WHEN CAN SCORES IMPROVE A VOTE?

Anonymous authors
Paper under double-blind review

ABSTRACT

A language model can answer a problem by generating several solutions and choosing the most common final answer. We characterize exactly when scores make voting errors decrease faster as more solutions are sampled. The key information is how scores distinguish the correct answer from its leading wrong rivals. We study independent samples for one fixed problem with finitely many possible answers, where errors occur but the correct answer is more likely than any individual wrong answer. Some fixed rule with positive, bounded score weights makes errors decrease faster than ordinary voting at an exponential rate if and only if the correct answer's score distribution cannot be reproduced by mixing those of the most frequent wrong answers. Two perfectly calibrated examples (whose scores match correctness frequencies) have identical score–correctness statistics and answer frequencies, yet differ in this capacity for improvement. Complementing this large-sample characterization, we identify the weights that preserve the long-run winner under calibration and give conditions for improving count ties safely at finite budgets. On a shared cohort of 252 competition-mathematics problems, score weighting improves eight-candidate accuracy by 2.78 percentage points for Math-7B (95% interval $[1.06, 4.63]$) and 2.91 for Phi-4 ($[1.19, 4.76]$). Count ties account for most of both gains. Further experiments evaluate scoring cost, transfer, and score-based diagnostics.

1 INTRODUCTION

Suppose a language model produces several solutions to the same mathematical problem. Different solution paths may lead to the same answer. One way to decide which answer to return is to count the answers and return the most frequent one. This is *plurality voting* where the winner need not receive more than half of the votes. Another way is to score each solution and give higher-scoring solutions more influence. Both are common in language-model reasoning systems (Wang et al., 2023; Cobbe et al., 2021; Lightman et al., 2024; Brown et al., 2024; Snell et al., 2025). A *process reward model* (PRM) scores the intermediate steps of a solution (Lightman et al., 2024; Zhang et al., 2025), then a system converts these step scores into one score per solution, here the minimum step score. This score is a quantity assigned to each candidate solution, not automatically a calibrated probability that its final answer is correct.

Selecting a solution and voting for an answer are different. Consider the following example pool of six solutions: three end with the answer 17 and score 0.45 each, two end with 42 and score 0.70 each, and one ends with 9 and a score of 0.95. Counting answers using plurality voting returns 17. Summing scores by answer returns 42 (1.40 against 1.35 and 0.95). Selecting the single highest-scoring solution returns 9. Suppose there is only one correct answer. This shows that a good rule for ranking distinct individual solutions may not be a good rule when given repeated answers.

What calibration misses. A score for a solution is *calibrated* when it agrees with correctness frequency: solutions assigned score 0.8 are correct 80% of the time. The calibration assumption in this paper applies separately to one problem and its sampling procedure. That is, for one fixed problem, candidates assigned score 0.8 must be correct 80% of the time. A pooled reliability plot does not establish this per-problem assumption. Suppose two equally sampled problems both produce score-0.8 candidates: one is correct 100% of the time, the other 60%. Together they look perfectly calibrated (80% correct) but neither satisfies the paper's assumption. Even perfect calibration leaves something unspecified. Among incorrect solutions, do high scores accumulate on the same wrong

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.

answer, or are they spread across many different mistakes? Concentrating them on one wrong answer can make that answer win a weighted vote; spreading them out may not. A diagnostic that records only scores and correct/incorrect labels discards this distinction.

Our contribution. We give a unified theoretical and empirical account of when scores add value beyond answer counts. Our results identify what makes improvement possible, what weighting can preserve, and where gains arise in practice.

For independent samples from a fixed distribution where errors occur but the correct answer is more frequent than every wrong answer, we characterize exactly when some bounded, strictly positive score weighting improves plurality's exponential error rate. The criterion depends on how scores distinguish the correct answer from its strongest wrong rivals, information that calibration alone omits. We also characterize which weights preserve a correct long-run winner under calibration and give conditions for safely resolving count ties.

Experiments locate a practical source of improvement across generator families: on the same 252 AIME problems, count ties account for 18 of 21 net gains with Math-7B and 17 of 22 with Phi-4. For Math-7B, weighting eight candidates also outperforms plurality with nine under recorded-time accounting (Section 6).

2 Setup: candidates, answers, scores, and budgets

A *candidate* to a problem is one generated solution, including its reasoning and final answer. A *pool* of candidates is the collection of candidates generated for one problem in one solution generation run. A candidate is *eligible* if it passes the fixed rule for inclusion in the vote; in our experiments, its final answer must parse as a boxed integer. Different candidate reasoning texts with the same parsed answer are part of the same answer group when voting.

The fixed-problem model. Let $\mathcal{A}$ be a finite set of possible answers and $a^\star$ its single correct answer. One eligible candidate produces a random answer $A \in \mathcal{A}$ and a score $V \in [0, 1]$. Their joint probability distribution is P. We draw N pairs (A_i, V_i) *independently and identically distributed* (iid): each pair follows the same P and does not depend on the others; the model and scorer are fixed.

A *weight function* h turns a score into voting weight. The indicator $\mathbf{1}\{A_i = a\}$ is one when candidate i has answer a and zero otherwise. For answer a, define its count and weighted total by

$$C_{a,N} = \sum_{i=1}^{N} \mathbf{1}\{A_i = a\}, \qquad W^h_{a,N} = \sum_{i=1}^{N} h(V_i)\mathbf{1}\{A_i = a\}. \tag{1}$$

Weighted plurality returns an answer with the largest $W^h_{a,N}$; write this answer as $\widehat{A}_{h,N}$. Taking $h(v) = 1$ for all v gives ordinary plurality, while $h(v) = v$ for all v gives raw-score weighting. A common positive multiplier on all weights does not change the winner in these cases. "Score-only" means applying the same h to every answer group. A bounded weight has a finite upper limit; a strictly positive weight never discards a candidate by assigning zero.

Ties and sample sizes. Unless stated otherwise, a tied vote returns the tied answer that appears first in the original candidate order (*first-occurrence tie-breaking*; not sorting by score). In our proofs, N counts eligible iid candidates. In the experiments, a stated budget counts generated candidates *before* eligibility filtering (which determines the eligible candidates that can participate in voting), so an eight-candidate run may have fewer than eight votes; empty pools retain the specified fallback and remain in the accuracy denominator. The theorems use exact ties. For each experimental decision rule, we specify the numerical tolerance for declaring a tie and how ties are broken.

Population frequencies are not sample counts. In this paper, we denote:

$$p_a = P(A = a), \qquad p = p_{a^\star}, \qquad q = \max_{a \neq a^\star} (p_a)\,.$$

Here p_a is the long-run frequency of answer a, not its frequency in one observed pool. When $p > q$, the correct answer is the unique *population mode*: more likely than every wrong answer separately.

This is weaker than a correct *majority*, $p > 1/2$; for example, frequencies $(0.34, 0.33, 0.33)$ have a correct mode but no majority. The main result assumes $p > q > 0$, so errors can occur. The *leading rivals* are the wrong answers with the largest population probability, the set $J = \{a \neq a^\star : p_a = q\}$. Sampling variation can make a different wrong answer appear most often in a finite pool.

What "faster error decay" means. Let $e_h(N) = P(\widehat{A}_{h,N} \neq a^\star)$ be the error probability over repeated pools of size N. Its *error exponent* is

$$I_h = \lim_{N\to\infty} -\frac{1}{N}\log(e_h(N)). \tag{2}$$

The logarithm is natural. A larger exponent means faster exponential decay of error as N grows. Under the theorem's assumptions, the limit exists and does not depend on how maximizing ties are resolved. Ordinary plurality has $I_1 = -\log(m_0)$ with $m_0 = 1 - p - q + 2\sqrt{pq}$. Thus $I_h > I_1$ implies greater accuracy at every sufficiently large budget. It does not by itself explain an advantage at eight votes or identify the first beneficial budget.

3 Which score information can improve voting?

3.1 Two calibrated populations with different possibilities

Table 1 describes two hypothetical populations. An entry of 12 means probability $12/200$, not 12 observations from a benchmark. In both populations, the correct answer occurs with probability $68/200 = 0.34$, and each wrong answer with probability $66/200 = 0.33$.

Population	Score	Correct answer $a^\star$	Wrong answer a_1	Wrong answer a_2
A	$1/5$	12	24	24
	$2/5$	56	42	42
B	$1/5$	12	1	47
	$2/5$	56	65	19

Table 1: Two hypothetical joint distributions of answers and scores. Divide each entry by 200 to obtain the probability of that answer–score pair. Within either population, 12 of the 60 low-score units are correct, and 56 of the 140 high-score units are correct. Calibration is therefore exact at both scores. Only the assignment of wrong answers to scores changes.

The complete relationship between score and correctness is identical in A and B, and so are their plurality and highest-score-selection accuracy curves at every budget under common score-blind and correctness-independent tie rules. Nevertheless, only A admits a positive score weight that improves plurality's error exponent. To see the relevant difference, compare scores *within each answer group*. Among correct candidates, the high-score fraction is $56/68$ in both populations. In A, it is only $42/66$ for either wrong answer. A small extra vote for a high score therefore favors the correct answer over *both* leading rivals. In B, the two wrong-answer fractions are $19/66$ and $65/66$, and the correct fraction lies between them. Rewarding high scores favors wrong answer a_1 over the correct answer; rewarding low scores favors a_2. The theorem below shows that no bounded, strictly positive score-only weight improves the exponent in B (Figure 4 in the appendix draws both cases).

3.2 Start with a small change to ordinary voting

A convenient way to describe weighting is to give every candidate one vote, then add a score-dependent *bonus*:

$$h_\eta(v) = 1 + \eta b(v), \qquad \eta > 0. \tag{3}$$

Here b is a fixed nonnegative bonus function with a finite upper bound, and η is its *intensity*: how strongly the bonus changes the vote. It is fixed as the number of candidates grows; a bonus can, for example, be one for a high score and zero otherwise. Let $\mu_a = \mathbb{E}[b(V) \mid A = a]$ be the mean bonus among candidates giving answer a; $\mathbb{E}$ denotes an expectation under P. The specified-bonus part of Theorem 1 says that a fixed bonus improves the error exponent at some positive intensity exactly when

its mean is higher for the correct answer than for every leading rival in J, $\mu_{a^\star} > \max_{a \in J}(\mu_a)$; every sufficiently small positive intensity then works, and when the condition fails no positive intensity of that bonus works. The comparison is not with the average of *all* incorrect candidates: a favorable average can hide a leading rival with an unfavorable gap. Nor does the result justify a particular conventional intensity. On two constructed calibrated laws that satisfy the criterion, intensity 0.25 of the bonus $(V - 1/2)_+$ improves both exponents. Intensity two gives weights proportional to the *half-floor*, $h_{1/2}(v) = \max(v, 1/2)$, which raises scores below $1/2$ to $1/2$; this rule improves one exponent and worsens the other (Table 8). A calibrated law with uniformly spread wrong answers on which the half-floor raises every mean margin yet lowers the exponent is given in Proposition 5.

3.3 Allow the bonus itself to be chosen

For each answer with $p_a > 0$, let $Q_a = \mathcal{L}(V \mid A = a)$ denote the *conditional score distribution*: the distribution of V among candidates whose answer is a. A mixture of the leading rivals' score distributions is formed by choosing a rival at random and then drawing a score from its distribution, $Q_\lambda = \sum_{a \in J} \lambda_a Q_a$ with λ in the probability simplex Δ_J. The collection of all such mixtures is their *convex hull*: a set of score distributions, not of answer frequencies. Total-variation distance $\mathrm{TV}(Q, Q')$ is the largest difference in probability that two distributions assign to the same score event.

Theorem 1 (Score information for improving plurality)**.** *Let P be as above, and for positive-mass answers let $Q_a = \mathcal{L}(V \mid A = a)$ be the conditional score law. With Δ_J the probability simplex on J and $\mathrm{TV}(Q, Q') = \sup_E |Q(E) - Q'(E)|$, let*

$$D = \min_{\lambda \in \Delta_J} \mathrm{TV}\Big(Q_{a^\star}, \sum_{a \in J} \lambda_a Q_a\Big).$$

For iid votes and any maximizing tie conventions:

Existence. *Some bounded, measurable, strictly positive score-only weight h has $I_h > I_1$ if and only if $D > 0$.*

A specified bonus. *Fix a bounded measurable bonus $b \geq 0$ and write $I_b(\eta) = I_{1+\eta b}$ and, for positive-mass answers, $\mu_a = \mathbb{E}[b(V) \mid A = a]$. The following are equivalent: $I_b(\eta) > I_1$ for some $\eta > 0$; the same holds for every sufficiently small $\eta > 0$; and $\mu_{a^\star} > \max_{a \in J}(\mu_a)$. Let $I_b'(0+)$ denote the right-hand derivative at zero: the initial rate of change of the error exponent as bonus intensity increases from zero. With $\theta_0 = \frac{1}{2}\log(p/q)$,*

$$I_b'(0+) = \frac{\theta_0 \sqrt{pq}}{m_0} \min_{a \in J}(\mu_{a^\star} - \mu_a), \qquad \max_{0 \leq b \leq 1}(I_b'(0+)) = \frac{\theta_0 \sqrt{pq}}{m_0} D. \tag{4}$$

The criterion identifies the score information that makes improvement possible. When the correct answer's score law lies outside the rivals' mixture set, a bounded bonus favors it over every leading rival; the distance D determines the largest initial rate gain among bonuses between zero and one. Inside the mixture set, every bonus gives the correct answer a weighted average of the rivals' means, so some rival matches or exceeds it. This characterization concerns a population law; learning useful weights from finite labeled pools is a separate estimation task (proof and sketch in Section B.10). The theorem allows uncalibrated scores and nonmonotone weights while requiring strictly positive weights, which retain every candidate. Filtering and zero-valued raw-score weights have separate scope. The calibrated examples in Section F.6 show how a nonmonotone bonus can help when every nondecreasing weight fails. A worked Math-7B problem applies the criterion to model outputs in Section F.7.

4 What calibration alone can guarantee

Now consider a different question: which weights work safely for *every* distribution that is calibrated? Let $R = \mathbf{1}\{A = a^\star\}$ indicate correctness. Exact, per-problem calibration means $P(R = 1 \mid V) = V$ under the same eligible-candidate distribution used by the vote.

4.1 PRESERVING THE LONG-RUN WINNER

After weighting, answer a has normalized population share

$$\nu_a^h = \frac{\mathbb{E}[h(V)\mathbf{1}\{A = a\}]}{\mathbb{E}[h(V)]}. \tag{5}$$

The denominator is the common expected weight. This reweighted distribution describes long-run vote totals; finite candidates are still sampled from P. A transform is *universally mode preserving* if every uniquely correct winner under p_a remains the unique winner under ν_a^h on every calibrated law, over arbitrary finite alphabets, arbitrary score laws on $[0, 1]$ and every allocation of wrong answers across rivals.

Theorem 2 (Characterization of calibrated mode preservation)**.** *A bounded measurable strictly positive transform h is universally mode preserving if and only if there is a $c_0 > 0$ such that*

$$h(v) = c_0 \quad (0 < v < \tfrac{1}{2}), \qquad h(0) \le c_0, \qquad h(v) \ge c_0 \quad (\tfrac{1}{2} \le v \le 1). \tag{6}$$

Every candidate law then has a positive finite normalizer.

Why the threshold of one half? At score v, calibration assigns fraction v to the correct answer, and a particular wrong answer can receive anywhere from zero to the entire remaining fraction $1 - v$, so the correct-minus-rival fraction ranges from $2v - 1$ to v and is guaranteed nonnegative only at $v \ge 1/2$ (Figure 1 in the appendix). Below that threshold a score alone does not tell which answer receives the stronger vote, so preserving weights must be constant there; above it they may be larger but not smaller. One preserving choice is the *half-floor*, $h_{1/2}(v) = \max(v, 1/2)$, which replaces scores below 0.5 by 0.5; it is not high-score *filtering*, since low-scoring candidates still vote. The appendix extends the theorem to nonnegative weights, to imperfect calibration, and to a bound on how much wrong mass any one rival may carry, which moves the threshold (Sections B.1 and B.2).

4.2 THE EXACT CLASS FOR PROTECTION AT EVERY BUDGET

Draw $(A_i, V_i)_{i=1}^N$ iid from P with joint distribution P^N, using first-occurrence ties for both votes. The finite threshold $N_0(P)$ below may depend on P.

Theorem 3 (Universal finite-vote protection)**.** *Let h be bounded, measurable and strictly positive, and use first-occurrence ties for both votes. The following are equivalent:*

1. *On every calibrated finite-answer law P with a uniquely correct population mode, weighted plurality is at least as accurate as ordinary plurality at every budget $N \ge 1$:*

$$\Pr_{P^N}(\widehat{A}_{h,N} = a^\star) \ge \Pr_{P^N}(\widehat{A}_{1,N} = a^\star). \tag{7}$$

2. *On every such law P, there is a finite $N_0(P)$ such that (7) holds for every $N \ge N_0(P)$.*

3. *For some $c_0 > 0$,*

$$h(v) = c_0 \quad (0 < v < 1), \qquad h(0) \le c_0, \qquad h(1) \ge c_0. \tag{8}$$

If the third condition fails, there is such a calibrated law that makes weighted plurality less accurate at every sufficiently large budget under every pair of maximizing tie rules: on that law its error has a strictly smaller exponent than plurality's (Sections B.6 and B.7).

The theorem gives an exact rule for protection across all calibrated laws: retain equal weights on $(0, 1)$, with freedom to downweight certainly wrong score-zero candidates and upweight certainly correct score-one candidates. It also separates protection at every budget from eventual improvement. On one calibrated three-answer law, the half-floor achieves 51.35% accuracy at three votes against plurality's 52.25%, with the loss entirely in a three-way count tie. Yet the correct answer remains the unique population winner under both rules, and the half-floor achieves a better error exponent and eventually better accuracy (Section B.4). The characterization is sharp even at large budgets: weights outside the class admit calibrated examples with eventual loss under every pair of tie rules. A cap on wrong-answer concentration leaves both finite-vote classes unchanged (Section B.7 and Proposition 4).

Additional structure gives a broad positive case. When wrong-answer identity is independent of score given incorrectness, calibration makes every nondecreasing bonus whose value $b(V)$ is not almost surely constant favor the correct answer in mean. With $p > q > 0$, every sufficiently weak positive intensity then improves the exponent (Section B.10).

5 What can be protected at a finite budget?

A score-based rule changes an ordinary vote in two ways: it can choose between answers tied for the highest count, or it can select an answer with a lower count. These need different guarantees.

5.1 Changing a count winner

For the first n eligible candidates, let P_n be the ordinary plurality winner and H_n the half-floor winner; these are answer-valued variables, not probability distributions. Let G_n be the ordinary winner's count minus the largest count of any other answer. For a fixed positive integer m, calibrated iid sampling gives

$$P(\text{for some } n \geq 1 : \; P_n = a^\star, \; G_n \geq m, \; H_n \neq a^\star) \leq e^{-2m}. \tag{9}$$

Each additional vote of lead strengthens this protection exponentially. For leads of one, two, and three votes, Equation (9) gives about 13.53%, 1.83%, and 0.248% (observed reversals: Tables 18 and 19). The guarantee applies across all sample sizes at once: it bounds the chance that the half-floor ever overturns a correct plurality winner with a lead of at least m votes. Under iid sampling, calibration suffices, whatever the correct answer's population frequency. The probability concerns the joint event of a correct count winner, a large lead, and a harmful reversal; conditioning on an observed lead requires a separate analysis, as do ties. Proposition 7 (Section B.9) states the result formally; its proof controls all rivals through one accumulated correct-minus-incorrect bonus, an exponential supermartingale and Ville's inequality (Howard et al., 2020).

5.2 Using scores only to break count ties

The *count-first* rule preserves every unique count winner. When several answers share the largest count, it chooses the one with the largest sum of scores or bonuses, then uses first occurrence for any remaining tie (see e.g. (Kokkinos & Margaritis, 2014)). Its guarantee needs more than a mean-score comparison: for every wrong answer and every threshold, the correct answer's conditional bonus distribution must be at least as likely to exceed that threshold (*first-order stochastic dominance*). Neither calibration nor a correct population mode is required.

Proposition 1 (Conditional bonus dominance)**.** *Fix $N \geq 1$ eligible iid candidates (A_i, V_i) under one fixed law, with $V \in [0, 1]$, a finite answer alphabet and one correct label $a^\star$. Fix a measurable bonus $0 \leq b(V) \leq B < \infty$. When $0 < \mathbb{P}(A = a^\star) < 1$, assume, for every positive-mass wrong label a and every $t \in \mathbb{R}$,*

$$\mathbb{P}(b(V) \leq t \mid A = a^\star) \leq \mathbb{P}(b(V) \leq t \mid A = a).$$

The wrong-label bonus laws may differ. Let P_N be ordinary plurality. Let Q_N maximize answer count, then summed bonus among count maximizers. Both resolve remaining ties by first occurrence in the original iid order. Write C for the vector of answer counts and $T(c) = \arg\max_a(c_a)$. For every c with $\mathbb{P}(C = c) > 0$,

$$\mathbb{P}(Q_N = a^\star \mid C = c) \geq \mathbb{P}(P_N = a^\star \mid C = c) = \frac{\mathbf{1}\{a^\star \in T(c)\}}{|T(c)|}.$$

Hence $\mathbb{P}(Q_N = a^\star) \geq \mathbb{P}(P_N = a^\star)$ unconditionally. If a candidate is correct with probability zero or one, both rules have that same accuracy, regardless of the scores.

The proposition gives count-first refinement a direct finite-budget guarantee: under the stated ordering, scores can improve tied decisions while preserving every unique count winner. Conditional on the counts, tied leaders contribute equally many bonuses and have uniform first-occurrence order; a stochastic-order coupling (Shaked & Shanthikumar, 2007) establishes the accuracy guarantee (Section B.12). For a nondecreasing bonus, calibration supplies the ordering when wrong-label identity is independent of score given incorrectness. Thus this family supports protected tie improvement even

Comparison	Change (pp)	95% interval
Math-7B, declared study: 252 problems, 756 eight-candidate decisions		
Raw-score weighting (primary)	$+2.78$	$[1.06, 4.63]$
Raw half-floor	$+2.25$	$[1.06, 3.70]$
DEV half-floor	$+1.72$	$[0.79, 2.78]$
Post hoc controls on the same 756 pools		
Raw weighting − highest-score selection	0.00	$[-1.19, 1.19]$
Raw weighting − random-tie plurality	$+2.20$	$[0.64, 3.83]$
Additional generators: same 252 problems, 756 decisions each		
Phi-4: raw weighting (primary)	$+2.91$	$[1.19, 4.76]$
Phi-4: raw half-floor	$+2.12$	$[0.93, 3.44]$
Math-1.5B: raw weighting (primary)	$+1.19$	$[-0.53, 2.91]$
Math-7B: 120 separate problems, 1,920 held-out decisions		
Count-first refinement − plurality	$+2.24$	$[1.25, 3.39]$
Count-first refinement − random-tie plurality	$+1.97$	$[1.10, 2.96]$

Table 2: Accuracy changes in percentage points relative to plurality unless another baseline is named; positive values favor scores. Intervals resample problems; retrospective comparisons are exploratory. Full comparisons: Tables 21, 24, 30 and 41.

though conventional half-floor weighting can worsen the exponent (Propositions 5 and 9). The benefit concerns finite-budget accuracy: strict wrong count leads remain unchanged, so with a uniquely correct population mode the error exponent stays equal to plurality's. The guarantee also admits a quantitative extension: Corollary 4 bounds the loss under imperfect ordering.

6 Experiments: When do the decisions change?

6.1 Questions, cohorts and comparisons

We test where scores help, what predicts their benefit, and whether useful bonuses can be learned or certified. Within each comparison, rules share candidate pools, isolating aggregation from generation.

Our declared Math-7B study tests Qwen2.5-Math-7B-Instruct (Yang et al., 2024) on 252 reserve AIME problems, with three eight-candidate pools per problem (756 decisions). Additional studies use Phi-4 (Abdin et al., 2024) and Math-1.5B on these same problems. Each new study fixed its rules and settings before generation, after observing the Math-7B result (Section F.1).

All studies use minimum step scores from Qwen2.5-Math-PRM-7B (Zhang et al., 2025), with raw weighting primary and the raw half-floor secondary. Math-7B also tests a development-mapped half-floor, fitted on 60 separate problems. Every pool remains in the denominator, including empty-pool fallbacks (first extractable answer or abstention). These paired readouts measure finite-budget effects; the theoretical guarantees retain their stated assumptions.

Differences are in percentage points; intervals resample problems, keeping their runs together. Reanalyses of inspected records are retrospective, including audit/evaluation splits and learned weights; full protocols and uncertainty are in Sections D, F, F.1 and F.5.

6.2 Count ties explain most of the Math-7B and Phi-4 gains

Section 5 separates choosing among tied count leaders from overturning a larger count. We classify every correctness change this way. For Math-7B, raw weighting adds 21 correct decisions, or 2.78 points. Tied leaders contribute 18 net gains (19 corrected errors, one new error), or 2.38 points; lower-count selections contribute three (nine gains, six losses), or 0.40 points. Phi-4 shows the same pattern: tied-leader changes contribute 17 of its 22 net gains (20 corrected errors, three new errors); count overrides contribute five (Table 25).

Using scores only at ties retains most of the Math-7B benefit: post hoc count-first refinement gains 2.38 points $[1.19, 3.84]$ while preserving every unique count winner. Thirteen of its 18 net gains involve all-distinct eligible answers, where count-first reduces to highest-score selection (Section G).

Raw weighting also exceeds expected uniform-random-tie plurality by 2.20 points. Against highest-score selection it gains 11 and loses 11: 0.00 points $[-1.19, 1.19]$. The two score rules have the same observed accuracy here (post hoc controls; Table 41).

6.3 Separate draws show tie gains and test their predictability

On 120 Math-7B problems disjoint from the declared 252, two rounds provide 256 audit candidates per problem; a third supplies 1,920 evaluation pools of eight. Count-first gains 2.24 points $[1.25, 3.39]$ over first-occurrence plurality and 1.97 $[1.10, 2.96]$ over random-tie plurality. All-distinct answers account for 70% of the net gain; repeated-answer ties contribute 30%.

Resampling eight candidates from each audit forecasts the gain on separate evaluation pools. These forecasts reduce squared error by 19.87 pp^2 $[9.15, 33.06]$ versus predicting no gain (Section F.4).

We also compare forecasts retaining score–wrong-answer pairing with forecasts redrawing wrong labels independently, preserving answer frequencies and score–correctness statistics. An audit-only screen selects forecast differences of at least .5 points. On the declared fallback of 18/120 screened Math-7B problems, retaining association lowers squared forecast error by 11.93 pp^2 $[.79, 24.42]$, conditional on the saved Monte Carlo forecasts (Section F.3). This tests the predictive information highlighted by Table 1.

6.4 Learning bonuses and certifying improvement

Theorem 1 separates bonus choice from intensity. Ranking score-bin weights on fitting problems and selecting on development problems, the declared secondary bootstrap screen selects Math-7B weights $(1, 1, 2, 3)$. Every eligible candidate keeps one vote, with larger bonuses in the upper bins. The gain is 2.92 points $[.21, 6.25]$ on 30 test-role problems. The primary uncertainty-bound screen retains plurality; Section F.8 gives the problem roles, shared preprocessing and all comparisons.

A separate audit tests whether the theorem's conditions can be certified despite sampling uncertainty. Accounting for data-chosen score bins, it succeeds on 2/120 Math-7B problems; both certified modes are correct (Section F.7).

A worked application of the new criterion. AIME 2006-II-3 asks for the largest k such that 3^k divides the product of the first 100 positive odd integers. The correct answer is $49 = 33 + 11 + 4 + 1$. Among 256 Math-7B development audit samples, 188 give 49 and 46 give the leading wrong answer, 48. Consider the bonus $b(v) = \mathbf{1}\{v \geq 1/2\}$: 171/188 correct-answer samples receive it, versus 7/46 samples giving 48. Their mean bonus gap is .7574.

The per-problem certificate establishes the correct population mode and leading rival, with a positive uncertainty-adjusted gap of .4063. Under the iid premises, on its confidence event, Theorem 1 therefore guarantees a better error exponent for weights $1 + \eta b(v)$ at every sufficiently small $\eta > 0$: each candidate keeps one vote, with a small bonus for scores at least one half. The answer–score table and certificate arithmetic are in Section F.7.

For a complementary finite-budget example, raw weighting recovers answer 98 on AIME 2011-I-15 from 90 for Math-7B and Math-1.5B, and from 100 for Phi-4. Math-7B resolves a tie; the others overcome one-vote leads. Section F.7 gives the scores and all runs for this retrospectively selected illustration.

6.5 Controlled laws explain why the bonus matters

Known laws make Theorem 1's criterion exactly computable. On one calibrated law, high scores favor the leading wrong answer; a bonus for the lower two bins instead favors the correct answer. At intensity .25, this nonmonotone bonus raises the exponent from .01716 to .02449, while no positive nondecreasing weight improves it. Further checks measure detection of known separation as audit size grows (Section F.6).

Grouping saved Math-7B decisions by plurality's count lead, both half-floor rules retain every correct winner leading by at least two votes. Observed harmful reversals occur closer to ties (Proposition 7 and Section E.1).

6.6 Transfer across generators and the value of scoring time

The same scorer improves Phi-4 accuracy from 37.57% to 40.48%, extending the positive weighting result beyond the Qwen generator family. Its 2.91-point gain retains a positive family-adjusted interval $[.79, 5.16]$. Math-1.5B increases from 37.96% to 39.15%; its smaller 1.19-point estimate is inconclusive (Table 2). Generator-specific prompts, horizons and paired comparisons are reported in Section F.1.

A post hoc Math-7B check on 140 further problems (420 eight-candidate prefixes) gains 3.33 points $[1.19, 5.71]$, with 19 gains and five losses (Section E.1). Scoring took 1.89% of generation time in the declared Math-7B study: generating and scoring eight candidates cost the equivalent of generating 8.15 for plurality. Giving plurality nine candidates in a declared extension still leaves eight-candidate weighting ahead by 2.65 points $[1.06, 4.37]$ under this recorded-time convention (Section F.2).

7 Related Work

Aggregation and voting limits. Self-consistency (Wang et al., 2023), reward aggregation (Uesato et al., 2022; Li et al., 2023; Luo et al., 2024), frequency–reward hybrids (Han et al., 2026; Di et al., 2026) and generative selection (Toshniwal et al., 2025; 2026) are established. CISC weights votes by confidence; empirically, calibration metrics poorly predict within-question discrimination (Taubenfeld et al., 2025); DeepConf combines confidence weighting with filtering (Fu et al., 2026); self-certainty scores best-of-N selection (Kang et al., 2025). RPC analyzes confidence-estimation squared error for likelihood–consistency aggregation (Zhou et al., 2025). Kuang et al. (2026) derive potentially signed MAP weights in a conditionally independent, symmetric generation/score-likelihood model. Plurality consistency under a strict correct population mode is classical (List & Goodin, 2001). Wu et al. (2025, Theorems 1–2) give base and reward-weighted modal convergence; Best-of-Infinity studies asymptotic voting and generator-mixture optimization (Komiyama et al., 2026). Yang et al. (2026) show power sampling can increase correct mass but reduce the strongest-rival margin, motivating importance-weighted consensus.

Process-verifier aggregation and calibration. Math-Shepherd uses sum-based aggregation (Wang et al., 2024, Section 3.4); the Qwen PRM report finds preliminary PRMs can underperform majority in selection (Zhang et al., 2025). RLCR trains calibrated verbalized confidence and reports gains from confidence-weighted voting (Damani et al., 2026). A common multiclass forecast admits distinct canonical, classwise and confidence calibration (Vaicenavicius et al., 2019; Kull et al., 2019). Scalar verifier calibration leaves answer-probability vectors and wrong-answer allocation unspecified. Identification from retained marginals is partial identification (Manski, 2003); our bounds quantify the missing allocation (Section C.1).

Weighted voting and error exponents. Classical votes use competence log-odds (Nitzan & Paroush, 1982); Berend & Kontorovich (2014) analyze error and unknown-competence estimation in a binary expert model. Gao et al. (2016) compare optimal crowdsourcing error exponents with majority voting. Confidence-rated boosting optimizes exponential loss (Schapire & Singer, 1999). For a fixed answer–score law, we characterize when a specified bonus at some positive intensity, or some bounded, strictly positive score-only weight, improves plurality's error exponent (Theorem 1).

8 Discussion and Limitations

Our results show how scores can add value beyond vote counts. We characterize when weighting makes errors decrease faster with more candidates, which weights preserve a correct long-run winner under calibration, and how to protect decisions at finite budgets. The Math-7B and Phi-4 gains make one opportunity concrete: most improvements come from resolving count ties.

Together, these findings guide scorer evaluation and voting-rule design: examine answer-specific score distributions, preserve count information, and measure gains at the intended budget. Extending these gains across benchmarks, candidate budgets, and scorers is an interesting direction for future work.

AI USE STATEMENT

The author originated the experimental and mathematical ideas. LLMs assisted with developing and checking proofs, implementing evaluations, verifying references, and writing. The checking implementations and reconstructions cited in the appendix were produced by AI agents in separate sessions. Independence refers to separately implemented calculations, not to human third-party validation or complete software independence: some checks share extraction or grading utilities or invoke producer routines for auxiliary checks or comparison. The authors are responsible for all claims, proofs, references, code and reported results.

REPRODUCIBILITY STATEMENT

Complete proofs are provided in the appendix. The supplement will include the experimental code, results, and reproduction instructions at https://anonymous.4open.science/r/scores-improve-votes-23EE/README.md.

REFERENCES

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024. URL https://arxiv.org/abs/2412.08905.

Leon Alaoglu. Weak topologies of normed linear spaces. *Annals of Mathematics*, 41(1):252–267, 1940. doi: 10.2307/1968829. URL https://www.jstor.org/stable/1968829.

Shubham Anand Jain, Rohan Shah, Sanit Gupta, Denil Mehta, Inderjeet J. Nair, Jian Vora, Sushil Khyalia, Sourav Das, Vinay J. Ribeiro, and Shivaram Kalyanakrishnan. PAC Mode Estimation using PPR Martingale Confidence Sequences. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 5815–5852. PMLR, 2022. URL https://proceedings.mlr.press/v151/anand-jain22a.html.

Ruth Ben-Yashar and Jacob Paroush. Optimal decision rules for fixed-size committees in polychotomous choice situations. *Social Choice and Welfare*, 18(4):737–746, 2001. doi: 10.1007/s003550000080. URL https://link.springer.com/article/10.1007/s003550000080.

Daniel Berend and Aryeh Kontorovich. Consistency of weighted majority votes. In *Advances in Neural Information Processing Systems*, volume 27, 2014. URL https://papers.nips.cc/paper_files/paper/2014/hash/3cf61323ed04cc4581f435dab1256ab8-Abstract.html.

Daniel Berend and Jacob Paroush. When is Condorcet's jury theorem valid? *Social Choice and Welfare*, 15(4):481–488, 1998. doi: 10.1007/s003550050118.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013. doi: 10.1093/acprof:oso/9780199535255.001.0001.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. Large Language Monkeys: Scaling Inference Compute with Repeated Sampling. arXiv preprint arXiv:2407.21787, 2024. URL https://arxiv.org/abs/2407.21787.

Lingjiao Chen, Jared Davis, Boris Hanin, Peter Bailis, Ion Stoica, Matei Zaharia, and James Zou. Are More LLM Calls All You Need? Towards the Scaling Properties of Compound AI Systems. In *Advances in Neural Information Processing Systems*, volume 37, pp. 45767–45790, 2024. doi: 10.52202/079017-1455. URL https://papers.nips.cc/paper_files/paper/2024/hash/51173cf34c5faac9796a47dc2fdd3a71-Abstract-Conference.html.

Herman Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, 23(4):493–507, 1952. doi: 10.1214/aoms/1177729330.

Nicolas Chopin and Omiros Papaspiliopoulos. *An Introduction to Sequential Monte Carlo*. Springer Series in Statistics. Springer, 2020. doi: 10.1007/978-3-030-47845-2. URL https://link.springer.com/book/10.1007/978-3-030-47845-2.

C. J. Clopper and E. S. Pearson. The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404. URL https://academic.oup.com/biomet/article-abstract/26/4/404/291538.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

Mehul Damani, Isha Puri, Stewart Slocum, Idan Shenfeld, Leshem Choshen, Yoon Kim, and Jacob Andreas. Beyond binary rewards: Training LMs to reason about their uncertainty. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/615675cc6e94ddb1a783904fb178b5f6-Abstract-Conference.html.

A. C. Davison and D. V. Hinkley. *Bootstrap Methods and their Application*. Cambridge University Press, 1997. doi: 10.1017/CBO9780511802843. URL https://doi.org/10.1017/CBO9780511802843.

A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 28(1):20–28, 1979. doi: 10.2307/2346806. URL https://academic.oup.com/jrsssc/article/28/1/20/6953573.

Morris H. DeGroot and Stephen E. Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 32(1–2):12–22, 1983. doi: 10.2307/2987588.

Amir Dembo and Ofer Zeitouni. *Large Deviations Techniques and Applications*, volume 38 of *Stochastic Modelling and Applied Probability*. Springer, 2 edition, 2010. doi: 10.1007/978-3-642-03311-7. URL https://link.springer.com/book/10.1007/978-3-642-03311-7.

Qiwei Di, Kaixuan Ji, Xuheng Li, Heyang Zhao, and Quanquan Gu. Best-of-Majority: Minimax-Optimal Strategy for Pass@k Inference Scaling. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/94cbc452f8fe10cc6350e3c06fe75171-Abstract-Conference.html.

Randal Douc, Olivier Cappé, and Eric Moulines. Comparison of Resampling Schemes for Particle Filtering. In *Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis*, pp. 64–69, 2005. URL https://arxiv.org/abs/cs/0507025.

Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, 27(3):642–669, 1956. doi: 10.1214/aoms/1177728174.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979. doi: 10.1214/aos/1176344552.

Bradley Efron. Better Bootstrap Confidence Intervals. *Journal of the American Statistical Association*, 82(397):171–185, 1987. doi: 10.1080/01621459.1987.10478410. URL https://doi.org/10.1080/01621459.1987.10478410.

Michael D. Ernst. Permutation methods: A basis for exact inference. *Statistical Science*, 19(4):676–685, 2004. doi: 10.1214/088342304000000396.

Austin Feng, Marius Alonso, Ambroise Odonnat, Vasilii Feofanov, and Ievgen Redko. Optimal Self-Consistency for Efficient Reasoning with Large Language Models. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2511.12309v2.

Shengyu Feng, Xiang Kong, Shuang Ma, Aonan Zhang, Dong Yin, Chong Wang, Ruoming Pang, and Yiming Yang. Step-by-Step Reasoning for Math Problems via Twisted Sequential Monte Carlo. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.01920.

Ronald A. Fisher. *The Design of Experiments*. Oliver and Boyd, Edinburgh, 1935.

András Frank and Kazuo Murota. Decreasing minimization on base-polyhedra: Relation between discrete and continuous cases. *Japan Journal of Industrial and Applied Mathematics*, 40(1): 183–221, 2023. doi: 10.1007/s13160-022-00511-4. URL https://link.springer.com/article/10.1007/s13160-022-00511-4.

