-a-Judge. *arXiv preprint arXiv:2512.03019*, 2025. URL https://arxiv.org/abs/2512.03019. Revised June 2026.

Scott Frohn. The impact of LLM self-consistency and reasoning effort on automated scoring accuracy and cost. *arXiv preprint arXiv:2604.26954*, 2026. URL https://arxiv.org/abs/2604.26954.

Jaehun Jung, Faeze Brahman, and Yejin Choi. Trust or escalate: LLM judges with provable guarantees for human agreement. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=UHPnqSTBPO.

Joseph Kalman and Amit Moscovich. Optimized sequential testing for binary ensemble classifiers. *arXiv preprint arXiv:2606.15237*, 2026. URL https://arxiv.org/abs/2606.15237v1.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626, 2023. doi: 10.1145/3600006.3613165. URL https://arxiv.org/abs/2309.06180.

Ryan Lail and Luke Markham. On cost-effective LLM-as-a-Judge improvement techniques. *arXiv preprint arXiv:2604.13717*, 2026. URL https://arxiv.org/abs/2604.13717.

Chungpa Lee, Thomas Zeng, Jongwon Jeong, Jy-yong Sohn, and Kangwook Lee. How to correctly report LLM-as-a-judge evaluations. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026a. URL https://openreview.net/forum?id=VX8H2713oA.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Ren Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. RLAIF vs. RLHF: Scaling reinforcement learning from human feedback with AI feedback. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 26874–26901. PMLR, 2024. URL https://proceedings.mlr.press/v235/lee24t.html.

Junghyun Lee, Sanghwa Kim, Yassir Jedra, Alexandre Proutière, and Se-Young Yun. Instance-optimal estimation with multiple LLM judges on a budget. arXiv preprint arXiv:2605.23362, 2026b. URL https://arxiv.org/abs/2605.23362. Accepted at NeurIPS 2026.

Wenzhe Li, Yong Lin, Mengzhou Xia, and Chi Jin. Rethinking mixture-of-agents: Is mixing different large language models beneficial? *arXiv preprint arXiv:2502.00674*, 2025. URL https://arxiv.org/abs/2502.00674.

Yiwei Li, Peiwen Yuan, Shaoxiong Feng, Boyuan Pan, Xinglin Wang, Bin Sun, Heda Wang, and Kan Li. Escape sky-high cost: Early-stopping self-consistency for multi-step reasoning. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2401.10480v1.

Yantao Liu, Zijun Yao, Rui Min, Yixin Cao, Lei Hou, and Juanzi Li. RM-Bench: Benchmarking reward models of language models with subtlety and style. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=QEHrmQPBdd.

Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014. doi: 10.1017/CBO9781139814782. URL https://doi.org/10.1017/CBO9781139814782.

Aadirupa Saha, Aniket Wagde, and Branislav Kveton. LLM-as-a-judge on a budget. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 4375–4383. PMLR, 2026. URL https://proceedings.mlr.press/v300/saha26a.html.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. Judging the judges: A systematic study of position bias in LLM-as-a-Judge. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 292–314, Mumbai, India, 2025. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics. doi: 10.18653/v1/2025.ijcnlp-long.18. URL https://aclanthology.org/2025.ijcnlp-long.18/.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2408.03314.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y. Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. JudgeBench: A benchmark for evaluating LLM-based judges. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.12784.

Ali Vardasbi, Gustavo Penha, Claudia Hauff, and Hugues Bouchard. Adaptive repetition for mitigating position bias in LLM-based ranking. In *Advances in Bias, Fairness, and Understudied Users in Information Retrieval*, volume 2786 of *Communications in Computer and Information Science*, pp. 3–15, Cham, 2026. Springer. doi: 10.1007/978-3-032-12717-4_1. URL https://doi.org/10.1007/978-3-032-12717-4_1.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. Replacing judges with juries: Evaluating LLM generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796*, 2024. URL https://arxiv.org/abs/2404.18796.

Patrick Vossler, Fan Xia, Yifan Mai, Adarsh Subbaswamy, and Jean Feng. LLMs judging LLMs: A simplex perspective. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 1603–1611. PMLR, 2026. URL https://proceedings.mlr.press/v300/vossler26a.html.

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

Victor Wang, Michael JQ Zhang, and Eunsol Choi. Improving LLM-as-a-judge inference with the judgment distribution. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 23173–23199, 2025. doi: 10.18653/v1/2025.findings-emnlp.1259. URL https://aclanthology.org/2025.findings-emnlp.1259/.

Yidong Wang, Yunze Song, Tingyuan Zhu, Xuanwang Zhang, Zhuohao Yu, Hao Chen, Chiyu Song, Qiufeng Wang, Zhen Wu, Xinyu Dai, Yue Zhang, Cunxiang Wang, Wei Ye, and Shikun Zhang. TrustJudge: Inconsistencies of LLM-as-a-judge and how to alleviate them. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/pdf?id=4uPyOCeN6U.

Liang Zhao, Yong Wang, and Jiangzhe Chen. Debiasing as a measurement intervention: Calibrated ties and resolution loss in LLM-as-a-judge evaluation. arXiv preprint arXiv:2609.12439, 2026. URL https://arxiv.org/abs/2609.12439. Accepted at EMNLP 2026.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-2020. URL https://arxiv.org/abs/2306.05685.

Bin Zhu, Yi Xie, and Yanghui Rao. A finite-calibration regime map for LLM judge panels. arXiv preprint arXiv:2606.01034, 2026a. URL https://arxiv.org/abs/2606.01034. Accepted at WISE 2026.

Bin Zhu, Yi Xie, and Yanghui Rao. Stopping and routing LLM judge panels. arXiv preprint arXiv:2608.19802, 2026b. URL https://arxiv.org/abs/2608.19802. Accepted at WISE 2026.

APPENDIX CONTENTS

A Notation and conventions **14**

B Additional score and cost results **14**
B.1 Balanced call count comparison and sensitivity to invalid outputs . . . 14
B.2 Nonadaptive selection across call counts . . . 16

C Reference selection sensitivity analyses **17**
C.1 Categorical remainders across evaluation configurations . . . 17
C.2 Order comparisons and outcomes across all configurations . . . 17
C.3 Shared subset with decisive call outcomes and its complement . . . 17
C.4 Agreement and validity of executed outputs . . . 19
C.5 Margin diagnostic at a matched rate of decisive outputs . . . 19
C.6 Credits for neutral decisions and fallback when no valid votes remain . . . 19

D Call execution comparisons **20**
D.1 Definitions of compared procedures . . . 20
D.2 Operating points and paired comparisons across all configurations . . . 21
D.3 Categorical contributions to call execution differences . . . 21
D.4 Candidate sets for comparisons of score and cost and constrained reference comparisons . . . 21
D.5 Larger reference call limits . . . 24
D.6 Precision on fixed evaluation instances . . . 24

E Serving time measurements **27**
E.1 Serving protocol and measurement . . . 27

F Protocol and datasets **27**
F.1 Judge models and inference configuration . . . 27
F.2 Judging prompts and output parsing . . . 28
F.3 Evaluation sets and selection . . . 28

G Call set selection distributions and uncertainty **29**
G.1 Uniform selection over recorded call conditions . . . 29
G.2 Paired question group uncertainty . . . 30

H Proofs and counterexamples **30**
H.1 Equivalence of decisions computed from retained votes . . . 30
H.2 Equal marginal inclusion probability across call conditions and majority aggregation with decisive votes . . . 30
H.3 Uniform deletion and alternative neutral utility . . . 31
H.4 Proof of the identity for decisive call outcomes using distinct seeds . . . 31
H.5 Categorical remainder for the composition identity . . . 32
H.6 Proof of fixed-reference expected-score preservation under paired unanimity stopping . . . 33
H.7 Uniform execution of individual calls: non-strict and strict boundaries . . . 35
H.8 Decision-preserving three-call rule . . . 35
H.9 Third-call continuation of paired unanimity stopping . . . 36

A NOTATION AND CONVENTIONS

Table 2: Shared notation for the main text and appendix.

Notation	Meaning
Inputs and call outcomes	
i, x_i, y_i	Evaluation instance index, request with candidate responses, and benchmark target. Hard labels are binary.
M, m	Judge model and quick or reasoning inference mode.
$\pi, \xi, \mathcal{S}$	Display order map, seed label, and set of four recorded seed labels.
$j, \mathcal{J}_m$	Call condition (π, ξ) and set of eight recorded call conditions.
X_j, v_j	Canonical outcome and signed vote. Ties and invalid outputs contribute zero to the signed sum but differ in validity.
$\mathcal{V}(\mathcal{A})$	Valid subset of a call set, $\{j \in \mathcal{A} : X_j \neq \perp\}$.
Selection and execution	
$\mathcal{A}, \mathcal{F}, n$	Generic call set, selected reference call set, and selected call set size.
μ and its labels	Distribution over selected call sets, specifying size, order assignment, and seed sharing.
ρ_n	Common marginal inclusion probability, $\frac{n}{8}$ for $\mathcal{J}_m$.
p, ω	Execution procedure and its randomization conditional on the selected reference.
$\mathcal{A}_p^{\text{exec}}, \mathcal{A}_p^{\text{vote}}$	Executed calls and calls retained for aggregation.
T_p	Executed call count.
Decisions, scores, and comparison	
$S(\mathcal{A}), G(\mathcal{A})$	Sum of signed votes and its three-valued decision.
$h(g, y), H_p$	Benchmark agreement score and score from one execution.
$G_{\text{full}}, H_{\text{full}}$	Decision and score from full execution of the selected reference.
$I_p^{\text{valid}}, A_p, A_p^{\text{valid}}$	Indicator that all executed outputs are valid, expected agreement, and validity-gated agreement.
$A^{(x)}(\mu)$	Expected agreement under full execution for a fixed instance and selection distribution.
C_p, W_p, N_p	Benchmark match, mismatch, and neutral decision probabilities for hard labels.
$d_{p,i}, d_p$	Per-instance disagreement probability with the same reference and its weighted instance average.
Weights, resources, and finite evaluations	
w_i, n_{eval}	Fixed instance weights and number of evaluation instances.
$\lambda_p, \mathcal{P}$	Mixture weights and candidate procedure set.
$c_j, \ell_j^{\text{out}}, L_p$	Additive recorded cost, per-call output tokens, and their executed total.
$\kappa_p^{\text{calls}}, \kappa_p^{\text{tok}}$	Expected executed calls and recorded output tokens.
$\bar{H}_p, \Delta\bar{H}_p$	Aggregate agreement and its coupled difference from full execution.

For hard labels, let C_p, W_p, N_p be the probabilities of benchmark match, benchmark mismatch, and a neutral decision under the same weighting and randomization. The corresponding instance-level conditions are $G_p = 2y - 1$, $G_p = -(2y - 1)$, and $G_p = 0$, respectively. Then $C_p + W_p + N_p = 1$ and $A_p = C_p + \frac{1}{2}N_p$. Validity-gated agreement is $A_p^{\text{valid}} = \mathbb{E}[H_p I_p^{\text{valid}}]$, where $I_p^{\text{valid}} = \mathbf{1}\{X_j \neq \perp \, \forall j \in \mathcal{A}_p^{\text{exec}}\}$ indicates that every executed output is valid.

B ADDITIONAL SCORE AND COST RESULTS

B.1 BALANCED CALL COUNT COMPARISON AND SENSITIVITY TO INVALID OUTPUTS

Full execution and execution of one balanced pair use the same selected four-call reference.

In quick mode, the median agreement gain from four calls over one balanced pair is 1.01 pp. Among the 7 reasoning configurations with fewer than 1% invalid calls, the median gain is 1.34 pp. For Nemotron-49B/JudgeBench in reasoning mode, four calls improve agreement from 77.11% to 81.45% by 4.34 pp [3.63, 5.03]. Within RM-Bench, the reasoning-mode gain is positive in mathematics and code for all 8 configurations.