Yichao Fu, Xuewei Wang, Hao Zhang, Yuandong Tian, and Jiawei Zhao. Deep Think with Confidence. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/98257285340854262185500e59bc0f28-Abstract-Conference.html.

Satoru Fujishige. Lexicographically optimal base of a polymatroid with respect to a weight vector. *Mathematics of Operations Research*, 5(2):186–196, 1980. doi: 10.1287/moor.5.2.186.

Chao Gao, Yu Lu, and Dengyong Zhou. Exact exponent in optimal rates for crowdsourcing. In *International Conference on Machine Learning*, pp. 603–611. PMLR, 2016.

Paul Gordan. Ueber die Auflösung linearer Gleichungen mit reellen Coefficienten. *Mathematische Annalen*, 6:23–28, 1873. doi: 10.1007/BF01442864.

Paul R. Halmos. *Measure Theory*. Van Nostrand, 1950. doi: 10.1007/978-1-4684-9440-2. Reprinted as Graduate Texts in Mathematics 18, Springer, 1974.

Jiuzhou Han, Wray Buntine, and Ehsan Shareghi. Uncertainty-Based Methods for Automated Process Reward Data Construction and Output Aggregation in Mathematical Reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(37):30925–30933, 2026. doi: 10.1609/aaai.v40i37.40351. URL https://ojs.aaai.org/index.php/AAAI/article/view/40351.

James A. Hanley and Barbara J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982. doi: 10.1148/radiology.143.1.7063747.

G. H. Hardy, J. E. Littlewood, and G. Pólya. *Inequalities*. Cambridge University Press, second edition, 1952.

Wassily Hoeffding. Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830. URL https://doi.org/10.1080/01621459.1963.10500830.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rygGQyrFvH.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020. doi: 10.1214/18-PS321. URL https://arxiv.org/abs/1808.03204.

Audrey Huang, Adam Block, Qinghua Liu, Nan Jiang, Akshay Krishnamurthy, and Dylan J. Foster. Is Best-of-N the Best of Them? Coverage, Scaling, and Optimality in Inference-Time Alignment. In *International Conference on Machine Learning*, 2025. URL https://arxiv.org/abs/2503.21878.

Baihe Huang, Shanda Li, Tianhao Wu, Yiming Yang, Ameet Talwalkar, Kannan Ramchandran, Michael I. Jordan, and Jiantao Jiao. Sample Complexity and Representation Ability of Test-time Scaling Paradigms. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/592729d610953abb6d16065773316d00-Abstract-Conference.html.

Qi Huangfu and J. A. Julian Hall. Parallelizing the dual revised simplex method. *Mathematical Programming Computation*, 10(1):119–142, 2018. doi: 10.1007/s12532-017-0130-5.

Zhewei Kang, Xuandong Zhao, and Dawn Song. Scalable Best-of-N Selection for Large Language Models via Self-Certainty. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://arxiv.org/abs/2502.18581.

Muhammad Khalifa, Rishabh Agarwal, Lajanugen Logeswaran, Jaekyeom Kim, Hao Peng, Moontae Lee, Honglak Lee, and Lu Wang. Process Reward Models That Think. *Transactions on Machine Learning Research*, 2026. URL https://arxiv.org/abs/2504.16828.

Yiannis Kokkinos and Konstantinos G. Margaritis. Breaking ties of plurality voting in ensembles of distributed neural network classifiers using soft max accumulations. In Lazaros Iliadis, Ilias Maglogiannis, and Harris Papadopoulos (eds.), *Artificial Intelligence Applications and Innovations*, volume 436 of *IFIP Advances in Information and Communication Technology*, pp. 20–28, Berlin, Heidelberg, 2014. Springer. doi: 10.1007/978-3-662-44654-6_2. URL https://link.springer.com/chapter/10.1007/978-3-662-44654-6_2.

Junpei Komiyama, Daisuke Oba, and Masafumi Oyamada. Best-of-Infinity: Asymptotic Performance of Test-Time LLM Ensembling. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2509.21091.

Peng Kuang, Yanli Wang, Xiaoyu Han, Yaowenqi Liu, Kaidi Xu, and Haohan Wang. Optimal Aggregation of LLM and PRM Signals for Efficient Test-Time Scaling. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.13918.

Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with Dirichlet calibration. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://papers.neurips.cc/paper_files/paper/2019/hash/8ca01ea920679a0fe3728441494041b9-Abstract.html.

Ludmila I. Kuncheva and Juan J. Rodríguez. A weighted voting framework for classifiers ensembles. *Knowledge and Information Systems*, 38(2):259–275, 2014. doi: 10.1007/s10115-012-0586-6. URL https://lucykuncheva.co.uk/papers/lkjrKAIS14.pdf.

Alexander K. Lew, Tan Zhi-Xuan, Gabriel Grand, and Vikash K. Mansinghka. Sequential Monte Carlo Steering of Large Language Models using Probabilistic Programs. *arXiv preprint arXiv:2306.03081*, 2023. URL https://arxiv.org/abs/2306.03081.

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. Making Language Models Better Reasoners with Step-Aware Verifier. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5315–5333, 2023. doi: 10.18653/v1/2023.acl-long.291. URL https://aclanthology.org/2023.acl-long.291/.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's Verify Step by Step. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2305.20050.

Christian List and Robert E. Goodin. Epistemic democracy: Generalizing the Condorcet jury theorem. *Journal of Political Philosophy*, 9(3):277–306, 2001. doi: 10.1111/1467-9760.00128. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-9760.00128.

Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Meiqi Guo, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, Jiao Sun, and Abhinav Rastogi. Improve Mathematical Reasoning in Language Models by Automated Process Supervision. *arXiv preprint arXiv:2406.06592*, 2024. URL https://arxiv.org/abs/2406.06592.

Charles F. Manski. *Partial Identification of Probability Distributions*. Springer Series in Statistics. Springer, New York, 2003. doi: 10.1007/b97478. URL https://link.springer.com/book/10.1007/b97478.

Maria Marina, Daniil Moskovskiy, Sergey Pletenev, Mikhail Salnikov, Alexander Panchenko, and Viktor Moskvoretskii. Boosting Self-Consistency with Ranking. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pp. 1017–1032. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-srw.89. URL https://aclanthology.org/2026.acl-srw.89/.

Andrés R. Masegosa, Stephan S. Lorenzen, Christian Igel, and Yevgeny Seldin. Second order PAC-Bayesian bounds for the weighted majority vote. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/386854131f58a556343e056f03626e00-Abstract.html.

Pascal Massart. The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283, 1990. doi: 10.1214/aop/1176990746.

Sascha Meyen, Frieder Göppert, Carina Schrenk, Ulrike von Luxburg, and Volker H. Franz. Information-theoretic measures of metacognition: Bounds and relation to group performance. *Open Mind*, 9:1728–1762, 2025. doi: 10.1162/OPMI.a.40. URL https://www.tml.cs.uni-tuebingen.de/team/luxburg/publications/Meyen_et_al_2025_Information-Theoretic_Measures.pdf.

Shmuel Nitzan and Jacob Paroush. Optimal Decision Rules in Uncertain Dichotomous Choice Situations. *International Economic Review*, 23(2):289–297, 1982. doi: 10.2307/2526438. URL https://www.jstor.org/stable/2526438.

Hirofumi Ota, Naoto Iwase, Yuki Ichihara, Junpei Komiyama, and Masaaki Imaizumi. CITE: Anytime-valid statistical inference in LLM self-consistency. In *The ICML 2026 Workshop on Hypothesis Testing*, 2026. URL https://openreview.net/forum?id=rLSNTN1LYy. Workshop poster; also arXiv:2605.05873.

Young-Jin Park, Kristjan Greenewald, Kaveh Alim, Hao Wang, and Navid Azizan. Know What You Don't Know: Uncertainty Calibration of Process Reward Models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://arxiv.org/abs/2506.09338.

John C. Platt. Probabilities for SV machines. In Alexander J. Smola, Peter L. Bartlett, Bernhard Schölkopf, and Dale Schuurmans (eds.), *Advances in Large-Margin Classifiers*, pp. 61–74. The MIT Press, 2000. doi: 10.7551/mitpress/1113.003.0008. Circulated in 1999 as "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods".

Dražen Prelec, H. Sebastian Seung, and John McCoy. A solution to the single-question crowd wisdom problem. *Nature*, 541:532–535, 2017. doi: 10.1038/nature21054. URL https://www.nature.com/articles/nature21054.

Isha Puri, Shivchander Sudalairaj, Guangxuan Xu, Abhishek Bhandwaldar, Kai Xu, and Akash Srivastava. Rollout Roulette: A Probabilistic Inference Approach to Inference-Time Scaling of LLMs using Particle-Based Monte Carlo Methods. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-5227. URL https://papers.nips.cc/paper_files/paper/2025/hash/e55c675d3230dbc3bf24c986d6685632-Abstract-Conference.html.

Ludger Rüschendorf. On the distributional transform, Sklar's theorem, and the empirical copula process. *Journal of Statistical Planning and Inference*, 139(11):3921–3927, 2009. doi: 10.1016/j.jspi.2009.05.030. URL https://www.sciencedirect.com/science/article/pii/S037837580900158X.

Robert E. Schapire and Yoram Singer. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37(3):297–336, 1999. doi: 10.1023/A:1007614523901. URL https://link.springer.com/article/10.1023/A:1007614523901.

Moshe Shaked and J. George Shanthikumar. *Stochastic Orders*. Springer Series in Statistics. Springer, New York, 2007. ISBN 978-0-387-34675-5. doi: 10.1007/978-0-387-34675-5. URL https://link.springer.com/book/10.1007/978-0-387-34675-5.

Lloyd Shapley and Bernard Grofman. Optimizing group judgmental accuracy in the presence of interdependencies. *Public Choice*, 43:329–343, 1984. URL https://sites.socsci.uci.edu/~bgrofman/65%20Shapley-Grofman-Optimizing%20group%20judgmental%20accuracy.pdf.

Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958. doi: 10.2140/pjm.1958.8.171. URL https://msp.org/pjm/1958/8-1/p14.xhtml.

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.

Amir Taubenfeld, Tom Sheffer, Eran Ofek, Amir Feder, Ariel Goldstein, Zorik Gekhman, and Gal Yona. Confidence Improves Self-Consistency in LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 20090–20111, 2025. doi: 10.18653/v1/2025.findings-acl.1030. URL https://aclanthology.org/2025.findings-acl.1030/.

Shubham Toshniwal, Ivan Sorokin, Aleksander Ficek, Ivan Moshkov, and Igor Gitman. GenSelect: A Generative Approach to Best-of-N. In *2nd AI for Math Workshop @ ICML 2025*, 2025. URL https://openreview.net/forum?id=8LhnmNmUDb. Workshop poster; also arXiv:2507.17797.

Shubham Toshniwal, Aleksander Ficek, Siddhartha Jain, Wei Du, Vahid Noroozi, Sadegh Mahdavi, Somshubra Majumdar, and Igor Gitman. Learning Generative Selection for Best-of-N. arXiv preprint arXiv:2602.02143, 2026. URL https://arxiv.org/abs/2602.02143.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving Math Word Problems with Process- and Outcome-Based Feedback. *arXiv preprint arXiv:2211.14275*, 2022. URL https://arxiv.org/abs/2211.14275.

Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas Schön. Evaluating model calibration in classification. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 3459–3467, 2019. URL https://proceedings.mlr.press/v89/vaicenavicius19a.html.

Jean Ville. *Étude critique de la notion de collectif*. Gauthier-Villars, Paris, 1939.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-Shepherd: Verify and Reinforce LLMs Step-by-step without Human Annotations. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 9426–9439, 2024. URL https://arxiv.org/abs/2312.08935.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2203.11171.

Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdú, and Marcelo J. Weinberger. Inequalities for the L1 Deviation of the Empirical Distribution. Technical Report HPL-2003-97 (R.1), HP Laboratories, 2003. URL https://shiftleft.com/mirrors/www.hpl.hp.com/techreports/2003/HPL-2003-97R1.pdf.

Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. Inference Scaling Laws: An Empirical Analysis of Compute-Optimal Inference for LLM Problem-Solving. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2408.00724.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. Qwen2.5-Math technical report: Toward mathematical expert model via self-improvement. arXiv preprint arXiv:2409.12122, 2024. URL https://arxiv.org/abs/2409.12122.

Haohui Yang, Jiaxing Sun, and Xiujun Ma. More correct mass, worse answers: Why power sampling can fail and how to fix it. arXiv preprint arXiv:2608.14420, 2026. URL https://arxiv.org/abs/2608.14420.

Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. The Lessons of Developing Process Reward Models in Mathematical Reasoning. In *Findings of the Association for Computational Linguistics: ACL*, 2025. URL https://arxiv.org/abs/2501.07301.

Stephen Zhao, Rob Brekelmans, Alireza Makhzani, and Roger Grosse. Probabilistic Inference in Language Models via Twisted Sequential Monte Carlo. In *International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2404.17546. arXiv:2404.17546.

Zhi Zhou, Yuhao Tan, Zenan Li, Yuan Yao, Lan-Zhe Guo, Yu-Feng Li, and Xiaoxing Ma. A theoretical study on bridging internal probability and self-consistency for LLM reasoning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2922. URL https://papers.nips.cc/paper_files/paper/2025/hash/7e9afa9a02857bce4515247842471444-Abstract-Conference.html.

A READER'S GUIDE AND NOTATION

The main text is intended to be read without prior knowledge of voting theory. The remaining sections supply the proofs, numerical examples, and details of the reported experiments. The accompanying artifact retains the complete source-study records. A reader interested in the empirical conclusions need not read the proofs first. The tables below provide routes through the material and identify notation that is local to a particular proof.

A.1 A ROUTE THROUGH THE RESULTS

Question	Where to read
Can any score weighting improve the rate at which voting errors disappear?	Theorem 1; proof in Section B.10.
Why is a well-calibrated score not enough?	The opening populations in Table 1 and Figure 4; their full calculation in Section C.1.
Which weights keep an already correct long-run winner?	Theorem 2; proof in Section B.1.
Can a population-safe weight still lose at a small budget?	The three-vote worked example in Section B.4.
Does that problem disappear with more votes?	Theorem 3; tie-independent constructions in Sections B.6 and B.7.
How often can the half-floor overturn a correct count lead?	Proposition 7; proof in Section B.9.
When is it safe to use scores only for ties?	Proposition 1; proof in Section B.12.
What were the candidate-generation and scoring procedures?	Sections D, F and F.1.
Does the benefit extend to other generators on the same problems?	Complete Phi-4 and Math-1.5B results in Section F.1.
Where are the additional comparisons reported?	Sections E, E.1 and G.
Was the central criterion tested using fitted weights?	Retrospective analysis in Section F.5; known-law simulations in Section F.6.
Does the score–rival association improve a forecast of the weighting effect?	Screened forecast comparison on separate draws in Section F.3.
Where does count-first refinement gain, and was the gain forecast?	Count strata and audit-law forecast in Section F.4.
Do the certificates survive selection of the score bins?	Selection-aware reanalysis in Section F.7.
Can a finite-budget weight be learned with separate problem roles?	Gates, nominees and test-role results in Section F.8.

A.2 BASIC PROBABILITY AND VOTING NOTATION

In this paper, a *law* is a probability distribution. A *population* means all possible candidate draws with their probabilities, rather than one saved list of model outputs. A *conditional law* restricts attention to a specified answer or score group. The notation $\mathcal{L}(X \mid A = a)$ means the distribution of X conditional on answer a.

Table 3: Reader's guide: probability and voting notation.

Symbol	Meaning
$\mathcal{A}$, a, $a^\star$	The finite answer alphabet, one answer label, and its single correct label. The star is part of the label notation, not an estimate.

Symbol	Meaning (continued)
A, V, R	One candidate's random answer, scalar score, and correctness indicator $R = \mathbf{1}\{A = a^\star\}$.
P, P^N	One-candidate probability law and the product law of N independent candidates. P^N does not mean raising a numerical probability to a power.
$\mathbb{P}(E)$, $\mathbb{E}X$	Probability of an event E, and the population mean (expectation) of a random quantity X. A bar, such as $\bar{X}$, denotes a sample average.
$\mathbf{1}\{E\}$	Indicator: one if E occurs, zero otherwise.
p_a, p, q	$p_a = P(A = a)$; $p = p_{a^\star}$; q is the largest individual wrong-answer frequency. These are population probabilities. A reported statistical p-value is a different quantity.
J	The set of wrong answers tied for the largest population frequency q. These are the leading rivals. In empirical records, J may instead locally denote a problem count; that use is specified there.
$h(v)$, $w(v)$	Score-to-vote weight. The proofs sometimes write w instead of h. "Score-only" means that the same function is applied to every answer group.
$b(v)$, B, η	A nonnegative bonus, its finite upper bound B, and its intensity in $h_\eta(v) = 1 + \eta b(v)$. Some constructions use the letter b for an upper score atom; that scalar is not a bonus function.
$h_{1/2}(v)$	The half-floor $\max(v, 1/2)$. It leaves low-score candidates in the vote with weight one half.
$g(v)$, S	A development-fitted score map and the score used in an audit, either $S = V$ or $S = g(V)$. DEV means development data, not proof of calibration.
$C_{a,N}$, $W^h_{a,N}$	The count and weighted total for answer a in N eligible candidates. Some local proofs abbreviate these as n_a and W_a.
$\widehat{A}_{h,N}$	The answer selected by the weighted vote; $\widehat{A}_{1,N}$ is ordinary plurality.
$\arg\max_a x_a$	The set of answers attaining the largest x_a. A tie rule chooses one member of this set.
N, n, empirical K	Eligible sample size in the theorems; a stream index in sequential bounds; or generated-candidate budget in experimental tables. Experimental budgets are before eligibility filtering.
ν^h_a, Z_h	Normalized weighted population share and its normalizer: $\nu^h_a = \mathbb{E}[h(V)\mathbf{1}\{A = a\}]/Z_h$, $Z_h = \mathbb{E}h(V)$. A "tilt" is this reweighting of a law.
$e_h(N)$, I_h	Error probability at budget N and its limiting exponential decay rate. The positive-bonus proof writes $I_\eta = I_{1+\eta b}$ and $I_0 = I_1$, where the last I_1 denotes ordinary plurality in the main text.
Q_a	In the central theorem, the score distribution conditional on answer a. In stochastic-order proofs, $Q_a(u)$ is locally a *quantile function*, not this probability measure.
μ_a	Mean bonus conditional on answer a, $\mathbb{E}[b(V) \mid A = a]$.
λ, Δ_J	Mixture weights over leading rivals and their probability simplex: all nonnegative vectors whose coordinates sum to one. A scalar λ in a moment bound is instead a positive exponential parameter.
TV, D	Total-variation distance, and the minimum such distance between the correct score law and a leading-rival mixture. An audit's saved point value estimates a separation against the observed possibly-leading rivals, not D itself.
θ_0, m_0	Constants $\theta_0 = \frac{1}{2}\log(p/q)$ and $m_0 = 1 - p - q + 2\sqrt{pq}$ in the plurality exponent and derivative formulas.
P_n, H_n, Q_N	Locally, ordinary, half-floor, and count-first *winning answers*. They are not the law P or the conditional score distributions Q_a.
G_n	Count lead of the ordinary winner over the next-largest count. $G_a(t)$ in the tie-ordering proof is instead a conditional cumulative distribution function.

Symbol	Meaning (continued)
$T(c)$, $\lvert T(c)\rvert$	Answers tied for the largest count in count vector c, and the number of such answers. Vertical bars on a set mean its size; on a number they mean absolute value.
$(x)_+$	The positive part $\max(x, 0)$. Thus $(v - 1/2)_+$ is zero below one half.
$F(x)$, $F(x-)$	A cumulative distribution function $P(X \le x)$ and its left limit $P(X < x)$. Their difference is the probability of an exact tie at x.
$\mathrm{Cov}(X, Y)$, $\mathrm{Var}(X)$	Covariance, $\mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)]$, and variance, $\mathbb{E}[(X - \mathbb{E}X)^2]$.
ϵ, Δ_a	Conditional calibration error and correct-minus-rival population margin. Their exact definitions are repeated in each robustness statement.
κ, τ	A bound on the fraction of incorrect mass any one wrong answer may receive, and the resulting floor threshold $\tau = \kappa/(1 + \kappa)$.
δ	Depending on the explicitly stated local result, an allowed probability of certificate failure, a score-law ordering deficit, or a ramp width. It is not one parameter shared across these results.
ΔL	In the forecast comparison, the squared-error difference between two forecasts of the same outcome, in percentage points squared (pp^2); it is not an accuracy gain.

A.3 How to read the guarantees

A mode is not a majority. The mode is an answer with the largest individual probability. A majority has probability greater than one half. Several proofs use a majority as an easier sufficient condition for a correct mode; those proofs do not cover every correct non-majority mode.

Universal does not mean typical. A universal statement must hold for every law in its stated class. One constructed counterexample is enough to disprove such a statement. The size or prevalence of the failure on a real benchmark is a separate empirical question.

Positive weights and filters have different scope. A bounded weight has a common finite upper bound. A strictly positive weight never assigns zero. A high-score filter gives zero weight to rejected scores, so an impossibility result limited to strictly positive weights does not automatically apply to it. A nondecreasing or monotone weight never decreases when its score increases.

A distribution is not an observed pool. Exact calibration of P does not require every finite sample to have exact calibrated frequencies. Similarly, a correct population mode can lose a realized vote. All accuracy probabilities average over possible pools under the specified law.

Conditional and joint guarantees differ. A bound on "the original vote is correct, its lead is large, and weighting makes it wrong" controls a joint event. It is not a bound conditional on seeing a large lead. Conditioning changes the probability distribution and would require another result.

A certificate and a diagnostic differ. A statistical certificate combines a stated sampling assumption with an error bound. A diagnostic is a measured quantity that may suggest whether an assumption is plausible. A positive empirical bonus gap, a good ranking score, or a non-rejected calibration test is not automatically a certificate. The earlier audit used a fixed-partition bound after choosing bins from the same data, yielding only *nominal* statuses. The selection-aware CDF bound covers data-chosen bins under its stated iid assumptions.

A fixed-problem theorem and a benchmark interval differ. The theory studies repeated draws for one fixed problem. Reported benchmark intervals resample problem clusters, keeping runs together. Their interpretation is described with the corresponding experiment; the number of candidate decisions is not the number of independent problems.

A.4 EXPERIMENTAL VOCABULARY

Table 4: Reader's guide: experimental vocabulary.

Term	Meaning in this manuscript
Readout	The final decision rule applied to an already generated candidate pool. Changing a readout need not change generation.
PRM	Process reward model: a model that scores solution steps. Here a complete candidate usually receives the minimum of its retained step scores.
SMC	Sequential Monte Carlo (Chopin & Papaspiliopoulos, 2020; Douc et al., 2005), applied to language-model generation with reward potentials (Lew et al., 2023; Zhao et al., 2024; Feng et al., 2025; Puri et al., 2025): generation with a population of partial solutions and repeated resampling. Copies may share ancestors, unlike independent draws. SMC-32 names a 32-particle generation procedure, not 32 independent votes.
Population in an SMC record	The finite set of particles at one generation stage. This differs from a probability "population law" in the theorems.
Boxed / extractable	A parsed answer in the requested `\boxed{}` format / an answer obtainable under a broader extraction rule. These define different candidate sets.
Plurality / BAP	Counting eligible final answers; BAP in the source records denotes boxed-answer plurality. This is an implementation name, not a separate theorem.
Highest-score selection / argmax	Return the answer of one highest-scoring eligible candidate. This is also called best-of-N when N candidates are scored. It does not sum scores of repeated answers.
Half-floor / DEV half-floor	Sum $\max(V, 1/2)$ / sum $\max(g(V), 1/2)$ by answer, using a score map fitted on development data.
Strict-high-score plurality	Discard candidates with score at most one half, then count retained answers. The empirical rule has specified fallbacks when nothing remains.
Count-first / count-then-score	Maximize answer count first; use the summed score only among tied count leaders. "Lexicographic" in a proof means applying these criteria in that order.
TEST / DEV	Evaluation problems / development problems used for fitting or configuration choices. A split does not by itself establish that the analysis was prospectively chosen.
Declared / frozen / sealed	The authors report locally fixing a plan, rule, or forecast at the stated time. These words do not imply an externally timestamped registration.
Post hoc / retrospective	Chosen after relevant outcomes had been inspected / reanalysis of previously collected records. Neither term means that the arithmetic is invalid; both limit the inference from selection.
Audit draws / evaluation draws	Candidate records used to fit or inspect a rule / separate records on which its decisions are evaluated. "Separate" must be checked for every fitting choice, including score-bin boundaries.
Audit law / plug-in forecast	The empirical answer–score distribution of a problem's audit draws / the Monte Carlo expectation of a readout's accuracy under independent pools drawn from that law. A forecast interval conditions on the saved forecasts and omits simulation and estimation uncertainty unless stated.
Audit-only screen	A subset selected using audit quantities rather than evaluation outcomes; this describes the computation, not whether outcomes had previously been inspected. A screened estimate does not extend to unscreened problems.

Term	Meaning (continued)
Gate / nominee	A declared selection rule applied to candidate weights on development-role problems / a weight that reaches the selection stage. A gate that "abstains" retains plurality as the voting rule; it does not abstain from answering.
Fixed-bin bound / variation bound	Concentration bounds for a certificate with score bins fixed in advance / a bound that is uniform over data-selected bin boundaries by penalizing the bonus's total variation. A fixed-bin bound applied to same-audit bins is invalid and is shown only for comparison.
Gains / losses / W-to-W	Wrong-to-correct changes / correct-to-wrong changes / a change from one incorrect answer to another. Only gains minus losses change accuracy.
Disc.	Number of distinct problems with a correctness disagreement. A problem can contribute gains and losses in different runs, so the number of problems is not the number of changed decisions.
Singleton tie	The maximum answer count is one: each eligible answer occurs once. Count-then-score then reduces to selecting the highest-scoring candidate, subject to the stated score-tie convention.
Repeated-answer tie	At least two answers tie at a maximum count greater than one. Their score *sums*, rather than single scores, compete.
Stratum contribution	A count stratum's share of a whole-cohort accuracy difference, using all cohort pools as the denominator; it is not the within-stratum difference.
pp / CI	Percentage points / confidence interval. A change from 36% to 39% is three percentage points. It is not a 3% relative increase.
Problem-cluster bootstrap	Resample problems, keeping all of each selected problem's runs together. The result does not treat every run as an independent problem.
Pointwise / family-adjusted	An interval for one specified comparison / an interval adjusted for a stated collection of comparisons. The adjustment does not automatically cover earlier research choices.
Nominal	The stated level of a method without a coverage claim for additional data-dependent choices outside its proof. The earlier same-audit-bin audit with a fixed-partition radius, and the retrospective bootstrap intervals, use this qualification; the selection-aware certificate states its separate coverage guarantee.
Fair-sign sensitivity	A test that randomly flips the signs of nonzero paired problem differences, conditional on their magnitudes. It needs its fair-sign null assumption; pairing alone does not justify it.
AUC	Area under the receiver-operating-characteristic curve. Here it ranks a correct candidate against an incorrect candidate, with half credit for tied scores, within the stated problem/pool comparisons. It measures discrimination, not calibration.
BCa	Bias-corrected and accelerated bootstrap interval, retained as a sensitivity analysis. Some sparse or degenerate source rows have no defined BCa interval.
DKW	Dvoretzky–Kiefer–Wolfowitz concentration inequality for the empirical cumulative distribution function; the manuscript uses its Massart bound.
Correct answer	The supplied answer-key label used for evaluation. An estimated audit mode can be incorrect.
M-split	The Math-7B audit/evaluation split: two generation rounds supply audit candidates and the third supplies evaluation pools; defined fully in Section F.5.
U / M in program tables	Unrestricted bonus / monotone bonus. A stage label U in SMC metadata instead means post-resampling offspring; the contexts are distinct.
SHA256 / replay	A file-content hash used to check file identity / recomputing a readout from saved records. A successful replay is not regeneration of model outputs.

A.5 WHICH PROBLEM COHORT DOES A TABLE DESCRIBE?

The same problem can appear in more than one sample collection. Repeated samples of a problem are not new independent benchmark problems. The labels A and B in the opening *constructed populations* are unrelated to the problem-number strata A and B in the experimental records.

Table 5: Reader's guide: which problem cohort each table describes.

Collection	Relationship to the other evidence
Inspected Math-7B	60 problems from the 200-problem TEST set, with three 128-candidate runs per problem. Earlier outcomes were inspected.
Remaining Math-7B	The other 140 problems of that TEST set, with three 20-candidate runs. The eight-candidate analysis uses their prefixes; the 60-candidate analysis combines the three runs.
Prospective Math-7B	252 additional problems, three eight-candidate runs, giving 756 decisions. The added-candidate comparison extends each original pool by one or two candidates. The post hoc count-then-score readout and the selection and random-tie controls reuse these same pools.
Additional generators	Phi-4 and Math-1.5B each generate three new eight-candidate pools per problem on the same 252-problem Math-7B reserve cohort (756 decisions each). These are new draws on previously evaluated questions.
Retrospective Math-7B audit cohort	120 problems (the 60 development and 60 test problems with repaired-scorer records), disjoint from the prospective 252: 256 audit candidates from two rounds and the third round split into held-out pools of 8, 16 or 32 (1,920 pools at eight). The audit, count-tie, association-forecast, selection-aware certificate and learned-weight analyses all reuse these records; none adds new draws.
Learned-weight problem roles	Training/development/test roles of 60/30/30 Math-7B problems, assigned within the 120-problem audit cohort. Fitting and selection use disjoint problems, but the inherited score-bin edges were estimated from audit scores pooled across all three roles.

Table 6: Theoretical guarantees and empirical counterparts. Readouts measure effects; the direct population-majority certificate is distinct from the calibration and bonus-order assumptions.

Result	Premise and guarantee	What the audit can and cannot check
Universal boundaries	Per-problem score-input calibration; exact preservation for population modes and finite votes under first-occurrence ties.	Paired readouts; robustness certificate covers unchanged decisions; the direct certificate includes gains and losses (Table 16).
Beneficial bonus	Uniquely correct base mode and bounded nonnegative bonus; higher mean than every most frequent wrong answer is necessary and sufficient for exponent improvement at some intensity.	Selection-aware audit certifies two Math-7B directions under iid premises; finite-budget effects are tested separately.
Count-margin protection	Calibrated iid stream; the joint event of a harmful half-floor reversal and a count lead $\geq m \geq 1$ has probability $\leq e^{-2m}$.	Complete count-lead cross-tabs; stopped exponential-moment audit.
Finite tie refinement	Correct-answer bonus distribution dominates wrong-answer distributions; count-then-bonus voting cannot lower iid accuracy.	Count-then-score readout separates singleton from repeated-answer ties; gains do not establish conditional dominance.

B Core Proofs and Precise Scope

Reading guide. The proofs below cover population preservation, finite-vote counterexamples, wrong-answer structure, the count-margin bound, the positive bonus criterion, and identification from observables. Section C.1 derives what evaluation summaries determine. Sections D and E give the candidate provenance, full paired comparisons and labelled diagnostics.

Notation across sections. P denotes a population law and P^h its score-weighted tilt; finite votes still use draws from P. The count-bound proof locally writes P_n, H_n for the two vote winners, while $H_\pm$ denote quantile integrals in the retained-summary result. $M_a(v)$ denotes a conditional correct-minus-rival margin; the mode-preservation proofs also write this as $d_a(v)$. The observable-identification result instead uses $d_a = \mathbb{E}|P(A = a \mid V) - V|$ for an integrated discrepancy. V is raw score and S is either V or its development map $g(V)$. Theoretical N counts eligible iid draws; n indexes a stream, and empirical N (or K in the scorer supplement) counts generated prefixes before eligibility filtering. D is reserved for the hull distance of Theorem 1; the appendix proofs write signed margins as $\Delta(t)$, vote increments as $X_{a,w}$, the explicit-intensity coefficient as A_b and the empirical derivative plug-in as $\widehat{I}_b'(0+)$, and displayed discordant-problem counts are labelled Disc. Populations A and B of Table 1 are the calibrated pair of Section C.1 and Figure 6.

B.1 Which Calibrated-Score Tilts Preserve a Correct Answer Mode?

For binary responses, Meyen et al. (2025, Theorem 8) relate calibrated confidence distributions to finite-group accuracy under conditional independence. Here we characterize population-mode preservation across multiclass laws with unspecified wrong-answer allocation.

The identification bounds fix a tilt and characterize the outcomes compatible with retained marginals. Here we ask a different, constructive question: which fixed score transforms preserve an initially correct unique mode for *every* calibrated population, including arbitrary concentration of wrong answers? The criterion is population mode preservation, not Bayes optimality under a model for wrong answers, finite-sample accuracy dominance, or superiority to signed calibrated aggregation (Kuang et al., 2026).

Setup. Fix one problem and one eligible proposal law P. There is one correct normalized answer $a^\star$, finitely many wrong answers, and $R = \mathbf{1}\{A = a^\star\}$. A score $V \in [0, 1]$ is calibrated when $\mathbb{E}[R \mid V] = V$, under this same law. Write $p_a = P(A = a)$ and, for a pointwise finite measurable $h : [0, 1] \to [0, \infty)$, define

$$Z_h = \mathbb{E}_P h(V), \qquad \nu_a^h = \frac{\mathbb{E}_P[h(V)\mathbf{1}\{A = a\}]}{Z_h} \quad \text{when } 0 < Z_h < \infty.$$

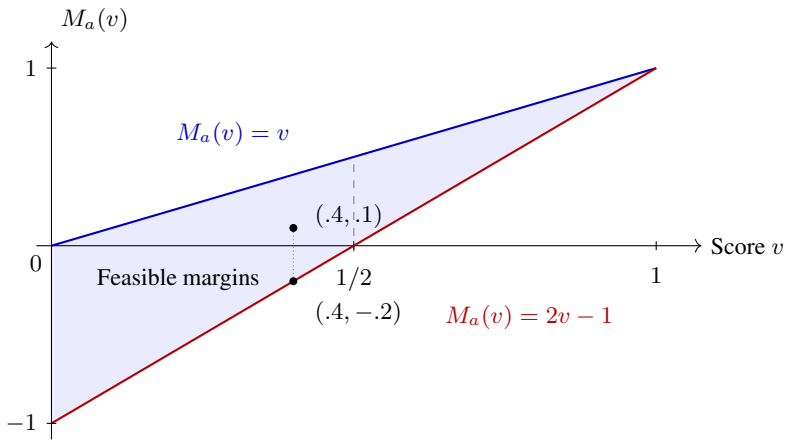


Figure 1: **Why one half is the population threshold.** Calibration fixes correct mass at v; a specified wrong answer can receive any share of $1 - v$, so the margin lies in $[2v - 1, v]$, whose lower edge becomes nonnegative at $1/2$. The marked points are the margins at score .4 when the wrong mass .6 is split equally between two rivals, $(.4, .1)$, or falls entirely on one rival, $(.4, -.2)$.

We call h *universally mode preserving* if every calibrated law with $p_{a^\star} > \max_{a \neq a^\star} p_a$ and $0 < Z_h < \infty$ also has $\nu^h_{a^\star} > \max_{a \neq a^\star} \nu^h_a$. A tie after tilting violates this strict criterion. The definition quantifies over arbitrary calibrated joint laws, arbitrary score distributions on $[0, 1]$ (including atoms at any point), and finite answer alphabets of arbitrary size. The necessity below already follows from laws with at most three answers.

Theorem 4 (Calibrated mode preservation for nonnegative weights)**.** *A pointwise finite measurable nonnegative transform h is universally mode preserving if and only if one of the following holds:*

1. *There is $c_0 > 0$ such that*

$$h(v) = c_0 \quad (0 < v < \tfrac{1}{2}), \qquad h(0) \le c_0, \qquad h(v) \ge c_0 \quad (\tfrac{1}{2} \le v \le 1). \tag{10}$$

2. *$h(v) = 0$ for every $0 \le v \le \frac{1}{2}$, with arbitrary nonnegative values above $\frac{1}{2}$.*

In the second case, every defined tilted law has a strictly correct majority; the transform may fail to define a target on a low-score problem. The identically zero transform satisfies the conditional statement vacuously. For bounded strictly positive h, only the first case is possible and every proposal law has a positive finite normalizer.

Proof of Theorem 2 and its nonnegative extension.

Proof. For a wrong answer a, let $q_a(v) = P(A = a \mid V = v)$. Calibration implies

$$d_a(v) := v - q_a(v) \in [2v - 1, v]. \tag{11}$$

Under the first case, $h(v)d_a(v) \ge c_0 d_a(v)$ pointwise: the weights agree on $(0, 1/2)$; at zero, $d_a(0) \le 0$ and $h(0) \le c_0$; on $[1/2, 1]$, $d_a(v) \ge 0$ and $h(v) \ge c_0$. Hence

$$\mathbb{E}[h(V)d_a(V)] \ge c_0(p_{a^\star} - p_a) > 0.$$

Dividing by Z_h proves strict preservation. Under the second case, all positive weighted mass is supported on $V > 1/2$. Whenever $0 < Z_h < \infty$,

$$2\nu^h_{a^\star} - 1 = \frac{\mathbb{E}[h(V)(2V - 1)]}{Z_h} > 0,$$

proving a strictly correct majority.

For necessity, first consider two distinct scores $0 < x < 1/2$ and $y > 0$. At score x, assign all wrong mass to a designated rival a_1; at score y, assign all wrong mass to a_2. Put probability t on the first score and $1 - t$ on the second. The correct-minus-a_1 margins before and after weighting are

$$\Delta(t) = (1-t)y - t(1-2x), \qquad \Delta_h(t) = h(y)(1-t)y - h(x)t(1-2x).$$

At $t_0 = y/(1-2x+y)$, the base pair ties. The three base masses are

$$p_{a^\star} = p_{a_1} = \frac{y(1-x)}{1-2x+y}, \qquad p_{a_2} = \frac{(1-2x)(1-y)}{1-2x+y}.$$

The third answer has smaller mass exactly when

$$y > f(x) := \frac{1-2x}{2-3x}. \tag{12}$$

If also $h(x) > h(y)$, then $\Delta_h(t_0) < 0$. Decreasing t sufficiently slightly makes the correct base mode unique while preserving the weighted reversal and the strict gap to the third answer. The tilted normalizer is positive: $t > 0$ and $h(x) > h(y) \geq 0$.

Suppose two scores $x, y \in (0, 1/2)$ satisfy $h(x) > h(y)$. Choose a third score $z < 1/2$ sufficiently close to $1/2$ that $z > f(x)$ and $y > f(z)$, and distinct from x, y. Such a choice exists since $f(x) < 1/2$ for $x > 0$ and $f(z) \to 0$. Either $h(x) > h(z)$, so the preceding construction applies to score pair (x, z), or $h(z) \geq h(x) > h(y)$, so it applies to (z, y). Universal preservation therefore forces $h(v) = c_0 \geq 0$ throughout $(0, 1/2)$.

If $c_0 > 0$ and some $y \in [1/2, 1]$ has $h(y) < c_0$, choose $x < 1/2$ close enough to $1/2$ that $y > f(x)$ and use the same construction. Consequently $h(y) \geq c_0$ on $[1/2, 1]$.

Next suppose $h(0) > c_0$. Put probability 0.1 at score zero, entirely on a_1, and at score 0.4 put joint masses

$$(P(A = a^\star, V = 0.4), P(A = a_1, V = 0.4), P(A = a_2, V = 0.4)) = (0.36,\ 0.26 - \delta,\ 0.28 + \delta).$$

This is calibrated. For $0 < \delta < 0.08$, the correct base mode is unique, whereas its weighted margin over a_1 equals $c_0\delta - 0.1(h(0) - c_0)$, which is negative for sufficiently small $\delta > 0$. Thus $h(0) \leq c_0$, including when $c_0 = 0$.

Finally, suppose $c_0 = 0$ and $h(1/2) > 0$. Assign joint masses $(0.4, 0.4, 0)$ to $(a^\star, a_1, a_2)$ at score $1/2$, and $(0.05, 0, 0.15)$ at score $1/4$. Calibration holds and the base masses are $(0.45, 0.40, 0.15)$. Tilting leaves a correct–a_1 tie, contradicting strict preservation. Hence $h(1/2) = 0$ in the zero-baseline case, completing the necessity proof. All constructions have finite score support and finite weights. If all three answers must have positive mass, the $y = 1$ endpoint construction can be perturbed by shifting an arbitrarily small amount of wrong mass at score x from a_1 to a_2; both strict margins survive. □

Why three answers matter. The necessity is different on a binary alphabet. If the correct answer is the unique base mode, then $p := \mathbb{E}V > 1/2$. Even the bounded strictly positive transform $h(v) = 1 + v$ preserves every such mode: calibration and Jensen's inequality give

$$\nu^h_{a^\star} = \frac{\mathbb{E}[(1+V)R]}{1+\mathbb{E}V} = \frac{p + \mathbb{E}[V^2]}{1+p} \geq p > \frac{1}{2}.$$

This transform violates (6). An unrestricted three-answer law is already enough to force that characterization.

A positive transform and two exact examples. The nonconstant transform

$$h_{1/2}(v) = \max\{v, 1/2\} \tag{13}$$

satisfies the theorem. More generally, among floors $h_\tau(v) = \max\{v, \tau\}$ with $0 \leq \tau \leq 1$, universal preservation holds exactly for $\tau \geq 1/2$. Any smaller floor leaves a nonconstant interval below $1/2$, violating the necessary condition. Two elementary consequences help interpret this choice. It is the unique continuous universally preserving transform that agrees with $h(v) = v$ on $[1/2, 1]$: continuity forces the lower constant and the zero endpoint to be $1/2$. Also, under calibration its monotonicity gives $\mathbb{E}[V h_{1/2}(V)]/\mathbb{E}[h_{1/2}(V)] \geq \mathbb{E}V$ by nonnegative covariance, so it does not reduce single-draw correctness while preserving a correct mode.

Population B of Table 1 (the perfectly calibrated construction) has joint masses, in units of $1/200$,

V	$a^\star$	a_1	a_2
$1/5$	12	1	47
$2/5$	56	65	19

Its base law is $(0.34, 0.33, 0.33)$. Raw-score tilting gives $(124, 131, 85)/340$ and a wrong mode. The half-floor is constant on these two scores and preserves the original correct mode exactly.

Preservation does not force the rule to leave every mode unchanged. Consider the following joint probabilities:

V	$a^\star$	a_1	a_2
$1/5$	.15	.40	.20
$4/5$	.20	0	.05

(14)

They are calibrated at both scores: $.15/.75 = .2$ and $.20/.25 = .8$. The base masses are $(.35, .40, .25)$, with wrong mode a_1. The half-floor gives unnormalized masses $(.235, .200, .140)$ and normalizer $.575$, making the correct answer uniquely modal. Hence the class can recover a wrong mode while preserving every already correct mode under the same calibration premise. For a benchmark calibrated separately at each problem, the indicator of a uniquely correct population mode cannot decrease under a transform in the theorem whenever all tilted laws are defined; this example shows strict increase is possible. This is a population statement. Consistent iid weighted votes inherit the unique limiting mode; no finite-sample accuracy dominance follows from the argument.

Sharp tolerance to conditional calibration error. The following proposition, formerly stated in the main text, quantifies how much conditional calibration error the half-floor's certificate tolerates; the main text points to it from Theorem 2.

Proposition 2 (Sharp tolerance to conditional calibration error)**.** *Under an arbitrary law, put* $r(v) = P(R = 1 \mid V = v)$ *and*

$$\epsilon = \mathbb{E}|r(V) - V|, \qquad \Delta_a = P(A = a^\star) - P(A = a).$$

For $h_{1/2}(v) = \max(v, 1/2)$*, every wrong answer satisfies*

$$\mathbb{E}\big[h_{1/2}(V)(\mathbf{1}\{A = a^\star\} - \mathbf{1}\{A = a\})\big] \geq \frac{\Delta_a - \epsilon}{2}. \tag{15}$$

Consequently the normalized tilted margin is at least $(\Delta_a - \epsilon)/(2\mathbb{E}h_{1/2}(V))$*. In particular, a correct base mode with minimum pairwise margin greater than* ϵ *is preserved. The coefficient of* ϵ *in* (15) *is sharp, even among three-answer laws with a uniquely correct base mode.*

The loss contributed by the excess weight $(v - 1/2)_+$ is at most $|r(v) - v|/2$ pointwise; integration proves the bound, and a score-one atom assigned to a wrong answer attains its error coefficient. A pooled reliability curve or AUC is not an upper bound on this per-problem conditional error.

Proof of Proposition 2.

Proof. Let $q_a(v) = P(A = a \mid V = v)$ and $d(v) = r(v) - q_a(v)$. Since $q_a(v) \leq 1 - r(v)$, one has $d(v) \geq 2r(v) - 1$. The excess over half the base pairwise margin is $\mathbb{E}[t(V)d(V)]$, where $t(v) = (v - 1/2)_+$. If $v \leq 1/2$, this excess is zero pointwise; if $r(v) \geq 1/2$, it is nonnegative. In the remaining case, put $b(v) = 1/2 - r(v)$. Then $t, b \in [0, 1/2]$ and

$$td \geq -2tb \geq -\frac{t + b}{2} = -\frac{|v - r(v)|}{2}.$$

The middle inequality follows from $t + b - 4tb = t(1 - 2b) + b(1 - 2t) \geq 0$. Integration proves the bound.

For sharpness, put probability e at score one entirely on a wrong answer a_1. At score $1/2$, put probability $1 - e$ with conditional answer probabilities $(1/2, 1/4, 1/4)$. For $0 < e < 1/5$, the correct base mode is unique. Calibration error is $\epsilon = e$, its base margin over a_1 is $(1 - 5e)/4$, and its unnormalized half-floor margin is $(1 - 9e)/8 = (\Delta_{a_1} - \epsilon)/2$, attaining equality. The normalizer is $Z_h = (1 + e)/2$. For $1/9 < e < 1/5$, this example has a correct base mode and a wrong tilted mode. □

What the premise does not establish. Calibration here concerns terminal correctness under the fixed problem and eligible proposal law. Neither a raw minimum of process scores, calibration pooled across problems, a fitted calibrator, nor calibration before changing eligibility establishes this premise. The error ϵ is the true conditional L^1 calibration error, not an automatically valid estimate from binned empirical calibration. We provide no new estimator of this conditional error; the labelled audit in Section E instead bounds a sufficient margin-minus-error condition. Multiple distinct correct answer groups require another statement: majority mass for their union need not make an individual correct answer the mode. Finally, a finite SMC population need not follow its target exactly, and using these terminal transforms as intermediate potentials is a different intervention. The theorem supplies no claim about the observed PRM's calibration, the registered system difference, or finite-particle improvement.

For a concrete finite-pool distinction, put equal mass on two calibrated score strata: at $V = .4$ the conditional answer law is $(.4, .3, .3)$, and at $V = .9$ it is $(.9, .1, 0)$. The correct base mass is .65. Nevertheless, an iid sample with three correct .4-score traces and two wrong .9-score traces has positive probability. Its unweighted vote is correct by 3 to 2, while the half-floor vote is wrong by 1.8 to 1.5. Population calibration is not exact calibration of every finite sample, so this does not contradict the theorem and rules out a pathwise finite-pool preservation claim.

Even expected accuracy at a fixed sample size need not improve. Consider the calibrated joint atoms (A, V) with probabilities

$$P(A = a^\star, V = .5) = .1, \qquad P(A = a_1, V = .5) = .1,$$
$$P(A = a^\star, V = .51) = .408, \qquad P(A = a_2, V = .51) = .392.$$

The correct answer is the unique base mode, with masses $(.508, .1, .392)$. For four iid draws and first-occurrence tie-breaking, exact enumeration of the 4^4 ordered samples gives plurality accuracy $22370923/39062500 \simeq .572696$ and half-floor accuracy $5253034/9765625 \simeq .537911$. Thus the population guarantee does not imply finite-sample expected-accuracy dominance either.

B.2 Bounded wrong-answer concentration

The main-text summary points here; the full statement is reproduced so that its displayed class and proof stay together.

Structure enlarges the preserving class. If each wrong answer satisfies $P(A = a \mid V) \le \kappa(1-V)$, with $0 < \kappa \le 1$, the threshold moves to $\tau = \kappa/(1+\kappa)$. Among bounded positive weights, universal preservation requires a positive constant below τ, no smaller weights at or above τ, and $h(0)$ no larger than that constant (Section B.2):

$$h(v) = c \ (0 < v < \tau), \qquad h(0) \le c, \qquad h(v) \ge c \ (\tau \le v \le 1). \tag{16}$$

Necessity allows arbitrary finite alphabets. A fixed minimal alphabet can force uniform wrong-answer allocation and admit additional weights.

Theorem 5 (Preservation under bounded wrong-answer concentration)**.** *Fix $0 < \kappa \le 1$ and $\tau = \kappa/(1+\kappa)$. Among calibrated laws on arbitrary finite answer alphabets satisfying $P(A = a \mid V) \le \kappa(1-V)$ for every wrong answer a, a bounded measurable strictly positive h preserves every uniquely correct population mode if and only if, for some $c > 0$, h satisfies (16).*

An embedding observation. Given a finite score law with $p = \mathbb{E}V > 0$ and a designated rival mass $0 \le q(v) \le \kappa(1-v)$, split the residual $1 - v - q(v)$ equally among M additional wrong answers. A finite integer $M \ge 1/\kappa$ with $M > (1-p)/p$ ensures that every additional answer respects the conditional cap and has unconditional mass below p. Consequently any strictly positive base margin $\mathbb{E}[V - q(V)]$ gives a uniquely correct base mode on this completed alphabet. This construction also applies at score zero.

Proof of Theorem 5. The conditional margin $d_a(v) = v - q_a(v)$ lies in $[(1+\kappa)v - \kappa, v]$. Under (16), $h(v)d_a(v) \ge c d_a(v)$ at every score: the weights agree on $(0, \tau)$; at zero the margin is nonpositive; and on $[\tau, 1]$ the margin is nonnegative. Taking expectations proves sufficiency.

For necessity, each score in $(0, \tau)$ allows either sign of a sufficiently small margin. If $h(u) > h(v)$ at two such scores, choose admissible margins $d(u) = -s$ and $d(v) = s + \delta$, with equally probable scores. For sufficiently small $s > 0$ and then $\delta > 0$, the base designated-rival margin is $\delta/2 > 0$, whereas the weighted margin is $[s(h(v) - h(u)) + h(v)\delta]/2 < 0$. The embedding observation gives a unique correct base mode, proving that $h = c > 0$ on $(0, \tau)$.

If $h(y) < c$ at some $y \in [\tau, 1]$, choose $u \in (0, \tau)$ with admissible margin $d(u) = -s < 0$, and set $q(y) = 0$, so $d(y) = y$. Putting probability $y/(s + y)$ at u yields zero base rival margin and weighted margin $sy(h(y) - c)/(s + y) < 0$. Slightly decrease that probability to obtain a positive base margin and retain the weighted reversal, then apply the embedding observation. This includes $y = \tau$ and $y = 1$.

Finally, if $h(0) > c$, use margin $-\kappa$ at score zero and margin u at $u \in (0, \tau)$. At score-zero probability $u/(\kappa + u)$, the base pair ties while its weighted margin is $\kappa u(c - h(0))/(\kappa + u) < 0$. Slightly reduce that probability and complete the alphabet. Hence $h(0) \le c$ is necessary. □

Alphabet scope. The necessity may require more than three answers. In particular, at $\kappa = 1/m$ with exactly m possible wrong answers, the cap forces their conditional allocation to be uniform. That smaller model admits additional preserving transformations, so the theorem does not assert a fixed-alphabet characterization.

Proposition 3 (Sharp robustness of the concentration-dependent floor)**.** *Let* $r(v) = P(A = a^\star \mid V = v)$*, suppose* $q_a(v) \le \kappa(1 - r(v))$ *for each wrong answer, and set* $\epsilon = \mathbb{E}|r(V) - V|$ *and* $\Delta_a = P(A = a^\star) - P(A = a)$*. For* $h_\tau(v) = \max(v, \tau)$*,*

$$\mathbb{E}[h_\tau(V)(\mathbf{1}\{A = a^\star\} - \mathbf{1}\{A = a\})] \ge \tau(\Delta_a - \epsilon).$$

The coefficient of ϵ *is sharp, including among laws whose base mode is uniquely correct.*

Proof. Set $d = r - q_a$. It suffices to show $(v - \tau)_+ d \ge -\tau|r - v|$ pointwise. The only nontrivial case is $v > \tau$ and $d < 0$. The cap gives $d \ge (1 + \kappa)(r - \tau)$ and hence $r < \tau$. Put $x = v - \tau \in (0, 1 - \tau]$ and $y = \tau - r \in (0, \tau]$. Since $xy/(x + y)$ increases with each argument,

$$-(v - \tau)d \le (1 + \kappa)xy \le \tau(x + y) = \tau(v - r).$$

Integrate and add $\tau\mathbb{E}d = \tau\Delta_a$.

For equality, put probability η at $V = 1$ and $1 - \eta$ at $V = \tau$, with $r(1) = 0$, $r(\tau) = \tau$, $q_a(1) = \kappa$, and $q_a(\tau) = 0$. Then $\epsilon = \eta$, $\Delta_a = (1 - \eta)\tau - \eta\kappa$, and equality holds. For $0 < \eta < 1/(2 + \kappa)$ this margin is positive. Split residual wrong mass among enough labels to make all their base probabilities smaller than $P(A = a^\star) = (1 - \eta)\tau$, while preserving the cap. Explicitly, put $p_{\text{true}} = \mathbb{E}r(V) = (1 - \eta)\tau$ and split $1 - r(v) - q_a(v)$ among M labels with $M \ge 1/\kappa$ and $M > (1 - p_{\text{true}})/p_{\text{true}}$. Each residual label then has conditional mass at most $\kappa(1 - r(v))$ and base mass below p_{true}. This uses the actual correctness probability r, not the miscalibrated score v. □

Score-independent wrong-answer allocation. Under the stronger model $q_a(v) = \pi_a(1 - v)$ with fixed probability vector π, define $p = \mathbb{E}V$ and $p_h = \mathbb{E}[h(V)V]/\mathbb{E}h(V)$. Exact calibration gives weighted probabilities p_h for the correct answer and $\pi_a(1 - p_h)$ for each rival. Thus the correct answer is the unique weighted mode exactly when $p_h > \pi_{\max}/(1 + \pi_{\max})$. For every bounded positive nondecreasing h,

$$p_h - p = \frac{\mathrm{Cov}(h(V), V)}{\mathbb{E}h(V)} \ge 0,$$

because twice the covariance equals $\mathbb{E}[(h(V) - h(V'))(V - V')] \ge 0$ for independent V, V'. This elementary structured-model consequence explains why monotone weighting is safe there; it does not establish finite-sample dominance.

The cap does not change the finite-vote classes. Theorem 5 enlarges the population-preserving class under the cap $P(A = a \mid V) \le \kappa(1 - V)$: its threshold $\kappa/(1 + \kappa)$ lies strictly below one half for $0 < \kappa < 1$ and equals it at $\kappa = 1$, where (6) is recovered. The finite-vote classes of Theorem 3 do not change, even when the wrong mass is spread uniformly over the rivals at every score. Throughout, the N votes are iid draws from P, $e_h(N)$ is the error probability of h-weighted plurality under a maximizing tie rule, $e_1(N)$ is plurality's, and I_h, I_1 are the exponents of Section 3.

Proposition 4 (Diffuse rivals do not enlarge the finite-vote class)**.** *Let h be bounded, measurable and strictly positive on $[0,1]$, and suppose that no $c>0$ satisfies (8). Fix $0<\kappa\le 1$. There are an integer $K=K(h,\kappa)\ge 1/\kappa$ and a law P with one correct answer, K wrong answers and a two-point score support such that:*

(i) P is exactly calibrated, and $P(A=a\mid V)=(1-V)/K$ almost surely for each wrong answer a; hence every wrong answer satisfies $P(A=a\mid V)\le\kappa(1-V)$, and the wrong mass at each score is shared equally by the K rivals;

(ii) the correct answer is the unique population mode and the unique weighted mode under h;

(iii) I_h and I_1 exist, do not depend on the maximizing tie rules, and $I_h<I_1$;

(iv) for every pair of maximizing tie rules, one for each vote, h-weighted plurality is strictly less accurate than plurality at every sufficiently large budget, and one threshold N_0 serves all tie rules.

Proof. Step 1: two scores with ordered weights. Since no $c>0$ satisfies (8), one of three cases holds. If h is not constant on $(0,1)$, pick $u,v\in(0,1)$ with $h(u)<h(v)$. Otherwise $h=c$ on $(0,1)$ for some $c>0$, and either $h(0)>c$, in which case put $v=0$ and $u=1/2$, or $h(0)\le c$ and $h(1)<c$, in which case put $u=1$ and $v=1/2$. In every case

$$u\in(0,1],\qquad v\in[0,1),\qquad u\ne v,\qquad 0<r:=h(u)<h(v)=:s<\infty.$$

Only these two values of h enter the construction, so measurability plays no further role.

Step 2: the law and its modes. Let V take the values u and v with probability $1/2$ each; given V, let $A=a^\star$ with probability V and $A=a_k$ with probability $(1-V)/K$ for each of the wrong answers $a_1,\dots,a_K$. The conditional masses sum to one, $P(A=a^\star\mid V)=V$ is exact calibration, and $P(A=a_k\mid V)=(1-V)/K\le\kappa(1-V)$ exactly when $K\ge 1/\kappa$. This is item (i) for every such K; it is an identity between conditional probabilities, so it covers the score-zero and score-one atoms without conditioning on a null event. The correct mass is $p=P(A=a^\star)=(u+v)/2$, with $0<p<1$ because $u>0$ and $v<1$; each rival has mass $q=(1-p)/K$, so the population mode is uniquely correct exactly when $K>(1-p)/p$. The normalizer $\mathbb{E}h(V)=(r+s)/2$ is positive and finite, and the unnormalized weighted masses are

$$\mathbb{E}[h(V)\mathbf{1}\{A=a^\star\}]=\frac{ur+vs}{2}\ \ge\ \frac{ur}{2}>0,\qquad \mathbb{E}[h(V)\mathbf{1}\{A=a_k\}]=\frac{(1-u)r+(1-v)s}{2K},$$

so the weighted mode is uniquely correct exactly when $K>K_w:=\{(1-u)r+(1-v)s\}/(ur+vs)$, a finite threshold. Item (ii) holds for $K>\max\{(1-p)/p,K_w\}$.

Step 3: a lower bound on the weighted pairwise moment. Fix one rival $a=a_k$ and the weighted correct-minus-rival increment $S=h(V)(\mathbf{1}\{A=a^\star\}-\mathbf{1}\{A=a\})$. It equals $r>0$ with probability at least $u/2>0$ and $-s<0$ with probability $(1-v)/(2K)>0$, and by Step 2 its mean is positive. Its negative-tilt moment is, for real t,

$$F_h(t)=\mathbb{E}e^{-tS}=(1-p-q)+\tfrac12\big(ue^{-rt}+ve^{-st}\big)+\tfrac{1}{2K}\big((1-u)e^{rt}+(1-v)e^{st}\big).$$

Dropping the two nonnegative terms $ve^{-st}/2$ and $(1-u)e^{rt}/(2K)$ gives

$$F_h(t)\ \ge\ (1-p-q)+g(t),\qquad g(t)=ae^{-rt}+\frac{d}{K}e^{st},\qquad a=\frac{u}{2}>0,\quad d=\frac{1-v}{2}>0. \tag{17}$$

The function g is strictly convex on $\mathbb{R}$ with $g'(t)=-rae^{-rt}+(sd/K)e^{st}$, which vanishes only at $t_\star=\log\{raK/(sd)\}/(r+s)$. Write $\gamma=r/(r+s)$, so that $0<\gamma<1/2$ because $r<s$. Substituting $e^{(r+s)t_\star}=raK/(sd)$,

$$\begin{aligned}\min_{t\in\mathbb{R}}g(t)=g(t_\star)&=a\Big(\frac{sd}{raK}\Big)^{\gamma}+\frac{d}{K}\Big(\frac{raK}{sd}\Big)^{1-\gamma}\\&=a^{1-\gamma}\Big(\frac{d}{K}\Big)^{\gamma}\Big(\frac{s}{r}\Big)^{\gamma}\Big(1+\frac{r}{s}\Big)=C_hK^{-\gamma},\qquad C_h=\frac{(r+s)\,a^{1-\gamma}d^{\gamma}}{r^{\gamma}s^{1-\gamma}}>0.\end{aligned}$$

Hence

$$m_h:=\inf_{t\in\mathbb{R}}F_h(t)\ \ge\ 1-p-q+C_hK^{-\gamma}. \tag{18}$$

Step 4: plurality's moment and the alphabet threshold. For the count increment $X = \mathbf{1}\{A = a^\star\} - \mathbf{1}\{A = a\}$ the moment is $F_1(t) = (1-p-q) + pe^{-t} + qe^{t}$, minimized at $e^{2t} = p/q > 1$, so

$$m_1 := \min_{t\in\mathbb{R}} F_1(t) = 1 - p - q + 2\sqrt{pq} = 1 - p - q + 2\sqrt{p(1-p)}\,K^{-1/2}.$$

The two minima share the term $1-p-q$. Since $\gamma < 1/2$, the bound (18) exceeds m_1 as soon as $C_h K^{-\gamma} > 2\sqrt{p(1-p)}K^{-1/2}$, that is, for

$$K > K_s := \Big(\frac{2\sqrt{p(1-p)}}{C_h}\Big)^{2(r+s)/(s-r)}.$$

Choose $K = K(h,\kappa)$ as any integer with $K \geq 1/\kappa$, $K > (1-p)/p$, $K > K_w$ and $K > K_s$. Then items (i) and (ii) hold and

$$0 < m_1 < m_h < 1, \tag{19}$$

where $m_h < 1$ because $F_h(0) = 1$ and $F_h'(0) = -\mathbb{E}S < 0$, and $m_1 > 0$ because $q > 0$; by strict convexity the minimizers of both moments are positive.

Step 5: from pairwise moments to the vote. Two facts about iid sums $S_N = \sum_{i\leq N} Y_i$ of a finitely supported increment Y with positive mean and both signs of positive probability are used; here $M(t) = \mathbb{E}e^{-tY}$ and $m = \min_{t\in\mathbb{R}} M(t)$, attained at some $t_* > 0$. First, Chernoff's bound (Chernoff, 1952; Boucheron et al., 2013): for every N,

$$\mathbb{P}(S_N \leq 0) \leq \inf_{t\geq 0} \mathbb{E}e^{-tS_N} = m^N. \tag{20}$$

Second, Cramér's lower bound (Dembo & Zeitouni, 2010) for the strict event: for $t > t_*$ the tilted one-draw law $Q_t(y) \propto P(Y = y)e^{-ty}$ has mean $\mu_t = -M'(t)/M(t) < 0$; the law of large numbers under Q_t and the change of measure give $\mathbb{P}(S_N < 0) \geq M(t)^N e^{Nt(\mu_t - \delta)} Q_t^N(|S_N/N - \mu_t| < \delta)$ for $0 < \delta < -\mu_t$, and letting $N \to \infty$, then $\delta \downarrow 0$ and $t \downarrow t_*$,

$$\liminf_{N\to\infty} \frac{1}{N} \log \mathbb{P}(S_N < 0) \;\geq\; \log m. \tag{21}$$

This is the argument behind (26); it needs no lattice or parity condition. Both increments S and X satisfy the hypotheses by Steps 2–4.

The law is symmetric in the K wrong answers, so every rival has the same pairwise moments F_h and F_1. Write $S_N^{(k)}$ for the weighted correct-minus-a_k total over N draws. If some rival's weighted total strictly exceeds the correct answer's, the correct answer is not maximizing and no tie rule selects it; if the correct answer is not selected, some rival's total is at least the correct answer's. Hence, for every maximizing tie rule,

$$\mathbb{P}\big(S_N^{(1)} < 0\big) \;\leq\; e_h(N) \;\leq\; \sum_{k=1}^{K} \mathbb{P}\big(S_N^{(k)} \leq 0\big) \;\leq\; K m_h^N, \tag{22}$$

and the same sandwich with the count totals and m_1 bounds $e_1(N)$. By (21) and (20), $N^{-1}\log e_h(N) \to \log m_h$ and $N^{-1}\log e_1(N) \to \log m_1$: both exponents exist, equal $I_h = -\log m_h$ and $I_1 = -\log m_1$ under every maximizing tie rule, and (19) gives $I_h < I_1$, item (iii). Both exponents are positive, so both accuracies tend to one. Finally $e_1(N) \geq q^N > 0$ (all N draws equal to a_1), and by (22)

$$\liminf_{N\to\infty} \frac{1}{N} \log \frac{e_h(N)}{e_1(N)} \;\geq\; \liminf_{N\to\infty} \frac{1}{N} \log \mathbb{P}\big(S_N^{(1)} < 0\big) - \log m_1 \;\geq\; \log \frac{m_h}{m_1} > 0.$$

The ratio therefore tends to infinity, so $e_h(N) > e_1(N)$ for all $N \geq N_0$. Neither bound in (22) depends on the tie rules, so N_0 serves every pair of them: item (iv). □

Corollary 1 (Finite-vote protection within the capped family)**.** *Fix $0 < \kappa \leq 1$ and let $\mathcal{P}_\kappa$ be the set of calibrated finite-answer laws with a uniquely correct population mode in which every wrong answer satisfies $P(A = a \mid V) \leq \kappa(1-V)$. For iid draws and first-occurrence ties in both votes, the following are equivalent for a bounded, measurable, strictly positive h: (1) on every $P \in \mathcal{P}_\kappa$, weighted plurality is at least as accurate as plurality at every budget $N \geq 1$; (2) on every $P \in \mathcal{P}_\kappa$ there is $N_0(P)$ after which it is at least as accurate; (3) h satisfies (8) for some $c_0 > 0$.*

Proof. (3)⇒(1): the every-budget sufficiency proof of Theorem 3 in Section B.3 is pathwise on every calibrated law under iid draws and first-occurrence ties, so it applies to every member of $\mathcal{P}_\kappa$. (1)⇒(2) with $N_0(P) = 1$. (2)⇒(3): if (3) fails, the law of Proposition 4 belongs to $\mathcal{P}_\kappa$ by its items (i) and (ii), and item (iv), applied to first-occurrence ties, makes weighted plurality strictly less accurate at every sufficiently large budget, contradicting (2). □

Scope. The cap therefore separates the two kinds of guarantee: for $0 < \kappa < 1$ it strictly enlarges the population-preserving class (Theorem 5), and at $\kappa = 1$ it recovers (6), whereas the every-budget and eventual finite-vote classes are the same constant-interior class at every κ. The witness alphabet $K(h, \kappa)$ may be large, in particular for nearly constant h; the proposition quantifies over arbitrary finite alphabets and says nothing about a fixed alphabet, about the calibration of a real scorer, or about accuracy at a given budget. On each fixed witness law the structured corollary of Section B.10 applies with $\pi_a = 1/K$: any nondecreasing bonus with $b(u) \neq b(v)$ improves the exponent at every sufficiently weak intensity, and the explicit sufficient intensity (32) is of order $1/\log K$ because $\theta = \frac{1}{2}\log\{pK/(1-p)\}$ there. That bound is sufficient, not an optimal rate; the exponent lesson is that a fixed intensity must be matched to the competition.

B.3 Which Calibrated Weights Preserve Finite-Vote Accuracy?

The population guarantee in Theorem 2 concerns the mode of a tilted law. A different question is whether weighted voting on finitely many samples can universally preserve the accuracy of ordinary plurality. Its exact answer is more restrictive.

Sampling and tie convention. For a finite answer alphabet, one correct answer $a^\star$, and a proposal law P with $\mathbb{E}[\mathbf{1}\{A = a^\star\} \mid V] = V$ and a uniquely correct base mode, draw $(A_i, V_i)_{i=1}^N$ iid from this original law P. For a bounded measurable transform $h : [0, 1] \to (0, \infty)$, let

$$\widehat{A}_{h,N} \in \arg\max_a \sum_{i=1}^N h(V_i)\mathbf{1}\{A_i = a\}.$$

Among maximizing answers, select the one appearing first in the sample. Use this same first-occurrence rule for unweighted plurality $\widehat{A}_{1,N}$. Both rules consume the same iid sample from P; this is not sampling from the tilted law or a finite SMC population.

Proof of Theorem 3, every-budget protection (equivalence of its first and third conditions).

Proof. For sufficiency, calibration implies $P(A = a^\star, V = 0) = P(A \neq a^\star, V = 1) = 0$. Almost surely, score-zero observations are wrong and score-one observations are correct. Write n_a for answer counts and W_a for total weights. Under (8),

$$W_{a^\star} \geq c_0 n_{a^\star}, \qquad W_a \leq c_0 n_a \quad (a \neq a^\star).$$

If unweighted plurality selects $a^\star$, it has at least as many votes as every rival, so its weighted total is also at least every rival's. Any rival still tied after weighting must have been tied before weighting: the weighted maximizing set is a subset of the original tied set and contains $a^\star$. First occurrence therefore still selects $a^\star$. This pathwise implication outside null events proves the accuracy inequality.

For necessity, first suppose h violates the population-preserving characterization in Theorem 2. The constructive necessity proof of that theorem supplies a calibrated finite-support three-answer law with a uniquely correct base mode and a rival whose weighted population mass is *strictly greater* than the correct mass. Strict positivity excludes the zero-baseline case, which was the only tie-only construction. On this finite-support witness, all weights are bounded, and the strong law of large numbers makes unweighted correctness probability tend to one and weighted correctness probability tend to zero. Thus (7) fails at some finite N. Consequently any universally preserving h must have the population form

$$h(v) = c_0 \ (0 < v < 1/2), \qquad h(0) \leq c_0, \qquad h(v) \geq c_0 \ (1/2 \leq v \leq 1).$$

It remains to rule out every interior boost. Suppose $h(b) > c_0$ for some $b \in [1/2, 1)$, and write $d = h(b)$. Put $a = 1/4$, $K = ab(1-b)^2$, and $\eta = K/16$. Since $0 < K \leq 1/32$, one has

$0 < \eta \le 1/512$. Construct the calibrated three-answer law

V	$a^\star$	a_1	a_2
a	ηa	$\eta(1-a)$	0
b	$(1-\eta)b$	0	$(1-\eta)(1-b)$

(23)

The correct-minus-a_2 base margin is $\eta a + (1-\eta)(2b-1) > 0$. Moreover, $p_{a^\star} \ge (1-\eta)/2 > 3\eta/4 = p_{a_1}$, so the correct base mode is unique.

Draw $N = 4$ observations and let L count low-score observations at $V = a$. When $L = 0$, every observation has weight d, so both votes agree. When $L = 1$ and the low-score observation is wrong, the majority answer among the three high-score observations wins both votes. When $L = 1$ and the low-score observation is correct, the only changed composition has one high-score correct answer and two high-score a_2 answers. Unweighted counts tie two versus two. Exchangeability makes the first occurrence correct with probability $1/2$, whereas weighted totals are $c_0 + d < 2d$, so the weighted vote is wrong. All other compositions with $L = 1$ give identical decisions.

Let Γ be weighted accuracy minus unweighted accuracy. The exact contribution from $L \le 1$ is therefore

$$-4\eta(1-\eta)^3\, a\, 3b(1-b)^2\, \frac{1}{2} = -6\eta(1-\eta)^3 K.$$

Every remaining event can improve correctness by at most one, and

$$\Pr(L \ge 2) = 6\eta^2 - 8\eta^3 + 3\eta^4 \le 6\eta^2.$$

Using $(1-\eta)^3 \ge 1/8$ and $\eta = K/16$ gives

$$\Gamma \le -6\eta(1-\eta)^3 K + 6\eta^2 \le -\frac{3K^2}{128} < 0. \tag{24}$$

This contradicts universal preservation for every interior boost, including an arbitrarily small one. Hence $h = c_0$ throughout $(0,1)$; the population theorem already supplies the two endpoint inequalities. □

Population and finite-vote guarantees differ. For continuous strictly positive weights, the finite-vote class consists only of constant transforms. By contrast, the continuous nonconstant half-floor $h(v) = \max(v, 1/2)$ preserves every correct population mode under calibration. These statements coexist because universal finite-vote accuracy is a stronger requirement. The four-sample necessity witness explicitly uses first-occurrence resolution of a 2–2 plurality tie. Corollary 2 strengthens necessity to all sufficiently large sample sizes, including odd sizes, under arbitrary maximizing tie rules; the stated convention still applies to sufficiency. The same expected-accuracy characterization also holds with uniform random choice among maximizing answers: endpoint changes only shrink a maximizing set containing the correct answer, and the four-sample witness again loses a correct-selection probability of $1/2$ on its decisive tie. The loss bound can be arbitrarily small, and establishes neither typical degradation nor its prevalence in a benchmark. Useful finite-budget improvements remain possible; calibration alone does not guarantee them uniformly.

B.4 A THREE-VOTE HALF-FLOOR REVERSAL WITH EXACT CALIBRATION

This exact example distinguishes population preservation, finite-vote accuracy, and the conditional bonus-order premise. It is a constructed probability law, not an estimate from the benchmark pools. Let the correct label be $a^\star$ and the two wrong labels be a_L, a_H. Draw iid from

(A,V)	$(a^\star, 2/5)$	$(a_L, 2/5)$	$(a^\star, 4/5)$	$(a_H, 4/5)$
$\mathbb{P}(A,V)$	$3/10$	$9/20$	$1/5$	$1/20.$

At the two occupied scores, the conditional correctness probabilities are $(3/10)/(3/4) = 2/5$ and $(1/5)/(1/4) = 4/5$, respectively. Thus calibration is exact. The answer probabilities are $(1/2, 9/20, 1/20)$, with a uniquely correct population mode. The conventional half-floor $h(v) = \max\{v, 1/2\}$ assigns weights $1/2$ and $4/5$. Its unnormalized answer masses are $(31/100, 9/40, 1/25)$, so the correct tilted mode also remains unique; its normalized mass is $62/115 > 1/2$. Throughout this example, both votes use first occurrence in the original iid order to resolve maximizing ties.

Why two votes cannot give this reversal. For any positive nondecreasing weight map h and two iid calibrated candidates, weighted voting returns the answer of a candidate with the larger weight, with first occurrence resolving equality. This description also holds when both answers coincide. Plurality returns the first candidate's answer. For iid scores V, V', calibration and exchangeability therefore give

$$\mathrm{Acc}_{h,2} - \mathrm{Acc}_{1,2} = \frac{1}{2}\mathbb{E}[(V - V')\,\mathrm{sign}(h(V) - h(V'))] \geq 0.$$

For the displayed law, the half-floor separates the two score atoms, and this difference is $(3/4)(1/4)(4/5 - 2/5) = 3/40$. The accuracies are $1/2$ and $23/40$: a gain of 7.5 percentage points.

The smallest harmful budget for this half-floor. At $N = 3$, every two-vote answer defeats any one-vote answer under the half-floor, because $2(1/2) > 4/5$. Thus every strict count winner is preserved. The rules can differ only when all three labels occur once, an event of probability

$$6(1/2)(9/20)(1/20) = 27/400.$$

Conditional on that event, plurality selects the correct label with probability $1/3$. The correct candidate has score $4/5$ with probability $(1/5)/(1/2) = 2/5$. Only then can it win the half-floor vote, by tying a_H and preceding it, which has probability $1/2$. The conditional weighted accuracy is consequently $1/5$, and

$$\mathrm{Acc}_{h,3} - \mathrm{Acc}_{1,3} = \frac{27}{400}\left(\frac{1}{5} - \frac{1}{3}\right) = -\frac{9}{1000}.$$

The probability of a correct strict majority is $1/2$, giving full accuracies $209/400 = 52.25\%$ for plurality and $1027/2000 = 51.35\%$ for the half-floor. At one vote the returned answer cannot change, and the preceding identity rules out half-floor harm at two votes. Hence three is the smallest harmful budget under calibrated iid sampling and this tie convention. Independent uniform resolution of maximizing ties gives the same expected accuracies; arbitrary fixed label priorities are not asserted to do so.

Extension to arbitrary intensity. Replacing the fixed half-floor with $h_\eta(v) = 1 + \eta(v - 1/2)_+$ gives weights 1 and $1 + 3\eta/10$. For any $0 < \eta < 10/3$, two weight-1 votes strictly defeat one weight-$(1 + 3\eta/10)$ vote, so the strict-count-winner preservation and single-occurrence conditional distribution remain identical to the fixed half-floor under first-occurrence ties. The correct answer's mean bonus is $3/25$, which exceeds zero for the unique most frequent wrong rival a_L. Thus Theorem 1 guarantees improvement for all sufficiently small $\eta > 0$. Intersecting these intervals yields a fixed rule that loses at $N = 3$ but has a strictly larger error exponent, and therefore greater accuracy at every sufficiently large budget.

A finite-budget certificate for the half-floor on this law. Let $e_0(N)$ and $e_2(N)$ be the error probabilities of plurality with first-occurrence ties and of the half-floor (weights $1/2$ and $4/5$) under any resolution of maximizing ties, on the law above, writing $c = a^\star$ for the correct answer, with label masses $\nu_c = 1/2$, $\nu_{a_L} = 9/20$, $\nu_{a_H} = 1/20$. We show $e_2(N) < e_0(N)$ for every $N \geq 396$.