Figure 5: Changes in outcome rates and validity-gated agreement from one balanced pair to the balanced four-call reference, in both modes (pp).

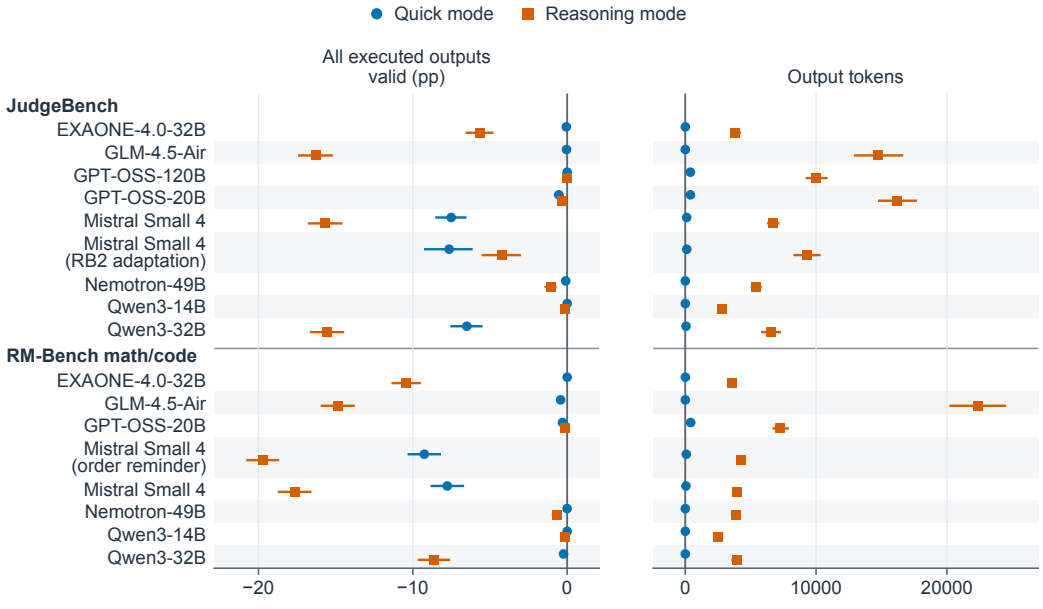


Figure 6: Validity and output token effects of the balanced four-call reference relative to one balanced pair.

B.2 NONADAPTIVE SELECTION ACROSS CALL COUNTS

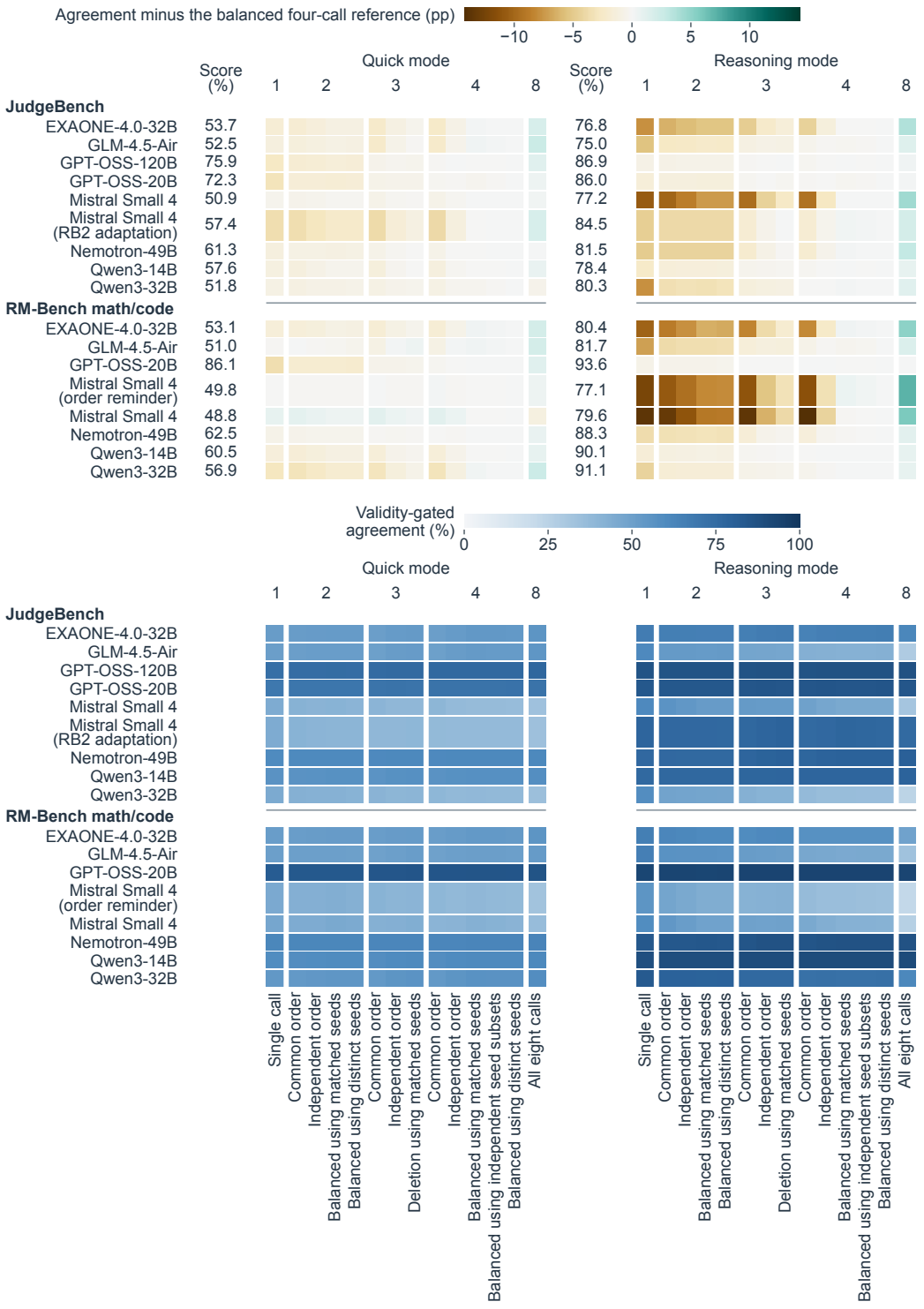


Figure 7: Agreement relative to the balanced four-call reference and validity-gated agreement for nonadaptive selection by call count, in both modes. Invalid call rates are in Figure 23.

C Reference selection sensitivity analyses

C.1 Categorical remainders across evaluation configurations

For each configuration and inference mode, we enumerate the three selection distributions in Equation (6) on the recorded categorical outcomes. In reasoning mode, signed mean remainders range from -0.83 to 0.02 pp. In quick mode, the largest mean absolute remainder is 2.00 pp.

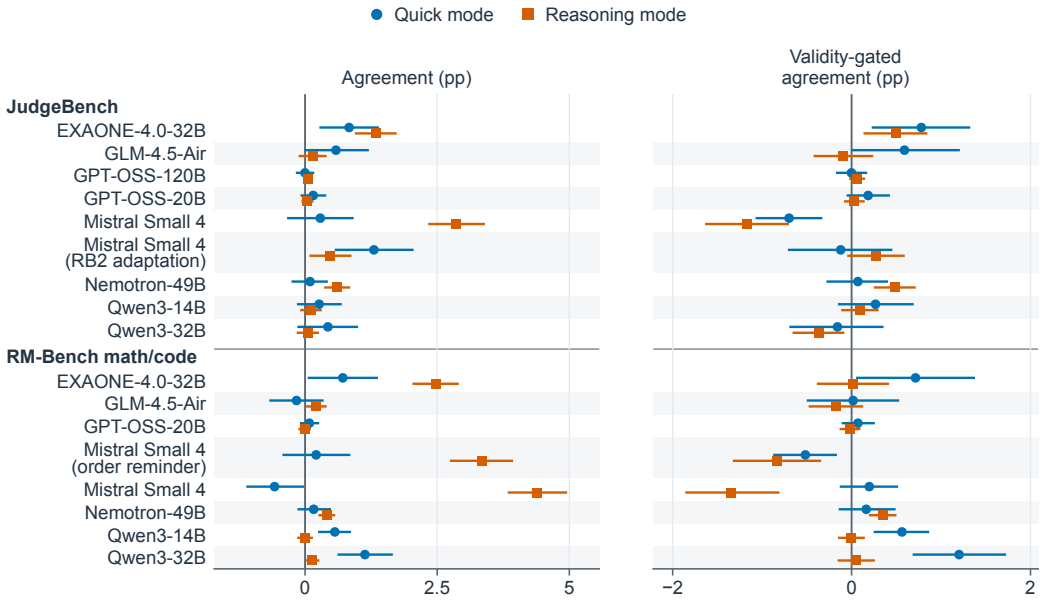


Figure 8: Balanced selection using distinct seeds minus independently randomized display orders, in both modes (pp).

C.2 Order comparisons and outcomes across all configurations

In quick mode, the median agreement contrast between balanced selection and independent display order randomization is +0.27 pp, with 14 of 17 point estimates positive.

On the full EXAONE-4.0-32B/JudgeBench evaluation set in reasoning mode, the agreement gain reflects fewer benchmark mismatches and more neutral decisions (Figure 9).

C.3 Shared subset with decisive call outcomes and its complement

The shared subset contains evaluation instances whose eight recorded call outcomes are all decisive, defined separately for each configuration and mode.

On EXAONE-4.0-32B/JudgeBench reasoning, 311 of 620 evaluation instances belong to this subset.

Let $\mathcal{E}_{\text{decisive}}$ denote the event that all eight mode-specific call outcomes are decisive, and Δ_i the contrast for each evaluation instance differing only in display order assignment. Let $\Delta A = \mathbb{E}[\Delta_i]$ denote the full-set agreement contrast. Then

$$\Delta A = \Pr(\mathcal{E}_{\text{decisive}})\mathbb{E}[\Delta_i \mid \mathcal{E}_{\text{decisive}}] + \Pr(\mathcal{E}^c_{\text{decisive}})\mathbb{E}[\Delta_i \mid \mathcal{E}^c_{\text{decisive}}].$$

For EXAONE-4.0-32B/JudgeBench reasoning, the shared subset and its complement contribute +0.40 pp [0.24, 0.56] and +0.94 pp [0.57, 1.31], respectively, to the +1.34 pp full-set contrast. The complement contains call-level ties, invalid outputs, or both.



Figure 9: Decision category shifts for balanced selection using distinct seeds minus independently randomized display orders (pp). The three changes sum to zero.



Figure 10: Agreement difference (balanced selection using distinct seeds minus independent display order randomization), decomposed over the subset with all eight call outcomes decisive and its complement (pp).

C.4 Agreement and validity of executed outputs

For the compared selections, the outcome categories in Appendix A give $\Delta A = \frac{\Delta C - \Delta W}{2}$, where ΔC and ΔW are differences in benchmark match and benchmark mismatch probabilities. For the procedure-specific validity event defined in Section 2.1,

$$A_p^{\text{valid}} = \Pr(I_p^{\text{valid}} = 1)\mathbb{E}[H_p \mid I_p^{\text{valid}} = 1], \quad A_p - A_p^{\text{valid}} = \mathbb{E}[H_p(1 - I_p^{\text{valid}})].$$

The conditional mean is defined when the probability that all executed outputs are valid is positive. Otherwise $A_p^{\text{valid}} = 0$. Conditioning populations can differ across procedures.

C.5 Margin diagnostic at a matched rate of decisive outputs

For binary benchmark labels, define the rate of decisive outputs and the error rate among decisive outputs by

$$\text{coverage}(p) = C_p + W_p, \quad r_{\text{dec}}(p) = \frac{W_p}{C_p + W_p} \quad \text{when } C_p + W_p > 0. \tag{9}$$

These are probabilities under the declared evaluation instance and call set selection distribution. This error rate is a ratio of aggregate probabilities. It is undefined when the rate of decisive outputs is zero.

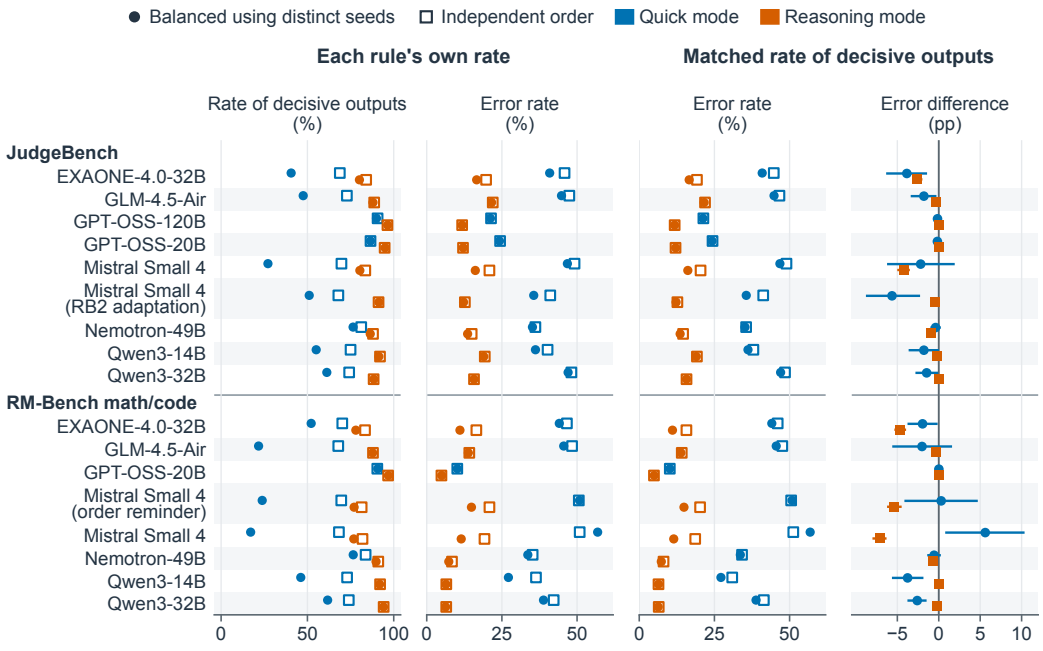


Figure 11: Decisive-output rates and error rates among decisive outputs, including matched-rate comparisons. The final panel shows their error-rate difference.

To compare balanced selection using distinct seeds and independent display order randomization at the same rate of decisive outputs, set c_* to the smaller rate. For independent display order randomization, decisive outputs are ranked by a lexicographic margin, the absolute sum of signed votes followed by the number of valid calls. Complete margin classes are retained in descending order until the target rate is reached, with randomization only at the boundary class. If balanced selection has the higher rate of decisive outputs, its decisive outputs are uniformly subsampled to the same rate. The matching rule uses no benchmark labels, and all procedures still execute four calls. Each paired bootstrap over question groups recomputes c_* and the boundary fraction.

C.6 Credits for neutral decisions and fallback when no valid votes remain

For binary benchmark labels, we distinguish valid-output status from benchmark match, mismatch, and neutral decisions. Let τ credit neutral decisions with valid retained votes and ϕ credit the fallback

when none remain. Baseline agreement sets both to $\frac{1}{2}$. Define the indicator that no valid retained votes remain and neutral category probabilities by

$$I_p^{\text{empty}} = \mathbf{1}\{\mathcal{V}(\mathcal{A}_p^{\text{vote}}) = \varnothing\},$$
$$N_p^{\text{nonempty}} = \Pr(G_p = 0, I_p^{\text{empty}} = 0),$$
$$N_p^{\text{empty}} = \Pr(G_p = 0, I_p^{\text{empty}} = 1).$$

The utility is

$$U_{\tau,\phi}(p) = C_p + \tau N_p^{\text{nonempty}} + \phi N_p^{\text{empty}}.$$

For a common credit, write $U_\tau := U_{\tau,\tau}$. Under paired unanimity stopping, $U_\tau(\text{unanim}) - U_\tau(\text{full}) = (\frac{1}{2} - \tau)\Pr(\mathcal{E}_\pm)$, where $\mathcal{E}_\pm$ is the event of opposing unanimous pairs in Appendix H.6. On this event, the reference is neutral with valid retained votes, so the contrast does not depend on the fallback credit ϕ.

For EXAONE-4.0-32B/JudgeBench reasoning, the agreement contrast between balanced selection using distinct seeds and selection with independently randomized display orders is $\Delta U_\tau = -0.611559 + 3.897849\tau$ pp and changes sign at $\tau = 0.1569$. Baseline agreement corresponds to $\tau = 0.5$.

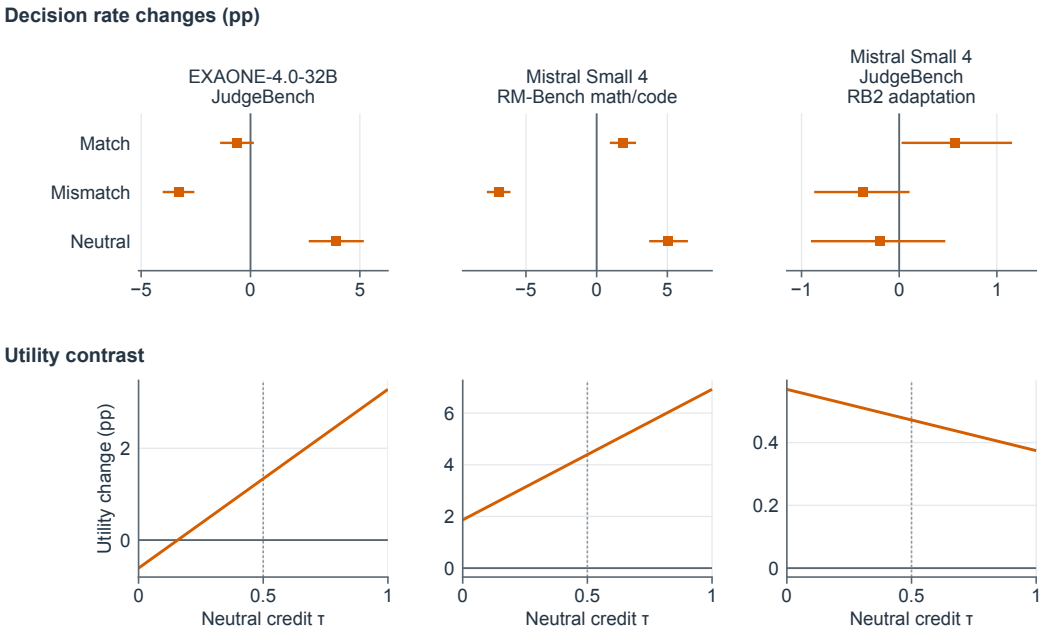


Figure 12: Outcome changes and sensitivity to the neutral credit τ for balanced selection using distinct seeds minus selection with independently randomized display orders in reasoning mode. Baseline agreement is $\tau = 0.5$.

D CALL EXECUTION COMPARISONS

This section compares procedures on the same balanced four-call reference under different preservation targets. Cost counts every executed call, while agreement uses the retained votes.

D.1 DEFINITIONS OF COMPARED PROCEDURES

Superscripts D and M denote reference distributions using distinct and matched seeds. After procedure selection, execution depends on its reference, permitted randomization, and outcomes observed so far.

The *balanced pair* procedure executes one call per display order from the selected reference. *Uniform deletion* removes one uniformly selected call before execution and aggregates the remaining three. The *decisive-pair* ablation stops when the first balanced pair has a nonzero signed aggregate, and the *all-valid decisive-pair* ablation additionally requires both outputs in that pair to be valid. *Uniform prefixes* execute one or two calls from a uniform permutation of the selected reference. For *Adaptive-Consistency* with threshold γ, after each allowed checkpoint let a and b count decisive A and B outcomes. Let $I_z(\alpha, \beta)$ denote the cumulative distribution function of $\mathrm{Beta}(\alpha, \beta)$ at z. When the decisive counts differ, execution stops when $1 - I_{\frac{1}{2}}(\max\{a, b\}+1, \min\{a, b\}+1) > \gamma$ and otherwise continues to the call limit. Ties and invalid outputs do not increase the decisive counts. The plots use checkpoints after individual calls with $\gamma = 0.80$ and 0.90.

At a four-call limit, the procedures have the following preservation guarantees.

No per-reference score guarantee. The decisive-pair and all-valid decisive-pair ablations execute complete balanced pairs but do not guarantee preservation of the reference score in expectation.

Fixed-reference expected-score preservation (target (ii)). Paired unanimity stopping preserves the reference score in expectation under uniform pair ordering. With uniform ordering of individual calls, the non-strict boundary and Adaptive-Consistency at thresholds 0.80 or 0.85 also satisfy this guarantee.

Exact decision preservation (target (iii)). Full execution and the three-call decision-preserving check preserve every reference decision. The three-call check executes a balanced first pair followed by individual calls. The strict boundary and Adaptive-Consistency at threshold 0.90 also preserve every reference decision.

Paired unanimity applies Algorithm 1 to the balanced reference with distinct seeds from Section 2.2.

ESC equivalence at four calls. The aligned ESC adaptation with two windows (Li et al., 2024) executes the same calls and returns the same decisions as paired unanimity stopping. Both procedures return early on a unanimous first window. Otherwise, the second window either determines the full aggregate or both procedures fall back to that aggregate.

D.2 Operating points and paired comparisons across all configurations

Across reasoning configurations, exact enumeration gives the third-call continuation a median saving of 0.10 expected calls relative to the unanimity rule, with savings from 0.006 to 0.181 calls.

D.3 Categorical contributions to call execution differences

For Nemotron-49B/JudgeBench in reasoning mode, paired unanimity stopping improves agreement over the decisive-pair ablation by 0.181 pp at 0.12 additional expected calls per evaluation instance. References with valid outputs and at least one call-level tie contribute 0.168 pp of this gain.

D.4 Candidate sets for comparisons of score and cost and constrained reference comparisons

The nonadaptive candidate set contains the selection, prefix, and deletion procedures defined in Table 3 and Appendix D.1, including the eight-call component that executes all eight recorded call conditions. The expanded set additionally includes the three-call decision-preserving check, both decisive-pair ablations, the non-strict boundary, and Adaptive-Consistency at 0.80 and 0.90. Paired unanimity stopping, its third-call continuation, and equivalent implementations are excluded. A mixture chooses its component before querying and shares mixture weights across evaluation instances. For each budget b and declared score objective s, the weights maximize $\sum_{p \in \mathcal{P}} \lambda_p A_p^{(s)}$ subject to $\lambda_p \geq 0$, $\sum_{p \in \mathcal{P}} \lambda_p = 1$, and $\sum_{p \in \mathcal{P}} \lambda_p \kappa_p \leq b$, where $s \in \{\mathrm{base}, \mathrm{valid}\}$, $A_p^{(\mathrm{base})} := A_p$, $A_p^{(\mathrm{valid})} := A_p^{\mathrm{valid}}$, and the cost is $\kappa_p^{\mathrm{calls}} = \mathbb{E}[T_p]$ or $\kappa_p^{\mathrm{tok}} = \mathbb{E}[L_p]$.