Lower bound for plurality. On $\{n_{a_L} > n_c\}$ plurality errs; on $\{n_{a_L} = n_c\}$ either a_H leads or the correct answer sits in an equal-count tie and, by exchangeability of the vote order, is selected with probability at most $1/2$. Hence $e_0(N) \geq \frac{1}{2}\mathbb{P}_\nu(S_N \leq 0)$ with $S_N = \sum_{i \leq N} Y(A_i)$, $Y = \mathbf{1}\{A = c\} - \mathbf{1}\{A = a_L\}$. Write $p = \nu_c$, $q = \nu_{a_L}$, $\theta = \frac{1}{2}\log(p/q)$ and $m_0 = 1 - p - q + 2\sqrt{pq}$, and let $\rho_b = \nu_b e^{-\theta Y(b)}/m_0$ be the tilted law, so $\rho_c = \rho_{a_L} = \sqrt{pq}/m_0$ and $\rho_{a_H} = \nu_{a_H}/m_0$. Under ρ, Y is symmetric about zero with variance $v = 2\sqrt{pq}/m_0$; symmetry gives $\mathbb{P}_\rho(S_N \leq 0) \geq \frac{1}{2}$ and Chebyshev gives $\mathbb{P}_\rho(S_N < -2\sqrt{Nv}) \leq \frac{1}{4}$, so $\mathbb{P}_\rho(-2\sqrt{Nv} \leq S_N \leq 0) \geq \frac{1}{4}$. On that event the likelihood ratio $d\nu^{\otimes N}/d\rho^{\otimes N} = m_0^N e^{\theta S_N}$ is at least $m_0^N e^{-2\theta\sqrt{Nv}}$. Therefore

$$e_0(N) \;\geq\; \tfrac{1}{8}\exp\big[-NI_0 - 2\theta\sqrt{Nv}\big], \qquad I_0 = -\log m_0.$$

Upper bound for the half-floor. With scaled weights $w(V) \in \{1, 8/5\}$ and weighted totals T_a, an error requires $T_a \geq T_c$ for some $a \neq c$ under every tie rule, so by the union bound and Markov's

inequality at tilt $t_a = 5 \log r_a$,

$$e_2(N) \;\le\; \sum_{a \neq c} \mathbb{P}(T_a \ge T_c) \;\le\; M_{a_L}^N + M_{a_H}^N, \qquad M_a = \mathbb{E}\, r_a^{5w(V)(\mathbf{1}\{A=a\} - \mathbf{1}\{A=c\})},$$

with the rational tilts $r_{a_L} = 103/100$ and $r_{a_H} = 23/20$; the moments are exact rationals, $M_{a_L} \approx 0.98834$ and $M_{a_H} \approx 0.81748$.

Comparison. The logarithm of the ratio of the upper to the lower bound has derivative in N at most $\max_a \log M_a + I_0 + \theta\sqrt{v/N}$, which is decreasing in N and negative at $N = 396$; and at $N = 396$ the bounds satisfy $M_{a_L}^{396} + M_{a_H}^{396} < 0.009607 < 0.009612 < \frac{1}{8}\exp[-396 I_0 - 2\theta\sqrt{396 v}]$ (constants evaluated in interval arithmetic and independently rechecked). Hence the analytic bounds give $e_2(N) < e_0(N)$ for all $N \ge 396$; this threshold is where the two bounds cross, not the earliest beneficial budget. Below it, exact enumeration of every count vector under first-occurrence ties, in exact rational arithmetic, gives a strictly positive accuracy difference in favour of the half-floor at every budget $4 \le N \le 395$ (and at $396 \le N \le 600$, consistent with the certificate; independently reproduced for $N \le 80$); together with the certificate, under first-occurrence ties the half-floor is less accurate than plurality only at $N = 3$. The finite verification below 396 is exact arithmetic, not a bound. The analytic certificate holds for every maximizing tie rule for the half-floor, despite its first-occurrence loss at three votes. No threshold for other intensities is asserted here.

Why the additional ordering premise matters. On this law at $N = 3$, the half-floor and the count-then-score policy with $b(v) = v$ coincide. The high-score wrong label has score $4/5$ with certainty, whereas the correct conditional score law has masses $3/5$ at $2/5$ and $2/5$ at $4/5$. Thus the correct bonus law does not stochastically dominate this rival. Protecting strict count winners alone does not supply the ordering premise in Proposition 1. The harmful count vector has three singleton labels: conditioning on answer composition changes which conditional score laws compete. The example neither establishes a prevalence of such reversals in real pools nor replaces the general necessity and eventual-loss proofs.

For the raw bonus $b(v) = v$, the one-sided conditional CDF deficits are $\delta_{a_L} = 0$ and $\delta_{a_H} = 3/5$. On the three-way singleton tie, the rank-integral bound of Corollary 4 is

$$\int_{3/5}^{1} u(u - 3/5)\, du = \frac{26}{375},$$

while the actual conditional correct-selection probability is $1/5$. The positive lower bound still permits a loss relative to plurality's $1/3$; it is not a guarantee of improvement. These deficits are computed from the fully specified example law, not estimated from benchmark data.

Relation to a general four-vote construction and verified large-budget losses. The half-floor already loses at four votes: one calibrated law gives 57.27% accuracy for plurality versus 53.79% for the half-floor under first-occurrence ties (exact fractions in Section B.1). The general necessity construction and its quantified scope remain unchanged; the three-vote law above is a simpler illustration for this particular weight map.

To confirm that half-floor loss persists at large finite budgets, an independently checked calibrated law with four atoms $(a, V, P(a, V))$ given by $(\text{correct}, 2/5, 1/10)$, $(a_1, 2/5, 3/20)$, $(\text{correct}, 11/20, 33/80)$, and $(a_2, 11/20, 27/80)$ separates population preservation from finite-budget protection. Under every pair of maximizing tie conventions, the half-floor loses against plurality by at least 0.003 percentage points at $N = 65$ and 0.169 percentage points at $N = 129$.

An attained ten-vote loss with nearly equal weights. Loss does not require a steep weight. Take scores $1/2$ and $501/1000$ with $P(V = 501/1000) = p = 696/881$, and joint masses $(\text{correct}, 1/2) = (a_1, 1/2) = (1 - p)/2$, $(\text{correct}, 501/1000) = 0.501\, p$ and $(a_2, 501/1000) = 0.499\, p$. Both score strata are exactly calibrated and the correct answer is the unique population mode (mass $0.5 + 0.001p$ against $0.499p$ and $(1 - p)/2$). The half-floor weights are 0.500 and 0.501. Exact enumeration of every count vector at $N = 10$ gives plurality accuracy 0.634130 and half-floor accuracy 0.593831 under first-occurrence ties, a loss of 4.03 percentage points, reproduced by two independent enumerators. The law was found by a bounded search over calibrated three-answer laws with two score values; it is an attained example, not a maximum over laws. At ten votes,

first-occurrence and independent uniform resolution of maximizing ties give the same expected accuracies for each voting rule.

B.5 Half-floor loss under exact calibration with score-independent uniform allocation

The negative results above quantify over adversarial wrong-answer allocation. The following family shows that the ordinary half-floor can lose even under the structured assumptions that make every monotone weight preserve the correct mode.

A useful direction is not a useful intensity. The criterion of Theorem 1 certifies weak intensities. On two constructed calibrated laws that satisfy it, intensity 0.25 of the bonus $(V - 1/2)_+$ improves both exponents, while the standard half-floor, intensity two, improves one and worsens the other (Table 8). Even under exact calibration and uniform wrong labels independent of score conditional on incorrectness, a weight can increase every normalized correct-minus-rival mean margin while reducing the error exponent. On the law of Proposition 5 below, the half-floor $\max(V, 1/2)$ equals V on its support and, normalized to mean one, raises every correct-minus-rival mean margin by $\varepsilon^2(1 + 1/K)/p$, yet for every sufficiently small $\varepsilon > 0$ its error exponent is strictly smaller than plurality's under every pair of maximizing tie conventions. The conditional score gap $\varepsilon^2/[p(1-p)]$ is positive, so the same bonus improves the exponent at every sufficiently weak intensity; the conventional intensity is too strong for this law. The necessity parts of Theorems 2 and 3 already hold over laws with at most three answers; this family instead uses $K = 256$ wrong answers, so that no single rival carries concentrated wrong mass.

Proposition 5 (Structured half-floor exponent loss)**.** *Fix $p = 2/3$, $K = 256$ wrong answers and $q = (1-p)/K$. For $0 < \varepsilon < 1/6$ let V take the values $p \pm \varepsilon$ with probability $1/2$ each, and given $V = v$ let $A = a^\star$ with probability v and $A = a$ with probability $(1-v)/K$ for each wrong answer a. This law is exactly calibrated; conditional on incorrectness, its wrong label is uniform and independent of V; its answer marginals are $(p, q, \dots, q)$ for every ε, and the half-floor equals V on the support. For every sufficiently small $\varepsilon > 0$ the half-floor vote has a strictly smaller error exponent than plurality, under every pair of maximizing tie conventions; hence it is less accurate at every sufficiently large budget.*

Proof. Normalize the half-floor weights to $w(V) = V/p$ (a common factor does not change the winner). For a fixed wrong answer a, let $X_{a,w} = w(V)(\mathbf{1}\{A = a^\star\} - \mathbf{1}\{A = a\})$; the expected normalized correct-minus-rival weight is $p - q + \varepsilon^2(1 + 1/K)/p > 0$, so the weighted correct mode is unique and the question is the convergence rate. With $\delta = \varepsilon/p$ the Chernoff moment is exactly

$$F_\varepsilon(t) = 1 - p - q + pe^{-t}\{\cosh(t\delta) - \delta\sinh(t\delta)\} + qe^{t}\{\cosh(t\delta) - \tfrac{p}{1-p}\,\delta\sinh(t\delta)\}, \qquad t \geq 0.$$

At $\varepsilon = 0$ the unique minimizer is $\theta = \frac{1}{2}\log(p/q)$ with minimum $m_0 = 1 - p - q + 2\sqrt{pq}$ and second derivative $2\sqrt{pq} > 0$; F_ε is even and analytic in ε, strictly convex and coercive in t, so by the implicit-function theorem the minimizer is $t_\varepsilon = \theta + O(\varepsilon^2)$ and its displacement changes the minimum only at order ε^4. Expanding at θ gives

$$m_\varepsilon := \min_{t \geq 0} F_\varepsilon(t) = m_0 + \frac{\sqrt{pq}\,\theta}{p^2}\Big(\theta - \frac{1}{1-p}\Big)\varepsilon^2 + O(\varepsilon^4),$$

and $\theta = \frac{9}{2}\log 2 > 3 = 1/(1-p)$ (since $\log 2 > 2/3$), so $m_\varepsilon > m_0$ for every sufficiently small $\varepsilon > 0$. Every wrong answer has the same pairwise law, the increment $X_{a,w}$ is bounded and takes both signs with positive probability, and the rivals are finitely many; the strict-defeat/weak-defeat sandwich of Section B.10 therefore gives the half-floor the full error exponent $-\log m_\varepsilon$ under any maximizing tie rule, while plurality's exponent is $-\log m_0$ (the unweighted rate of Theorem 1). Hence $I_{\text{half-floor}} < I_{\text{plurality}}$; both exponents are positive, so both accuracies tend to one and the half-floor is less accurate at every sufficiently large N. □

Relation to the positive criterion. On this family the conditional score gap is $\mathbb{E}[V \mid A = a^\star] - \mathbb{E}[V \mid A = a] = \varepsilon^2/[p(1-p)] > 0$, so Theorem 1 guarantees that the bonus $(V - 1/2)_+$ improves the exponent at every sufficiently small intensity, while the conventional intensity two loses. The useful intensity depends on the law; as ε decreases the law changes, so this is not a statement about arbitrarily small intensities on one fixed law. No numerical ε or crossover budget is asserted.

B.6 Proof of the Tie-Independent Error-Rate Separation

Proof of Corollary 2: the interior-boost case of the eventual-loss clause of Theorem 3 (the remaining case is in Section B.7). The following statement, formerly a main-text corollary, gives the explicit law and the error-ratio limit behind the final sentence of Theorem 3.

Corollary 2 (Eventual finite-vote loss independent of ties: a quantitative refinement of Theorem 3)**.** *Let h be bounded, measurable, strictly positive, and population mode preserving as in Theorem 2, with baseline c_0 on $(0, 1/2)$. If $h(b) > c_0$ for some $b \in [1/2, 1)$, there is a calibrated three-answer law P with uniquely correct base and tilted modes such that weighted plurality is less accurate than unweighted plurality for every sufficiently large sample size. Samples are iid from the original P, and either rule may resolve ties arbitrarily among maximizing answers. For the explicit law and constants $0 < m_1 < m_\rho < 1$ below, the respective error probabilities satisfy*

$$e_h(N), e_1(N) \longrightarrow 0, \qquad \lim_{N\to\infty} \frac{1}{N} \log \frac{e_h(N)}{e_1(N)} = \log \frac{m_\rho}{m_1} > 0. \tag{25}$$

The limit holds over all positive integers, including odd sample sizes. This strengthens the interior-boost necessity in Theorem 3; it does not change its preserving class or the tie convention used for sufficiency.

For the dominant rival, the weighted negative-tilt moment function strictly exceeds the unweighted one at every positive tilt. Their minima $m_\rho > m_1$ give the distinct error exponents. Strict and non-strict tail events have the same exponential rate, covering every maximizing tie convention.

Proof. Put $a = 1/4$, $\eta = (1 - b)/2$, and $\rho = c_0/h(b) \in (0, 1)$. Use the joint law

V	$a^\star$	a_1	a_2
a	ηa	$\eta(1-a)$	0
b	$(1-\eta)b$	0	$(1-\eta)(1-b)$

Both score strata are calibrated. Write $u = \eta a$, $v = (1 - \eta)b$, $p = u + v = P(A = a^\star)$, $q = (1 - \eta)(1 - b) = P(A = a_2)$, and $r = \eta(1 - a) = P(A = a_1)$. All four nonzero joint atoms are positive, and

$$p - q = \eta/4 + (1 - \eta)(2b - 1) > 0, \qquad q - r = \frac{(1 - b)(1 + 4b)}{8} > 0.$$

Thus $p > q > r > 0$, so the correct base mode is unique. After dividing all weights by $h(b)$, the weighted correct-minus-a_2 mean is $\rho u + v - q > 0$; the correct-minus-a_1 mean is at least $\rho(p - r) > 0$. The tilted mode is also uniquely correct.

The one-observation weighted correct-minus-a_2 increment has values $\rho, 1, 0, -1$ with probabilities u, v, r, q. Its negative-tilt moment function is

$$M_\rho(t) = r + ue^{-\rho t} + ve^{-t} + qe^{t}.$$

Its positive mean, strictly positive second derivative, and $q > 0$ imply a unique finite minimizer $t_\rho > 0$ over $t \geq 0$. Define $m_\rho = M_\rho(t_\rho) < 1$. For ordinary correct-minus-a_2 counts,

$$M_1(t) = r + pe^{-t} + qe^{t}, \qquad m_1 = \min_{t\geq 0} M_1(t) = r + 2\sqrt{pq}.$$

Since $u > 0$ and $\rho < 1$, one has $M_\rho(t) > M_1(t)$ for every $t > 0$. Consequently

$$m_\rho = M_\rho(t_\rho) > M_1(t_\rho) \geq m_1.$$

The corresponding minimum for ordinary correct-minus-a_1 counts is $m_{\text{other}} = q + 2\sqrt{pr}$. Moreover,

$$m_1 - m_{\text{other}} = (\sqrt{q} - \sqrt{r})(2\sqrt{p} - \sqrt{q} - \sqrt{r}) > 0.$$

Hence $0 < m_{\text{other}} < m_1 < m_\rho < 1$. These strict comparisons apply at $b = 1/2$ as well as above it.

For completeness, consider any finite-support increment Y with positive mean and both positive and negative values of positive probability. Set $M(t) = \mathbb{E}e^{-tY}$, let $t_* > 0$ minimize M, and put $m = M(t_*)$. For iid sums S_N, the exact logarithmic tail rate is

$$\lim_{N\to\infty} \frac{1}{N} \log \Pr(S_N < 0) = \lim_{N\to\infty} \frac{1}{N} \log \Pr(S_N \leq 0) = \log m. \tag{26}$$

Indeed, Chernoff's inequality gives $\Pr(S_N \le 0) \le m^N$. For a lower bound, fix $t > t_*$ and change the one-draw law to $Q_t(Y = y) = P(Y = y)e^{-ty}/M(t)$. Its mean $\mu_t = -M'(t)/M(t) < 0$. For $0 < \delta < -\mu_t$, the law of large numbers gives $Q_t^N(E_N) \to 1$ for $E_N = \{|S_N/N - \mu_t| < \delta\} \subset \{S_N < 0\}$. Changing measure yields

$$\Pr(S_N < 0) \ge M(t)^N e^{Nt(\mu_t - \delta)} Q_t^N(E_N).$$

Taking the lower logarithmic limit, then $\delta \downarrow 0$ and $t \downarrow t_*$, proves the matching bound in (26). This argument holds for every integer N without a lattice or parity condition.

Let $S_{1,N}$ and $S_{\rho,N}$ be the unweighted and normalized weighted correct-minus-a_2 totals. A strict rival lead forces an error, whereas any error requires a rival to tie or exceed the correct total. Independently of tie handling,

$$\Pr(S_{1,N} < 0) \le e_1(N) \le m_1^N + m_{\text{other}}^N.$$

For the other rival, the normalized weighted sample margin satisfies

$$W_{a^\star} - W_{a_1} = \rho(n_{a^\star} - n_{a_1}) + (1 - \rho)n_{a^\star,b} \ge \rho(n_{a^\star} - n_{a_1}),$$

where $n_{a^\star,b}$ counts correct observations at score b. Thus a weighted a_1 tie or win implies an unweighted a_1 tie or win, giving

$$\Pr(S_{\rho,N} < 0) \le e_h(N) \le m_\rho^N + m_{\text{other}}^N.$$

Applying (26) to both sandwiches gives $N^{-1} \log e_1(N) \to \log m_1$ and $N^{-1} \log e_h(N) \to \log m_\rho$, proving (25). Both errors are positive at every N, since drawing only a_2 has probability $q^N > 0$. Their ratio therefore diverges. The same strict lower and non-strict upper bounds work for every maximizing tie rule, so the eventual comparison is uniform over tie handling. □

The threshold sample size may depend on the boost, and need not remain bounded as $h(b)/c_0 \downarrow 1$. Both absolute errors and their difference vanish; the result supplies no benchmark prevalence or practical effect size. The endpoint $b = 1$ is excluded necessarily, since boosting a certainly correct score can be harmless.

Magnitude and sensitivity. Nearly equal weights can still produce a substantial finite-budget loss. On an explicit calibrated law, half-floor weights of 0.500 and 0.501 lower accuracy by 4.03 percentage points at ten votes under first-occurrence ties. On a different law, the ordinary half-floor loses at least 0.169 points at 129 votes for every pair of maximizing tie rules. Steeper population-preserving weights yield larger losses: a constructed step loses at least 6.09 points at 65 votes and 10.35 at 129, and a continuous ramp approaches these losses. The appendix gives the distinct laws and complete signed sensitivity grids. The same interior restriction is necessary even for eventual nondegradation (Theorem 3).

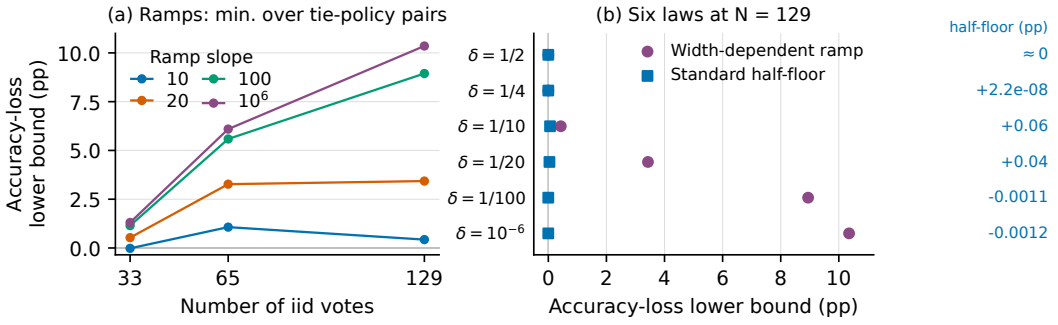


Figure 2: **Finite-vote loss on calibrated populations with correct base and weighted modes.** Each point is plurality's strict-correct probability minus the weighted vote's weak-correct probability, in percentage points. Positive values imply a loss for every pair of maximizing tie rules; nonpositive values give no such guarantee. (a) Continuous ramps at three budgets. (b) Six laws at $N = 129$, indexed by ramp width δ. Within each row, the ramp and standard half-floor use the same law; the right column gives the half-floor value. Rows are categorical, and the horizontal axis measures loss. The law changes across rows. Full signed results and first-occurrence outcomes appear in the appendix.

B.7 EVENTUAL NONDEGRADATION AND QUANTITATIVE COUNTEREXAMPLES

Proof of Theorem 3, eventual protection (equivalence of its second and third conditions). Eventual nondegradation has the same class as protection at every budget: with the first-occurrence tie convention for both votes and a bounded, measurable, strictly positive h, the existence of $N_0(P)$ after which weighted plurality is at least as accurate on every calibrated finite-answer law P with a uniquely correct base mode is equivalent to the class (8). If that class condition fails, a counterexample is eventually worse under every pair of maximizing tie rules; the stated first-occurrence convention is retained for the sufficiency direction.

Proof. The class condition preserves every correct plurality decision at every N by the every-budget part of Theorem 3, so take $N_0(P) = 1$. Conversely, if h is outside the positive population-preserving class, the necessity constructions in Theorem 2 give a finite-support law with a uniquely correct base mode and a rival whose weighted mass is *strictly* greater. Strict positivity excludes the zero-baseline construction, which is the only tie-only case in that proof. The strong law then sends plurality accuracy to one and weighted accuracy to zero, regardless of ties. If h is population preserving but violates the class condition, it has an interior boost; apply Corollary 2. Its strict-versus-weak error envelopes ensure eventual loss under any maximizing tie rules. □

Quantitative losses independently of tie handling. The four-draw half-floor example loses 3.48 percentage points under first-occurrence ties, gains 1.49 when both votes resolve ties against the correct answer, and loses 8.45 when both favor it. Its sign therefore depends on ties. Larger losses need not. On the calibrated law

V	$a^\star$	a_1	a_2
$1/4$	$1/16$	$3/16$	0
$1/2$	$3/8$	0	$3/8$

the population-preserving step transform $h_s(v) = 1 + (s-1)\mathbf{1}\{v \geq 1/2\}$ has a correct population mode for every finite $s \geq 1$. Nevertheless, complete multinomial enumeration gives the losses below. The last column is the plurality accuracy with every correct tie counted as an error minus the weighted accuracy with every correct tie counted as correct; a positive value therefore holds for *every* pair of maximizing tie rules.

Transform	N	First-occurrence loss	Minimum over ties
Step, $s = 2$	65	9.74	6.09
Step, $s = 2$	129	12.67	10.35
Step, $s = 10$	129	25.47	23.85
Raw score	33	1.19	-3.04
Raw score	65	7.01	3.55
Raw score	129	16.03	13.60
Raw score	257	27.63	25.89

All entries are percentage points. The raw-score rows use calibrated joint atoms $(a^\star, .45), (a_1, .45), (a^\star, .2), (a_2, .2)$ with respective probabilities $(.279, .341, .076, .304)$: the correct base mode becomes wrong after weighting, so the loss tends to 100 points. Continuity does not remove the population-preserving obstruction: moving the step law's high score to $1/2 + 10^{-6}$ and replacing the step by a linear ramp preserves losses of at least 6.08 and 10.33 points at $N = 65, 129$, respectively (Proposition 6). These are constructed lower bounds on worst-case loss, not minimax optima or empirical prevalence estimates. They do not establish a large tie-independent loss for the half-floor, whose earlier examples and full budget grid are retained.

A reproducible lower envelope. For the four ordered atoms $(a^\star, v_1), (a_1, v_1), (a^\star, v_2), (a_2, v_2)$, let $p = (p_0, p_1, p_2, p_3)$ and $w = (w_0, w_1, w_2, w_3)$. For $n \in \mathbb{N}_0^4$ with $\sum_i n_i = N$, define

$$B(n) = \mathbf{1}\{n_0 + n_2 > \max(n_1, n_3)\}, \qquad H_w(n) = \mathbf{1}\{w_0 n_0 + w_2 n_2 \geq \max(w_1 n_1, w_3 n_3)\}.$$

Then

$$L_N(p, w) = \sum_{\sum_i n_i = N} \frac{N!}{\prod_i n_i!} \prod_i p_i^{n_i} [B(n) - H_w(n)] \tag{27}$$

is the minimum plurality-minus-weighted accuracy over all pairs of maximizing tie rules: it counts every plurality tie against the correct answer and every weighted tie in its favor. In particular, $L_N > 0$ gives a tie-independent loss. An ordered first-occurrence tie selects the correct answer with conditional probability $(n_0 + n_2)/\sum_{a \in T} n_a$ when it belongs to the maximizing answer set T, and zero otherwise. This is generally not uniform choice among tied answer labels.

All displayed rows enumerate the complete count support with integer winner comparisons. An independent implementation sums their probabilities exactly using integer common denominators, including first-occurrence probabilities; the largest discrepancy from float64 summation is 3.2×10^{-14}. The fixed laws and budgets were fixed before this reproduction. Every declared row, including the negative lower envelope at $N = 33$, is retained. The earlier 216-row half-floor grid remains unchanged.

Proposition 6 (Continuous population-preserving weights can have large finite losses)**.** *For $0 < \delta < 1/2$, define*

$$h_\delta(v) = 1 + \min\{1, \max\{0, (v - 1/2)/\delta\}\}.$$

This continuous transform is universally population mode preserving. There is a calibrated three-answer law for which, at $\delta = 10^{-6}$, plurality is more accurate under every maximizing tie rule by at least 0.08 percentage points at $N = 65$ and 10.33 points at $N = 129$, although both population modes are uniquely correct.

Proof. The transform equals one below $1/2$ and is at least one above, so Theorem 2 applies. Put joint masses $(1/16, 3/16, 0)$ at score $1/4$ and $(3/8 + 3\delta/4, 0, 3/8 - 3\delta/4)$ at score $1/2 + \delta$. Both score strata are calibrated. The correct-minus-a_2 base and unnormalized weighted margins are $1/16 + 3\delta/2$ and $1/16 + 3\delta$, respectively; both also strictly exceed a_1. Compared on the four-atom index space with the $\delta = 0$ step-law construction, the one-draw total variation distance is $3\delta/4$, and the atomwise weights remain $(1, 1, 2, 2)$. The integrand in (27) lies in $[-1, 1]$. The product-law total variation bound consequently gives

$$L_N(P_\delta, h_\delta) \geq L_N(P_0, h_2) - 2N(3\delta/4).$$

Exact rational enumeration gives $L_{65}(P_0, h_2) > 0.0609398$ and $L_{129}(P_0, h_2) > 0.1035300$. With $\delta = 10^{-6}$ the respective lower bounds exceed 0.0608423 and 0.1033365, proving the claim. Independent exact enumeration of the perturbed laws additionally yields 0.0609393754 and 0.1035287626. □

The single value $\delta = 10^{-6}$ was fixed after reproducing the step examples, to make this continuity argument quantitative; no parameter grid was searched. Its steep transition is deliberate. A continuous ramp with the *original* two score values $1/4, 1/2$ and unequal weights could not belong to the population-preserving class, because continuity forces $h(1/2)$ to equal the lower baseline. The construction above moves the upper score, rather than making that invalid inference.

Magnitude and scope of the original witness. The universal four-sample witness of Theorem 3 can be tiny. For $h(b)/c_0 = 1.1$, its first-occurrence losses in probability units are 3.6264×10^{-4}, 5.0883×10^{-5}, 1.8706×10^{-6}, and 2.2575×10^{-10} at $b = 1/2, 3/4, 9/10, 99/100$, respectively. These are particular constructions, not upper bounds on the theorem's reach or on worst-case loss for a specified transform. Likewise, neither the old half-floor grid nor these new examples proves a uniform small-loss bound for the half-floor.

For the raw-score example the base masses are $(.355, .341, .304)$ and the unnormalized weighted masses are $(.14075, .15345, .0608)$. Thus a strictly correct base mode becomes a strictly wrong weighted mode, and the accuracy loss tends to one under every maximizing tie rule. On this occupied score support, the continuous strictly positive transform $h(v) = \max(v, .01)$ gives exactly the same votes as raw-score weighting; its behavior elsewhere does not affect the example. These results establish quantitative lower bounds, not a minimax characterization or the frequency of failures in actual verifiers.

How does population preservation translate into finite-budget accuracy? We compare the half-floor with plurality on two fully specified calibrated laws, using the same iid sample and first-occurrence ties. Enumerating every count vector at each displayed budget gives each rule's accuracy,

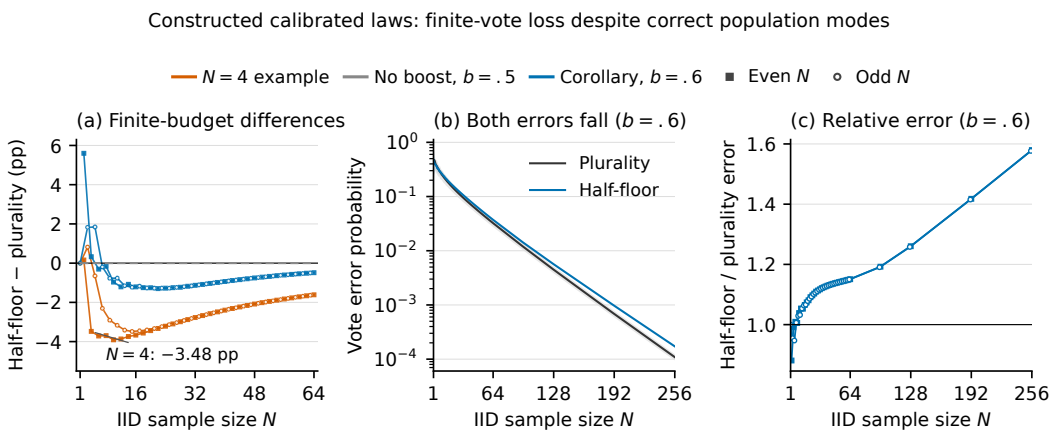


Figure 3: **Constructed calibrated laws: population preservation does not imply finite-vote accuracy dominance.** Both rules use the same iid sample and first-occurrence ties, with $h(v) = \max(v, 1/2)$. All displayed laws have uniquely correct base and weighted population modes. (a) Accuracy differences for the existing four-draw law and the corollary's $b = .6$ law; $b = .5$ gives an exactly constant-weight control. (b) Error probabilities for $b = .6$; shaded bands span all maximizing tie rules. (c) Their error ratio, with odd and even budgets marked separately. We enumerate the complete multinomial support using integer winner comparisons and floating-point probability sums; the $N = 4$ result is also verified by independent rational enumeration. Budgets are $1, \ldots, 64$ and the pairs $95/96$, $127/128$, $191/192$, $255/256$; lines connect evaluated points. The examples allow small-budget improvements, and both absolute errors vanish asymptotically. They make no empirical calibration or prevalence claim.

so the difference directly illustrates the distinction in Section 4: both population winners can be correct while their finite-budget accuracies differ (Figure 3). The existing four-draw example loses 3.48 percentage points at $N = 4$, but improves at $N = 2, 3$. For the construction in Corollary 2 with $b = .6$, the half-floor improves at $N = 2, \ldots, 5$ and loses at every evaluated budget from $N = 6$ onward, including odd sizes. At $N = 256$, its error is 1.58 times plurality's, yet the absolute errors are only 0.01722% and 0.01090%, respectively. Both errors and their difference tend to zero; the asymptotic theorem concerns their relative rate. These are constructed calibrated laws, not estimates of benchmark prevalence or evidence of empirical score calibration.

B.8 Disclosed ramp-width sensitivity and a half-floor comparator

How sensitive is the finite-budget separation to the steepness of the weight map, and does it also occur for the standard half-floor? This calculation extends the distinction in Section 4 by comparing the continuous ramp and half-floor with plurality on the same calibrated law at each grid point. Exact enumeration gives the accuracy loss under first-occurrence ties and its minimum and maximum over all maximizing tie rules. A positive minimum therefore establishes a loss independently of tie handling.

The exploratory grid was declared after the original continuous counterexample was observed and before these outcomes were computed. It fixes $\eta = a = 1/4$, $b = 1/2 + \delta$ and calibrated atom probabilities $(1/16, 3/16, 3b/4, 3(1-b)/4)$ with labels $(a^\star, a_1, a^\star, a_2)$. We enumerate every count vector for $\delta \in \{1/2, 1/4, 1/10, 1/20, 1/100, 10^{-6}\}$ and $N \in \{33, 65, 129\}$. The ramp $h_\delta(v) = 1 + \min\{1, \max\{0, (v - 1/2)/\delta\}\}$ has occupied weights $(1, 1, 2, 2)$ and slope $1/\delta$; the rescaled half-floor has weights $(1, 1, 1 + 2\delta, 1 + 2\delta)$. Both population modes remain correct, and the transforms coincide at $\delta = 1/2$.

Table 7 retains every signed outcome. At slope 10 and $N = 65$, the ramp's minimum loss is 1.069 points. The half-floor's minimum loss is $.05938$ points at $\delta = .1$, $N = 129$. At $\delta = 10^{-6}$, its first-occurrence losses arise entirely from changed plurality-tie decisions because $2\delta N < 1$. The grid varies the law together with the transform and describes these constructed cases; it is neither a

δ	Slope	N	Ramp loss (pp)	Half-floor loss (pp)
1/2	2	33	-0.002301 [-0.002301, -0.002301]	-0.002301 [-0.002301, -0.002301]
1/2	2	65	-6.1e-07 [-6.1e-07, -6.1e-07]	-6.1e-07 [-6.1e-07, -6.1e-07]
1/2	2	129	-5.751e-14 [-5.751e-14, -5.751e-14]	-5.751e-14 [-5.751e-14, -5.751e-14]
1/4	4	33	-0.04294 [-0.1548, 0.07373]	-0.07363 [-0.1711, 0.02691]
1/4	4	65	0.0006059 [-0.0006102, 0.001854]	-0.0002818 [-0.001259, 0.0007047]
1/4	4	129	5.584e-07 [3.5e-07, 7.754e-07]	1.515e-07 [2.227e-08, 2.826e-07]
1/10	10	33	2.468 [-0.01845, 5.072]	0.8926 [-0.9376, 2.773]
1/10	10	65	1.827 [1.069, 2.609]	0.4634 [-0.01607, 0.9441]
1/10	10	129	0.5347 [0.4295, 0.6441]	0.1192 [0.05938, 0.1792]
1/20	20	33	4.673 [0.5303, 9.01]	1.936 [-1.128, 5.079]
1/20	20	65	5.391 [3.271, 7.575]	1.213 [-0.1031, 2.529]
1/20	20	129	4.226 [3.43, 5.049]	0.4854 [0.03901, 0.9322]
1/100	100	33	6.383 [1.147, 11.86]	2.763 [-1.088, 6.717]
1/100	100	65	8.987 [5.588, 12.48]	2.09 [-0.1124, 4.292]
1/100	100	129	10.95 [8.943, 13.03]	1.205 [-0.001131, 2.411]
10^{-6}	10^6	33	6.729 [1.306, 12.4]	2.909 [-1.086, 7.011]
10^{-6}	10^6	65	9.745 [6.094, 13.5]	2.278 [-0.1138, 4.67]
10^{-6}	10^6	129	12.67 [10.35, 15.05]	1.425 [-0.001158, 2.851]

Table 7: All 36 declared sensitivity comparisons. Entries are first-occurrence loss followed by [minimum, maximum] over pairs of maximizing tie rules, in percentage points; positive loss means weighting is less accurate than plurality. Scientific notation preserves the signs of tiny results. All count vectors are summed exactly; unrounded values and exact fractions accompany the source.

worst-case optimization nor a benchmark prevalence estimate. Exact rational enumeration agrees with independent floating-point sums within 1.7×10^{-14}.

B.9 Half-Floor Reversals and the Ordinary Count Margin

Proposition 7 (A count-margin bound for the half-floor)**.** *Fix a finite answer alphabet with one correct label $a^\star$ and at least one other label. Let $(A_i, V_i)_{i\ge1}$ be iid under one eligible law, with a fixed score map $V \in [0,1]$ satisfying $\mathbb{P}(A = a^\star \mid V) = V$. Let P_n and H_n be ordinary and $h(v) = \max(v, 1/2)$ weighted plurality on the same first n draws, with any maximizing tie conventions. Write $C_{a,n} = \sum_{i=1}^{n} \mathbf{1}\{A_i = a\}$ and $G_n = C_{P_n,n} - \max_{a\neq P_n} C_{a,n}$, assigning zero counts to absent labels. For every fixed integer $m \ge 1$,*

$$\mathbb{P}(\exists n \ge 1 :\ P_n = a^\star,\ G_n \ge m,\ H_n \neq a^\star) \le e^{-2m}. \tag{28}$$

No correct population mode or population majority is required.

We prove Proposition 7 using a scalar exponential supermartingale and Ville's inequality (Ville, 1939; Howard et al., 2020). The concentration machinery is standard; the reduction below compares half-floor voting with a correct ordinary plurality without a union bound over wrong labels.

Proof of Proposition 7. Let $R_i = \mathbf{1}\{A_i = a^\star\}$, $b(v) = (v - 1/2)_+$, and $\mathcal{F}_n = \sigma((A_i, V_i) : 1 \le i \le n)$, with $\mathcal{F}_0$ trivial. Define

$$T_0 = 0, \qquad T_n = \sum_{i=1}^{n} b(V_i)(2R_i - 1), \qquad M_n = e^{-4T_n}.$$

All one-step exponential factors are bounded, hence integrable. At $v \le 1/2$, the factor is identically one. At $v = 1/2 + b$ with $0 < b \le 1/2$, calibration gives

$$\begin{aligned}\mathbb{E}\Big[e^{-4b(V_i)(2R_i-1)} \mid V_i = v\Big] &= (1/2+b)e^{-4b} + (1/2-b)e^{4b} \\ &= \cosh(4b) - 2b\sinh(4b) \le 1.\end{aligned}$$

Indeed, for $b > 0$ the last inequality is equivalent to $\tanh(2b) \le 2b$; at $b = 0$ it holds with equality. These conditional statements hold almost surely under the score law, including any atoms.

Independence of the next pair from $\mathcal{F}_{n-1}$ therefore yields

$$\mathbb{E}[M_n \mid \mathcal{F}_{n-1}] = M_{n-1}\mathbb{E}\Big[e^{-4b(V_n)(2R_n-1)}\Big] \le M_{n-1}.$$

Thus $(M_n)_{n\ge 0}$ is a nonnegative supermartingale with $M_0 = 1$.

For an answer a, write its accumulated bonus and half-floor weight as

$$U_{a,n} = \sum_{i=1}^{n} b(V_i)\mathbf{1}\{A_i = a\}, \qquad W_{a,n} = \tfrac{1}{2}C_{a,n} + U_{a,n}.$$

Since there is one correct label, $T_n = U_{a^\star,n} - \sum_{a\neq a^\star} U_{a,n}$. Bonuses are nonnegative, so every wrong answer satisfies

$$U_{a^\star,n} - U_{a,n} \ge T_n.$$

Suppose $P_n = a^\star$ and $G_n \ge m$. Then $C_{a^\star,n} - C_{a,n} \ge m$ for every wrong answer a. If also $H_n \neq a^\star$, its maximizing label $a = H_n$ obeys $W_{a,n} \ge W_{a^\star,n}$, even if the weighted decision is tied. Consequently

$$0 \ge W_{a^\star,n} - W_{a,n} = \tfrac{1}{2}(C_{a^\star,n} - C_{a,n}) + U_{a^\star,n} - U_{a,n} \ge m/2 + T_n.$$

The harmful-reversal event at any n therefore implies $M_n \ge e^{2m}$. This implication holds pathwise for every pair of maximizing tie conventions; it does not require their agreement or a union bound over possible winning rivals. Ville's inequality gives

$$\mathbb{P}(\exists n \ge 1 :\ P_n = a^\star,\ G_n \ge m,\ H_n \neq a^\star) \le \mathbb{P}\Big(\sup_{n\ge 0} M_n \ge e^{2m}\Big) \le e^{-2m},$$

proving the proposition. The margin threshold m is fixed in this statement; it is not selected after observing the evaluated sequence. □

Fixed-budget accuracy regret. For every fixed $n \ge 1$ and fixed integer $m \ge 1$, under the same assumptions and arbitrary maximizing tie conventions,

$$\mathbb{P}(P_n = a^\star) - \mathbb{P}(H_n = a^\star) \le \mathbb{P}(P_n = a^\star,\ G_n < m) + e^{-2m}. \tag{29}$$

To see this, subtract the two correctness indicators in expectation:

$$\begin{aligned}\mathbb{P}(P_n = a^\star) - \mathbb{P}(H_n = a^\star) &= \mathbb{P}(P_n = a^\star, H_n \neq a^\star) - \mathbb{P}(P_n \neq a^\star, H_n = a^\star)\\ &\le \mathbb{P}(P_n = a^\star, G_n < m) + \mathbb{P}(P_n = a^\star, G_n \ge m, H_n \neq a^\star).\end{aligned}$$

The last term is at most e^{-2m} by (28). For example, $m = 3$ gives $e^{-6} \simeq 0.00248$ in probability units for the joint harmful-reversal event with that lead. This is neither a conditional error rate given the observed margin nor an estimate of how often the small-margin term occurs. The proposition needs no assumption that the correct answer is a base population mode; it does not assert a positive benefit from weighting.

A scalar-moment extension. For the same iid stream and fixed score map, exact calibration can be replaced by the sufficient condition

$$\psi_\lambda := \mathbb{E}\exp\{-\lambda(V - 1/2)_+(2R - 1)\} \le 1 \quad \text{for a fixed } \lambda > 0.$$

Then $e^{-\lambda T_n}$ is a nonnegative supermartingale, and the right-hand side of (28) becomes $e^{-\lambda m/2}$; the regret inequality changes accordingly. If only $\psi_\lambda \le e^\eta$ is known for a fixed real η, Markov's inequality instead bounds the fixed-n joint harmful-reversal probability by $\min\{1, e^{n\eta - \lambda m/2}\}$. Calibration implies the first condition at $\lambda = 4$ by the proof above. These are conditions on the true problem-specific eligible law, not certificates supplied by a pooled empirical moment or a fitted score map. For a dependent stream, the supermartingale argument requires the corresponding conditional moment bound given the past; marginal calibration alone is insufficient.

Log-odds-dominated bonuses. The same argument applies to a fixed measurable weight $h(v) = c_0 + b(v)$, where $c_0 > 0$, b is bounded and nonnegative, $b(v) = 0$ on $[0, 1/2]$, and $\lambda b(v) \le \log(v/(1-v))$ on $(1/2, 1)$ for a fixed $\lambda > 0$. Any finite $b(1) \ge 0$ is allowed. For $1/2 < v < 1$, put $x = e^{\lambda b(v)} \ge 1$. Calibration gives

$$\mathbb{E}[e^{-\lambda b(V)(2R-1)} \mid V = v] = v/x + (1-v)x \le 1 \quad \Longleftrightarrow \quad 1 \le x \le v/(1-v).$$

On $[0, 1/2]$ the exponential factor is one; at score one it is $e^{-\lambda b(1)} \le 1$ almost surely. Thus, for the same iid law, $e^{-\lambda T_n^b}$ is a nonnegative supermartingale starting at one, where $T_n^b = \sum_{i \le n} b(V_i)(2R_i - 1)$. Write $U_{a,n}^b = \sum_{i \le n} b(V_i)\mathbf{1}\{A_i = a\}$ and let H_n^h denote the maximizing vote under this h, so its answer totals are $c_0 C_{a,n} + U_{a,n}^b$. Nonnegative bonuses imply $U_{a^\star,n}^b - U_{a,n}^b \ge T_n^b$ for every wrong label. A reversal from $P_n = a^\star$ to $H_n^h \ne a^\star$ at ordinary count lead $G_n \ge m$ therefore forces $T_n^b \le -c_0 m$. Ville's inequality gives the corresponding bound for this weight:

$$\mathbb{P}\big(\exists n \ge 1 :\ P_n = a^\star,\ G_n \ge m,\ H_n^h \ne a^\star\big) \le e^{-\lambda c_0 m}. \tag{30}$$

The same fixed-law, fixed-threshold and arbitrary maximizing-tie scope applies. For the half-floor, $c_0 = 1/2$ and $b(v) = (v - 1/2)_+$; the inequality $\log(v/(1-v)) \ge 4(v - 1/2)$ follows by differentiation from equality at $1/2$, since $1/[v(1-v)] \ge 4$. Taking $\lambda = 4$ recovers (28). This sufficient envelope uses the competence log-odds familiar from classical binary voting (Nitzan & Paroush, 1982); it does not assert Bayes optimality under unrestricted wrong-answer allocation. Related work characterizes Condorcet consistency for heterogeneous independent voters (Berend & Paroush, 1998) and optimal committee rules for polychotomous choices (Ben-Yashar & Paroush, 2001).

B.10 Exact criterion for a beneficial score bonus

We prove the specified-bonus part of Theorem 1 (this appendix writes I_η for $I_b(\eta)$ and I_0 for I_1), then its existence part and the total-variation formula, and give an explicit sufficient intensity under a structured calibrated law. The argument uses classical Chernoff exponents, conditional Jensen and differentiation of the minimizing moment; see classical weighted-vote and exponential-loss analyses (Berend & Kontorovich, 2014; Schapire & Singer, 1999).

Proof sketch: why the criterion has this form. For each positive-mass wrong answer a, define

$$X_{a,\eta} = (1 + \eta b(V))(\mathbf{1}\{A = a^\star\} - \mathbf{1}\{A = a\}), \qquad F_a(\eta, t) = \mathbb{E}e^{-tX_{a,\eta}}.$$

Strict and weak pairwise defeats have the same exponent, so $I_b(\eta) = \min_{a \ne a^\star :\, p_a > 0}\{-\log \inf_{t \ge 0} F_a(\eta, t)\}$. If a leading rival has $\mu_a \ge \mu_{a^\star}$, conditional Jensen gives, for every $\eta, t \ge 0$,

$$\begin{aligned} F_a(\eta, t) &\ge 1 - p - q + pe^{-t(1+\eta\mu_{a^\star})} + qe^{t(1+\eta\mu_a)} \\ &\ge 1 - p - q + 2\sqrt{pq} = m_0. \end{aligned}$$

The last inequality uses the ordering of the means and $pe^{-s} + qe^{s} \ge 2\sqrt{pq}$. Thus $I_b(\eta) \le I_1$ at every intensity, including when the mean gap is zero. At zero intensity, only the leading rivals attain plurality's exponent. Positive mean gaps increase all these active rates, giving (4); the finitely many remaining rates retain their positive baseline gap for sufficiently small intensity.

Allowing the bonus to vary turns this condition into separation from the rivals' mixture set: inside it, the correct answer's mean of any score-only weight is a mixture of its means on the leading rivals, so it cannot exceed them all; outside it, a bounded bonus separates the correct answer from every leading rival, and a minimax exchange makes the best normalized separation exactly the total-variation distance (proved below).

Pairwise exponents and arbitrary ties. Omit zero-mass labels. For a wrong label a, set $q_a = P(A = a) > 0$ and

$$\begin{aligned} X_{a,\eta} &= (1 + \eta b(V))(\mathbf{1}\{A = a^\star\} - \mathbf{1}\{A = a\}), \\ F_a(\eta, t) &= \mathbb{E}e^{-tX_{a,\eta}}, \qquad m_a(\eta) = \inf_{t \ge 0} F_a(\eta, t). \end{aligned}$$

The argument below is written for $w = 1 + \eta b$, but it uses only that the increment is bounded and takes both signs with positive probability, so it applies verbatim to every bounded strictly positive weight;

the existence corollary's proof invokes it in that form. For each fixed η, the increment is bounded and takes both positive and negative values with positive probability. If its mean is positive, Cramér's theorem (Dembo & Zeitouni, 2010) gives the same exponent $-\log m_a(\eta)$ for $\sum_{i\le N} X_{a,\eta,i} < 0$ and for $\sum_{i\le N} X_{a,\eta,i} \le 0$: zero is in the interior of the convex support, and the rate function is continuous there. This also handles lattice-valued increments and every integer sample size. If the mean is negative, both events tend to probability one by the law of large numbers. At zero mean their exponent is zero by the same large-deviation lower bound (or the central limit theorem, since the variance is positive). In both cases this agrees with $m_a(\eta) = 1$, the convex moment's minimum over $t \ge 0$.

The vote's error event contains the union of strict pairwise defeats and is contained in the union of weak defeats. There are finitely many rivals, so these two unions have the same exponential rate. Consequently

$$I_\eta = \min_{a \ne a^\star} \{-\log m_a(\eta)\}, \tag{31}$$

independently of deterministic, randomized or sample-dependent maximizing tie choices. Absent labels cannot beat the positive weight of an observed label and need not be included.

Discontinuity in the law. Because J uses exact equality, D and the right derivative in (4) can jump when rival frequencies tie. A rival just below q is absent from J but can still limit the improving intensity. A positive initial derivative therefore need not give an intensity range bounded away from zero as the law varies. With a single leading rival, the existence condition reduces to $Q_{a^\star} \ne Q_{a_1}$. The second calibrated law in Section B.11 illustrates how a lower-frequency rival can limit the useful intensity.

Derivative and local sufficiency. At $\eta = 0$ the unique minimizing positive tilt and minimum are

$$\theta_a = \tfrac{1}{2}\log(p/q_a), \qquad m_a(0) = 1 - p - q_a + 2\sqrt{pq_a}.$$

Boundedness permits differentiation of F_a in a neighborhood of $(0, \theta_a)$, and its second derivative in t is strictly positive. The implicit function theorem therefore gives a smooth local minimizer; convexity makes it the global minimizer. The envelope calculation yields

$$\partial_\eta F_a(0, \theta_a) = \theta_a\sqrt{pq_a}(\mu_a - \mu_{a^\star}), \qquad \frac{d}{d\eta}[-\log m_a(\eta)]\Big|_0 = \frac{\theta_a\sqrt{pq_a}}{m_a(0)}(\mu_{a^\star} - \mu_a).$$

Since $m_a(0)$ strictly increases with $q_a < p$, precisely the rivals in J attain the minimum baseline exponent. The other finitely many rivals are separated by a positive exponent gap. The right derivative of the minimum is therefore the minimum derivative over J, proving the first formula in (4); the second is proved below. If every active mean gap is positive, all sufficiently small positive intensities improve the full exponent.

Global necessity. Suppose some $a \in J$ has $\mu_a \ge \mu_{a^\star}$. Conditional Jensen and $\eta, t \ge 0$ give

$$\begin{aligned} F_a(\eta, t) &\ge 1 - p - q + pe^{-t(1+\eta\mu_{a^\star})} + qe^{t(1+\eta\mu_a)} \\ &\ge 1 - p - q + pe^{-t(1+\eta\mu_{a^\star})} + qe^{t(1+\eta\mu_{a^\star})} \\ &\ge 1 - p - q + 2\sqrt{pq} = m_0. \end{aligned}$$

Hence $I_\eta \le -\log m_a(\eta) \le -\log m_0 = I_0$ at every positive intensity. This proves necessity even when an active mean gap is zero, completing the equivalence. Whenever $I_\eta > I_0$, the exact rates imply $\log(e_\eta(N)/e_0(N))/N \to -(I_\eta - I_0) < 0$, where $e_\eta(N)$ is the error probability. Thus weighted accuracy is strictly better at every sufficiently large N, for each fixed qualifying intensity. No finite crossover bound is implied by the criterion itself; Section B.4 gives one for a specific law by direct bounds.

Structured calibrated corollary. Assume additionally $P(A = a^\star \mid V) = V$ and $P(A = a \mid V) = \pi_a(1 - V)$ for fixed wrong-answer probabilities $\pi_a \ge 0$ summing to one. Then $p = \mathbb{E}V$ and, for $\pi_a > 0$,

$$\mu_{a^\star} - \mu_a = \frac{\mathbb{E}[Vb(V)]}{p} - \frac{\mathbb{E}[(1-V)b(V)]}{1-p} = \frac{\mathrm{Cov}(V, b(V))}{p(1-p)}.$$

For nondecreasing b, an independent copy V' gives $2\,\mathrm{Cov}(V, b(V)) = \mathbb{E}[(V - V')(b(V) - b(V'))] \geq 0$. This is strictly positive if $b(V)$ is not almost surely constant: two distinct bonus values then have strictly ordered score values with positive probability. With $b(v) = (v - 1/2)_+$, every $1 + \eta b(v)$ belongs to Theorem 2's preserving class. The additional structure and a nonconstant bonus therefore give sufficiently weak members a strictly better error exponent. This conclusion fails to follow from a concentration cap alone and does not prescribe the conventional half-floor intensity $\eta = 2$.

Explicit sufficient intensity under this structure. Choose $B > 0$ such that $0 \leq b(v) \leq B$ for all v. Let $\pi = \max_a \pi_a$, $q = \pi(1 - p)$, $C_b = \mathrm{Cov}(V, b(V)) > 0$, $\theta = \frac{1}{2}\log(p/q)$ and

$$A_b = \theta\sqrt{\frac{\pi}{p(1-p)}}C_b, \qquad L = (1 + \mathrm{e})\theta^2 B^2\sqrt{pq}.$$

Here $B > 0$, $\theta > 0$ and $A_b > 0$. Every intensity

$$0 < \eta \leq \min\left\{\frac{1}{\theta B}, \frac{A_b}{L}\right\} \tag{32}$$

improves the error exponent. To prove this, keep the tilt $t = \theta$ fixed for a top rival. Its derivative at zero is $-A_b$. For $0 \leq \eta \leq 1/(\theta B)$, its second derivative in intensity is bounded by

$$\theta^2 B^2\{pe^{-\theta} + qe^{\theta}\mathrm{e}\} = (1 + \mathrm{e})\theta^2 B^2\sqrt{pq} = L.$$

Taylor's bound gives $F_{\mathrm{top}}(\eta, \theta) \leq m_0 - A_b\eta + L\eta^2/2 \leq m_0 - A_b\eta/2 < m_0$. For any rival the structured moment is

$$F_a(\eta, t) = \mathbb{E}[Ve^{-th_\eta(V)}] + (1 - p) + \pi_a\mathbb{E}[(1 - V)(e^{th_\eta(V)} - 1)].$$

It increases with π_a for $t \geq 0$, so the top rival dominates all pairwise moments and the bound improves the full exponent. This is a population-parameter guarantee. Estimating the mean gaps, rival set and intensity from labelled candidates requires accounting for their estimation error and selection; substituting empirical estimates alone does not give a deployment certificate.

The existence part of Theorem 1 is restated next in the form its proof uses; it adds no claim beyond the main-text statement.

Corollary 3 (Existence part of Theorem 1, restated)**.** *Under the assumptions of Theorem 1, write $Q_a(\cdot) = \mathbb{P}(V \in \cdot \mid A = a)$ for the conditional score laws and J for the most frequent wrong answers. The following are equivalent: (i) some bounded, pointwise strictly positive measurable weight $w(V)$ has error exponent $I_w > I_0$; (ii) some bounded measurable bonus $b \geq 0$ satisfies $\mathbb{E}[b(V) \mid A = a^\star] > \max_{a \in J}\mathbb{E}[b(V) \mid A = a]$; (iii) $Q_{a^\star} \notin \mathrm{conv}\{Q_a : a \in J\}$.*

Proof. (iii)⇒(ii). Let $k = |J|$ and let $L \subseteq \mathbb{R}^k$ be the linear subspace of vectors $(\int f\,dQ_{a^\star} - \int f\,dQ_a)_{a \in J}$ over bounded measurable f. If no such vector has all coordinates positive, L misses the open positive orthant, so by Gordan's alternative (Gordan, 1873) some $y \geq 0$, $y \neq 0$, is orthogonal to L; taking f to be indicators gives $Q_{a^\star} = \sum_a (y_a/\sum_{a'} y_{a'})Q_a$, contradicting (iii). Hence some bounded f has all gaps positive, and $b = f - \inf f \geq 0$ has the same gaps. (ii)⇒(i). Theorem 1 with $w = 1 + \eta b$ at a sufficiently small $\eta > 0$. not (iii)⇒not (i). If $Q_{a^\star} = \sum_{a \in J}\lambda_a Q_a$, then for every bounded $w > 0$, $\mathbb{E}[w \mid a^\star] = \sum_a \lambda_a \mathbb{E}[w \mid a]$, so some $a \in J$ has $\mathbb{E}[w \mid a] \geq \mathbb{E}[w \mid a^\star] =: m$. With $X_{a,w} = w(V)(\mathbf{1}\{A = a^\star\} - \mathbf{1}\{A = a\})$, $S_{a,w,N} = \sum_{i=1}^N X_{a,w,i}$ for the iid copies, and $t \geq 0$, conditional Jensen gives $\mathbb{E}e^{-tX_{a,w}} = 1 - p - q + p\,\mathbb{E}[e^{-tw} \mid a^\star] + q\,\mathbb{E}[e^{tw} \mid a] \geq 1 - p - q + pe^{-tm} + qe^{tm} \geq 1 - p - q + 2\sqrt{pq} = m_0$. The error contains the strict defeat $\{S_{a,w,N} < 0\}$ under every maximizing-tie rule; $X_{a,w}$ is bounded and takes both signs because $w > 0$ and $p, q > 0$, so the strict and weak tails share the Cramér rate $\sup_{t \geq 0}(-\log \mathbb{E}e^{-tX_{a,w}}) \leq -\log m_0 = I_0$. Hence $\limsup_N -N^{-1}\log \mathbb{P}(\widehat{A}_{w,N} \neq a^\star) \leq I_0$, and $I_w \leq I_0$. (The pairwise-exponent argument at the start of this appendix applies verbatim to the increments $X_{a,w}$: they are bounded, take both signs, and there are finitely many rivals, so I_w exists and is the same under every maximizing tie convention.) □

In the opening populations A and B of Table 1 the score space has two points, so each conditional law is determined by its high-score probability: in population A the correct answer's value $56/68$

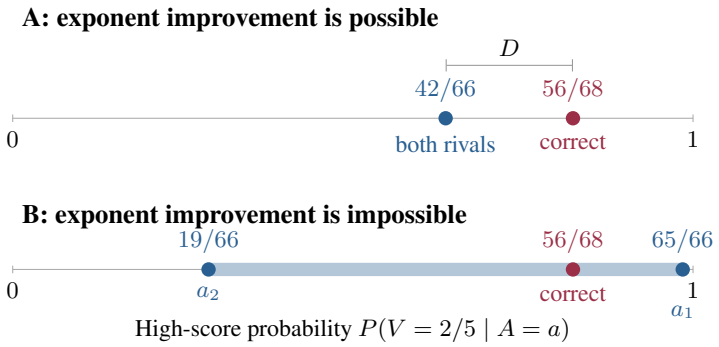


Figure 4: The criterion resolves the opening example. With two score values, a conditional score law is determined by its high-score probability. Blue marks the leading rivals' score laws and their mixtures; red marks the correct answer's law. In A the correct law lies outside the mixture set, so $D > 0$ and some bounded strictly positive score-only weight improves plurality's error exponent. In B it lies inside, so $D = 0$ and no such weight improves the exponent. These are the same calibrated populations as in Table 1.

differs from the common rival value $42/66$ (outside the single-point hull), while in population B the value $56/68$ lies between the rival values $19/66$ and $65/66$ (inside the interval); Figure 4 draws both cases.

For normalized bonuses $0 \le b \le 1$, the derivative formula also gives an exact measure of the best initial improvement. Write $c_{\exp} = \theta_0\sqrt{pq}/m_0$, using the quantities defined in Section 3 and Theorem 1, and let $Q_\lambda = \sum_{a\in J} \lambda_a Q_a$ for $\lambda \in \Delta_J$. Then

$$\max_{0\le b\le 1} I'[b](0+) = c_{\exp} \min_{\lambda\in\Delta_J} \mathrm{TV}(Q_{a^\star}, Q_\lambda),$$

and both optima are attained.

Proof. Let $\nu = Q_{a^\star} + \sum_{a\in J} Q_a$ and $\mathcal{B} = \{b \in L^\infty(\nu) : 0 \le b \le 1\}$. This set is convex and weak-* compact: it is a weak-* closed subset of the dual unit ball, which is compact by the Banach–Alaoglu theorem (Alaoglu, 1940). Each signed measure $Q_{a^\star} - Q_a$ has an $L^1(\nu)$ density. Consequently

$$F(b, \lambda) = \int b\, d(Q_{a^\star} - Q_\lambda)$$

is weak-* continuous and affine in b, and continuous and affine in λ on the finite-dimensional compact simplex Δ_J. The derivative formula and Sion's minimax theorem (Sion, 1958) therefore yield

$$\begin{aligned}\frac{1}{c_{\exp}} \max_{b\in\mathcal{B}} I'[b](0+) &= \max_{b\in\mathcal{B}} \min_{a\in J} \int b\, d(Q_{a^\star} - Q_a) \\ &= \max_{b\in\mathcal{B}} \min_{\lambda\in\Delta_J} F(b, \lambda) \\ &= \min_{\lambda\in\Delta_J} \max_{b\in\mathcal{B}} F(b, \lambda).\end{aligned}$$

For fixed λ, the signed measure $Q_{a^\star} - Q_\lambda$ has total mass zero. Its positive part has mass $\mathrm{TV}(Q_{a^\star}, Q_\lambda)$, attained by taking b to be the indicator of a positive set in a Hahn decomposition (Halmos, 1950, Section 29). This identifies the inner maximum and proves the formula.

The minimum of finitely many continuous gap functionals is upper semicontinuous on compact $\mathcal{B}$, so a maximizing bonus exists. The total-variation objective is continuous in λ, so a minimizing mixture exists as well. Representatives of the maximizing $L^\infty(\nu)$ class can be chosen in $[0, 1]$ everywhere without changing any gap. □

The optimization is over the right derivatives of fixed bonuses. It does not differentiate an intensity-optimized envelope or select an optimal finite intensity. Although the inner maximum for a fixed mixture has an indicator solution, a bonus maximizing the minimum gap across all rivals need not be an indicator.

B.11 THREE FINITE LAWS ILLUSTRATING THE POSITIVE CRITERION

Table 8: Computed error exponents for $h_\eta(V) = 1 + \eta(V - 1/2)_+$ on two calibrated laws. The endpoints $\eta = 0$ and $\eta = 2$ correspond to plurality and the standard half-floor, respectively. A beneficial weak intensity need not make the standard half-floor beneficial.

Constructed calibrated law	$\eta = 0$	$\eta = 0.25$	$\eta = 2$
Three-vote law (Section B.4)	0.001318	0.002159	0.011776
Second calibrated law (Section B.11)	0.085299	0.089990	0.077827

These calculations ask three questions about the criterion in Section 3 and Theorem 1: does a positive bonus gap yield an improved exponent at a small intensity, why must the gap hold against every leading rival, and why can a larger intensity lose the improvement? Three fully specified laws isolate these cases. Because each answer–score law is known, we can calculate the conditional bonus means and compare plurality's error exponent directly with each weighted rule's exponent, without fitting parameters to model outputs. In the calibrated binary law, $V \in \{0.6, 0.9\}$ equiprobably and $P(A = a^\star \mid V) = V$. For $b(V) = (V - 1/2)_+$, the conditional bonus means are 0.28 for the correct answer and 0.16 for its sole rival. The explicit sufficient interval in (32), using the support bound $B = 0.4$, is $0 < \eta \leq 0.3672$. The fixed intensity $\eta = 0.25$ therefore qualifies; its error exponent is 0.15159, compared with plurality's 0.14384. Direct finite-support convolution evaluates every integer budget $N = 1, \ldots, 256$ with uniform random tie resolution, at fixed intensities $0, 0.25, 2$. The complete curves in Figure 5 show nonnegative accuracy differences on this law over that range; this is not a finite-budget dominance theorem for other laws.

The uncalibrated law in Figure 5(d) has answer masses $(0.4, 0.3, 0.3)$ and deterministic bonuses $(0.2, 0.3, 0)$ for the correct answer and two rivals respectively. Although the correct answer's mean bonus exceeds the aggregate wrong-answer mean 0.15, one tied top rival has the larger mean 0.3. Theorem 1 rules out exponent improvement at *every* positive intensity. The two pairwise rates move in opposite directions, demonstrating why the criterion must examine every tied top rival rather than an aggregate correct–wrong score comparison.

A separate calibrated construction (the second law in Table 8) limits the safe intensity. Its four actual atoms $(a, V, P(a, V))$ are $(\text{correct}, 9/20, 9/50)$, $(a_1, 9/20, 11/50)$, $(\text{correct}, 13/20, 39/100)$, and $(a_2, 13/20, 21/100)$. The correct answer's overall mass is 0.57. The correct answer's conditional mean bonus is approximately 0.102632, above zero for the uniquely most frequent wrong answer a_1 but below 0.15 for a_2. The mean-gap criterion therefore guarantees improvement at sufficiently small intensity. At $\eta = 2$, the computed exponent is below plurality's as a_2 becomes the limiting rival. The displayed main-table exponent rows for both laws are covered by an independent high-precision receipt; a separate float64 check covers $0, 0.05, 2$.

B.12 TIE REFINEMENT UNDER CONDITIONAL BONUS DOMINANCE

The statement below extends the main-text Proposition 1 by the uniform-tie alternative and the weight representation; one proof covers both.

Proposition 8 (Conditional bonus dominance: full statement of Proposition 1)**.** *Fix a finite answer alphabet with one correct label $a^\star$, an integer $N \geq 1$, and an iid eligible stream $(A_i, V_i)_{i=1}^N$ under one fixed law, with $V \in [0, 1]$. Fix a measurable bonus $b : [0, 1] \to [0, B]$, where $B < \infty$. For each positive-mass label, let $G_a(t) = \mathbb{P}(b(V) \leq t \mid A = a)$. When $0 < \mathbb{P}(A = a^\star) < 1$, suppose*

$$G_{a^\star}(t) \leq G_a(t) \quad \text{for every } t \in \mathbb{R} \text{ and every wrong } a \text{ with } \mathbb{P}(A = a) > 0. \tag{33}$$

No ordering condition is needed when the correct label has probability zero or one.

Write $C_a = \sum_{i=1}^N \mathbf{1}\{A_i = a\}$ and $S_a = \sum_{i:A_i=a} b(V_i)$. Ordinary plurality P_N maximizes C_a. Define Q_N by maximizing C_a first, then S_a among the maximum-count labels. Both rules resolve any remaining ties by first occurrence in the original iid order. Alternatively, each chooses uniformly among its remaining maximizers using auxiliary randomness independent of the sample. For every count vector c of positive probability, with $T(c) = \arg\max_a c_a$,

$$\mathbb{P}(Q_N = a^\star \mid C = c) \geq \mathbb{P}(P_N = a^\star \mid C = c) = \frac{\mathbf{1}\{a^\star \in T(c)\}}{|T(c)|}. \tag{34}$$

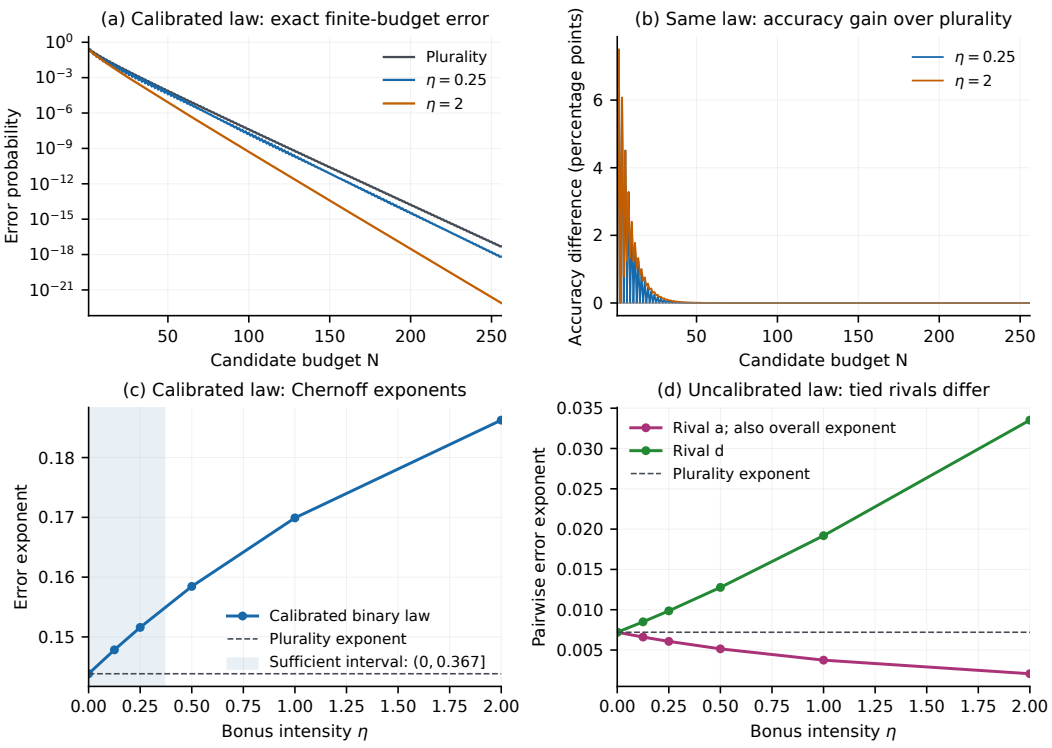


Figure 5: Two prescribed laws with bonus $b(v) = (v - 1/2)_+$. (a,b) Exact finite-support distributions, evaluated in floating point, give errors and accuracy differences at all $N = 1, \ldots, 256$ for the calibrated binary law $V \in \{0.6, 0.9\}$ equiprobably, $P(A = a^\star \mid V) = V$, with uniform random ties. (c) Its Chernoff exponent at the complete fixed grid $\eta \in \{0, 0.125, 0.25, 0.5, 1, 2\}$; shading is the explicit sufficient intensity interval, not an estimated optimum. (d) For the uncalibrated law with masses $(0.4, 0.3, 0.3)$ and scores $(0.7, 0.8, 0.5)$, the larger-bonus top rival determines the overall rate. Lines between exponent markers guide the eye between the six evaluated intensities. Both rival rates and all declared settings are retained.

Hence the inequality also holds unconditionally. Calibration, identical wrong-answer bonus laws, and a correct population mode are not required.

For every deterministic $\eta > 0$ satisfying $\eta NB < 1$, weighting by $h_\eta(v) = 1 + \eta b(v)$ has the same maximizing set as this lexicographic policy. Thus it gives the same first-occurrence decision, or the same decision law under uniform ties.

Proof of Propositions 1 and 8. Put $p_a = \mathbb{P}(A = a)$. If $p_{a^\star} = 0$, the correct label never appears. Since $N \geq 1$, the maximum count is positive and both policies are wrong almost surely. No conditional law given the absent correct label is needed. If $p_{a^\star} = 1$, every candidate has the correct answer and both policies are correct almost surely. These conclusions do not require any restriction on the scores. Henceforth assume $0 < p_{a^\star} < 1$; labels with $p_a = 0$ have zero counts almost surely and can be omitted.

Grouped bonuses and original-order priority. Fix a count vector c of positive probability. Conditional on $C = c$, the answer sequence is uniform over the arrangements of this multiset: every arrangement has the same unconditioned answer probability $\prod_a p_a^{c_a}$. One can construct the conditional law by drawing c_a independent bonuses from G_a for each label a, independently across labels, and independently drawing a uniform arrangement of the answer multiset. Place each label's bonuses into its successive occurrences. Given an answer sequence, the original iid law also has independent bonuses with these conditional laws, so the construction has exactly the required conditional distribution.

Consequently the grouped sums S_a are independent across labels and independent of the answer arrangement. Let $T = T(c)$, $k = |T|$, and $r = \max_a c_a \geq 1$. Every label in T has the same count r. Permuting the names of labels within T preserves the uniform multiset-arrangement law, so their relative first-occurrence order is uniform over its $k!$ permutations. This order, denoted Π, is independent of the grouped bonus sums. In particular, conditional on any collection of sums, first occurrence chooses uniformly within each nonempty subset of T that those sums make tied maximizers. The independent-uniform tie convention has the same conditional selection probabilities. For the proof it can therefore also be represented by an independent uniform priority Π; this representation concerns its decision law and does not require the two original rules to share their auxiliary randomness.

If $a^\star \notin T$, both policies are wrong, proving (34) for this count vector. If $a^\star \in T$ and $k = 1$, both are correct. Assume below that $a^\star \in T$ and $k \geq 2$. Ordinary plurality then has conditional correctness probability $1/k$.

Increasing each rival to the correct conditional law. For a positive-mass label, let its increasing generalized bonus quantile be

$$Q_a(u) = \inf\{t \in [0, B] : G_a(t) \geq u\}, \qquad 0 < u < 1.$$

If $B = 0$, this quantile is identically zero. More generally, (33) implies $Q_a(u) \leq Q_{a^\star}(u)$ for every such u and every wrong positive-mass label a. No continuity or density is required.

For each rival $a \in T \setminus \{a^\star\}$, draw independent $U_{a,1}, \ldots, U_{a,r}$ uniformly from $(0, 1)$, independently across rivals. Define

$$Y_a = \sum_{j=1}^{r} Q_a(U_{a,j}), \qquad Z_a = \sum_{j=1}^{r} Q_{a^\star}(U_{a,j}).$$

Then $Y_a \leq Z_a$ almost surely. The Y_a have precisely the independent conditional rival-sum laws; those laws need not be identical. The Z_a are independent copies of the correct r-draw sum law. Independently draw X from that correct-sum law and draw the priority Π independently of all these variables. Thus $(X, (Y_a)_{a \neq a^\star}, \Pi)$ reproduces the conditional distribution of all quantities relevant to the actual tie-refinement decision, while X and all the Z_a are iid.

For a fixed priority Π, let $f_\Pi(x, (y_a))$ be the indicator that the correct label wins when its sum is x, the rival sums are (y_a), and Π resolves tied largest sums. Increasing any rival sum cannot change this indicator from zero to one. This monotonicity includes exact ties and arbitrary atoms. Therefore

$$f_\Pi(X, (Y_a)) \geq f_\Pi(X, (Z_a)) \quad \text{almost surely.}$$

The right-hand comparison consists of k iid sums and an independent uniform priority, so symmetry gives each label winning probability $1/k$, including when the largest sums tie. Taking expectations proves

$$\mathbb{P}(Q_N = a^\star \mid C = c) = \mathbb{E} f_\Pi(X, (Y_a)) \geq \mathbb{E} f_\Pi(X, (Z_a)) = \frac{1}{k}.$$

Together with the preceding cases this proves the conditional claim; averaging over count vectors proves the unconditional one. If all bonuses are zero, or more generally constant, each top-count label has the same bonus sum and the conclusion holds with equality.

Representation by sufficiently weak positive weights. For $\eta > 0$ and $\eta N B < 1$, the weighted total is $W_a = C_a + \eta S_a$. Whenever $C_a > C_d$,

$$W_a - W_d \geq C_a - C_d - \eta C_d B \geq 1 - \eta N B > 0.$$

Hence only maximum-count labels can maximize W_a. Within this set the common count cancels, and positive η makes maximizing W_a equivalent to maximizing S_a. The maximizing sets therefore coincide, as do first-occurrence decisions or uniform-tie decision laws. This argument also covers $B = 0$. At $\eta = 0$ the weights instead give ordinary plurality; equivalence with a nontrivial bonus refinement is not asserted at zero intensity. □

Scope and relation to the calibrated specialization. The assumption compares the full conditional bonus distributions, not only their means. It allows each wrong label to have a different distribution and does not require calibration, a monotone score map, or a correct population mode. It is

sufficient, not claimed necessary. Under the calibrated score-independent-allocation assumptions of Proposition 9, every nondecreasing bonus satisfies this stochastic order, as the conditional-law and covariance argument in that proposition's retained proof shows. Its $\eta = 0$ case is ordinary plurality. That proposition and its full proof remain a self-contained sufficient specialization.

The guarantee assumes the same fixed eligible iid law and fixed bonus throughout. A fit using independent training data may be conditioned on only if the stated conditional-order premise holds for the realized bonus on the evaluated law. Marginal score diagnostics, pooled calibration, and observed accuracy differences do not establish that premise. Arbitrary fixed label priorities and assigning priority after sorting candidates by score are outside the tie conventions proved here. The budget counts eligible draws; empty-pool fallback behavior is outside the $N \geq 1$ eligible-stream statement.

Opportunity and error-exponent limitations. Only ties for the largest count can change a decision. The conditional gain is zero when the correct label is not among those tied maxima or when the maximum is unique; otherwise it is at most $1 - 1/|T|$. Under the dominance premise,

$$0 \leq \mathbb{P}(Q_N = a^\star) - \mathbb{P}(P_N = a^\star) \leq \mathbb{E}\left[\mathbf{1}\{a^\star \in T(C),\ |T(C)| \geq 2\}\left(1 - \frac{1}{|T(C)|}\right)\right].$$

Strict gain is not guaranteed; constant bonuses and tie-free count vectors give equality. With a binary alphabet and odd N, the two policies coincide for every sample.

Even without the dominance premise, any policy that always selects an answer attaining the largest count has the same error exponent as ordinary plurality when the correct population mode is unique. In particular, suppose

$$p = \mathbb{P}(A = a^\star) > q = \max_{a \neq a^\star} \mathbb{P}(A = a) > 0.$$

At every N, the lexicographic rule's error event is sandwiched between

$$\{\max_{a \neq a^\star} C_a > C_{a^\star}\} \ \subseteq\ \{Q_N \neq a^\star\} \ \subseteq\ \{\max_{a \neq a^\star} C_a \geq C_{a^\star}\}.$$

For each positive-mass rival, the strict and weak pairwise count tails have the same Chernoff exponent. Taking the finite union over rivals and applying the argument in Section B.10 gives

$$\lim_{N \to \infty} -\frac{1}{N} \log \mathbb{P}(Q_N \neq a^\star) = -\log\bigl(1 - p - q + 2\sqrt{pq}\bigr),$$

the ordinary-plurality exponent. If $q = 0$, every draw is correct and both error probabilities are zero. Thus this finite-budget sufficient guarantee does not replace the exact criterion for strict improvement of the error exponent in Theorem 1. A positive fixed intensity satisfying that theorem may override count winners; the lexicographic rule never does. Its representation by $1 + \eta b$ permits an N-dependent intensity. This is a different quantifier from the fixed-transform characterization in Theorem 3.