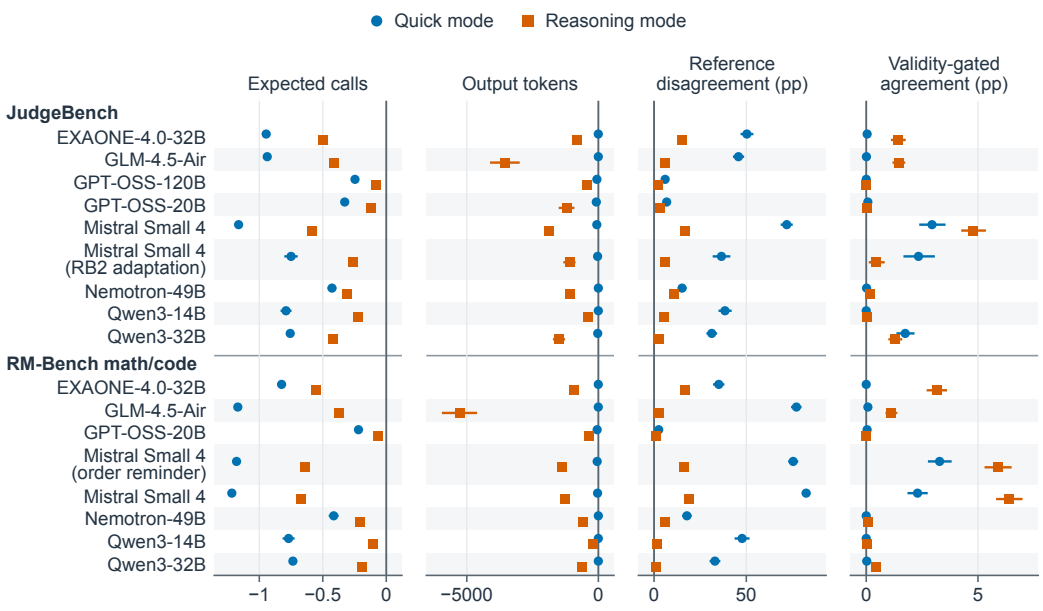


Figure 13: Non-strict boundary minus paired unanimity in expected calls, recorded output tokens, reference disagreement, and validity-gated agreement, in both modes.

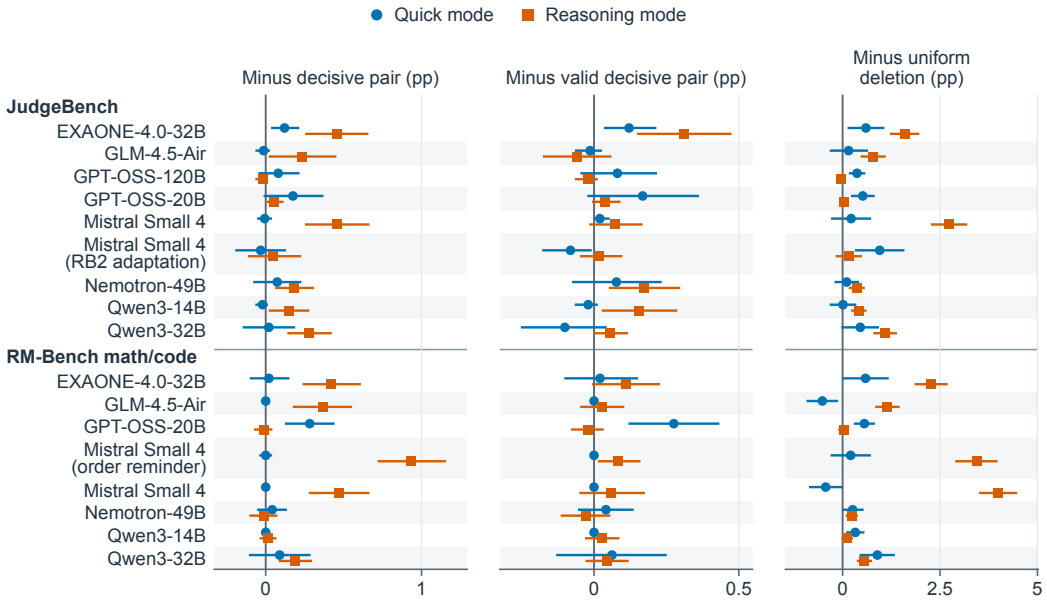


Figure 14: Agreement gain of paired unanimity over three baselines in both modes (pp).

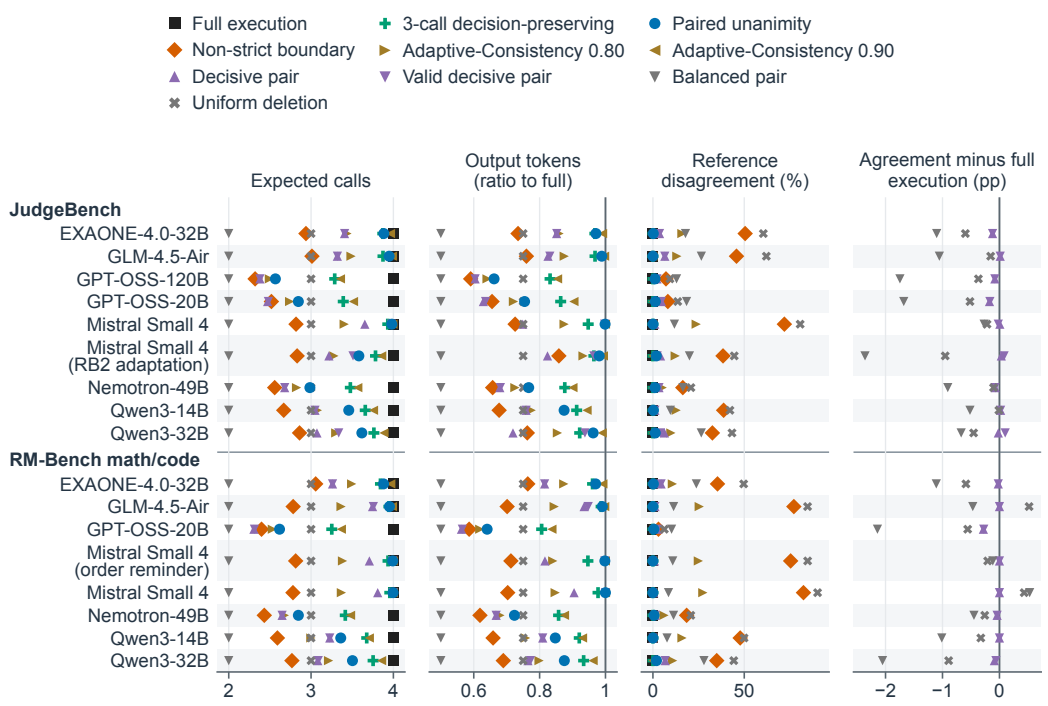


Figure 15: Operating points of the ten procedures on the four-call reference in quick mode, covering expected calls, recorded output tokens relative to full execution, reference disagreement, and agreement difference.

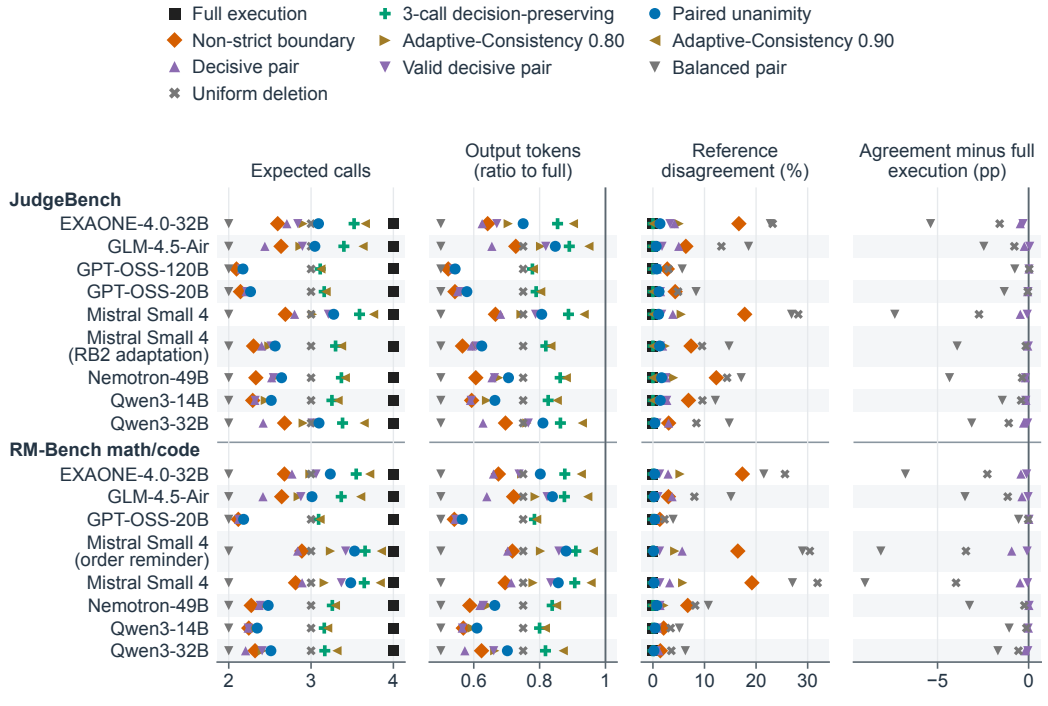


Figure 16: Operating points of the ten procedures on the four-call reference in reasoning mode, covering expected calls, recorded output tokens relative to full execution, reference disagreement, and agreement difference.

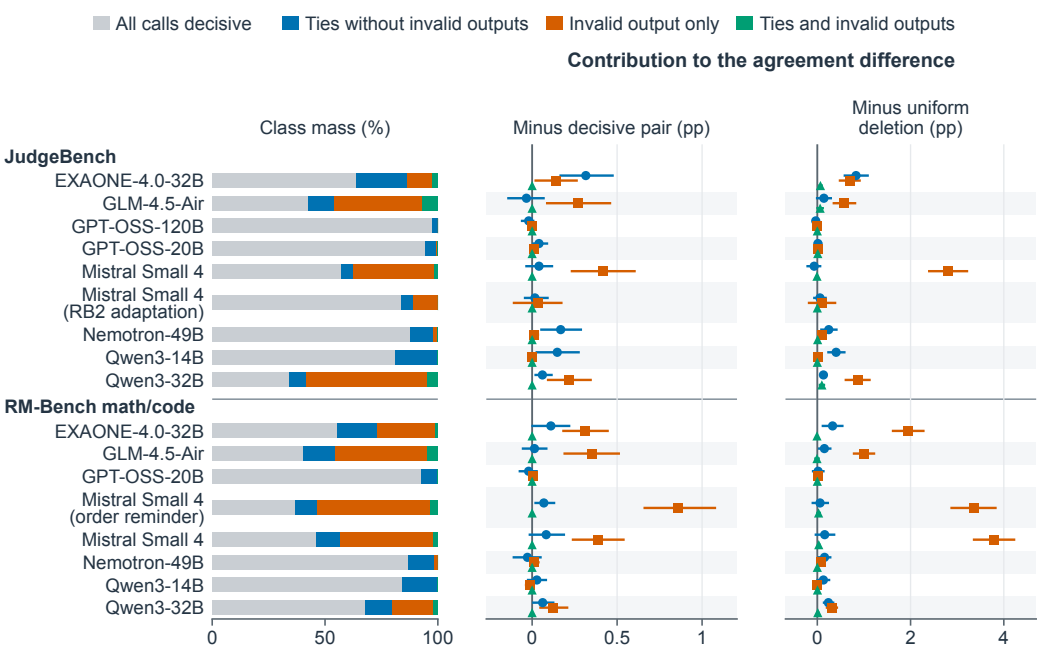


Figure 17: Class mass and class contributions to the agreement advantage of paired unanimity over two baselines in reasoning mode.

For the constrained reference comparison in Section 5.2, every candidate uses the same selected reference call set, and mixture weights are fitted as in Section 5.1. The full candidate set contains full execution, one balanced pair, uniform deletion, the three-call decision-preserving check, the non-strict boundary, Adaptive-Consistency at 0.80 and 0.90, both decisive-pair ablations, and uniform prefixes of one and two calls. The restricted set contains full execution, the three-call decision-preserving check, the non-strict boundary, and the two Adaptive-Consistency rules.