B.13 A RANK-INTEGRAL BOUND FOR IMPERFECT CONDITIONAL BONUS ORDER

Corollary 4 (A rank-integral bound under imperfect bonus order)**.** *Use the fixed eligible iid law, bounded measurable bonus and count-first policies P_N, Q_N of Proposition 1, without assuming conditional stochastic dominance. Both policies use first occurrence in the original iid order, or independent uniform choices among their remaining maximizers. Suppose first that $0 < p_\star := \mathbb{P}(A = a^\star) < 1$. For each positive-mass label let $G_a(t) = \mathbb{P}(b(V) \leq t \mid A = a)$, and let $F_{a,r}$ be the CDF of a sum of $r \geq 1$ independent draws from this conditional bonus law. For every positive-mass wrong label define*

$$\delta_a = \sup_t [G_{a^\star}(t) - G_a(t)]_+,$$
$$d_{a,r} = \sup_t [F_{a^\star,r}(t) - F_{a,r}(t)]_+,$$
$$e_{a,r} = 1 - (1 - \delta_a)^r.$$

For a positive-probability count vector c, put $T(c) = \arg\max_a c_a$, $k(c) = |T(c)|$, $r(c) = \max_a c_a$ and

$$L_c(\boldsymbol{z}) = \int_0^1 \prod_{a \in T(c) \setminus \{a^\star\}} (u - z_a)_+ \, du. \tag{35}$$

An empty product is one. Then $d_{a,r} \le e_{a,r}$, and if $a^\star \in T(c)$, with $r = r(c)$,

$$\mathbb{P}(Q_N = a^\star \mid C = c) \ge L_c(\boldsymbol{d}_r) \ge L_c(\boldsymbol{e}_r). \tag{36}$$

In particular, if $\delta \in [0,1]$ and $\delta_a \le \delta$ for all rivals in this tie, then

$$\mathbb{P}(Q_N = a^\star \mid C = c) \ge \frac{(1-\delta)^{r(c)k(c)}}{k(c)}. \tag{37}$$

If $a^\star \notin T(c)$, both policies are wrong; a unique count winner is unchanged. Consequently,

$$\begin{aligned} &\mathbb{P}(P_N = a^\star) - \mathbb{P}(Q_N = a^\star) \\ &\quad \le \mathbb{E}\left[\mathbf{1}\{a^\star \in T(C),\ k(C) \ge 2\}\left\{\frac{1}{k(C)} - L_C(\boldsymbol{e}_{r(C)})\right\}\right]. \end{aligned} \tag{38}$$

The loss bound inside the indicated event lies in $[0, 1/k(C)]$. If $p_\star$ is zero or one, the policies have equal correctness probabilities without defining an absent conditional law, and the regret bound is zero.

At zero deficits this recovers the exact conditional nondegradation guarantee. For common single-draw deficit $\delta < 1$, the conditional lower bound remains positive. It controls possible loss; it does not assert a positive gain or certify any population deficit from data.

Proof of Corollary 4. If $p_\star = 0$, the correct label never appears and cannot attain the positive maximum count for $N \ge 1$, so both accuracies are zero. If $p_\star = 1$, all answers are correct and both accuracies are one. Neither case requires constant scores. Assume henceforth that $0 < p_\star < 1$ and omit zero-mass wrong labels, which have zero counts almost surely.

The conditional comparison and its tie priorities. Fix a count vector c of positive probability. Conditional on $C = c$, the grouped bonuses can be drawn independently from their label-specific laws, followed by an independent uniform arrangement of the answer multiset. This is the conditional iid construction in Section B.12. The top-count labels all have $r = r(c)$ draws, so their relative first-occurrence order is uniform and independent of all grouped sums. If the correct label is outside $T = T(c)$, both policies are wrong. If it is the unique count maximizer, both are correct and the empty-product integral is one.

Consider $a^\star \in T$ with $k = |T| \ge 2$. Let X be its bonus sum and Y_a the sum of rival $a \in T \setminus \{a^\star\}$. Write $F = F_{a^\star,r}$ and $F_a = F_{a,r}$. These sums are mutually independent. Represent their uniform relative priority by iid $U_\star, (U_a)_a \sim \mathrm{Unif}(0,1)$, independent of all sums, with larger priority winning a tied sum. This has the same winning-label distribution as original first occurrence or independent uniform maximizer ties. It does not assert that literal first-occurrence positions are independent uniforms.

Given $(X, U_\star) = (x, u)$, the correct label defeats rival a with probability

$$q_a(x,u) = F_a(x-) + u\{F_a(x) - F_a(x-)\} = (1-u)F_a(x-) + uF_a(x).$$

The rival events are conditionally independent, because their pairs (Y_a, U_a) are independent. The exact conditional win probability is therefore $\prod_a q_a(x,u)$. The deficit inequality gives $F_a(x) \ge F(x) - d_{a,r}$ and, by taking increasing limits, $F_a(x-) \ge F(x-) - d_{a,r}$. Since probabilities are nonnegative,

$$q_a(x,u) \ge \left[F(x-) + u\{F(x) - F(x-)\} - d_{a,r}\right]_+.$$

Multiplication preserves these inequalities because every factor is nonnegative.

Randomized ranks handle score-sum atoms. The distributional transform

$$Z = F(X-) + U_\star\{F(X) - F(X-)\}$$

is uniform on $(0,1)$, including for discrete or mixed F (Rüschendorf, 2009). For completeness, fix $z \in (0,1)$ and let x be its generalized F-quantile, so $F(x-) \le z \le F(x)$. An atom of size $F(x) - F(x-) > 0$ distributes its rank uniformly over the interval from $F(x-)$ to $F(x)$, giving

$$\mathbb{P}(Z \le z) = F(x-) + \{F(x) - F(x-)\}\frac{z - F(x-)}{F(x) - F(x-)} = z.$$

If there is no jump at this quantile, then $F(x-) = F(x) = z$ and $\mathbb{P}(Z \le z) = \mathbb{P}(X \le x) = z$; flat intervals of the CDF carry no probability mass. This also covers a point-mass F, for which $Z = U_\star$ almost surely. It is the same independent $U_\star$ used to resolve the correct label's score-sum ties; no additional tie randomization or independence between Z and X is asserted. Integrating the product bound yields

$$\mathbb{P}(Q_N = a^\star \mid C = c) \ge \mathbb{E} \prod_{a \in T \setminus \{a^\star\}} (Z - d_{a,r})_+ = L_c(\boldsymbol{d}_r).$$

This argument uses no coupling-success event and remains valid with arbitrary score and bonus atoms.

From single-draw deficits to sum deficits. We use a standard stochastic-order coupling argument (Shaked & Shanthikumar, 2007), with an explicit construction here. Let $Q_a, Q_\star$ be increasing generalized quantiles of the single-draw conditional bonus laws. The definition of δ_a implies

$$Q_a(u) \le Q_\star(u + \delta_a) \quad (0 < u < 1 - \delta_a).$$

Indeed, if $x = Q_\star(u + \delta_a) < Q_a(u)$, right continuity gives $G_{a^\star}(x) \ge u + \delta_a$ and $G_a(x) < u$, contradicting $G_{a^\star}(x) \le G_a(x) + \delta_a$.

For $U \sim \mathrm{Unif}(0,1)$, define $R_\delta(U) = (U + \delta) \bmod 1$, assigning the arbitrary value $1/2$ when $U = 1 - \delta \in (0,1)$. This null-point convention avoids evaluating a quantile at zero. The rotated variable is again uniform; at $\delta = 0$ or 1 the map is the identity almost surely. Couple a rival bonus and a correct-law bonus as $Q_a(U)$ and $Q_\star(R_{\delta_a}(U))$. On $U < 1 - \delta_a$, the rival bonus is no larger. Taking r independent copies, let Y and X' be their respective sums. They have CDFs $F_{a,r}$ and $F_{a^\star,r}$, and

$$\mathbb{P}(Y \le X') \ge (1 - \delta_a)^r = 1 - e_{a,r}.$$

For every t, the event $\{X' \le t,\ Y \le X'\}$ is contained in $\{Y \le t\}$. Hence

$$F_{a^\star,r}(t) - F_{a,r}(t) \le \mathbb{P}(Y > X') \le e_{a,r},$$

proving $d_{a,r} \le e_{a,r}$. The argument includes deficits zero and one. It does not assume that a coupling-success event is independent of a winning event.

The integral L_c is nonincreasing in each deficit, so $L_c(\boldsymbol{d}_r) \ge L_c(\boldsymbol{e}_r)$. If all $\delta_a \le \delta$, then $e_{a,r} \le e := 1 - (1 - \delta)^r$, giving

$$L_c(\boldsymbol{e}_r) \ge \int_e^1 (u - e)^{k-1}\, du = \frac{(1-e)^k}{k} = \frac{(1-\delta)^{rk}}{k}.$$

This proves the conditional bounds. The unique-count case already gives equality one and also satisfies the displayed common-deficit lower bound.

Averaging the conditional regret. For a correct-containing top tie, ordinary plurality has conditional correctness $1/k$. Every deficit lies in $[0,1]$, so $0 \le L_c(\boldsymbol{e}_r) \le \int_0^1 u^{k-1}\, du = 1/k$. Its conditional accuracy minus the refined accuracy is consequently at most $1/k - L_c(\boldsymbol{e}_r)$. It is zero when the maximum count is unique or the correct label is outside the maximizing set. Averaging proves (38). □

Relation to the subtraction bound. The integral uniformly dominates the weaker subtraction guarantee. For a correct-containing tie, abbreviate $e_a = e_{a,r(c)}$ and put

$$\rho(c) = 1 - \prod_{a \in T(c) \setminus \{a^\star\}} (1 - e_a) = 1 - \prod_{a \in T(c) \setminus \{a^\star\}} (1 - \delta_a)^{r(c)}. \tag{39}$$

For $m = k - 1 \ge 1$, $\varphi(u) = u^m - \prod_a (u - e_a)_+$ is nondecreasing on $[0,1]$. Below $\max_a e_a$ the product is zero; above it,

$$\varphi'(u) = m u^{m-1} - \sum_a \prod_{b \ne a} (u - e_b) \ge 0.$$

Continuity handles the boundary, since each summand is at most u^{m-1}. Thus $\varphi(u) \le \varphi(1) = \rho(c)$ and

$$L_c(\boldsymbol{e}_r) = \frac{1}{k} - \int_0^1 \varphi(u)\, du \ge \max\{0, 1/k - \rho(c)\}.$$

This comparison is not a second assumption or a separate policy. For a hypothetical three-way tie with $r = 1$ and all $\delta_a = 0.2$, the new bound is $0.8^3/3 = 64/375 \approx 0.1707$, while the subtraction lower bound is zero. These are arithmetic values under hypothetical population deficits, not empirical estimates or evidence of an accuracy gain.

Conditional attainment for one draw per tied label. For $r = 1$, the integral can be attained for every prescribed deficit vector $(d_a)_a \in [0,1]^{k-1}$ at this general ordering-deficit scope. Let the correct bonus be uniform on $[0, 1/2]$. Independently give rival a density 2 on $[d_a/2, 1/2]$ and an atom of mass d_a at 1. Its density has total mass $1 - d_a$, so this is a probability law, including when $d_a = 0$ or 1. On $0 \le t \le 1/2$, the correct CDF is $F(t) = 2t$ and the rival CDF is $F_a(t) = (2t - d_a)_+$; on $1/2 < t < 1$, they equal 1 and $1 - d_a$. Consequently $\sup_t [F(t) - F_a(t)]_+ = d_a$. The correct bonus is continuous and cannot equal a rival's atom at 1, so score ties have probability zero. Independence gives

$$\mathbb{P}(\text{correct sum wins}) = \mathbb{E}\prod_a F_a(X) = \int_0^1 \prod_a (u - d_a)_+ \, du.$$

This comparison is realizable within the eligible iid model: take exactly k labels with any positive answer probabilities p_a, set $b(v) = v$ with the above label-conditional score laws, and use $N = k$. The all-singleton count vector has probability $k! \prod_a p_a > 0$ and has exactly the independent conditional laws just constructed. Thus the claim concerns sharpness conditional on that count vector for $r = 1$, without a calibration premise. It does not establish sharpness under the $r > 1$ iid-convolution restriction or for unconditional regret.

Interpretation and limits. The deficits concern the true conditional bonus laws under one fixed eligible iid law and a fixed bonus. The corollary estimates none of them. A statistical certificate would require separate control of conditional sample sizes, labels, thresholds and any choice of bonus; rare answers can make such estimation uninformative. The stochastic-order and distributional-transform tools are classical. The result controls possible finite-budget harm of the existing tie-refinement rule, not a positive benefit, a new algorithm, or an improvement of its plurality error exponent. The exact criterion for fixed-intensity exponent improvement remains Theorem 1.

A calibrated sufficient specialization. The following proposition obtains the conditional bonus ordering in Proposition 1 from calibration and score-independent wrong-answer allocation.

B.14 A FINITE-BUDGET GUARANTEE FOR REFINING PLURALITY TIES

The structured allocation assumption also gives a finite-budget result without a correct population-mode premise. Its scope is only count ties: confidence-sum tie refinement is an existing ensemble rule (Kokkinos & Margaritis, 2014). The proposition supplies a sufficient guarantee under the present calibrated-law assumptions; we make no claim that the tie-refinement algorithm is new.

Proposition 9 (Finite-budget nondegradation within count ties)**.** *Fix a finite answer alphabet with correct label $a^\star$, an integer $N \ge 1$, and an iid eligible stream $(A_i, V_i)_{i=1}^N$ under one fixed law, with $V \in [0,1]$. Suppose, almost surely,*

$$\mathbb{P}(A = a^\star \mid V) = V, \qquad \mathbb{P}(A = a \mid V) = \pi_a(1 - V) \quad (a \ne a^\star), \qquad \pi_a \ge 0, \quad \sum_{a \ne a^\star} \pi_a = 1.$$

Fix a measurable nondecreasing bonus $b : [0,1] \to [0, B]$, with $B < \infty$, and a deterministic $\eta \ge 0$ satisfying $\eta N B < 1$. Let P_N be ordinary plurality and $Q_{\eta,N}$ be plurality weighted by $h_\eta(v) = 1 + \eta b(v)$. Both rules resolve maximizing ties by the first occurrence in the original iid order. Alternatively, each uses a uniform choice among its maximizing labels, with independent auxiliary randomness. Then, for every answer-count vector c of positive probability,

$$\mathbb{P}(Q_{\eta,N} = a^\star \mid C = c) \ge \mathbb{P}(P_N = a^\star \mid C = c), \tag{40}$$

and hence the same inequality holds unconditionally. Here $C_a = \sum_{i=1}^N \mathbf{1}\{A_i = a\}$. No assumption that $a^\star$ is a population mode is required.

Proof. Write $S_a = \sum_{i:A_i=a} b(V_i)$ and $W_a = C_a + \eta S_a$. If $C_a > C_d$, nonnegativity and boundedness give

$$W_a - W_d \geq C_a - C_d - \eta C_d B \geq 1 - \eta N B > 0.$$

Thus only labels in $T = \arg\max_a C_a$ can win the weighted vote. If $\eta = 0$ or $B = 0$, the two rules have identical correctness probabilities. Otherwise the weighted rule maximizes S_a within T, then applies its stated tie convention.

Let $p = \mathbb{P}(A = a^\star) = \mathbb{E}V$. If $p = 0$, calibration implies $V = 0$ almost surely and the correct label never occurs; both accuracies are zero. If $p = 1$, $V = 1$ almost surely and all draws are correct; both accuracies are one. Assume $0 < p < 1$, and let F be the marginal score law. The conditional score laws for the correct label and any positive-mass wrong label are, respectively,

$$F_1(dv) = \frac{vF(dv)}{p}, \qquad F_0(dv) = \frac{(1-v)F(dv)}{1-p}.$$

The factor π_a cancels in the second expression. Zero-mass labels have zero counts almost surely and can be omitted. For every bounded nondecreasing ϕ,

$$\mathbb{E}_{F_1}\phi(V) - \mathbb{E}_{F_0}\phi(V) = \frac{\mathrm{Cov}_F(V, \phi(V))}{p(1-p)} \geq 0,$$

because, for an independent copy V', $2\,\mathrm{Cov}_F(V, \phi(V)) = \mathbb{E}[(V - V')(\phi(V) - \phi(V'))] \geq 0$. Consequently the correct conditional bonus law G_1 stochastically dominates the common wrong conditional bonus law G_0. Their quantiles coupled through one uniform variable are ordered almost surely; independent copies of this coupling show that the sum of r G_1 bonuses dominates the sum of r G_0 bonuses for every $r \geq 1$. This argument requires no densities and allows arbitrary score or bonus atoms.

Condition on $C = c$. An exact construction of the conditional iid law is to draw, independently for each label, its c_a scores from its conditional score law, draw a uniform arrangement of the answer multiset independently of those scores, and place each label's scores in its successive occurrences. Indeed, all arrangements have the same product of answer probabilities, and, given an arrangement, scores are independent with the stated conditional laws. Hence grouped bonus sums are independent across labels and independent of the arrangement. All $k = |T|$ top-count labels have a common positive count r. Their relative first-occurrence order is uniform over its $k!$ permutations by relabeling symmetry, and is independent of the sums. Thus, conditional on the sums, first occurrence selects uniformly from any tied maximizing subset of T. The alternative uniform tie convention has this same property directly.

If $a^\star \notin T$, both conditional accuracies are zero. If $a^\star \in T$ and $k = 1$, both are one. Otherwise let X be its bonus sum, and let $Y_1, \ldots, Y_{k-1}$ be the rival top-count sums. The Y_j are iid sums of r wrong-law bonuses, independent of X, and X stochastically dominates their common law. For fixed rival sums y, the conditional probability of a correct weighted win is

$$f(x; y) = \begin{cases} 0, & x < \max_j y_j, \\ (1 + \#\{j : y_j = x\})^{-1}, & x = \max_j y_j, \\ 1, & x > \max_j y_j. \end{cases}$$

This is nondecreasing in x, including at atoms. Replacing X by an independent X_0 with the common wrong-sum law therefore gives

$$\mathbb{E}f(X; Y) \geq \mathbb{E}f(X_0; Y) = 1/k.$$

The equality follows from symmetry of k iid sums and uniform resolution of their maximizing ties. Ordinary plurality also has conditional correctness $1/k$. This proves (40); averaging over count vectors proves the unconditional assertion. □

Scope of the guarantee. Every positive η satisfying $\eta N B < 1$ gives the same rule: maximize count, then bonus sum, then apply the stated tie convention (with the same auxiliary random draw when comparing randomized rules). It never changes a unique plurality winner, so any accuracy advantage is bounded by the probability of a top-count tie. Strict gain is not guaranteed: constant bonuses give equality, as does a binary alphabet with odd N. The tie rule may instead be defined lexicographically

at every budget; its numerical weight representation then uses a budget-dependent intensity. This is not a finite-budget guarantee for the usual half-floor intensity $\eta = 2$ with $b(v) = (v - 1/2)_+$.

This procedure cannot improve the error exponent either. To see this, suppose the correct population mode is unique, with $p > q = \max_{a \neq a^\star} \mathbb{P}(A = a) > 0$, and apply the lexicographic rule at each N. Its error event lies between

$$\{\max_{a \neq a^\star} C_a > C_{a^\star}\} \quad \text{and} \quad \{\max_{a \neq a^\star} C_a \geq C_{a^\star}\}.$$

The strict and weak pairwise count events have the same Chernoff rate, and finite unions preserve the minimum rate, as shown in Section B.10. Thus its exponent equals ordinary plurality's, $-\log(1 - p - q + 2\sqrt{pq})$, in contrast to the fixed-intensity improvement in Theorem 1. If $q = 0$, every draw is correct and both error probabilities are zero.

Arbitrary fixed label priorities are outside the proposition, as is assigning first-occurrence priority after sorting by score. The assumptions concern the same fixed eligible iid law; neither pooled calibration nor the empirical gain decompositions establish score-independent allocation or the required per-problem calibration. The statement is a structured sufficient result, not a guarantee for unrestricted calibrated weighting.

B.15 Observable-law identification and finite-sample selectors

Scope of the observable-law selector. Given the full answer–score law, conditional calibration error at most ϵ and $c = \mathbb{E}(2V - 1)_+ > 2\epsilon$, plurality among scores above one half identifies the correct label with exponentially decreasing error. This selector uses zero weights and need not dominate finite plurality (Proposition 10 and Section B.15).

Identification from all observables. With the full law of (A, V), the unknown correct label can instead be restricted by the calibration premise. Defining its compatible set is immediate; the substantive question is when that set contains one label and a finite-sample selector can recover it.

Proposition 10 (Calibration-compatible answers)**.** *Let $d_a = \mathbb{E}|P(A = a \mid V) - V|$ and $c = \mathbb{E}(2V - 1)_+$. Suppose the only restriction on the unknown correct label is the per-problem error bound $d_{a^\star} \leq \epsilon$, without a base-mode assumption. Then the identified set for the correct answer is exactly $\mathcal{I}_\epsilon = \{a : d_a \leq \epsilon\}$. If $c > 2\epsilon$, this set is a singleton, and plurality among the N iid candidates satisfying $V > 1/2$ has error at most $\exp[-N(c - 2\epsilon)^2/2]$, with arbitrary ties.*

For distinct labels, $d_a + d_b \geq c$; moreover the correct label has a strict majority of high-score mass, so Hoeffding's inequality (Hoeffding, 1963) gives the error bound (proof below). This elementary selector can recover a nonmodal correct answer; it uses zero weights below the threshold and does not uniformly dominate finite plurality. Exact calibration also implies $p_{a^\star} = \mathbb{E}V$, motivating frequency matching when no rival shares that marginal. These observations scope our impossibility to fixed positive weighting rules. All guarantees assume a calibration-error bound; they do not estimate it from pooled data.

Proof of identification from all observables. The marginal bounds retain less information than the observable law of (A, V). The following elementary complement makes the distinction precise. It also gives a selector outside the fixed-positive-weight restriction.

Proposition 11 (Observable identification and frequency matching)**.** *Fix a finite answer alphabet of size $K \geq 2$. For the observable law P, let $r_a(v) = P(A = a \mid V = v)$, $d_a = \mathbb{E}|r_a(V) - V|$, and $\mu = \mathbb{E}V$. Suppose the only restriction on the unknown correct label is $d_{a^\star} \leq \epsilon$. Its identified set is exactly*

$$\mathcal{I}_\epsilon(P) = \{a : d_a \leq \epsilon\}.$$

For distinct a, b, $d_a + d_b \geq c := \mathbb{E}(2V - 1)_+$. Thus $c > 2\epsilon$ guarantees uniqueness whenever the model is nonempty.

For N iid observations, define

$$\widehat{m}_a = \left| \frac{1}{N} \sum_{i=1}^{N} (\mathbf{1}\{A_i = a\} - V_i) \right|, \qquad m_a = |p_a - \mu|.$$

If $\gamma = \min_{a \neq a^\star} m_a - m_{a^\star} > 0$*, then any minimizer* $\widehat{a}$ *of* $\widehat{m}_a$ *satisfies*

$$P(\widehat{a} \neq a^\star) \leq \min\{1, 2K \exp(-N\gamma^2/8)\}.$$

For $\delta \in (0,1)$*, the set*

$$\widehat{\mathcal{C}}_\delta = \left\{ a : \widehat{m}_a \leq \epsilon + \sqrt{\frac{2\log(2/\delta)}{N}} \right\}$$

contains $a^\star$ *with probability at least* $1-\delta$*. Consequently, outputting a label only when this set is a singleton has probability at most* δ *of issuing a wrong certificate.*

Proof. Any label in $\mathcal{I}_\epsilon$ can be designated correct by defining $R = \mathbf{1}\{A = a\}$, without changing the observable law; any label outside violates the assumed error bound. Since $r_a + r_b \leq 1$, $|r_a - v| + |r_b - v| \geq (2v-1)_+$ pointwise. Integrate for the separation. It is sharp: $V = 1/2 + \epsilon$ and equiprobable labels a, b have $d_a = d_b = \epsilon$, for $0 \leq \epsilon \leq 1/2$.

For each label, $X_i(a) = \mathbf{1}\{A_i = a\} - V_i \in [-1, 1]$. Hoeffding's inequality (Hoeffding, 1963) and a union bound give

$$P\left(\max_a |\overline{X}(a) - \mathbb{E}X(a)| \geq t\right) \leq 2K \exp(-Nt^2/2).$$

The absolute-value map is 1-Lipschitz, so errors below $\gamma/2$ make the correct empirical residual uniquely smallest. For coverage of $\widehat{\mathcal{C}}_\delta$, apply the one-label bound to $a^\star$ and use $m_{a^\star} \leq d_{a^\star} \leq \epsilon$. This last argument requires no union over labels. □

Scope and abstention. The result assumes a known exhaustive answer alphabet; unobserved possible labels cannot be discarded when claiming a singleton. The certificate controls the unconditional probability of a wrong output and may abstain frequently; it does not assert conditional accuracy among issued outputs. Its assumed ϵ is a per-problem conditional error bound, not a quantity estimated by this procedure. Exact calibration gives $m_{a^\star} = 0$, but frequency matching can fail when a wrong label shares the same marginal, even if the full-law identified set is a singleton. For example, let $V = .2, .6$ equiprobably and let the conditional answer laws be $(.2, .4, .4)$ and $(.6, .4, 0)$. The first label is uniquely calibrated, but its marginal $.4$ equals the second label's and $\mathbb{E}V$.

A selector under the high-score separation condition. The full-law sufficient condition also yields a simple finite-sample rule. Let $\widehat{a}_+$ maximize $\sum_i \mathbf{1}\{V_i > 1/2, A_i = a\}$, with arbitrary ties. If $g = c - 2\epsilon > 0$, then

$$P(\widehat{a}_+ \neq a^\star) \leq \exp(-Ng^2/2).$$

Indeed, for each wrong a,

$$\begin{aligned}&\mathbb{E}[\mathbf{1}\{V > 1/2\}(\mathbf{1}\{A = a^\star\} - \mathbf{1}\{A = a\})] \\ &\quad \geq \mathbb{E}[\mathbf{1}\{V > 1/2\}(2r_{a^\star}(V) - 1)] \geq c - 2\epsilon.\end{aligned}$$

More sharply, $Z_i = \mathbf{1}\{V_i > 1/2\}(\mathbf{1}\{A_i = a^\star\} - 1/2)$ lies in $[-1/2, 1/2]$ and has mean at least $g/2$. Whenever $\sum_i Z_i > 0$, the correct answer is a strict majority among the retained samples. One-sided Hoeffding gives the displayed bound without an alphabet-size factor. This rule does not require a correct base mode, allows zero weights below the threshold, and has no asserted finite-sample dominance over plurality.

Calibration across problems is insufficient. Consider two equally likely problems with $V = 2/3$ always. On the first, a is correct and generated answers are a with probability $1/3$ and b with probability $2/3$. On the second, b is correct and always generated. Pooled correctness is $2/3$, perfectly matching the score. On the first problem frequency matching nevertheless converges to wrong b: its per-problem calibration premise fails. Nor does a finite-bin reliability curve (DeGroot & Fienberg, 1983) automatically certify the conditional L^1 error. If $V = 1/2 - d, 1/2 + d$ equiprobably and conditional correctness is $1, 0$, respectively, the single-bin residual is zero while that error is $1/2 + d$.

C IDENTIFICATION CONSTRUCTIONS AND SHARP BOUNDS

What evaluation summaries identify. A labelled report retaining only the correctness–score law and answer frequencies can leave the weighted winner undetermined. Sharp bounds on the largest compatible wrong-answer mass characterize possible and necessary correct modes; full labelled records determine the vote directly (Section C.1 and Figure 6).

Retained evaluation summaries. Weighted-vote outcomes compatible with retained marginals form a partial-identification problem (Manski, 2003). Our two-score bounds use marginal intersections; the arbitrary-score result combines rearrangement and simultaneous allocation across rivals. These bounds quantify missing information in labelled summaries.

C.1 WHAT BINARY VERIFIER DIAGNOSTICS IDENTIFY

Fix one problem with one correct answer $a^\star$, let $R = \mathbf{1}\{A = a^\star\}$, and make every trace boxed. The following population statements distinguish knowing the score–correctness law from knowing which wrong answer receives each score.

Proposition 12 (Non-identification from binary diagnostics and answer marginals)**.** *Two laws can have identical joint distributions of (R, V) and identical marginal distributions of A, yet different correctness of the unique answer mode under $P_1 \propto VP$.*

Proof. Let a_1, a_2 be wrong. The entries below are joint probability masses in units of $1/200$ for a rank-inverted pair of laws, R_A and R_B (distinct from the calibrated Populations A and B of Table 1):

Law	V	$a^\star$	a_1	a_2
R_A	$1/6$	81	99	0
R_A	1	5	0	15
R_B	$1/6$	81	84	15
R_B	1	5	15	0

Both laws have answer marginal $(.430, .495, .075)$, score probabilities $(.9, .1)$, and correctness posteriors $(.45, .25)$ at the two scores. Thus their entire (R, V) laws agree and are strictly rank-inverted. At exponent one the common normalizer is $Z = .25$, giving

$$\nu_1^{R_A} = (.370, .330, .300), \qquad \nu_1^{R_B} = (.370, .580, .050).$$

R_B never has a correct mode under any V^β tilt: the unnormalized a_1-minus-$a^\star$ mass is $.015(1/6)^\beta + .05 > 0$. Nevertheless, R_A and R_B have the same single-draw correctness under every such tilt, since their (R, V) laws agree. □

A perfectly calibrated example. Non-identification does not require inversion or miscalibration. Consider Populations A and B of Table 1, with joint masses again in units of $1/200$ and coordinates $(a^\star, a_1, a_2)$:

Population	$V = 1/5$	$V = 2/5$
A	$(12, 24, 24)$	$(56, 42, 42)$
B	$(12, 1, 47)$	$(56, 65, 19)$

Both have answer marginal $(.34, .33, .33)$, score probabilities $(.30, .70)$, and exact calibration $\mathbb{E}[R \mid V] = V$. A score-only tilt preserves this conditional equality. With $r = 2^\beta$, single-draw correctness is $.2 + .2\,[7r/(3 + 7r)]$, strictly increasing in β. At exponent one, $Z = .34$ and

$$\nu_1^{\mathrm{A}} = (124, 108, 108)/340, \qquad \nu_1^{\mathrm{B}} = (124, 131, 85)/340.$$

After removing the common positive factor $(1/5)^\beta/200$, the correct-minus-wrong mass differences are $(-12 + 14r, -12 + 14r)$ in A and $(11 - 9r, -35 + 37r)$ in B. Thus A retains its correct unique mode for every $\beta \geq 0$, whereas B ties at $\beta = \log_2(11/9)$ and has a wrong unique mode above it;

(a) Same marginals, different association

	World A		World B	
	Low	High	Low	High
$a^\star$	0.060	0.280	0.060	0.280
a_1	0.120	0.210	0.005	0.325
a_2	0.120	0.210	0.235	0.095

Entries are joint probability masses.

Both: $P(A) = (.34, .33, .33)$

Calibrated: $P(A = a^\star \mid V) = V, \quad V \in \{.2, .4\}$

(b) Different modes after the same tilt

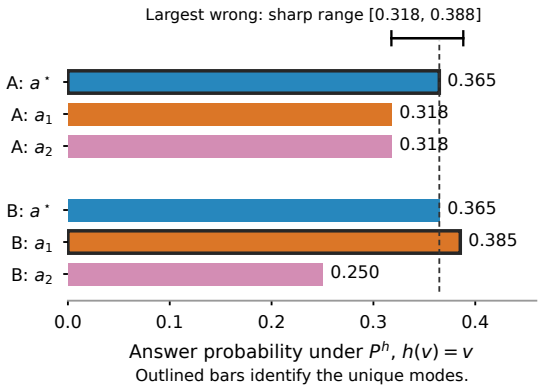


Figure 6: **Binary calibration does not determine the weighted answer mode.** The panels' World A and World B are Populations A and B of Table 1: each has one correct answer $a^\star$, two wrong answers, all candidates eligible, and scores $0.2, 0.4$ satisfying $P(A = a^\star \mid V) = V$. Their complete correctness–score and base answer laws agree. Under raw-score weighting, correct mass rises from 0.34 to $31/85 \approx 0.365$ in both; only A retains the correct mode, while B's wrong mode leads by $7/340 \approx 0.021$. The sharp largest-wrong range is $[27/85, 33/85] \approx [0.318, 0.388]$; A attains its lower endpoint, whereas B lies $1/340$ below its upper endpoint (Section C.1 and Lemma 1). This exact population example is not an empirical estimate or a finite-vote guarantee.

its exponent-one wrong-mode margin is $7/340$ (about 2.06 percentage points). Under iid base-law candidates and correctness-independent score-tie handling, score-argmax accuracy also increases with K, equaling $.4 - .2(.30)^K$. Perfect calibration is not perfect discrimination: both base-law binary AUCs, with half-credit ties, are $111/187 \approx .594$. This strengthens the same population non-identification statement; it establishes neither failure of calibrated aggregation generally nor a practical decision rule.

Consequence for beneficial weighting. Populations A and B also differ in whether any fixed bounded strictly positive score-only weight can improve plurality's error exponent (Theorem 1). Up to a positive factor, any nonconstant positive weight on the two scores is plurality plus a positive indicator bonus on the high score or on the low score. Equal weights retain plurality's error exponent. In population A the high-score bonus has conditional mean $56/68$ on the correct answer and $42/66$ on each most frequent wrong answer, so every sufficiently small positive intensity strictly improves the exponent. In population B the high-score bonus has mean $65/66$ on wrong answer 1 and the low-score bonus has mean $47/66$ on wrong answer 2, each above the correct answer's $56/68$ and $12/68$, so no positive intensity of either bonus improves the exponent. Both opening scores are below one half, so the ordinary half-floor is constant there and is not population A's improving bonus. Identical score–correctness laws and answer frequencies therefore do not determine whether there exists an exponent-improving fixed weighting; the allocation of scores among wrong answers does. The raw answer–score pairs distinguish the populations. Failure to improve the exponent does not rule out finite-budget gains.

Theorem 6 (Sharp arbitrary-score identification bounds)**.** *Work on a common eligible population with one correct answer and finitely many wrong answers of fixed base masses $p_1, \dots, p_m > 0$, where $m \geq 1$ and $W = \sum_j p_j$. Fix the complete score–correctness law and a measurable nonnegative tilt weight $w(V)$. Let $q(u)$, $0 < u < W$, be the increasing, mass-parameterized quantile of $w(V)$ among incorrect traces: its underlying measure has mass W, before weighting. Assume*

$$T = \int_0^W q(u)\, du < \infty, \qquad C = \mathbb{E}[w(V)\mathbf{1}\{A = a^\star\}] < \infty, \qquad Z = C + T > 0.$$

Define the lower and upper quantile integrals

$$H_-(s) = \int_0^s q(u)\, du, \qquad H_+(s) = \int_{W-s}^W q(u)\, du \quad (0 \leq s \leq W).$$

Sort $p_{(1)} \geq \cdots \geq p_{(m)}$ and put $P_k = \sum_{j=1}^{k} p_{(j)}$. Over all compatible population score–answer couplings, allowing fractional allocation of score atoms to answer labels, the largest wrong-answer probability ranges over exactly $[L_\star/Z, U_\star/Z]$, where

$$L_\star = \max_{1 \leq k \leq m} \frac{H_-(P_k)}{k}, \qquad U_\star = H_+(p_{(1)}). \tag{41}$$

The correct answer is the unique mode for every compatible law iff $C > U_\star$, and for at least one compatible law iff $C > L_\star$. Every compatible law has a wrong answer with mass strictly exceeding the correct mass iff $C < L_\star$. In particular, $L_\star < C < U_\star$ permits both a correct unique mode and a wrong answer strictly exceeding the correct one. Endpoint equalities require explicit tie handling.

Sharp bounds without discretizing scores. The two-score calculation extends to an arbitrary nonnegative tilt weight. The optimization uses classical rearrangement (Hardy et al., 1952, Chapter X) and minimax allocation principles, related to lexicographic optimization over base polyhedra (Fujishige, 1980; Frank & Murota, 2023); the application here is identification from verifier marginals and answer frequencies.

Proof of Theorem 6.

Proof. An allocation is represented by measurable functions $f_j : [0, W] \to [0, 1]$ with $\sum_j f_j = 1$ almost everywhere and $\int_0^W f_j = p_j$. The resulting unnormalized wrong masses are $z_j = \int_0^W q(u) f_j(u)\, du$. Such allocations represent compatible couplings: score atoms may be split between labels, while their total mass and the correct-score measure remain fixed. The allocation set is nonempty, convex and weak-star compact in a finite product of L^∞ unit balls by the Banach–Alaoglu theorem (Alaoglu, 1940) and weak-star closedness of its constraints. The constraints and these integrals are weak-star continuous since $1, q$ are integrable. Thus the set $\mathcal{Z}$ of achievable vectors z is nonempty, compact and convex.

Rearrangement bounds answer j's mass by $H_+(p_j)$. This bound is nondecreasing in p_j, and allocating the largest $p_{(1)}$ mass of weights to a largest answer attains $U_\star$.

For the lower endpoint, convex minimax (Sion, 1958) gives

$$\min_{z \in \mathcal{Z}} \max_j z_j = \max_{\lambda \in \Delta_m} \min_{z \in \mathcal{Z}} \sum_j \lambda_j z_j, \qquad \Delta_m = \{\lambda \geq 0 : \textstyle\sum_j \lambda_j = 1\}.$$

For fixed λ, order labels by $\lambda_{\pi(1)} \geq \cdots \geq \lambda_{\pi(m)}$, let $c_k = \sum_{j=1}^{k} p_{\pi(j)}$, and set $\lambda_{\pi(m+1)} = 0$. Rearrangement minimizes the weighted sum by allocating the smallest quantiles to the largest coefficients, in consecutive blocks of these capacities. Summation by parts yields

$$\min_{z \in \mathcal{Z}} \lambda^\top z = \sum_{k=1}^{m} (\lambda_{\pi(k)} - \lambda_{\pi(k+1)}) H_-(c_k).$$

The coefficients $k(\lambda_{\pi(k)} - \lambda_{\pi(k+1)})$ are nonnegative and sum to one. Hence this expression is a convex combination of $H_-(c_k)/k$, bounded above by

$$\max_{\varnothing \neq S \subseteq \{1,\ldots,m\}} \frac{H_-(\sum_{j \in S} p_j)}{|S|}.$$

Conversely, choosing λ uniform on any such S and zero elsewhere attains that subset's ratio in the inner minimization. Since $q \geq 0$, H_- is nondecreasing, so among subsets of size k the largest k capacities maximize the ratio. This proves the formula for $L_\star$. Compactness gives attainment of both extrema; continuity of $\max_j z_j$ on the connected set $\mathcal{Z}$ gives the entire interval. Division by the fixed Z and comparison with C prove the mode statements. □

Population sharpness and finite candidates. These endpoints are pointwise in the chosen tilt; an extremizing coupling need not be shared across exponents. They require correctness-labelled population marginals and do not supply a label-free decision rule or a finite-sample confidence guarantee. With two weights $a \leq b$ and wrong mass t at weight b,

$$H_-(s) = as + (b - a) \max\{0, s - (W - t)\}, \qquad H_+(s) = as + (b - a) \min\{s, t\},$$

recovering the sharp interval in Lemma 1 below.