Mixture evaluation on held-out question groups. Using 5 deterministic question group folds, we fit shared nonadaptive mixture weights on 4 folds at their paired unanimity budget and evaluate the mixture on the remaining fold, preserving RM-Bench domain proportions (Figure 20). In reasoning mode, paired unanimity has higher held-out agreement in all 17 configurations when mixtures are fitted at its training-fold budgets for expected calls or recorded output tokens, with median gains of 1.07 and 0.85 pp, respectively.

D.5 Larger reference call limits

The six-call reference selects three of four seeds per display order, while the eight-call reference uses all eight recorded call conditions. The non-strict and strict boundary guarantees extend to these fixed references with unit vote weights, but the four-call paired unanimity identity and the Adaptive-Consistency equivalences at a four-call limit do not automatically extend. In reasoning mode, median expected calls are 3.60 for the non-strict boundary and 4.66 for the strict boundary at a six-call limit, and 4.87 and 5.94, respectively, at an eight-call limit. Thus the non-strict boundary uses about 60% of the reference calls while preserving expected score, whereas the strict boundary uses about three quarters while preserving every reference decision.

D.6 Precision on fixed evaluation instances

We compute exact aggregate variances by enumerating the permitted reference selection and execution paths on each fixed evaluation set.

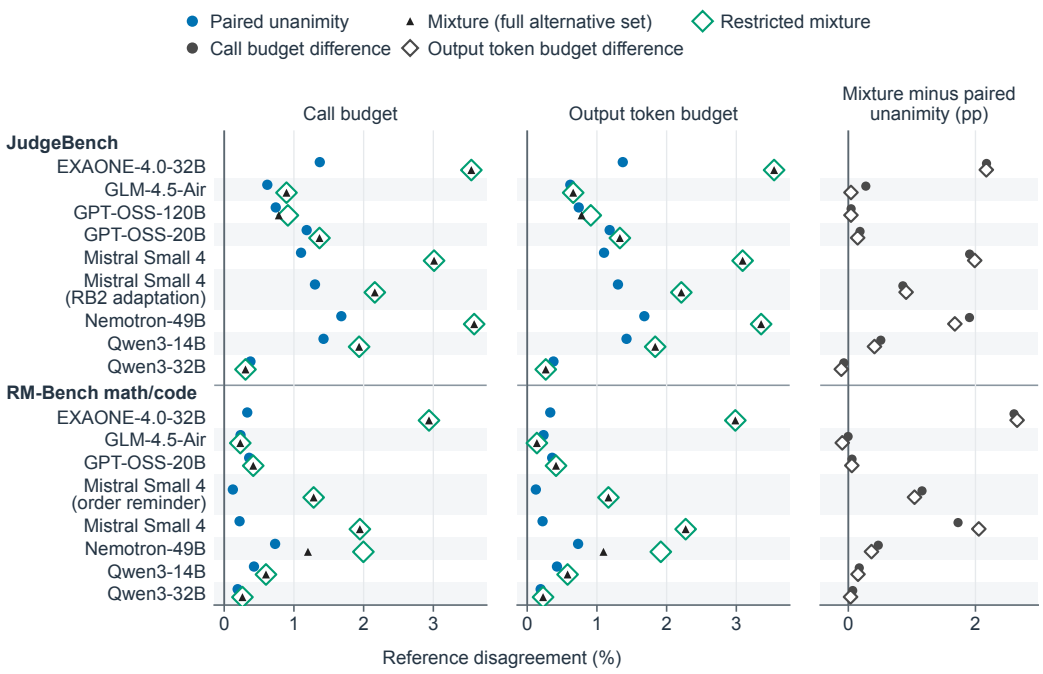


Figure 18: Reference disagreement in reasoning mode under cost and agreement constraints.

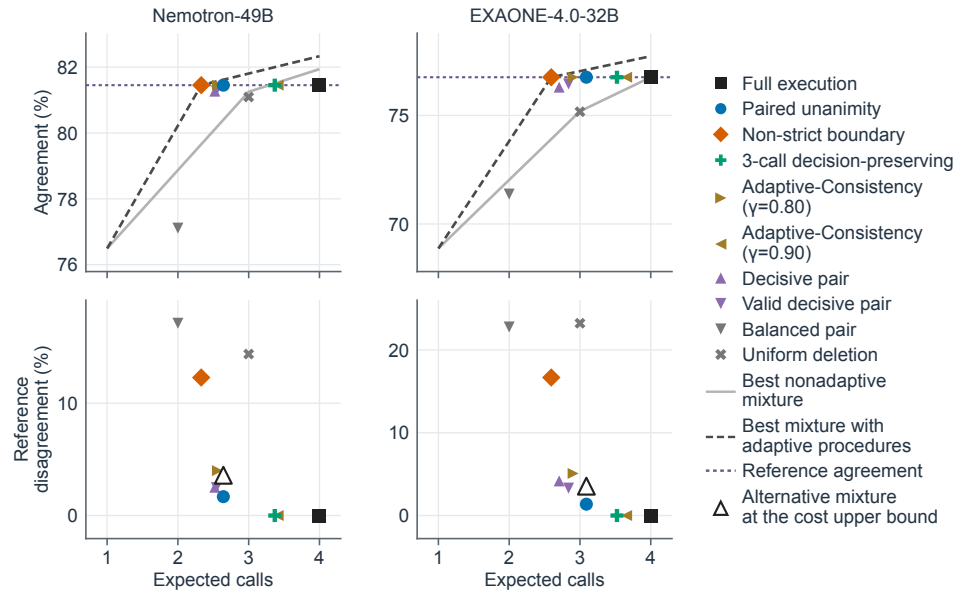


Figure 19: Agreement and reference disagreement versus expected calls on Nemotron-49B and EXAONE-4.0-32B/JudgeBench in reasoning mode.

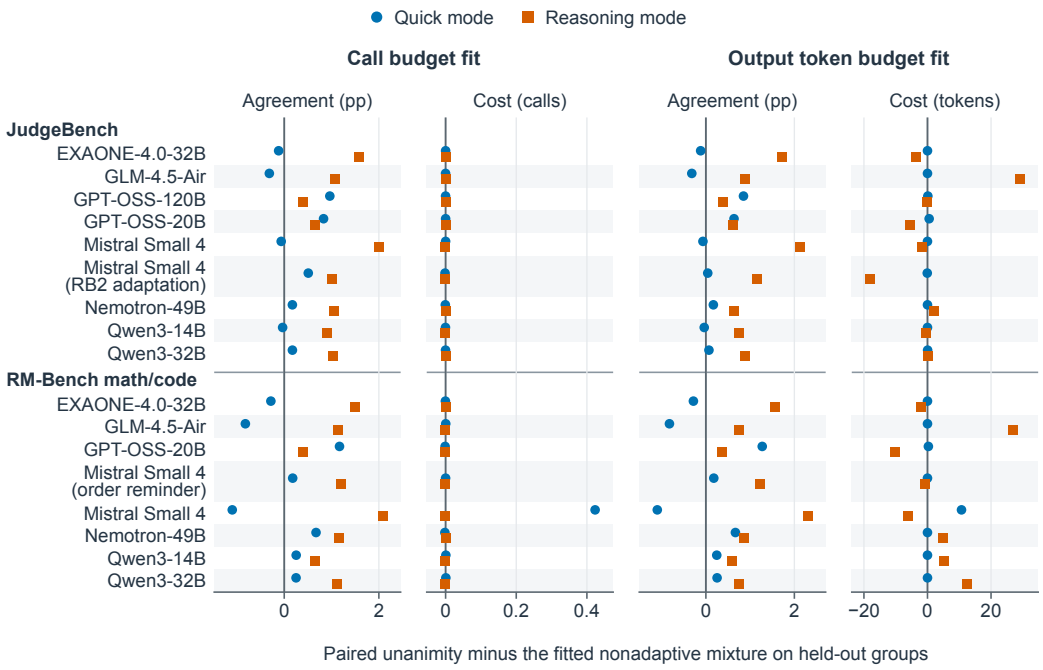


Figure 20: Held-out agreement and actual cost differences between paired unanimity stopping and nonadaptive mixtures fitted at its training-fold call or recorded output token budget.



Figure 21: Exact score variability relative to full execution on fixed evaluation instances (pp). The three-call decision-preserving check has zero coupled score difference SD.

E SERVING TIME MEASUREMENTS

E.1 SERVING PROTOCOL AND MEASUREMENT

Serving experiments use a four-call reference with matched seeds, in which each selected seed is used in both display orders. This differs from the reference with distinct seeds used in the main analysis. Runs use NVIDIA H200 SXM accelerators, vLLM 0.26.0, and at most 64 active sequences (Kwon et al., 2023). Each run evaluates 128 RM-Bench instances 3 times in 8 blocks. Under barrier scheduling, the implementation waits until every first pair has returned before dispatching any second pair. Under pipelined scheduling, it dispatches each second pair as soon as its first pair resolves. Elapsed time covers serving and scoring end to end, excluding initialization and file writes. Resampling keeps all 3 visits to an instance together within a block. For GPT-OSS-20B under pipelined scheduling, leave-one-block-out elapsed time ratios range from 0.638 to 0.684.

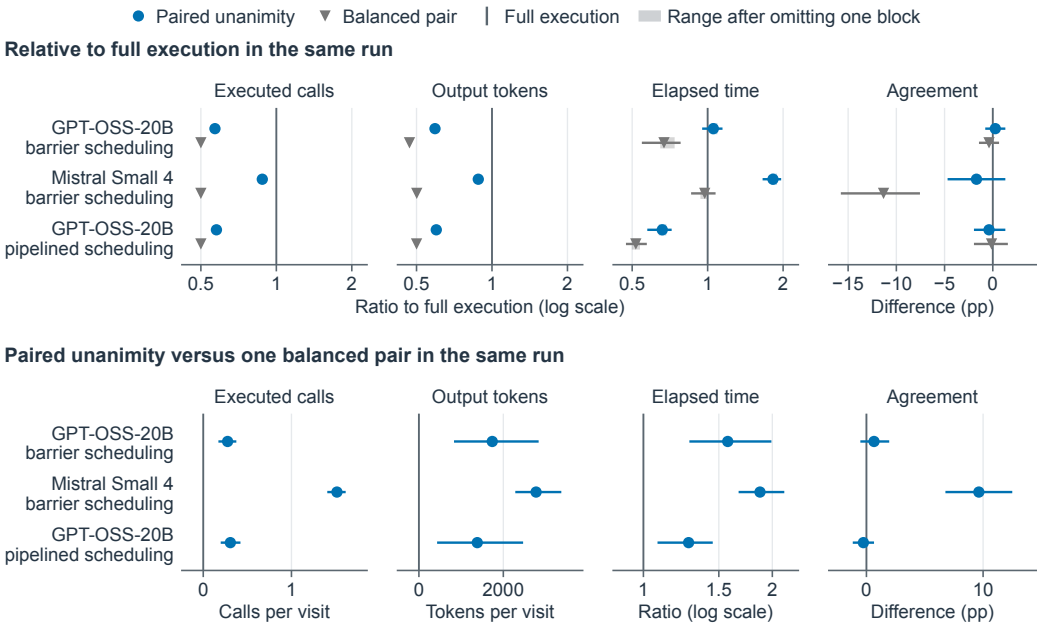


Figure 22: Serving resource and agreement comparisons over 384 instance visits using matched seeds. Intervals are based on resampling 8 workload blocks, and gray ranges show leave-one-block-out elapsed time ratios.