For a fixed finite list of equally weighted, indivisible traces, the lower endpoint is a relaxation of the minimum over score permutations. For example, two wrong traces with base mass $1/4$ each, tilt weights $0.1, 0.3$, and distinct answer labels have $L_\star = 0.05$ after fractional allocation, whereas every indivisible assignment has maximum wrong mass 0.075. The upper endpoint remains attainable by assigning the largest weights to a largest answer group. An unresolved finite-pool certificate therefore does not prove ambiguity: two feasible assignments with opposite strict correctness outcomes are sufficient witnesses. The full observed answer–score pairing determines the actual weighted decision; these bounds characterize the information lost when only the stated marginals are retained.

Lemma 1 (Sharp two-score bounds at a fixed exponent)**.** *Let the finite wrong-answer alphabet have known base masses $p_1, \dots, p_m > 0$, with $m \geq 1$, $W = \sum_i p_i \leq 1$, and $P = \max_i p_i$. Let the two positive tilt weights be $0 < a \leq b$. The binary score–correctness law fixes the wrong mass $t \in [0, W]$ in the b-weighted stratum and the correct unnormalized tilted mass C. Put*

$$T = aW + (b-a)t, \qquad Z = C + T, \qquad \eta = \inf\left\{ u \geq 0 : \sum_i \min(bp_i, u) \geq T \right\}.$$

Over all compatible score–answer couplings, the largest wrong-answer probability ranges over exactly $[L/Z, U/Z]$, where

$$L = \max\{aP, \eta\}, \qquad U = aP + (b-a)\min(P, t). \tag{42}$$

The correct answer is the unique mode for every compatible law iff $C > U$, and for at least one compatible law iff $C > L$. Every compatible law has a wrong answer whose mass strictly exceeds the correct mass iff $C < L$. When $L < C < U$, both a correct unique mode and a mode strictly exceeding the correct answer are possible; equality cases require explicit tie handling.

Proof. Write x_i for wrong answer i's mass in the b-weighted stratum. Compatibility is exactly $0 \leq x_i \leq p_i$, $\sum_i x_i = t$, and the unnormalized wrong masses are $z_i = ap_i + (b-a)x_i$. Since $x_i \leq \min(p_i, t)$ and the resulting upper bound is nondecreasing in p_i, the largest possible $\max_i z_i$ is U. It is attained by assigning $\min(P, t)$ to an answer of base mass P; the remaining assignment is feasible because $t \leq W$.

For $b > a$, the equivalent constraints are $ap_i \leq z_i \leq bp_i$ and $\sum_i z_i = T$. A proposed maximum u is feasible exactly when $u \geq aP$ and $\sum_i \min(bp_i, u) \geq T$: the lower endpoints sum to $aW \leq T$, and the capped upper endpoints must accommodate T. Hence the smallest maximum is L. For $b = a$, every coupling gives $z_i = ap_i$, and the formulas give $L = U = aP$ without division by $b - a$. Finally, the continuous function $\max_i z_i$ maps the nonempty compact connected coupling polytope onto the full interval between its extrema. Comparing that interval with C proves the mode statements. □

For the rank-inverted laws R_A, R_B at exponent one, $a = 1/6$, $b = 1$, $t = .075$, $T = .1575$, and $C = .0925$. Thus $L = .0825$, $U = .145$, and $Z = .25$: R_A and R_B attain the exact endpoints $[.33, .58]$, while correct mass is .37. For Populations A and B, $a = .2$, $b = .4$, $t = .42$, $T = .216$, $C = .124$, and $Z = .34$, giving $[L/Z, U/Z] = [27/85, 33/85]$. Population A attains the lower endpoint; population B's largest wrong mass is $1/340$ below the upper endpoint. These bounds are pointwise in the exponent; a lower-extremizing coupling need not be common to all exponents. They require population marginals including correctness and are neither a label-free decision rule nor a finite-sample certificate.

D Dataset and Score Provenance

The empirical studies use Qwen2.5-Math-7B-Instruct and Qwen2.5-Math-1.5B-Instruct (Yang et al., 2024), Phi-4 (Abdin et al., 2024), and one fixed process reward model. Their cohorts serve distinct purposes. The initial 60 test problems and 60 disjoint development problems provide large candidate pools for rule comparisons, score fitting and mechanism audits. A further 140 test problems support the matching-budget analysis in Section E.1. The prospective replication uses a separate reserve of 252 problems to test the selected eight-candidate rules on new problems for this configuration (Section F). The additional generators use new draws on those same 252 problems (Section F.1). Each comparison applies its voting rules to identical candidates.

The AIME 2000–2017 problems come from the curated Hugging Face compilation qq8933/AIME_1983_2024. The artifact records its cached source digest, row filters and exclusions. The original manifest has two problem-number strata, 1–8 and 9–15, with 100 test and 30 disjoint development problems per stratum. The initial Math study uses 30 test problems per stratum and three independent 128-candidate runs per problem. The other 140 test problems have three 20-candidate runs. The mechanism studies also use three 128-candidate runs on each of the 60 development problems; their audit/evaluation splits are specified in Section F.5. These reused records are not additional independent problem samples.

Math-7B generation uses temperature .7, top-p .95 (Holtzman et al., 2020), the native chat template and a 13×256-token horizon. The problem is the user turn; the system prompt is "Please reason step by step, and put your final answer within `\boxed{}`."

Scores and exclusions. V is the minimum Qwen2.5-Math-PRM-7B (Zhang et al., 2025) step score under the fixed 4096-token suffix policy, retaining the last step even if over budget. The scorer's model card describes per-step scores; the minimum-step scalarization and the 4096-token suffix policy are implementation choices of this manuscript, not procedures established by the scorer's report, whose own evaluation multiplies step scores (Zhang et al., 2025, Section 2.2). Minimum-step scoring also appears in Math-Shepherd (Wang et al., 2024, Section 3.4), which does not identify this code or calibrate its scores. All voting results use the pinned scorer revision and verified loading procedure; scores from a defective or unverified load are excluded. Math-7B's development within-problem AUC is 0.724, with one-sided 95% lower bound 0.674. This gate tests whether scores rank correct candidates above incorrect candidates. It does not assume that scores are calibrated probabilities: the per-problem calibration checks in Section E assess that separate property. AUC (Hanley & McNeil, 1982) averages defined runs within each problem, then problems; single-class pools are omitted from this discrimination measure. The development probability map is a logistic function of clipped $\mathrm{logit}(V)$ (Platt, 2000), fitted only on the 60 disjoint development problems. Its exact parameters and the voting rules appear in Section E.

Sampling and inference. Rules share nested launch-order prefixes, parsed boxed eligibility, first-occurrence ties and empty-set fallbacks; empty outcomes remain in the denominators. Paired comparisons weight problems equally after averaging their three runs, with stratified problem-cluster resampling (percentile bootstrap intervals; Efron, 1979; Davison & Hinkley, 1997). The original inspected-cohort intervals are pointwise and descriptive; later sensitivity and transfer adjustments retain their explicitly limited families. Labelled population bounds assume iid eligible draws from fixed laws, including the same law across runs. Same-pool certificate groups are not independent validation; non-rejection does not establish calibration.

Math-7B source pools. The Math-7B source data include 60 inspected TEST problems and the remaining 140 TEST problems (70 per stratum). The latter supply three independent 20-candidate runs per problem under the same frozen generator, prompt, scorer and sampling settings; they were previously unrun with Math-7B. The Math-7B configuration was selected using a development discrimination gate: within-problem AUC 0.724, with one-sided 95% lower bound 0.674. This gate assesses ranking performance. The fixed-pool comparisons below evaluate how the scores affect voting accuracy. The accompanying artifact also contains complete guided-generation, high-score filtering, scorer-transfer and gated-policy experiments. This appendix focuses on the fixed-pool voting comparisons and their theoretical mechanisms; the additional studies retain their full protocols and outcomes in the artifact.

E SAVED-POOL EVALUATION OF SCORE-WEIGHTED VOTING

This analysis asks when replacing answer counts with score-weighted totals changes accuracy. Each rule receives the same saved candidates, and budgets use their first 8, 16, 32, 64 or 128 draws. The paired difference from plurality therefore measures the effect of the final voting rule, while the budget grid shows how that effect varies with sample size. Raw weighting and the two half-floors probe the distinction between finite accuracy and preservation of the long-run winner in Sections 4 and 5. The analysis is exploratory: earlier outcomes informed the configurations and readouts.

Cohort, scores and shared readout conventions. The exploratory comparison uses 60 test problems, three independent runs per problem and 128 candidates per run. The rule combinations were fixed before their calculation on these previously inspected pools. We take the first $N \in \{8, 16, 32, 64, 128\}$ launch-order candidates, then restrict to parsed boxed answers. Raw V is the minimum step score under the scoring policy in Section D. Every rule uses the same eligible candidates and first-occurrence tie priority. Answers within 10^{-12} of the maximum normalized answer mass are treated as tied. An empty eligible pool returns the first extractable answer, or abstains if none exists; every decision remains in the denominator. No candidate is generated or rescored for this analysis.

Development-set score maps. The common rules are plurality, reward-sum raw V, $\max(V, 1/2)$, and $\max(g(V), 1/2)$, where the newly fitted probability map is

$$g(v) = \mathrm{sigmoid}\big(a \,\mathrm{logit}(\mathrm{clip}(v, 10^{-6}, 1 - 10^{-6})) + b\big) .$$

The fixed objective is development binary cross entropy, equal-weighted over nonempty problem/run pools and uniform within each pool, plus $10^{-6}(a^2 + b^2)/2$, with $a \in [0, 20]$, $b \in [-20, 20]$. The fit converges to $(a, b) = (0.735564, -0.792496)$ using 18,536 eligible development candidates. No development pool is empty, and test labels enter no fit.

For the fitted Math-7B map, $g(v)$ is at most one half whenever v is at most one half. Thus $\max(g(v), 1/2)$, viewed as a transform of V, has the population-preserving shape in Theorem 2. This supplies a sufficient route under calibration of V, separate from applying the theorem to the half-floor of $S = g(V)$ under calibration of S. Neither calibration premise is established here.

Complete comparison and uncertainty. Table 9 reports all three weighting rules at all five budgets. Paired pointwise intervals use 10000 percentile-bootstrap replicates (Davison & Hinkley, 1997), seed 28391, resampling sorted problem IDs within frozen A/B strata and retaining all runs. They are descriptive, without multiplicity adjustment. These intervals treat the observed problem clusters as draws from comparable problems within the stated strata, retaining their fixed stratum weights and conditioning on the observed runs. They do not quantify new-generation uncertainty on the exact fixed benchmark or arbitrary task transfer. Nested prefixes are not independent confirmations. Sparse same-sign problem differences can force a percentile endpoint above zero: the Math-7B raw-score $N = 128$ gain rests on only four discordant problems. Its interval describes a post hoc contrast on the inspected problems; it does not establish improvement on other problems. The same caution applies to negative contrasts. Table 10 gives the signed problem support for every nonbaseline comparison and marks endpoints at zero in either direction. A zero-width interval reflects identical observed problem-level differences, not population equality. At $N = 128$, raw half-floor weights vary in 172/180 pools and change ten predictions, producing six gains and no losses. The DEV half-floor changes six predictions, producing three gains and no losses. The distinction between population preservation and finite-pool improvement is therefore substantive, rather than a consequence of constant weights.

Sparse paired support. This sensitivity asks whether a reported gain is spread across problems or rests on a few changed answers. We retain every signed problem-level contrast, then compare the original bootstrap intervals with conditional sign-flip and BCa calculations. The number and direction of discordant problems reveal the support behind an effect; the alternative calculations show how its uncertainty depends on the stated sampling or sign assumptions.

The sensitivity keeps runs clustered within problems and retains the original 87-contrast multiplicity correction from the complete artifact analysis, so the reported adjusted result is unchanged. Conditional sign-flip tests (Fisher, 1935; Ernst, 2004) enumerate independent fair signs for nonzero problem sums given their magnitudes; a mean-zero null alone does not justify this assumption. Math raw weighting has 14 gains on 12 problems at $N = 8$ ($p = 1/2048$), versus eight on four at $N = 128$ ($p = 1/8$). The former has Bonferroni-87 $p = 0.04248$, adjusting only this retained family, not earlier research choices. The original percentile intervals remain unchanged. BCa sensitivities (Efron, 1987) use the same 10,000 stratified resamples and a fixed midrank convention for discrete ties; degenerate rows have no defined BCa interval. Neither procedure repairs sparse support. At $N = 128$, one A and three B problems supply all Math raw gains. The exact empirical-bootstrap probability of missing all four is $(29/30)^{30}(27/30)^{30} = 0.01533$. Because all observed differences are nonnegative, this

Config. / N	Plurality	Raw V	Half-floor	DEV half-floor
Math-7B / 8	33.89	$+7.78$ $[+3.89, +12.78]$	$+4.44$ $[+1.67, +7.78]$	$+3.33$ $[+1.11, +6.11]$
Math-7B / 16	37.22	$+3.89$ $[+0.56, +8.33]$	$+2.22$ $[+0.00, +5.00]$	$+2.22$ $[+0.00, +5.00]$
Math-7B / 32	38.33	$+3.89$ $[+0.56, +8.33]$	$+1.67$ $[+0.00, +4.44]$	$+0.56$ $[+0.00, +1.67]$
Math-7B / 64	37.78	$+5.00$ $[+0.56, +10.56]$	$+2.22$ $[+0.00, +5.56]$	$+1.67$ $[+0.00, +4.44]$
Math-7B / 128	38.33	$+4.44$ $[+0.56, +9.44]$	$+3.33$ $[+0.00, +7.78]$	$+1.67$ $[+0.00, +4.44]$

Table 9: Math-7B on 60 test problems at five candidate budgets. Plurality is accuracy (%); other cells are paired differences from plurality in percentage points with pointwise 95% problem-cluster intervals. All three weighting rules use the same candidates.

Config. / N	Raw V	Half-floor	DEV half-floor
Math-7B / 8	$12:12/0$	$7:7/0$	$6:6/0$
Math-7B / 16	$4:4/0^{\dagger}$	$3:3/0^{\ddagger}$	$3:3/0^{\ddagger}$
Math-7B / 32	$5:5/0^{\dagger}$	$2:2/0^{\ddagger}$	$1:1/0^{\ddagger}$
Math-7B / 64	$4:4/0^{\dagger}$	$2:2/0^{\ddagger}$	$2:2/0^{\ddagger}$
Math-7B / 128	$4:4/0^{\dagger}$	$3:3/0^{\ddagger}$	$2:2/0^{\ddagger}$

Table 10: Problem support for the 15 weighting contrasts in Table 9. Each cell is Disc.: g/l: problems with any paired correctness discordance, followed by problems with at least one gain or loss. Runs remain clustered. † marks raw-score contrasts whose positive lower endpoint rests on four or five discordant problems; ‡ marks eight half-floor contrasts whose lower endpoint equals zero.

forces a positive ideal-percentile lower endpoint; it gives no exact population coverage guarantee. Complete sensitivities, including all signs and unavailable cases, accompany the analysis package.

Coarsened residuals do not estimate an upper calibration error. The calibration diagnostics ask whether the probabilities needed by Section 4 are compatible with the saved scores on each problem. AUC addresses a different question: whether correct candidates rank above incorrect candidates. Here we compare scores with observed correctness, report coarse calibration residuals, and use a simultaneous mean-residual bound to reject calibration where possible. This separates score ranking, calibration rejection and the sufficient preservation certificate defined next.

Per-problem diagnostics use all eligible samples, up to 384 pooled across three runs. For $S = V$ and $S = g(V)$ we record the empirical answer margin and ten equal-width-bin residuals. Population conditional Jensen gives

$$\sum_{j=1}^{10} |\mathbb{E}[(R - S)\mathbf{1}\{S \in B_j\}]| \le \mathbb{E}|\mathbb{E}[R \mid S] - S| =: \varepsilon_S.$$

The empirical binned residual is noisy in either direction relative to that left side; it is not an upper bound on ε_S and cannot be substituted into $\Delta > \varepsilon_S$.

A conservative, label-dependent population certificate. This audit asks which problems have enough labelled evidence to verify the robustness condition in Proposition 2: the correct answer's population margin exceeds its calibration error. We lower-bound their difference using all eligible candidates for each problem. A positive bound certifies preservation of the correct population winner by the half-floor under the stated sampling assumptions; finite-pool accuracy is measured separately.

The readout, score and original calibration-failure summaries were visible when this post-hoc certificate was devised; its counts were computed afterward. The original fixed-map Hoeffding justification alone did not cover that choice. The uniform argument below instead covers the full stated class of monotone maps and summaries, under a fixed raw scorer and candidate laws. For

$S = g(V)$, conditional Jensen and the correct-majority lower bound give

$$\varepsilon_S \le \mathbb{E}|R - S|, \qquad \Delta := p_{a^\star} - \max_{a \neq a^\star} p_a \ge 2\Pr(R = 1) - 1, \qquad \Delta - \varepsilon_S \ge \mathbb{E}X_S,$$

where $X_S := 2R - 1 - |R - S| \in [-2, 1]$. With $Q = 320$, $\alpha = .05$ and the radius e_q below, use

$$C_q(S) = \overline{X}_S - 3e_q, \tag{43}$$

$$L_q(S) = [|\overline{R - S}| - 3e_q]_+. \tag{44}$$

On the simultaneous event, $\Delta - \varepsilon_S \ge C_q(S)$ and $\varepsilon_S \ge L_q(S)$. If $C_q(S) > 0$, the correct base and half-floor population modes are certified, with every unnormalized half-floor gap at least $C_q(S)/2$. This label-dependent offline condition is sufficient, uses a majority lower bound, and implies no finite-vote improvement. Nonpositive values are inconclusive. Positive L_q rejects exact calibration but supplies no calibration-error upper bound. The DEV fit has positive slope and hence belongs to the monotone class. The original $Q = 320$ confidence-family size is retained for these Math-only summaries, so no confidence bound is tightened after selecting the reported cohort.

Uniform inference for monotone score maps. The analysis is post hoc. To account for the choice of score map and the following certificate summaries, we use one uniform event rather than conditioning coverage on selecting a formula before its count was computed. Fix the Q problem/cohort laws, the raw scalar score $V \in [0, 1]$, the generated sample counts, and candidate-local eligibility independently of outcomes. Within each row the $n_q > 0$ eligible observations (R_i, V_i) are assumed iid conditional on their count. Under iid generation and a fixed generated count, conditioning on the eligibility indicators leaves the retained observations iid under the eligible law. Zero-eligible rows remain uncertified and retain their allocation in Q. Different rows may overlap. Let $g : [0, 1] \to [0, 1]$ range over all nondecreasing measurable maps, including data-chosen maps, and set

$$X_g = Rg(V) - (1 - R)(1 + g(V)), \qquad Y_g = \max(g(V), 1/2)(2R - 1), \qquad e_q = \sqrt{\frac{\log(4Q/\alpha)}{2n_q}}.$$

With probability at least $1 - \alpha$, simultaneously for every row and map,

$$|\overline{X}_g - \mathbb{E}X_g| \le 3e_q, \qquad |\overline{Y}_g - \mathbb{E}Y_g| \le 2e_q, \tag{45}$$

$$|\overline{R} - \mathbb{E}R| \le e_q, \qquad |\overline{R - g(V)} - \mathbb{E}[R - g(V)]| \le 3e_q. \tag{46}$$

For a data-chosen map, each population expectation evaluates the realized map against the original law, equivalently on a new independent draw.

To prove these classical empirical-process bounds, define $W_1 = V$ on correct observations and -1 otherwise, and $W_0 = V$ on wrong observations and -1 otherwise. The two-sided Dvoretzky–Kiefer–Wolfowitz–Massart inequality (Dvoretzky et al., 1956; Massart, 1990) and a union bound give uniform CDF error at most e_q for all $2Q$ extended scalar laws. On this event every nondecreasing $f : [-1, 1] \to [0, B]$ with $f(-1) = 0$ has empirical-mean error at most Be_q: integrate its upper level sets, using CDF left limits when needed. Apply this fact to the extensions of g, $1 + g$, the half-floor, and the constant one on $[0, 1]$, whose ranges are respectively $1, 2, 1, 1$. This proves the first three bounds; adding the base-probability bound to the two class-weighted g bounds proves the last. The event is uniform over all such functions, so subsequent map or summary selection within this family incurs no additional penalty.

For $Q = 320$ and $\alpha = .05$, the X radius retains the factor $4Q = 1280$. The uniform calibration-error lower bound is $[|\overline{R - g(V)}| - 3e_q]_+$, since $\epsilon_g \ge |\mathbb{E}[R - g(V)]|$. Earlier, smaller fixed-map residual radii are retained in the source artifacts; they are not part of this uniform claim. This argument does not protect changes to the raw scorer, candidate law, cohort universe, or outcome-dependent stopping, and does not establish that the observed candidates satisfy the iid model. The original 640- and 1280-procedure fixed-map results, and the later 2240-procedure comparison, remain in the audit artifacts. The radius for the robustness condition equals the original one; its task counts are unchanged. Calibration rejection below uses the wider uniform residual radius, not the earlier fixed-map radius.

Config.	Score	Defined / total	Mean binned residual	Calibration rejected	Robustness certified
Math-7B	raw	60/60	0.198	8/60	7/60
Math-7B	DEV	60/60	0.168	3/60	5/60

Table 11: Math-7B diagnostics using all eligible draws, up to 384 per problem. Calibration rejection and robustness certification use the uniform monotone-map event under iid eligible sampling. Rejection detects calibration error; the robustness certificate establishes population preservation, whose finite-vote consequences are measured separately.

Configuration	Score	Occupied wrong bins	Median wrong count	Median concentration
Math-7B	raw	524	12.0	0.444
Math-7B	DEV	443	12.0	0.500

Table 12: Wrong-answer concentration within each problem and one of ten fixed score bins. An occupied wrong bin contains at least one incorrect candidate. Concentration is the largest wrong-answer count divided by all wrong counts in that bin. Medians are across occupied bins; a concentration of one can result from a single observation. Counts and full per-bin rows accompany the artifact, so these sparse empirical fractions are not interpreted as population concentration bounds.

Wrong-answer concentration and interpretation. This diagnostic asks whether wrong candidates at a given score tend to repeat one answer or spread over many answers. That is the information used by the concentration-dependent guarantees in Section B.2. Counting labels within fixed score bins gives a descriptive view of this structure, alongside the bin sample sizes needed to interpret concentrated or sparse observations.

Within each problem and score bin we count wrong candidates, distinct wrong labels and the fraction assigned to the largest wrong label (Table 12). Full per-bin records retain empty bins and denominators. Sparse bins can have empirical concentration one from a single observation; these are not population concentration caps. The certificates concern population modes on the robustness-certified subset; the finite-vote comparison concerns the full evaluated Math cohort. The cross-tabs below show that every observed half-floor decision change occurs outside the robustness-certified subset. The artifact retains declarations, exact fits, input hashes, predictions, intervals, all diagnostic and certificate rows, and independent verification.

Certificate overlap and simpler labelled baselines. This comparison asks how many problems the robustness certificate covers, and whether simpler labelled bounds already certify them. We apply the robustness, score-free and direct weighted-majority conditions to the same problems with a common confidence event, then cross-tabulate their coverage and the observed voting changes. The comparison separates the robustness condition's interpretation from the number of tasks it certifies.

This additional analysis was declared after viewing the readout, score and certificate summaries, before computing their intersections or baseline counts. All original inputs remain unchanged. We retain both Math score maps and all five budgets; these cross-tabs are descriptive. For $h(S) = \max(S, 1/2)$, the score-free sufficient condition $p > 2/3$ follows from $\mathbb{E}[h(S)(\mathbf{1}\{A = a^\star\} - \mathbf{1}\{A = a\})] \geq p/2 - (1-p)$. The direct labelled comparator uses $Y = h(S)(2R-1)$: positive weighted-majority and base-majority lower bounds certify preservation. With the common uniform radius e_q, these are $\overline{Y} - 2e_q > 0$ and $\widehat{p} - e_q > 1/2$; the range-only rule uses $\widehat{p} - e_q > 2/3$. Pointwise, $Y \geq X$ and $X \leq 2R - 1$. Thus $\overline{X} - 3e_q > 0$ implies both direct conditions: the calibration-error certificate supplies no incremental preservation certificates over this direct baseline. Its additional interpretation is the sufficient inequality $\Delta > \varepsilon$, not stronger labelled modal certification. No robustness-certified Math-7B task changes its half-floor decision at any of the five budgets under either score map. The next comparison examines the larger set certified by the direct labelled bound.

Finite decisions inside the wider direct certificate. Here we ask whether preserving the population winner also makes every finite-pool decision identical. We inspect the paired predictions inside the

Config.	Score	Total	X certificate	$p > 2/3$	Direct preservation
Math-7B	raw	60	7	7	14
Math-7B	DEV	60	5	7	14

Table 13: Labelled sufficient certificates on the same uniform event. The direct certificate subsumes X analytically. The range-only rule uses neither scores nor calibration.

Config.	Score	robustness-certified	Eligible count	Empirical correctness
Math-7B	raw	7/60	351–384	84.2–98.2%
Math-7B	DEV	5/60	351–384	91.5–98.2%

Table 14: Regime covered by the robustness certificate. Eligible counts and correctness are ranges over robustness-certified tasks, using all certification-pool samples. They are descriptive empirical frequencies, not confidence limits. Across every fixed budget, robustness-certified tasks have zero changed half-floor decisions, zero gains and zero losses.

direct certificate's larger covered set. Changed decisions in this set illustrate the distinction between long-run preservation and finite-budget behavior in Sections 4 and 5.

The direct certificate requires both $\widehat{p} - e_q > 1/2$ and $\overline{Y} - 2e_q > 0$, so it covers correct base and weighted population majorities on the same uniform event. These direct certificates require no calibration assumption. Their masks use the same full labelled pools containing the evaluated shorter prefixes; the cross-tab is not independent subgroup validation. The larger directly certified set has the following nonzero finite-decision changes. The sufficient X rule cannot certify a correct non-majority mode and adds no preservation certificates beyond the direct labelled comparison. Its zero-change subset and the wider direct certificate must not be conflated. The uniform monotone-map event addresses choices within its proved class; it does not cover arbitrary cohort or scorer selection or establish iid sampling. The earlier, separately declared fixed-2240-procedure sensitivity remains in the artifact, together with all original confidence-family results.

E.1 Transfer and count margins

The 140 remaining Math-7B test problems provide three independent 20-candidate runs each. The eight-candidate prefix comparison asks whether the selected voting rule helps on additional problems at the same budget. The count-margin audit then connects decisions on these pools to the protection mechanism in Proposition 7. These analyses were specified after the initial cohort's outcomes and the remaining cohort's plurality and filtering outcomes were known. The complete larger-budget efficacy comparisons remain in the artifact.

A separately specified matching-budget check. This follow-up asks whether the eight-candidate gain also appears on the remaining 140 Math-7B problems when the voting budget matches the initial study. Using each saved pool's first eight candidates holds the budget and readout protocol fixed while changing the problem cohort. Paired accuracy against plurality measures transfer of the selected rule at that budget; the post hoc timing is recorded below.

After the inspected $N = 8$ and remaining-problem $N = 20/60$ outcomes were known, we specified one additional readout of the first eight generated candidates of each remaining-problem pool, then applied eligibility. This is an outcome-informed, post hoc transfer check; no new generation, scoring, map fit or change to the earlier declarations was made. All 600 source-pool hashes verify against the recorded inputs; 46 scientific configuration fields match across the inspected and remaining Math cohorts. Generated pool size differs (128 versus 20), while both readouts use their first eight draws. There is no standalone equal-time cost measurement at this budget.

Plurality is correct on 151/420 decisions. Raw weighting, the sole primary of this addendum, gains 19 decisions and loses five: $+3.33$ points, 95% interval $[1.19, 5.71]$. Its 20 discordant problems have 16 positive and four negative net differences; one problem contains both a gain and a loss. The conditional sign-flip value is $4393/524288 = 0.008379$, or 0.02514 after Bonferroni adjustment for the three rules. The same stratified cluster procedure gives the individual 98.33% intervals in

Config.	Score	$N = 8$	16	32	64	128
Math-7B	raw	16/8/0	10/4/0	11/3/0	8/4/0	10/6/0
Math-7B	DEV	10/6/0	6/4/0	5/1/0	6/3/0	6/3/0

Table 15: All changed predictions / correctness gains / correctness losses in defined tasks outside the robustness-certified subset. The robustness-certified and undefined groups have zero in every cell. Prefixes are nested and are not independent observations. Full artifacts retain all group denominators, task and run counts, empirical margin and correctness ranges, and the other declared certificate rules.

Configuration	Map / budget	Tasks	Runs	Changes	Gains	Losses
Math-7B	raw / 8	14	42	1	1	0
Math-7B	DEV / 8	14	42	1	1	0

Table 16: Nonzero changed-prediction cells among Math-7B tasks covered by the direct-preservation certificate. Every other map/budget cell has zero changes. Both rows concern the same task/run. This illustrates a finite-vote gain while the correct population winner is preserved.

Table 17. Its sampling interpretation is the comparable-problem, observed-run model described in Section E. The fair-sign assumption remains essential, and neither adjustment accounts for the outcome-informed nomination. Both half-floors remain secondary. Raw half-floor's positive adjusted bootstrap interval coexists with adjusted sign-flip $p = 0.05273$; DEV half-floor has only seven nonzero problems. These procedures use different assumed sampling/null constructions, and neither is selected by its conclusion. This outcome-informed matching-budget result is positive on these saved pools. The separate prospectively declared comparison on 252 further problems appears in Section F.

Count margins. This audit checks where harmful reversals occur, as motivated by the count-lead protection in Proposition 7. We group paired decisions by the count winner's lead before scores are applied, then count gains, losses and wrong-to-wrong changes for each rule. This tests whether observed losses concentrate near ties; estimating the theorem's probability bound would additionally require its population premises.

For each evaluated prefix, we tabulate ordinary plurality's top-two eligible-answer count lead before weighting. A singleton answer has lead equal to its eligible count; empty pools form a separate stratum. Tables 18 and 19 retain all fixed cohort, budget, rule and lead cells. Across both half-floors and both Math cohorts and budgets in this count-margin audit, every harmful reversal occurs at lead zero or one. Higher-lead changes include gains and wrong-to-wrong changes. The raw-weighted rule does have harmful changes at larger leads. These are descriptive cross-tabs of dependent prefixes, with no within-stratum test. They confront the bound's near-tie mechanism but do not establish its calibration premise or its joint-event probability, and do not estimate conditional error given an observed lead.

Execution checks. Independent reconstruction checked source-pool hashes, recomputed the voting decisions and reproduced the declared inference calculations. Counts and decisions matched exactly;

Table 17: Post hoc, matching-budget transfer on the remaining 140 Math-7B problems: 420 shared eight-candidate prefixes. Differences from plurality in percentage points; all three fixed rules retained. The individual 98.33% intervals and adjusted conditional sign-flip values cover only this three-rule family. G/L counts decisions; Disc. counts distinct discordant problems (output field `distinct_problems_any_discordance`).

Rule	Correct	Change [95% CI]	[98.33% CI]	Adjusted p	G/L	Disc.
Raw V (primary)	165/420	$+3.33$ $[1.19, 5.71]$	$[0.71, 6.35]$	0.02514	19/5	20
Raw half-floor	161/420	$+2.38$ $[0.71, 4.29]$	$[0.48, 4.76]$	0.05273	11/1	10
DEV half-floor	160/420	$+2.14$ $[0.71, 3.81]$	$[0.48, 4.29]$	0.04688	9/0	7

N	Rule	Lead 0	Lead 1	Lead 2	Lead ≥ 3	Empty
8	Raw	47:8/0/26	48:3/0/8	15:1/0/1	66:2/0/1	4:0/0/0
8	Floor	47:7/0/8	48:0/0/0	15:1/0/0	66:0/0/0	4:0/0/0
8	Mapped floor	47:6/0/4	48:0/0/0	15:0/0/0	66:0/0/0	4:0/0/0
16	Raw	31:4/0/20	33:1/0/7	18:0/0/3	97:2/0/1	1:0/0/0
16	Floor	31:4/0/6	33:0/0/0	18:0/0/0	97:0/0/0	1:0/0/0
16	Mapped floor	31:4/0/2	33:0/0/0	18:0/0/0	97:0/0/0	1:0/0/0
32	Raw	16:1/0/13	25:2/0/10	16:0/0/6	122:4/0/4	1:0/0/0
32	Floor	16:1/0/7	25:0/0/1	16:0/0/0	122:2/0/0	1:0/0/0
32	Mapped floor	16:0/0/3	25:0/0/1	16:0/0/0	122:1/0/0	1:0/0/0
64	Raw	16:2/0/10	15:1/0/9	10:1/0/1	139:5/0/8	0:0/0/0
64	Floor	16:2/0/2	15:0/0/2	10:0/0/0	139:2/0/0	0:0/0/0
64	Mapped floor	16:1/0/2	15:0/0/1	10:0/0/0	139:2/0/0	0:0/0/0
128	Raw	9:1/0/8	12:1/0/8	7:1/0/2	152:5/0/6	0:0/0/0
128	Floor	9:1/0/2	12:1/0/1	7:1/0/0	152:3/0/1	0:0/0/0
128	Mapped floor	9:1/0/2	12:0/0/1	7:0/0/0	152:2/0/0	0:0/0/0

Table 18: *Post hoc transfer analysis; rules fixed before computing; plurality and thresholded-rule outcomes on these pools were already known.* Inspected Math-7B: all count-margin cells. Each entry is decisions:G/L/W, where G, L and W are correctness gains, losses and wrong-to-wrong changes relative to plurality. The number of changed decisions is their sum. No within-stratum inference is made; budgets and rules share candidates.

N	Rule	Lead 0	Lead 1	Lead 2	Lead ≥ 3	Empty
20	Raw	65:4/3/34	86:3/0/34	34:2/1/7	232:1/1/3	3:0/0/0
20	Floor	65:4/3/14	86:1/0/2	34:1/0/1	232:0/0/0	3:0/0/0
20	Mapped floor	65:2/2/3	86:0/0/0	34:1/0/0	232:0/0/0	3:0/0/0
60	Raw	11:1/0/7	16:0/1/5	10:2/0/2	102:1/1/4	1:0/0/0
60	Floor	11:1/0/2	16:0/0/0	10:0/0/0	102:0/0/1	1:0/0/0
60	Mapped floor	11:1/0/1	16:0/0/0	10:0/0/0	102:0/0/1	1:0/0/0

Table 19: *Post hoc transfer analysis; rules fixed before computing; plurality and thresholded-rule outcomes on these pools were already known.* Remaining 140 Math-7B problems: all count-margin cells. Each entry is decisions:G/L/W, where G, L and W are correctness gains, losses and wrong-to-wrong changes relative to plurality. The number of changed decisions is their sum. No within-stratum inference is made; budgets and rules share candidates.

accuracy arithmetic differed only by floating-point roundoff. The complete nine-output reconstruction is retained in the artifact.

F Prospective replication at eight candidates

This experiment tests whether the eight-candidate weighting gain found in the earlier Math-7B pools repeats on new problems for that configuration. We use the available disjoint reserve of 252 problems and generate three fresh eight-candidate pools for each one. Within every pool, raw weighting and the two fixed half-floors compete with plurality on identical candidates. Their paired accuracy differences therefore test the benefit of scoring at a fixed small budget. The count-tie and count-override decomposition then locates the gain using the distinction in Section 5.

Declared comparison and cohort. The raw-V rule and $N = 8$ budget were selected using earlier inspected outcomes; this replication was declared before its own generation (a plan saved locally, without external timestamp). The plan specified neither a target effect size nor a power calculation. The inspected-cohort estimate that informed that selection, $+7.78$ points $[3.89, 12.78]$ on 60 problems at $N = 8$ (Table 20), exceeds the declared replication's $+2.78$ $[1.06, 4.63]$, whose point estimate lies below that interval; inspected-cohort intervals are selection-informed and are not replication forecasts. It uses a disjoint AIME 2000–2017 cohort of 252 problems, previously unrun under this Math-7B configuration: 140 in stratum A (problem numbers 1–8) and 112 in B (9–15). This is the

Measurement	Math-7B, 60 test problems
Plurality accuracy, $N = 128$ (%)	38.33
Correct answer is empirical mode	23/60
Exact raw calibration rejected	8/60
Robustness bound, raw / mapped	7 / 5
Score-free bound ($p > 2/3$)	7
Direct preservation bound, raw / mapped	14 / 14
Raw V change, $N = 8$ (pp)	$+7.78$ $[3.89, 12.78]$
Raw V change, $N = 32$ (pp)	$+3.89$ $[0.56, 8.33]$
Raw V change, $N = 128$ (pp)	$+4.44$ $[0.56, 9.44]$
Raw half-floor change, $N = 128$ (pp)	$+3.33$ $[0.00, 7.78]$
Mapped half-floor change, $N = 128$ (pp)	$+1.67$ $[0.00, 4.44]$
Raw half-floor changes / gains / losses, $N = 128$	10 / 6 / 0
Half-floor changes under the robustness bound	0

Table 20: Math-7B score diagnostics, preservation certificates and paired voting effects on the initial test cohort. Certificate counts retain the original confidence-family sizes; intervals are pointwise 95% problem-cluster intervals. Calibration rejection checks a necessary mean condition. The prospective replication below tests the selected eight-candidate rules on separate problems.

audited available reserve, not a random sample of mathematics problems. The existing missing-figure exclusion was retained; one context-overflow problem and two known aliases of MATH-500 questions (Lightman et al., 2024) were excluded before generation, while 13 questions with Asymptote markup were retained. These 252 problems are separate from the 140-problem post hoc $N = 8$ prefix analysis; their outcomes are not pooled.

Frozen protocol and complete endpoint. Three independently generated iid pools per problem contain eight candidates each: 756 pools and 6,048 candidates, with all 756 decisions retained for every rule. The generator is Qwen2.5-Math-7B-Instruct (revision ef9926d75ab1d54532f6a30dd5e760355eb9aa4d); the scorer is Qwen2.5-Math-PRM-7B (revision 0610740060112df12585d00a1c5f4624d2f59051), with the verified `repaired_v1` load and native self-test. Settings are `qwen25_math_cot`, temperature 0.7, top-p 0.95, thirteen 256-token blocks (3,328 tokens), terminal minimum-step PRM aggregation, nominal 4,096-token scorer input budget, and LM/PRM chunks 32/16. Saved metadata flag PRM truncation for 22 of 6,048 candidates; all are retained. The original DEV/MECH horizon reference set is preserved. Seeds use the fixed problem/run/N derivation. Standalone $N = 8$ changes the seed and batch size relative to an eight-candidate prefix of $N = 20$; it uses the same sampling settings, not the same samples.

Raw V versus boxed plurality is the sole primary; the raw half-floor $\max(V, 1/2)$ and $\max(g(V), 1/2)$ with the frozen Math-7B DEV map are fixed secondaries. The map is never refit. All rules take the first eight generated candidates before eligibility, with the frozen extraction, absorption, log-score-authoritative raw-V arithmetic and common fallback. Answers within 10^{-12} of the maximum normalized answer mass are tied, with first-occurrence priority. Seven decisions on five problems have no eligible boxed integer answer and abstain under every rule; they remain incorrect in all denominators. There are no missing cells or algorithm-failure decisions. The declaration, cohort, source closure and DEV map were frozen at 02:35:16 UTC on 22 September 2026, before pilot submission at 02:44:04 and the first saved population at 02:45:32. Four technical pilot cells were retained unchanged. The complete-grid analysis ran once at 15:47:29; independent reconstruction from raw pools cleared the predictions, statistics and protocol. The audit found no artifact trace of interim correctness inspection; artifacts cannot exclude an unlogged read.