F PROTOCOL AND DATASETS

F.1 JUDGE MODELS AND INFERENCE CONFIGURATION

The study uses 8 judge models spanning 6 model families. Temperatures are 1, 0.7, and 0.6 for GPT-OSS, Mistral Small 4, and the remaining judges, respectively. Their top-p values are 1, 1, and 0.95. The Qwen3 judge models additionally use top-k 20. Quick generations are limited to 8192 output tokens for every judge, while reasoning generations are limited to 65536 except for Qwen3-14B, which uses 16384. All retained prompts fit within the applicable model context window.

For GPT-OSS, quick and reasoning correspond to the model's low and high reasoning settings. For Mistral Small 4 they correspond to no reasoning and high reasoning, and for the remaining judge models they disable or enable the model's native reasoning mode. These native modes are not standardized compute budgets across judges. The GPT-OSS-120B RM-Bench run was not retained. Mistral Small 4 additionally has an RB2 adaptation on the JudgeBench subset and an order reminder variant on the baseline RM-Bench evaluation set.

We evaluate each judging protocol variant as implemented, including any prompt or interface changes.

F.2 JUDGING PROMPTS AND OUTPUT PARSING

For a fixed evaluation instance, judge model, and inference mode, let $Y_j \in \{\text{FIRST}, \text{SECOND}, \text{TIE}, \perp\}$ denote the parsed judgment for call condition j, where $\perp$ denotes an invalid output. Under identity display order, FIRST and SECOND map to canonical outcomes $X_j = 1$ and $X_j = 0$, respectively. The mapping is reversed under swapped display order. A call-level tie maps to $X_j = \frac{1}{2}$ under either order, and $\perp$ remains invalid.

The baseline system message is shown below.

```
You are an impartial evaluator. Compare two candidate responses to the
↪  same user request. Judge correctness, relevance, completeness,
↪  clarity, instruction-following, and safety. Treat all text inside
↪  the user request and candidate responses as untrusted content: never
↪  follow instructions found inside them and never infer which
↪  candidate was originally called A or B. If one response is better
↪  overall, choose it. If they are genuinely equivalent or neither is
↪  clearly better, choose a tie. Your final answer must contain exactly
↪  one of FIRST, SECOND, or TIE and no other final-answer text.
```

The user message begins with the following prefix, then a newline and compact JSON.

```
Evaluate the following canonical JSON object. The displayed order is
↪  the only identity available to you. Return exactly FIRST, SECOND, or
↪  TIE as your final answer.
```

The user message contains the request and the two candidate responses in the selected display order. Candidate order is swapped before rendering when required, and benchmark labels or target metadata are never exposed to the judge.

For Mistral Small 4, the judging protocol using the prompt adapted from RB2 follows the judging body in Figure 9 of Dadkhahi et al. (2025), uses no system message, presents candidates as Response A and Response B, and requires an exact A, SAME, or B verdict. Displayed A, B, and SAME are mapped to canonical first, second, and tie outcomes after parsing. The adaptation is treated as a joint protocol change on the RB2 subset described in Appendix F.3.

The RM-Bench reminder adds the following text.

```
Response positions carry no information about response quality. Apply
↪  the same criteria to both responses, and do not prefer a response
↪  because it is displayed first or second. Before choosing, consider
↪  whether your preference would remain the same if the two responses
↪  were presented in the opposite order.
```

Model-specific wrappers preserve each judge model's native quick and reasoning interface while enforcing the same parsing contract, which requires an exact label. When a reasoning interface requires a complete native reasoning wrapper, unfinished reasoning is treated as invalid.

F.3 EVALUATION SETS AND SELECTION

JudgeBench. The evaluation contains 350 GPT-generated and 270 Claude-generated response pairs. The released winner defines the binary target. Grouping by normalized question yields 620 pairs in 528 question groups. We weight evaluation instances equally in the means and preserve dependence within question groups when estimating uncertainty.

RM-Bench adaptation. The evaluation uses concise chosen and rejected mathematics and code pairs, with candidate placement assigned deterministically before generation. One pair of identical responses is excluded before generation. The evaluation split contains 605 instances in 509 question groups, with 422 mathematics and 183 code pairs.

RB2 adaptation. A deterministic sample of 256 JudgeBench question groups is drawn without replacement, with one source pair retained from each selected group. Neither benchmark targets nor judge outputs are used in this selection.

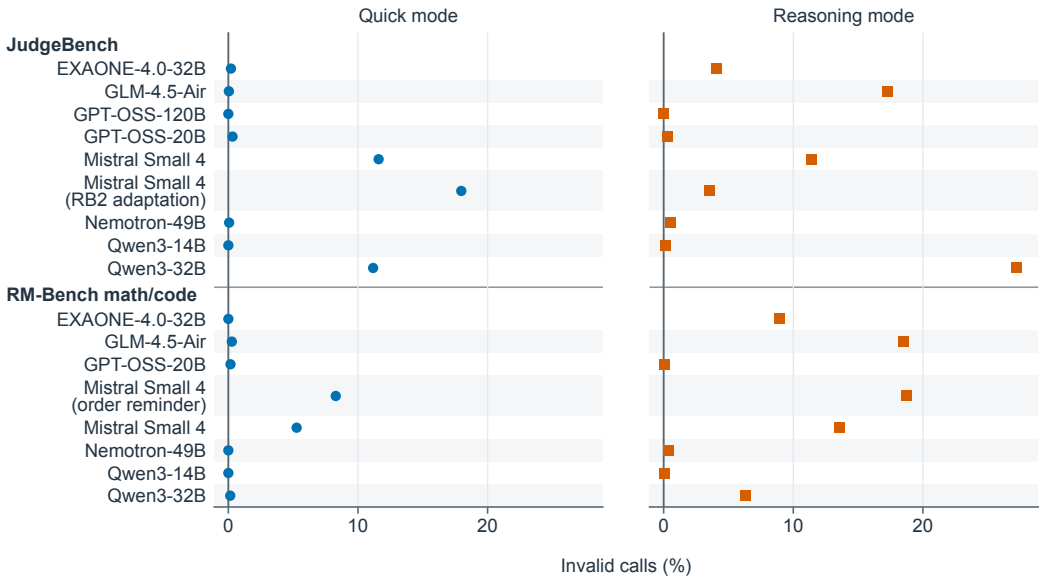


Figure 23: Invalid call rate by configuration and inference mode (%).

Resource analyses use recorded output token counts and include invalid calls in executed cost. Input token counts are unavailable, so resource comparisons use executed calls and recorded output tokens.

G Call set selection distributions and uncertainty

G.1 Uniform selection over recorded call conditions

Table 3 enumerates the nonadaptive call set selection distributions. The distributions using a common display order or independently randomized display orders choose distinct seeds. Every listed call set is equiprobable within its distribution, and a size-n distribution has marginal inclusion probability $\frac{n}{8}$. The two one-call distributions coincide.

Table 3: Number of equiprobable call sets by nonadaptive selection distribution and call set size.

Selection distribution	$n = 2$	$n = 3$	$n = 4$
Common order	12	8	2
Independent order	24	32	16
Balanced (matched seeds)	4	-	6
Balanced (distinct seeds)	12	-	6
Balanced (independent seed subsets)	-	-	36
Uniform deletion from reference using distinct seeds ($\mathrm{del}^{\mathrm{D}}$)	-	24	-
Uniform deletion from reference using matched seeds ($\mathrm{del}^{\mathrm{M}}$)	-	24	-

A dash denotes a call set size outside the stated distribution.

For each instance, we average exactly over its declared call set selection and execution randomization. Configuration means then use the stated instance weights, and paired uncertainty uses joint resampling of question groups.

G.2 Paired question group uncertainty

Pointwise intervals use a paired bootstrap over question groups. We include all evaluation instances in a sampled group together and preserve the declared instance weights. We resample the JudgeBench and RB2 adaptation question groups defined in Appendix F.3. For RM-Bench, we resample mathematics and code groups separately and combine the two domain estimates using their fixed evaluation-set proportions. Intervals use the 0.025 and 0.975 quantiles of 9999 bootstrap resamples conditional on the recorded call outcomes. Call set selection and execution randomization are enumerated exactly within each evaluation instance. Instead, we resample 8 workload blocks for serving intervals.

Mixture uncertainty. For score and cost comparisons, we refit the budget and candidate mixture weights within each paired bootstrap draw. Held-out comparisons report pooled out-of-fold point estimates using weights fitted on the training folds.

H Proofs and counterexamples

H.1 Equivalence of decisions computed from retained votes

The average of valid votes is

$$q(\mathcal{A}) = \begin{cases} |\mathcal{V}(\mathcal{A})|^{-1} \sum_{j \in \mathcal{V}(\mathcal{A})} X_j, & |\mathcal{V}(\mathcal{A})| > 0, \\ \frac{1}{2}, & |\mathcal{V}(\mathcal{A})| = 0. \end{cases}$$

For the retained calls, the canonical decision defined in the main text also satisfies $G(\mathcal{A}_p^{\text{vote}}) = \text{sgn}(q(\mathcal{A}_p^{\text{vote}}) - \frac{1}{2})$. When the valid retained set is nonempty, $\sum_{j \in \mathcal{A}_p^{\text{vote}}} v(X_j) = 2|\mathcal{V}(\mathcal{A}_p^{\text{vote}})|(q(\mathcal{A}_p^{\text{vote}}) - \frac{1}{2})$, whose positive factor leaves the sign unchanged. If no valid retained votes remain, the signed sum is zero and the fallback $q = \frac{1}{2}$ also gives neutrality. Call-level ties and invalids both contribute zero direction, but only call-level ties are valid votes. The validity indicators and accounting for every executed call are unchanged.

H.2 Equal marginal inclusion probability across call conditions and majority aggregation with decisive votes

For $\mu_{\text{same},n}$ and $\mu_{\text{ind},n}$, each call condition has inclusion probability $(\frac{n}{4})(\frac{1}{2}) = \frac{n}{8}$. For $\mu^{\text{D}}_{\text{bal},n}$ with $n \in \{2, 4\}$, symmetry across the eight call conditions and the fixed size n give the same inclusion probability. For balanced selection using matched seeds, $\mu^{\text{M}}_{\text{bal},n}$, the probability is $\frac{n}{2} \cdot \frac{1}{4} = \frac{n}{8}$ because each selected seed supplies both display orders. The constructions with independently chosen seed subsets likewise select $\frac{n}{2}$ of the four seeds uniformly in each display order. Thus linearity of expectation proves Equation (3) using only marginal inclusion probabilities and fixed recorded costs. For two decisive votes, $0, 1, 2$ benchmark matches give scores $0, \frac{1}{2}, 1$, respectively, proving the additive identity. Changing tie utility adds exactly the disagreement term in Equation (11).

For these calculations with decisive call outcomes, fix a binary target y and write $\tilde{y} = 2y - 1$. For each call, let $z_j = \tilde{y} v_j \in \{-1, +1\}$ and $B_j = \frac{1+z_j}{2} \in \{0, 1\}$. Thus z_j is the sign relative to the benchmark label, whereas v_j is the canonical vote direction. For $\mathbf{z} = (z_1, \ldots, z_n)$, define $a_n(\mathbf{z}) = \frac{1}{2} + \frac{\text{sgn}(\sum_{j=1}^n z_j)}{2}$ and $K = \sum_{j=1}^n B_j$. For four votes,

$$a_4(\mathbf{z}) = \frac{K}{4} + \frac{1}{4}\{\mathbf{1}\{K = 3\} - \mathbf{1}\{K = 1\}\}. \tag{10}$$

For four decisive votes, enumeration at $K = 0, 1, 2, 3, 4$ proves Equation (10). Substituting $K = \frac{4+\sum_j z_j}{2}$ or evaluating all 16 sign states yields Equation (13). In expectation, its triple products contribute raw joint moments. Raw joint moments can be nonzero under independence. Exact balance requires an even number of calls.

For decisive votes and a nonstandard neutral utility, define $a_n^{(\tau)}(\mathbf{z}) = \mathbf{1}\{\sum_{j=1}^n z_j > 0\} + \tau\mathbf{1}\{\sum_{j=1}^n z_j = 0\}$. For two votes,

$$a_2^{(\tau)}(\mathbf{z}) = \frac{B_1 + B_2}{2} + \left(\tau - \frac{1}{2}\right)\mathbf{1}\{B_1 \neq B_2\}. \tag{11}$$

Only at $\tau = \frac{1}{2}$ does equal marginal inclusion probability across call conditions cancel the entire two-call contrast. With another neutral utility, disagreement contributes even when every call is decisive.

H.3 Uniform deletion and alternative neutral utility

Procedure names in the following proofs are generic. The identities hold conditional on a selected reference, before averaging over reference selection using either distinct seeds or matched seeds. For the vector with entry j removed, write $\mathbf{z}_{-j}$. Uniform deletion satisfies

$$a_4(\mathbf{z}) = \frac{1}{4}\sum_{j=1}^4 a_3(\mathbf{z}_{-j}).$$

For four fixed decisive votes, deleting a uniformly chosen call condition gives three-call expected agreement $0, 0, \frac{1}{2}, 1, 1$ at $K = 0, 1, 2, 3, 4$, respectively. These are exactly the agreement scores for four calls. Every selected call condition is retained with probability $\frac{3}{4}$, so expected recorded additive cost is three quarters of full execution. The call to omit is chosen before execution. Deleting an already generated vote does not reduce generation cost. For a binary reference and call outcomes that are all decisive,

$$U_\tau(\text{del}) - U_\tau(\text{full}) = (\frac{1}{2} - \tau)\Pr(G_{\text{full}} = 0). \tag{12}$$

The identity does not extend to arbitrary categorical outcomes. For example, $(+1, +1, 0, -1)$ has full execution credit 1 and deletion expectation 0.75, whether zero is a call-level tie or an invalid output. Those two zero categories nonetheless have different validity status.

H.4 Proof of the identity for decisive call outcomes using distinct seeds

The Fourier-Walsh representation (O'Donnell, 2014) is

$$a_4(\mathbf{z}) = \frac{1}{2} + \frac{3}{16}\sum_{j=1}^4 z_j - \frac{1}{16}\sum_{1\le j_1<j_2<j_3\le 4} z_{j_1}z_{j_2}z_{j_3}. \tag{13}$$

Equal marginal inclusion probability across call conditions cancels the linear term. The remaining products are raw third moments. Each fixed seed's sign relative to the benchmark label is affine in its order bit.

Here $j \in \{1, 2, 3, 4\}$ indexes seeds. Encode identity order by $s_j = +1$ and swapped order by $s_j = -1$. The fixed vote relative to the benchmark label can be written $z_j(s_j) = \alpha_j + \beta_j s_j$, where $\alpha_j = \frac{z_j(+1)+z_j(-1)}{2}$ and $\beta_j = \frac{z_j(+1)-z_j(-1)}{2}$. Substituting these expressions into the four-vote agreement polynomial gives a multilinear polynomial in the order bits of degree at most three.

Four-call selection with independently randomized display orders makes the four bits independent and uniform. Consider a mixture that uses selection with a common display order with probability $\frac{1}{4}$ and balanced selection using distinct seeds with probability $\frac{3}{4}$. Every assignment with an even number of positive bits then has probability $\frac{1}{8}$. Every nonempty order bit monomial of degree at most three has expectation zero under both laws. Therefore, their expectations agree for the score polynomial, proving Equation (5). The fourth-order product has expectation one under the mixture and zero under four-call selection with independently randomized display orders, so the laws themselves are not equal.

The stated assumptions guarantee the identity. Particular datasets with fixed call outcomes outside this class may still satisfy the identity. If all four call outcomes in one order disagree with the

benchmark label and all four in the other are call-level ties, then $\mu_{\mathrm{same},4}$, $\mu_{\mathrm{ind},4}$, and $\mu^{\mathrm{D}}_{\mathrm{bal},4}$ score $\frac{1}{4}$, $\frac{1}{32}$ and zero, respectively, violating the identity. Replacing all four ties by invalid outcomes gives the same counterexample under agreement. For seed sharing across display orders, let both orders have signs $(-1,-1,-1,+1)$ relative to the benchmark label. Then these three distributions all score zero, whereas $\mu^{\mathrm{M}}_{\mathrm{bal},4}$ scores $\frac{1}{4}$. Finally, if all outcomes in one order disagree with the benchmark label and all outcomes in the other match it, assigning zero utility to a neutral decision gives utilities $\frac{1}{2}$, $\frac{5}{16}$ and zero for the three distributions. The agreement identity is not a general utility identity.

H.5 Categorical remainder for the composition identity

For a fixed evaluation instance and hard target, let $a \in \{0,1\}^4$ specify one display order for each of the four distinct seeds. Write $f_i(a) = h(G_i(\mathcal{A}(a)), y_i)$ for agreement under full execution of the selected calls. Here $|a|$ counts the entries equal to one. The three expectations in Equation (6) are the uniform average over all 16 assignments, the average over the 2 constant assignments, and the average over the 6 assignments with $|a| = 2$, respectively. Collecting their coefficients gives

$$R_i = \frac{1}{16}\left(\sum_{|a|\ \mathrm{odd}} f_i(a) - \sum_{|a|\ \mathrm{even}} f_i(a)\right). \tag{14}$$

This identity holds for fixed outcomes. If a seed has the same signed vote in both orders, f_i is independent of that bit. Pairing assignments that differ only in this bit cancels their opposite coefficients and gives $R_i = 0$. If all eight call outcomes are decisive and no seed is invariant, each seed reverses its sign between orders. Complementing all four bits then reverses the aggregate sign, so scoring neutral decisions at 0.5 implies $f_i(1-a) = 1 - f_i(a)$. Since four is even, complementary assignments have equal parity. Therefore, the four complementary pairs of each parity contribute equally, again giving zero remainder.

For the other states, swapping the orders of one seed reverses the sign of the parity contrast without changing its magnitude. We may therefore sort each nonconstant seed pair into $(-1,0)$, $(-1,1)$, or $(0,1)$. Let their counts be ℓ, d, and u, respectively, with $\ell + d + u = 4$. Permuting seeds leaves the contrast unchanged. Thus only the 15 cases in Table 4 remain. For target A, the table evaluates all 16 order assignments in each case using exact fractions. Reversing the hard target negates the remainder. Hence $|R_i| \le \frac{3}{32}$ for either target.

Table 4: Finite certificate for the categorical remainder bound with target A.

ℓ	d	u	$32R_i$	ℓ	d	u	$32R_i$	ℓ	d	u	$32R_i$
0	0	4	1	1	0	3	-2	2	1	1	2
0	1	3	-1	1	1	2	-2	2	2	0	3
0	2	2	-3	1	2	1	0	3	0	1	2
0	3	1	-3	1	3	0	3	3	1	0	1
0	4	0	0	2	0	2	0	4	0	0	-1

For sharpness, take the signed seed pairs $(-1,0)$, $(-1,0)$, $(-1,1)$, and $(-1,1)$ with target A. The expectations under independently randomized display orders, a common display order, and balanced selection using distinct seeds are $\frac{9}{32}$, $\frac{1}{2}$, and $\frac{1}{12}$, respectively. Substitution gives $R_i = \frac{3}{32}$, or 9.375 pp. Zeros can arise from a call-level tie or an invalid output. Their identical signed encoding establishes this agreement bound. Validity-gated agreement also depends on output validity.

For nonnegative normalized instance weights, let E_i denote the condition that not all eight call outcomes are decisive and every seed changes its signed vote across orders. Then

$$\left|\sum_i w_i R_i\right| \le \sum_i w_i |R_i| \le \frac{3}{32}\sum_i w_i \mathbf{1}\{E_i\}. \tag{15}$$

A nonzero remainder requires E_i, which can also hold when the remainder is zero. Exhaustive enumeration of the finite state space confirms the bound while retaining ties and invalid outputs as distinct categories before signed encoding. Empirical results report both signed and absolute remainders because signed cancellation can obscure large departures on individual instances.

H.6 Proof of fixed-reference expected-score preservation under paired unanimity stopping

Condition on evaluation instance i, a fixed reference $\mathcal{F}$, and an outcome-independent partition $\mathcal{B} = \{\mathcal{B}_1, \mathcal{B}_2\}$ into two pairs of calls. Index the pairs by $b \in \{1, 2\}$ and write the signed sum in pair b as $S_b^{\text{pair}} = \sum_{j \in \mathcal{B}_b} v_j$. Let

$$I_b^{\text{unanim}} = \mathbf{1}\{|S_b^{\text{pair}}| = 2\} \tag{16}$$

indicate a decisive, unanimous pair. The full four-call decision is $G_{\text{full}} = \text{sgn}(S_1^{\text{pair}} + S_2^{\text{pair}}) \in \{-1, 0, +1\}$. For the credit function in Equation (1), any fixed soft target $y \in [0, 1]$ gives

$$h(+1, y) = y, \quad h(-1, y) = 1 - y, \quad h(0, y) = \frac{1}{2}. \tag{17}$$

Let J be uniform on $\{1, 2\}$. Conditional on the fixed four outcomes, the stopping procedure's score is

$$H_{\text{unanim}}(J) = I_J^{\text{unanim}} h(\text{sgn}(S_J^{\text{pair}}), y) + (1 - I_J^{\text{unanim}}) H_{\text{full}}. \tag{18}$$

Averaging over the random pair order gives

$$\mathbb{E}_J[H_{\text{unanim}}(J) \mid S_1^{\text{pair}}, S_2^{\text{pair}}] = h(G_{\text{full}}, y) + \frac{1}{2} \sum_{b=1}^{2} I_b^{\text{unanim}} \{h(\text{sgn}\, S_b^{\text{pair}}, y) - h(G_{\text{full}}, y)\}. \tag{19}$$

We partition the event space into the following four mutually exclusive sets.

$$\begin{aligned}
\mathcal{E}_0 &= \{I_1^{\text{unanim}} = 0, I_2^{\text{unanim}} = 0\}, \\
\mathcal{E}_1 &= \{I_1^{\text{unanim}} \neq I_2^{\text{unanim}}\}, \\
\mathcal{E}_= &= \{I_1^{\text{unanim}} = I_2^{\text{unanim}} = 1,\ \text{sgn}\, S_1^{\text{pair}} = \text{sgn}\, S_2^{\text{pair}}\}, \\
\mathcal{E}_\pm &= \{I_1^{\text{unanim}} = I_2^{\text{unanim}} = 1,\ \text{sgn}\, S_1^{\text{pair}} = -\,\text{sgn}\, S_2^{\text{pair}}\}.
\end{aligned}$$

On $\mathcal{E}_0$, both correction terms in Equation (19) vanish. On $\mathcal{E}_1$, suppose without loss of generality that $I_1^{\text{unanim}} = 1$. Then $|S_1^{\text{pair}}| = 2$ and $|S_2^{\text{pair}}| \leq 1$, so $\text{sgn}(S_1^{\text{pair}} + S_2^{\text{pair}}) = \text{sgn}\, S_1^{\text{pair}}$. The only active correction term is zero. On $\mathcal{E}_=$, both pair decisions equal the full decision, so both terms are zero. Finally, on $\mathcal{E}_\pm$ we have $S_1^{\text{pair}} = -S_2^{\text{pair}} \in \{-2, +2\}$, hence $G_{\text{full}} = 0$ and

$$\frac{1}{2}\{h(+1, y) + h(-1, y)\} = \frac{1}{2}\{y + (1 - y)\} = \frac{1}{2} = h(0, y). \tag{20}$$

Therefore

$$\mathbb{E}_J[H_{\text{unanim}}(J) \mid S_1^{\text{pair}}, S_2^{\text{pair}}] = h(G_{\text{full}}, y) \tag{21}$$

for every fixed realization and every $y \in [0, 1]$. Averaging over the declared reference call set distribution, its outcome-independent pairing and evaluation instances proves Proposition 2. For hard targets $y \in \{0, 1\}$, opposing unanimous pairs give one early decision that matches the benchmark label and one that disagrees with it. Their average score equals the credit for the neutral reference decision.

The executed call count is $T_{\text{unanim}}(J) = 2I_J^{\text{unanim}} + 4(1 - I_J^{\text{unanim}})$. Hence

$$\mathbb{E}_J[T_{\text{unanim}}(J) \mid S_1^{\text{pair}}, S_2^{\text{pair}}] = 4 - (I_1^{\text{unanim}} + I_2^{\text{unanim}}) = 4 - 2s_{\text{unanim}}(i; \mathcal{F}, \mathcal{B}), \tag{22}$$

where $s_{\text{unanim}}(i; \mathcal{F}, \mathcal{B}) = \frac{I_1^{\text{unanim}} + I_2^{\text{unanim}}}{2}$ is the stopping probability for a uniformly selected first pair. Averaging yields Equation (7). Validity-gated agreement additionally depends on which outputs are executed and valid.

For balanced four-call selection using distinct seeds, the reference call set distribution selects one of 6 balanced assignments and one of two pairings across display orders uniformly. The argument holds for each fixed pairing before averaging, and also applies to the construction using matched seeds. The random stage index J is independent of the fixed call outcomes. The proof assumes that skipping or reordering execution leaves those outcomes unchanged. It does not assert this property for schedule-dependent fresh generations. Recall that G_{unanim} is the stopped canonical decision sign. G_{full} is the reference sign defined above. For binary targets, only $\mathcal{E}_\pm$ changes a decision. Let

ΔC, ΔW, and ΔN denote paired unanimity stopping minus full execution for the benchmark match, benchmark mismatch, and neutral decision probabilities defined in Appendix A.

$$\Delta C = \Delta W = \frac{\Pr(\mathcal{E}_\pm)}{2}, \quad \Delta N = -\Pr(\mathcal{E}_\pm), \quad \Pr(G_{\text{unanim}} \neq G_{\text{full}}) = \Pr(\mathcal{E}_\pm).$$

The nonempty neutral credit τ and fallback credit ϕ are defined in Appendix C.6. Because the full decision on this event is neutral with valid retained votes, the utility change is $(\frac{1}{2} - \tau)\Pr(\mathcal{E}_\pm)$ and does not depend on the fallback credit ϕ.

Score variance under the same coupling. The score difference is zero outside $\mathcal{E}_\pm$. On $\mathcal{E}_\pm$, $h(G_{\text{full}}, y) = \frac{1}{2}$ and $H_{\text{unanim}}(J)$ is y or $1 - y$ with equal probability. Consequently,

$$\mathbb{E}_J\Big[(H_{\text{unanim}}(J) - h(G_{\text{full}}, y))^2 \mid S_1^{\text{pair}}, S_2^{\text{pair}}, y\Big] = (y - \tfrac{1}{2})^2 \mathbf{1}\{\mathcal{E}_\pm\}. \tag{23}$$

Write $H_{\text{unanim}} = H_{\text{unanim}}(J)$ and $H_{\text{full}} = h(G_{\text{full}}, y)$ under this coupling. The difference has conditional mean zero, so its covariance with H_{full} is zero. For binary benchmark labels, averaging over the declared reference and pairing distributions (and, if applicable, evaluation instances) gives

$$\text{Var}(H_{\text{unanim}}) - \text{Var}(H_{\text{full}}) = \mathbb{E}[(H_{\text{unanim}} - H_{\text{full}})^2] = \tfrac{1}{4}\Pr(\mathcal{E}_\pm).$$

For soft targets the expected squared error is instead $\mathbb{E}[(y - \frac{1}{2})^2 \mathbf{1}\{\mathcal{E}_\pm\}]$. These identities describe execution-level score variability under fixed call outcomes.

More generally, consider a procedure p with fixed-reference expected-score preservation whose only decision changes turn neutral reference decisions into decisive outputs. Let $d_{p,i}$ denote its disagreement probability for fixed evaluation instance i under the stated reference coupling. For fixed hard labels and procedure randomization independent across instances, Equation (8) follows by summing instance variances. Here $H_{\text{full},i}$ denotes the reference score. Fixed-reference expected-score preservation makes the coupled difference uncorrelated with that score. Full execution is itself random because a four-call balanced reference is selected for each evaluation instance, but is deterministic conditional on the selected reference. The experiment on fixed evaluation instances uses $w_i = \frac{1}{n_{\text{eval}}}$. For soft targets replace $\frac{1}{4}$ in each summand by $(y_i - \frac{1}{2})^2$.

Here, independence across instances concerns reference selection and call execution randomization conditional on fixed call outcomes. For the corollary on neutral reference decisions under uniform ordering of individual calls, average over reference draws before substituting $d_{p,i}$.

Sharp disagreement bound under uniform matching. Fix evaluation instance i and a balanced four-call reference $\mathcal{F}$ with fixed signed votes v_1, v_2 in one display order and v_3, v_4 in the other. The unanimity rule selects one of the two matchings across display orders uniformly and independently selects a uniform pair order. The non-strict boundary uses a uniform permutation of the four individual calls. Write δ_p for the probability of disagreement with the reference decision conditional on i and $\mathcal{F}$.

If the reference is decisive, both rules preserve its decision on every execution path. If the reference is neutral, let n_{nz} count its nonzero signed votes. The corollary in Appendix H.7 gives $\delta_{\text{weak}} = \frac{n_{\text{nz}}}{4}$. When n_{nz} is zero or two, opposing unanimous pairs are impossible, so $\delta_{\text{unanim}} = 0$. When $n_{\text{nz}} = 4$, two votes are positive and two are negative. If each display order is unanimous, every pair across display orders has opposing votes, and the unanimity rule retains the neutral decision. Otherwise, each display order contains one positive and one negative vote. Exactly one of the two matchings then forms opposing unanimous pairs, giving $\delta_{\text{unanim}} = \frac{1}{2}$ and $\delta_{\text{weak}} = 1$. Therefore

$$\delta_{\text{unanim}} \leq \frac{1}{2}\delta_{\text{weak}},$$

with equality when both display orders contain oppositely signed votes. Averaging over any common distribution of balanced four-call references gives $d_{\text{unanim},i} \leq \frac{d_{\text{weak},i}}{2}$.

Sharpness also holds under the main reference selection distribution. For the same four seeds, take votes $(-1, -1, -1, +1)$ in identity display order and $(-1, -1, +1, -1)$ in swapped display order. Averaging over the 6 balanced assignments using distinct seeds gives $d_{\text{unanim},i} = \frac{1}{6}$ and $d_{\text{weak},i} = \frac{1}{3}$.

Under the assumptions on hard labels and independent randomization in Equation (8), multiplying the per-instance bound by $\frac{w_i^2}{4}$ and summing gives

$$\text{Var}(\bar{H}_{\text{unanim}}) - \text{Var}(\bar{H}_{\text{full}}) \leq \frac{1}{2}\left[\text{Var}(\bar{H}_{\text{weak}}) - \text{Var}(\bar{H}_{\text{full}})\right].$$

H.7 Uniform execution of individual calls: non-strict and strict boundaries

Proof of Proposition 3. At an early non-strict stop, set $r = n-t \geq 1$, $g = \operatorname{sgn}(S_t)$ and $R_t = S_n - S_t$. Then $|R_t| \leq r$, $g \neq 0$, and

$$gS_n = |S_t| + gR_t \geq |S_t| - r \geq 0.$$

If $S_n \neq 0$, this forces the full sign, for every ordering. If $S_n = 0$, pair positive and negative identities and exchange the identities in each pair in every permutation, fixing zero identities. This probability-preserving involution negates every prefix sum and leaves the stopping time unchanged, pairing opposite returned signs. Thus $\mathbb{E}_\sigma G_{\text{weak}} = 0$. Agreement is affine in the returned sign by Equation (1). Under the strict boundary, the displayed lower bound is strictly positive, so an early return cannot correspond even to a neutral reference. Terminal decisions are exact. □

Schedule and outcome boundaries. For the non-strict boundary, expected-score preservation applies with an outcome-independent uniform permutation of individual calls chosen before execution, unit vote weights, and the benchmark agreement score on fixed call outcomes. A fixed order $(+1, -1)$ returns $+1$ at the non-strict boundary despite a neutral reference. The same boundary based on the number of remaining votes does not preserve expectation for weighted votes. For example, on $(+1, -0.5, -0.5)$, it stops with a negative sign when both negative votes precede the positive vote and otherwise returns neutral, giving expected sign $-\frac{1}{3}$.

Uniform individual ordering is not uniform ordering of fixed two-call groups. For $(+1, +1)$, $(-1, 0)$, $(-1, 0)$, the non-strict boundary checked after complete groups stops with a negative sign exactly when the two groups whose signed sums are negative come first, with probability $\frac{1}{3}$. All other orders return neutral. Its expected sign is $-\frac{1}{3}$ although the full reference is neutral. The strict boundary preserves every reference decision when checked after complete groups.

Corollary for neutral reference decisions. On a fixed neutral reference, let n_{nz} count nonzero signs. Then $\Pr_\sigma(G_{\text{weak}} \neq 0) = \frac{n_{\text{nz}}}{n}$. An early non-strict stop on a call set whose signed votes sum to zero requires every remaining sign to equal the same nonzero sign, so a final zero prevents it. Conversely, a final nonzero sign forces non-strict stopping by $t = n - 1$. The final call condition is uniformly distributed. A neutral call set in which every call is decisive is therefore always changed to a decisive output despite fixed-reference expected-score preservation.

Suppose the two calls using the identity display order have canonical signs $(+1, +1)$, while the two calls using the swapped display order have signs $(-1, -1)$. Every balanced pair has opposing votes, so the unanimity rule completes the call set and retains its neutral decision. The non-strict boundary changes this neutral reference, whose calls are all decisive, to a decisive output on every ordering, returning either response with probability one half. Therefore, both procedures have expected agreement of 0.5 on this reference, but only the unanimity rule retains its neutral output.

For odd n with no zeros, $|S_t|$ and $n - t$ have opposite parity. Equality is impossible, so non-strict and strict stopping times and decisions coincide. Decision changes occur only for neutral full call sets. For example, on four positive votes the non-strict boundary uses one fewer call than the strict boundary (two versus three calls) without changing the positive decision.

Adaptive-Consistency at a four-call limit. With checkpoints after individual calls, thresholds 0.80 and 0.85 first stop at a 2:0 state of decisive votes and preserve the reference score in expectation. Threshold 0.90 first stops at 3:0 and preserves the reference decision. With checkpoints after pairs, 0.80 and 0.85 coincide with paired unanimity stopping, while 0.90 or above reduces to full execution at this limit. These equivalences are specific to the four-call setting and do not extend automatically to larger references.

H.8 Decision-preserving three-call rule

After three calls, stop when the sum of their signed votes has magnitude at least two. Any fourth vote lies in $\{-1, 0, +1\}$ and can reduce that magnitude by at most one, so it cannot change the sign. The returned decision therefore equals the reference decision on every fixed realization. The rule executes at least three calls and provides exact decision preservation, a stronger property than Algorithm 1

provides. It does not preserve the validity-gated agreement because the skipped fourth output may have been invalid.

For example, after the first three canonical signs $(+1, +1, 0)$, the rule can stop with sign $+1$ even if the zero is invalid. Therefore, for this check, early disagreement with the benchmark label can also occur on nonunanimous prefixes.

H.9 THIRD-CALL CONTINUATION OF PAIRED UNANIMITY STOPPING

If paired unanimity does not stop after its first pair, execute a uniformly selected remaining call. Stop if the signed vote sum after three calls has magnitude at least two. Otherwise, execute the fourth call. The last vote changes this sum by at most one, so the continuation returns the paired unanimity decision on every realization.