Estimand, inference and support. For each rule we average the three paired correctness differences within each problem, then weight the 252 problems equally. The primary mixture is therefore 140/252 A and 112/252 B. The shared A/B-stratified problem-cluster bootstrap retains all three runs and draws 140 A and 112 B problems per replicate: 10,000 draws using NumPy `default_rng(28391)`, A then B, lexicographically sorted string IDs, and linear quantiles. Table 22 reports the central pointwise 95% and individual 98.333% intervals (the fixed Bonferroni-three family sensitivity). These intervals assume independent comparable problem/run clusters; 756 decisions are not 756 independent clusters,

and the intervals do not give distribution-free repeated-generation guarantees on these fixed questions. The exact two-sided fair-sign sensitivity uses the absolute sum of per-problem integer run differences, conditioning on their magnitudes. It requires independent conditionally fair signs; a zero-mean null alone does not suffice. The table retains both secondaries, all signed support and all adjusted p values.

Declared descriptive diagnostics and interpretation. The A/B and equally weighted 50/50 effects in Table 22 are mixture descriptives, with no separate efficacy claims. Table 23 retains all count-lead categories, including undefined pools. The count lead is the plurality winner's count minus its strongest rival's count (zero for the rival if only one answer is represented). Maximum counts are tied in 241/756 decisions (31.9%). Of 28 raw-V gains and seven losses, 19 gains and one loss change the answer within a tied maximum; nine gains and six losses select an answer below the maximum count. Eight of these 73 count overrides occur at lead zero, including one gain and one loss. This is a fixed-configuration $N = 8$ result; it establishes neither calibration nor the theory's population premises, and supports no large-N, equal-cost or broader generalization claim. These comparisons use the same eight-candidate pools. The subsequent generation-budget extension compares weighting on these pools with plurality after one or two additional draws (Section F.2).

F.1 Additional generators on the same 252 problems

Question and design. Does the weighting benefit extend to other generators on the same problems? We generate three new eight-candidate pools per problem with Phi-4 (Abdin et al., 2024) and Qwen2.5-Math-1.5B-Instruct (Yang et al., 2024), using the exact 252-problem Math-7B cohort: 140 problems in stratum A and 112 in B. Each generator contributes 756 decisions and 6,048 generated candidates. Within each generator, plurality, raw weighting and the raw half-floor use identical candidates and the same Qwen2.5-Math-PRM-7B scorer. Raw weighting is the sole primary comparison; the raw half-floor is the fixed secondary. Neither new generator uses a development-fitted score map.

The Math-1.5B configuration matches Math-7B's prompt, temperature .7, top-p .95, 13×256-token generation horizon and 4,096-token context. Phi-4 uses its native chat template with the AIME prompt, the same temperature and top-p, a 32×256-token horizon and a 16,384-token context. All studies use the same minimum-step scorer and 4,096-token scoring suffix policy. Thus each within-generator contrast isolates the voting rule; cross-generator comparisons describe these specified configurations on a shared problem cohort. They do not isolate parameter count, prompt or generation horizon as a causal factor.

Complete results. Phi-4's raw-weighting accuracy is 40.48%, versus 37.57% for plurality: a 2.91-point gain with a positive problem-cluster interval. The raw half-floor gains 2.12 points. Math-1.5B has smaller positive point estimates, 1.19 and .93 points, whose intervals include or touch zero. Table 24 reports both rules and all signed changes. Abstentions remain in the denominator: nine Phi-4 decisions on four problems and 13 Math-1.5B decisions on eight problems. Neither study records an algorithm failure.

Uncertainty and paired comparisons. The estimand gives equal weight to problems after averaging their three paired correctness differences. As in Section F, the 10,000-draw stratified bootstrap resamples problems, retaining all runs, and the fair-sign sensitivity conditions on problem-level magnitudes. Its conditional symmetry assumption is separate from a zero-mean null. The confidence intervals have the problem-cluster interpretation stated there; 756 decisions are not 756 independent problems. Phi-4 conservatively retains the parent's factor-three adjustment, although it tests two rules. The promised factor-two supplement gives 97.5% intervals $[.93, 5.03]$ for raw weighting and $[.79, 3.70]$ for the half-floor, with adjusted fair-sign $p = .002198$ and $.000977$. Both adjustments give positive Phi-4 intervals. Math-1.5B uses its declared factor-two adjustment throughout.

For the cross-generator comparison, one shared resampled problem index is applied to both generators. The raw-weighting gain differs from Math-7B's by $+.13$ points $[-2.38, 2.51]$ for Phi-4 and -1.59 $[-3.84, .53]$ for Math-1.5B. The corresponding secondary half-floor differences are $-.13$ $[-1.85, 1.72]$ and -1.32 $[-2.91, .13]$. All four nominal 95% paired intervals contain zero; these comparisons do not establish different weighting effects across the configurations. The within-generator intervals in Table 24 come from each study's primary analysis.

Contrast	Change (pp)	95% interval
Declared study: 252 problems, eight candidates		
Raw weighting − plurality (primary)	$+2.78$	$[1.06, 4.63]$
Raw half-floor − plurality	$+2.25$	$[1.06, 3.70]$
DEV half-floor − plurality	$+1.72$	$[0.79, 2.78]$
Post hoc controls on the same pools		
Highest-score selection − plurality	$+2.78$	$[0.66, 5.03]$
Raw weighting − highest-score selection	0.00	$[-1.19, 1.19]$
Random-tie plurality − plurality	$+0.58$	$[-0.33, 1.44]$
Raw weighting − random-tie plurality	$+2.20$	$[0.64, 3.83]$
Declared added-candidate extension		
Raw weighting ($N = 8$) − plurality ($N = 9$)	$+2.65$	$[1.06, 4.37]$
Raw weighting ($N = 8$) − plurality ($N = 10$)	$+1.46$	$[-0.40, 3.31]$

Table 21: Math-7B paired accuracy differences on 252 problems and 756 decisions. The declared comparison, post hoc controls and added-candidate extension are shown separately. Intervals resample problem clusters within the two problem-number strata. DEV denotes the fixed development-fitted probability map; complete family-adjusted inference appears in the corresponding tables below.

Table 22: Complete prospective $N = 8$ comparison on 252 problems and 756 decisions per rule. Plurality is correct in 274/756 decisions (36.24%). Effects and intervals are percentage points versus plurality. All seven abstentions on five problems remain in every denominator. The 98.333% intervals and adjusted p values use the declared three-rule Bonferroni family.

Quantity	Raw V	Raw half-floor	DEV half-floor
Role	Sole primary	Fixed secondary	Fixed secondary
Correct decisions	295/756	291/756	287/756
Accuracy (%)	39.02	38.49	37.96
Difference (pp)	$+2.78$	$+2.25$	$+1.72$
Pointwise 95% CI	$[1.06, 4.63]$	$[1.06, 3.70]$	$[0.79, 2.78]$
Family 98.333% CI	$[0.79, 5.03]$	$[0.79, 3.97]$	$[0.66, 3.04]$
Exact fair-sign p	$43073/16777216$	$3/4096$	$1/1024$
Fair-sign p (decimal)	0.002567	0.000732	0.000977
Bonferroni-three p	0.007702	0.002197	0.002930
Statistic T / nonzero problems	21/29	17/15	13/11
Changed decisions / problems	245/133	69/52	34/28
Gains / losses / wrong-to-wrong	28/7/210	18/1/50	13/0/21
Any-gain / any-loss problems	22/7	14/1	11/0
Discordant problems	29	15	11
Net-positive / negative / zero	22/7/223	14/1/237	11/0/241
Both-gain-and-loss / cancellation	0/0	0/0	0/0
Changed, both correct	0	0	0
Stratum A effect (descriptive)	$+3.10$	$+2.62$	$+2.38$
Stratum B effect (descriptive)	$+2.38$	$+1.79$	$+0.89$
50/50 effect (descriptive)	$+2.74$	$+2.20$	$+1.64$

Support counts refer to paired comparisons with plurality. A discordant problem has at least one gain or loss across its three runs; a nonzero problem has a nonzero net sum. Cancellation counts discordant problems whose net sum is zero. The net-zero row includes problems without discordance.

Table 23: Declared descriptive raw-V decomposition. A tied-maximum change selects another answer with the same maximum ordinary count; a count override selects an answer with a lower count than that maximum. Each parenthesis gives gains/losses/wrong-to-wrong changes. There are no additional subgroup efficacy tests.

Count lead	Unchanged	Tied-maximum change	Count override	Undefined
0	61	172 (19/1/152)	8 (1/1/6)	0
1	99	0	49 (6/3/40)	0
2	63	0	11 (2/1/8)	0
≥ 3	281	0	5 (0/1/4)	0
Undefined	0	0	0	7
Total	504	172 (19/1/152)	73 (9/6/58)	7

	Phi-4		Math-1.5B	
Quantity	Raw V	Half-floor	Raw V	Half-floor
Role	Primary	Secondary	Primary	Secondary
Plurality correct	284/756 (37.57%)		287/756 (37.96%)	
Correct decisions	306/756	300/756	296/756	294/756
Accuracy (%)	40.48	39.68	39.15	38.89
Gain (pp)	$+2.91$	$+2.12$	$+1.19$	$+.93$
Pointwise 95% CI	$[1.19, 4.76]$	$[.93, 3.44]$	$[-.53, 2.91]$	$[0, 1.85]$
Declared family level	98.333% (factor 3)		97.5% (factor 2)	
Declared family CI	$[.79, 5.16]$	$[.79, 3.84]$	$[-.66, 3.04]$	$[0, 1.98]$
Exact fair-sign p	.001099	.000488	.214964	.091797
Family-adjusted p	.003297	.001465	.429927	.183594
Statistic / nonzero problems	22/29	16/16	9/24	7/10
Changed decisions / problems	241/133	60/50	190/106	32/29
Gains / losses	30/8	17/1	20/11	9/2
Wrong-to-wrong changes	203	42	159	21
Net-positive / negative / zero problems	24/5/223	15/1/236	16/8/228	8/2/242
Stratum A effect (pp)	$+2.86$	$+1.90$	$+.95$	$+.71$
Stratum B effect (pp)	$+2.98$	$+2.38$	$+1.49$	$+1.19$
50/50 effect (pp)	$+2.92$	$+2.14$	$+1.22$	$+.95$

Table 24: Complete comparisons on the same 252 problems, with 756 decisions per generator and rule. Gains and intervals are percentage points versus that generator's plurality. The family corrections apply within the respective study; they do not adjust selection across generators. Stratum and 50/50 effects are descriptive. Math-1.5B's half-floor family lower endpoint rounds to zero.

Where the decisions change. Phi-4 repeats the pattern seen with Math-7B: most of its net gain comes from choosing among tied count leaders. Those changes correct 20 errors and introduce three, accounting for 17 of the 22 net additional correct decisions. Count overrides contribute the remaining five (10 gains, five losses). For Math-1.5B, tied-maximum changes contribute three net correct decisions (seven gains, four losses), and overrides contribute six (13 gains, seven losses). Table 25 retains every count-lead category. These are descriptive decompositions of the fixed raw-weighting rule, not additional subgroup efficacy tests.

Count lead	Unchanged	Tied-maximum change	Count override	Undefined
Phi-4				
0	55	135 (20/3/112)	9 (0/0/9)	0
1	124	0	80 (7/3/70)	0
2	53	0	12 (2/1/9)	0
≥ 3	274	0	5 (1/1/3)	0
Undefined	0	0	0	9
Math-1.5B				
0	58	89 (7/4/78)	14 (1/1/12)	0
1	99	0	64 (7/5/52)	0
2	82	0	20 (5/1/14)	0
≥ 3	314	0	3 (0/0/3)	0
Undefined	0	0	0	13

Table 25: Declared raw-weighting decomposition on the same 252 problems and 756 decisions per generator. Parentheses give gains/losses/wrong-to-wrong changes. A tied-maximum change selects a different answer with the same maximum count; an override selects a lower-count answer. Tied count leaders occur in 199/756 Phi-4 pools and 161/756 Math-1.5B pools.

Predictions and chronology. The generators were selected after observing the Math-7B result, to test predictions about where weighting would help. Both plans fixed the cohort, rules, settings and predictions locally before generation. These are new draws on previously evaluated problems, not new unseen problem cohorts. Phi-4's predicted gain was $+.5$ to $+2.0$ points; the observed $+2.91$ exceeds that range. Its predicted tie share (20–35%) and plurality accuracy (25–40%) contain the observed 26.3% and 37.57%. Math-1.5B's $+1.19$ is below its predicted $+2$ to $+5$ range; its 21.3% tie share is below the predicted 32% threshold, and its paired effect difference has the opposite sign from the prediction. The plans assigned subjective probabilities .6 and .75 to a positive cohort mean. Both means are positive; two outcomes do not assess probability calibration. We report each prediction component, rather than treating a positive sign as validation of the full prediction.

Eight Phi-4 and six Math-1.5B interim inspections were requested during generation and recorded. They changed no setting or efficacy stopping rule; both studies reached all 756 intended decisions before their complete analysis. A preceding matched Qwen3.5 run was stopped at 131/756 pools; its partial records remain archived and are not a complete-study comparison. Phi-4 required a documented analysis-specification repair: generator-identity fields had been copied from that preceding study. The repair was staged during generation and activated after completion, before the full analysis. It preserved every generated record and the original generation identity, rules, estimand and uncertainty procedure. Independent reconstruction checks the final decisions, signed summaries and paired comparisons from the saved pools.

Execution used mixed hardware, recorded per pool: Phi-4 has 330 B200 and 426 RTX PRO 6000 Blackwell pools; Math-1.5B has 68 and 688. Model and scorer revisions, source identities, per-pool configurations, inspection logs and the analysis repair are retained in the evidence records. These experiments support finite-budget benefits for the specified generators and scorer. The population criterion is illustrated separately by the certified examples in Section F.7.

F.2 Generation-equivalent-budget plurality baseline

Declared question and budget convention. This extension asks whether the gain survives spending the scoring budget on more candidates instead. We keep each rule's original eight candidates and give plurality one or two additional independent draws on the same problems. The primary accuracy

contrast is weighted voting with eight candidates versus plurality with nine; recorded generation and scoring seconds justify the budget comparison. This adds the practical cost axis to the fixed-candidate voting comparison in Section 2.

The prospective study (Section F) compared rules on identical candidates and permitted no equal-compute claim, because the weighted rules pay for scoring while plurality does not. Its complete execution record fixes that cost: over all 756 cells, terminal PRM scoring took 380.2 s against 20,149.1 s of generation, a ratio of 0.01887; the generation-equivalent plurality budget is therefore $8 \times (1 + 0.01887) = 8.15$ candidates. The convention counts recorded generation and scoring seconds only; it does not establish matched end-to-end wall-clock or GPU time. The declared primary baseline is plurality at nine candidates (a 10.4% budget advantage to plurality); plurality at ten is a declared secondary.

Generation and pools. For every prospective task we generated one additional iid population of exactly two candidates under the identical generator, prompt, decoding, scorer and cohort, in a separate run namespace with the standard seed derivation at $N = 2$, so no seed coincides with any earlier population. The nine-candidate pool is the eight frozen candidates in launch order followed by the first extra candidate; the ten-candidate pool appends both. Pools are taken before eligibility, and every weighted rule uses only the frozen eight. The extended pools mix two independently seeded draws of the same sampler. The plan, analysis specification, launcher and consumer were internally declared and frozen before the first extra candidate was generated (freeze digest 5b8966746c52, recorded in the repository; no external timestamp); a four-cell technical pilot inspected no correctness; all 756 extension cells completed with 0 algorithm failures.

Results. Table 26 reports every declared contrast with pointwise 95% and family 98.33% problem-cluster intervals (the prospective study's bootstrap specification) and the exact conditional fair-sign p-value. The sole primary, raw-V weighting at eight candidates minus plurality at nine, is $+2.65$ points $[1.06, 4.37]$ (family $[0.66, 4.76]$; fair-sign $p = 0.00247$), which is positive. Against plurality at ten candidates the contrast is $+1.46$ $[-0.40, 3.31]$ (family $[-0.79, 3.70]$), positive in sign but inconclusive. The extra candidates raise plurality by $+0.13$ $[-0.40, 0.66]$ and $+1.32$ $[0.40, 2.38]$ points. The raw and DEV half-floors at eight candidates minus plurality at nine are $+2.12$ $[0.93, 3.44]$ and $+1.59$ $[0.66, 2.65]$. The prospective primary is reproduced exactly from the parent pools as a consistency check (+2.78 $[1.06, 4.63]$). An independent implementation that imports nothing from the producer rebuilt every contrast, interval and sign test from the raw populations (bootstrap index digest equal; all endpoints equal to 10^{-9}).

Table 26: Generation-equivalent-budget extension: every declared contrast on 252 problems $\times$ 3 runs (756 paired decisions). Pointwise 95% and family 98.33% problem-cluster intervals; exact conditional fair-sign two-sided p.

Contrast	Role	Δ (pp)	95%	98.33%	G/L	p
raw V N8 − plurality N9	primary	+2.65	[1.06, 4.37]	[0.66, 4.76]	27/7	0.00247
raw V N8 − plurality N10	secondary	+1.46	[-0.40, 3.31]	[-0.79, 3.70]	25/14	0.153
raw half floor N8 − plurality N9	secondary	+2.12	[0.93, 3.44]	[0.66, 3.70]	18/2	0.00106
dev half floor N8 − plurality N9	secondary	+1.59	[0.66, 2.65]	[0.40, 2.91]	13/1	0.00317
plurality N9 − plurality N8	descriptive	+0.13	[-0.40, 0.66]	[-0.53, 0.79]	3/2	1
plurality N10 − plurality N8	descriptive	+1.32	[0.40, 2.38]	[0.26, 2.51]	12/2	0.0107
raw V N8 − plurality N8 reproduced from parent pools	consistency	+2.78	[1.06, 4.63]	[0.79, 5.03]	28/7	0.00257

F.3 Association-sensitive forecasts on separate draws

Does knowing which wrong answer receives each score improve forecasts of voting accuracy? The paired populations in Section 3 and the identification result in Proposition 12 show why this information can matter even when the correctness–score law and answer frequencies are fixed. Here we compare forecasts that retain or remove that association, and measure how well each predicts the raw-weighting-minus-plurality accuracy difference at $N = 8$ on separate candidate draws. Lower forecast error measures the predictive value of the retained association for these finite decisions.

This retrospective analysis reuses the Math-7B audit/evaluation records of Section F.5: 120 problems with 256 audit candidates and sixteen evaluation pools each. Audit and evaluation use different

Table 27: Association-retaining versus association-removed forecasts of the eight-candidate raw-weighting effect. Positive mean ΔL favors retaining association. Units are pp^2; intervals are nominal 95% problem-bootstrap intervals conditional on the saved forecasts and selected sets. Every Math-7B screen and the full cohort are retained. The .5-point screen is the declared fallback.

Configuration	Audit screen	Problems	Mean ΔL	95% interval
Math-7B	Full cohort	120/120	$+1.43$	$[-.36, 3.55]$
Math-7B	.5 points (fallback)	18/120	$+11.93$	$[.79, 24.42]$
Math-7B	1.0 point	11/120	$+20.64$	$[4.25, 38.16]$

draws of the same problems. Earlier aggregate outcomes were known. The local plan fixed the loss, audit-only screen and fallback before this readout; it was not a new prospective generation study.

Forecasts and comparison. The joint forecast samples the labelled audit records with replacement. The association-removed forecast independently redraws each wrong eligible candidate's label from the empirical wrong-label distribution. It preserves the correctness–score law and answer frequencies at the distribution level while removing score–wrong-label association. Correct and ineligible records are unchanged; the intervention describes a synthetic law, not a change to the generator. Both forecasts use 20,000 eight-candidate draws, sharing candidate indices before wrong labels are redrawn. The sealed protocol described random permutations of a finite audit bank as equivalent; the executed code instead uses independent label resampling. These are different joint laws for a candidate pool. The preserved protocol and implementation reconciliation record this discrepancy; the results below concern the executed product coupling.

For problem i, let f_i^{joint} and f_i^{ind} be the saved forecasts and y_i the observed mean raw-weighting accuracy minus plurality accuracy on its evaluation pools, all in percentage points. We compare

$$\Delta L_i = (f_i^{\text{ind}} - y_i)^2 - (f_i^{\text{joint}} - y_i)^2.$$

Positive ΔL means lower squared forecast error when association is retained; its units are percentage points squared, not an accuracy gain. The screen uses only audit forecasts, $|f_i^{\text{joint}} - f_i^{\text{ind}}| \geq .5$ points. This Math-7B comparison was the declared fallback in the original broader protocol, retained with its complete results in the artifact. It remains a fallback analysis here. The one-point screen and full-cohort comparison are secondary comparisons. Each interval bootstraps problems 10,000 times, holding the saved Monte Carlo forecasts and screen membership fixed. These nominal 95% intervals do not propagate simulation error, audit-law estimation error or screening instability. Per-law Monte Carlo standard errors are recorded separately; they do not supply that missing propagation.

Results. In the Math-7B fallback, 18/120 problems pass; joint and association-removed squared losses average 32.29 and 44.22 pp^2, respectively. Their paired difference is 11.93 $[.79, 24.42]$ pp^2. The full-cohort comparison remains inconclusive (Table 27). The mean absolute forecast separation on the screened Math subset is 2.29 points; the mean joint-forecast Monte Carlo standard error is .23 points.

The selected Math comparison shows the practical value of the information highlighted by Proposition 12: keeping score–rival association improves prediction of these observed finite-vote effects, conditional on the estimated forecasts. The result concerns forecast accuracy on the selected Math subset; the population exponent criterion in Theorem 1 addresses a different endpoint.

Moment diagnostic. A secondary diagnostic asks whether the score advantage estimated from an audit persists on separate candidate draws. The conditional bonus gaps are the quantities in Theorem 1; the fixed-tilt exponential moments are ingredients of its error-exponent calculation (Section B.10). We choose a binned bonus and intensity on the audit law, then compare these quantities between audit and evaluation draws. This checks the stability of the estimated direction and moment, separately from the forecast-accuracy comparison above. The audited mode passes the frequency-interval gate on 34 Math-7B problems with a positive estimated direction; it matches the correct answer on 26 of them. Gaps are observable on only 196/482 possibly-leading rival pairs, because both labels must occur in the evaluation sample. The audit-to-evaluation mean gap is $.445 \to .425$; the mean moment is $.464 \to .501$ across the 482 pairs. The full pairwise values and within-problem bootstrap

Table 28: All rule comparisons in the count-tie study. Accuracies are percentages; effects are percentage points with nominal 95% problem-cluster intervals (10,000 bootstrap draws). The two baseline columns use first-occurrence and expected uniform-random tie-breaking, respectively. Half-floor denotes both proportional implementations.

Rule	Accuracy	Versus first occurrence	Versus random tie
Math-7B: 120 problems, 1,920 pools			
Plurality	37.03	0	$-.27$ $[-.71, .17]$
Random-tie plurality	37.30	$+.27$ $[-.17, .71]$	0
Selection	39.06	$+2.03$ $[.00, 4.22]$	$+1.76$ $[-.25, 3.95]$
Raw weighting	40.10	$+3.07$ $[1.30, 5.10]$	$+2.80$ $[1.07, 4.82]$
Count-first	39.27	$+2.24$ $[1.25, 3.39]$	$+1.97$ $[1.10, 2.96]$
Half-floor	39.06	$+2.03$ $[1.09, 3.07]$	$+1.76$ $[.86, 2.80]$

outputs are retained in the records. Pairs share problems, and the gap bootstrap omits draws missing either label. A moment below one does not establish improvement over plurality's exponent; these descriptives certify neither that comparison nor the unobserved rivals.

F.4 Predicting and locating the count-tie effect

Can scores improve decisions by breaking count ties while keeping every unique count winner? Section 5 and Proposition 1 identifies this route to finite-budget improvement under conditional bonus dominance. We evaluate the count-first rule against both first-occurrence and uniform-random tie-breaking, partition its changes by count pattern, and ask whether an audit-law forecast predicts the observed gain better than a forecast of zero. The rule comparisons measure the gain, the partition locates it, and the forecast assesses whether separate answer–score records anticipate it.

We use the audit/evaluation split of Section F.3: 120 Math-7B problems with 256 audit candidates and sixteen evaluation pools of eight. This retrospective analysis reuses previously inspected evaluation records. In particular, the Math-7B $+2.24$-point count-first contrast is the one already reported in Section F.5; the additions are its audit-law forecast, decomposition and alternative comparators. All pools, including those with no eligible candidate, remain in the denominators.

Comparators and count strata. Count-first refinement maximizes answer count, then summed scores among the tied leaders, then first occurrence. Random-tie plurality assigns expected correctness $\mathbf{1}\{a^\star \in T\}/|T|$ to a pool with tied maximum-count set T; it requires no extra random draw. We also evaluate highest-score selection, raw weighting and the half-floor. The two recorded half-floor forms satisfy $1 + 2(V - 1/2)_+ = 2\max(V, 1/2)$ and have identical saved correctness, so Table 28 reports them together. The implementation uses the existing 10^{-12} mass-tie tolerance and standardized fallback. Count-first preserves unique count leaders; raw weighting and the half-floor can overturn them.

Singleton ties have at least two eligible candidates and no repeated answer; repeated ties have at least two answers attaining a maximum count of two or more. The remaining pools have a unique count leader or no eligible candidate. In each stratum, the count-first contribution is its net correctness change divided by *all* evaluation pools. These contributions add to the whole-cohort effect; their intervals are not intervals for conditional accuracy within the selected stratum (Table 29).

On Math-7B, 30 of the 43 net additional correct decisions arise in singleton ties, where count-first and selection have identical correctness on all 460 pools. The other 13 arise in 133 repeated-tie pools: their whole-cohort contribution is $+.68$ $[.21, 1.20]$ points. Against selection in these repeated ties, count-first has nine gains and six losses, only three net additional correct decisions; no saved interval establishes superiority for that contrast. The correct answer is among the count leaders in 82 singleton and 48 repeated Math-7B ties. Count-first exceeds random-tie plurality by 1.97 $[1.10, 2.96]$ points.

This decomposition locates the Math gain in the tied decisions singled out by Proposition 1, and the random-tie comparison shows that it is not specific to the first-occurrence baseline. Favorable conditional score ordering is a possible population explanation, as described by the proposition; the observed gain alone does not determine whether that condition holds.

Table 29: Count-first minus first-occurrence plurality by count pattern. The observed contribution and its nominal 95% interval use the whole-cohort denominator; the forecast predicts that same contribution. The final three columns are within-stratum point effects in percentage points for count-first (CF), raw weighting and selection (Sel.).

Stratum	Pools	Observed contribution	Forecast	CF	Raw	Sel.
Math-7B						
No eligible	26	0	0	0	0	0
Unique leader	1,301	0	0	0	+1.31	−.08
Singleton tie	460	+1.56 [.68, 2.55]	+.76	+6.52	+6.52	+6.52
Repeated tie	133	+.68 [.21, 1.20]	+.49	+9.77	+9.02	+7.52

Forecast test. For each problem, 20,000 iid Monte Carlo pools of eight from the full empirical audit law estimate the count-first gain f and its stratum contributions. Let y be the gain on the separate evaluation draws, both measured in percentage points. We test

$$\Delta L = y^2 - (f - y)^2,$$

so a positive mean indicates lower squared error than forecasting zero gain. This explicit formula is the producer's convention; the plan's verbal subtraction order is reversed. On Math-7B, the mean forecast is 1.25 points versus 2.24 observed, and mean ΔL is 19.87 pp^2 $[9.15, 33.06]$. The mean per-problem Monte Carlo standard error is .076 points. Bootstrap intervals condition on the saved forecasts; they omit Monte Carlo and audit-law estimation uncertainty. The secondary comparisons are not adjusted for multiplicity. The test evaluates the full joint-law forecast, including tie frequencies and all leaders, rather than a two-label statistic alone.

Two-label score-sum diagnostic. This diagnostic separates two notions of useful score information. Theorem 1 compares conditional mean bonuses for an error-exponent gain, whereas count-tie refinement compares sums of an equal number of scores (Proposition 1). We measure both on the same conditional empirical laws to see when their signs agree. For the correct label and the most frequent observed wrong label in the audit, let $S_c^{(r)}$ and $S_a^{(r)}$ be sums of r independent scores from their respective conditional empirical laws. We compute

$$T_r = \mathbb{P}(S_c^{(r)} > S_a^{(r)}) + \tfrac{1}{2}\mathbb{P}(S_c^{(r)} = S_a^{(r)}) - \tfrac{1}{2}, \qquad r \in \{1, 2\},$$

using the recorded 10^{-12} tie tolerance. This measures the two-leader refinement advantage; it need not have the sign of the mean-score gap used in the exponent criterion. Among 92 Math-7B problems where both labels appear, three have a positive mean gap but $T_1 \le 0$, three have $T_1 > 0$ but a nonpositive gap, 58 have both positive and 28 have neither positive. Positive T_2 occurs on 62 problems. These comparisons do not verify stochastic dominance over every rival or the exponent-improvement premises. Under exchangeable iid sampling, first-occurrence and uniform-random tie-breaking have equal expected correctness; the small simulated differences between those baselines are Monte Carlo error.

F.5 Score directions estimated on separate draws

Can a bonus estimated from one set of candidates improve voting on separate draws of the same problems? Theorem 1 identifies an improving population direction through separation of conditional score laws. Here we fit a binned version of that direction from audit records, choose its intensity using only those records, and compare its accuracy with plurality on evaluation pools. We also measure agreement with the audit's most frequent answer, so we can distinguish improving that fitting target from improving correctness.

This retrospective analysis reuses Math-7B AIME records; earlier aggregate outcomes were already available. The analysis plan and code were sealed before this readout, not before the evaluation records were generated or first inspected.

Samples and estimands. For Math-7B, 120 problems with repaired-PRM records (60 development, 60 test) supply 256 audit candidates from rounds r0–r1; round r2 is divided into consecutive pools

of 8, 16 or 32 (*M split*). These problems are disjoint from the prospective Math-7B cohort. No answer-key labels enter direction or intensity fitting. We report accuracy against the answer key and, separately, agreement with the estimated audit mode. The estimated audit mode matches the correct answer on 51/120 Math-7B problems; it need not be the theorem's correct population mode.

Audit frequencies and conditional score laws use boxed candidates with non-null answers. Budgets count candidates before eligibility. All pools remain in the accuracy denominator, including the 26 eight-candidate pools with no eligible candidate. Plurality, selection, count-first refinement, ties and fallbacks use the recorded readout. The Math-7B contrasts were secondary or descriptive in the original broader reanalysis and retain that status here; the complete protocol and results remain in the artifact. Intervals resample problems 10,000 times; they are nominal 95% percentile intervals without adjustment across the secondary grid.

Computing a direction. The audit uses four score bins with edges $(.08, .35, .89)$, obtained from pooled Math-7B audit-score quartiles. Let $\hat{a}$ be the audit mode and $\widehat{J}$ contain each observed rival whose upper frequency bound reaches the largest rival lower bound. The bounds are simultaneous Clopper–Pearson intervals (Clopper & Pearson, 1934) with error allocation $\delta/2$ across 1001 classes (integers 0–999 and one merged remainder), with $\delta = .05$. The finite-bin program is

$$\widehat{G} = \max_{0 \le b_s \le 1} \min_{a \in \widehat{J}} \sum_s b_s \{\widehat{Q}_{\hat{a}}(s) - \widehat{Q}_a(s)\}.$$

It separates the estimated mode from the possibly leading observed rivals. Its expanded rival set differs from the exact leading-rival set defining D in Theorem 1. A monotone program adds $b_1 \le \cdots \le b_k$. An estimated direction is available only for a unique audit mode and positive program value; otherwise the rule uses plurality. We report intensity $\eta = 1$ separately from an intensity chosen from $\{.25, .5, 1, 2, 4, 8\}$ by agreement with the audit mode on 2,000 audit subsamples of eight. If fewer than eight audit candidates are eligible, the default is $\eta = 1$. Evaluation uses weights $1 + \eta b(V)$; no evaluation outcome selects the reported intensity.

Accuracy on separate draws. Count-first refinement on Math-7B gains 2.24 points $[1.25, 3.39]$ at $N = 8$: 754 versus 711 correct decisions among 1,920 pools, with 27 net-positive and five net-negative problems (fair-sign $p = .000113$). At $N = 16$ its gain is 1.15 $[.21, 2.29]$; the fair-sign sensitivity is $p = .0703$. At $N = 32$ the difference, $+.62$ $[-.21, 1.67]$, is inconclusive. Table 30 reports the Math-7B comparisons at all three budgets; the records contain the full rule grid.

The unrestricted Math-7B bonus at $\eta = 1$ changes accuracy against the answer key by $+.10$ $[-.36, .52]$, inconclusive, and audit-mode agreement by $+.89$ $[.31, 1.51]$. Selection changes the same outcomes by $+2.03$ $[0.00, 4.22]$ and -4.22 $[-7.24, -1.41]$, respectively. These contrasting endpoints show why the reference answer matters: the fitted bonus targets the audit mode, which matches the correct answer on only 51/120 Math-7B problems, whereas Theorem 1 concerns the correct population mode. The descriptive Spearman correlation between $\widehat{G}$ and the eight-candidate agreement difference is $-.10$ on M split.

The small observed effects also distinguish a positive estimated gap from an effective finite-budget rule. The theory concerns population gaps and a sufficiently weak intensity in the large-budget error rate; estimating the reference mode and score laws, choosing intensity, and obtaining decision changes at eight votes are separate steps. These are possible reasons a fitted positive direction need not yield a detectable eight-vote gain here, rather than evidence against the population criterion.

Audit status and coverage. A separate question is whether the fitted direction remains positive after allowing for sampling uncertainty in the separation criterion of Theorem 1. We apply a nominal certification screen: subtract uncertainty allowances, require a uniquely modal reference answer, and check every possibly leading rival. The pass/fail counts locate the remaining uncertainty. Because these bins were selected from the same scores, the outputs below are nominal statuses. The radius-adjusted program subtracts $r_{\hat{a}} + r_a$ from each constraint, where

$$r_a = \min\left\{1, \sqrt{\frac{\log((2^k - 2)/\delta_t)}{2n_a}}\right\}, \qquad \delta_t = \frac{\delta/2}{1 + |\widehat{J}|},$$

Table 30: Retrospective evaluation of estimated score directions. Entries are accuracy differences from plurality in percentage points, with nominal 95% problem-cluster bootstrap intervals (10,000 draws); the plurality row gives accuracy. All reported Math-7B contrasts are secondary or descriptive. Radius-screened rules use the recorded nominal statuses; real-data coverage is not established (Section F.5).

Math-7B (audit 256 draws)			
Rule	r2, $N{=}8$	r2, $N{=}16$	r2, $N{=}32$
Pools	1920	960	480
Plurality (accuracy %)	37.03	39.58	41.67
Highest-score selection	$+2.03$ $[0.00, 4.22]$	$+1.46$ $[-1.35, 4.38]$	$+2.71$ $[-0.62, 6.25]$
Count-first refinement	$+2.24$ $[1.25, 3.39]$	$+1.15$ $[0.21, 2.29]$	$+0.62$ $[-0.21, 1.67]$
Estimated bonus, $\eta = 1$	$+0.10$ $[-0.36, 0.52]$	$+0.00$ $[-0.31, 0.31]$	$+0.21$ $[-0.42, 0.83]$
Estimated bonus, audit η	$+0.10$ $[-0.62, 0.73]$	$+0.31$ $[-0.10, 0.83]$	$+0.62$ $[-0.21, 1.88]$
Monotone bonus, audit η	$+0.42$ $[-0.05, 1.09]$	$+0.42$ $[0.00, 0.94]$	$+1.04$ $[0.00, 2.50]$
Radius-screened bonus	$+0.00$ $[0.00, 0.00]$	$+0.00$ $[0.00, 0.00]$	$+0.00$ $[0.00, 0.00]$

with $k \geq 2$ the number of score bins and $n_a > 0$ the conditional sample count for answer a. The factor $2^k - 2$ counts the nontrivial subsets of the fixed partition: the signed difference between the empirical and true bin distributions has total mass zero, so their total variation distance equals $\max_S\{\widehat{Q}_a(S) - Q_a(S)\}$ over subsets S of bins, and one-sided Hoeffding (Hoeffding, 1963) for each nontrivial S gives $\mathbb{P}(\mathrm{TV}(\widehat{Q}_a, Q_a) > r) \leq (2^k - 2)e^{-2n_a r^2}$, which is the displayed radius; Weissman et al. (2003, Theorem 2.1) state a distribution-dependent version in L_1 distance, twice this total variation, and substituting $\epsilon = 2r$ there with their $\varphi \geq 2$ recovers the same bound. It returns a positive status only if the adjusted value is positive, the mode's lower frequency bound exceeds every rival upper bound, and an unseen rival cannot be leading. For a fixed partition, the radius controls all bonuses in $[0, 1]$ simultaneously, conditional on the label sequence. The real-data partitions were selected from those same audit scores; this fixed-partition argument does not establish coverage for that selection. We therefore report their outputs as *nominal statuses*, not population certificates.

There is one nominal positive among 120 M-split problems; it changes no evaluation correctness. On 65 problems the audit fails to establish a unique mode. Table 31 reports every status. These counts describe this procedure and these samples; they are neither population impossibility results nor lower bounds on required audit size.

F.6 Constructed-law tests of the score-law programs

How do the score-law programs behave when the population law is known? We use five specified finite laws to test the separation criterion in Theorem 1, including the distinction between unrestricted and nondecreasing bonuses. Exact program values identify which weight classes admit exponent improvement, and optimized exponents quantify it. Repeated audit samples measure positive-certificate frequencies against this known truth at each sample size. Laws with no such direction expose false positive outputs.

Unlike the real-data audit, the score partitions are fixed before sampling. Each of the 17 law–sample-size settings uses 200 independent multinomial samples. The frequency intervals allocate $\delta/2$ over the law's known finite answer alphabet, and the conditional-law radii use the other half as above. A positive adjusted value certifies separation, at nominal error level .05, on the simultaneous frequency-and-score-law event. It concerns exponent improvement at a sufficiently weak intensity, not accuracy at a specified budget or intensity. The observed success frequencies describe these five laws and sample sizes.

Laws and the role of score ordering. The first two laws are Populations A and B of Table 1. A third, separated law gives scores .15 and .85 equal probability, correctness probability equal to the score, and divides incorrect mass equally between two rivals. It has $D = .7$ and plurality exponent .04384, compared with 7.46×10^{-5} in A and B.

Table 31: Audit program outputs and fixed-law certification experiments at $\delta = .05$. Left: nominal unrestricted statuses on the AIME audit records (monotone counts agree); data-selected bins prevent a fixed-partition coverage claim. Right: positive statuses among 200 independent samples per fixed-law setting, unrestricted (U) and monotone (M). Population values are $D = .187$ (A), 0 (B), .7 (separated), .526 with monotone value zero (calibrated nonmonotone law), and .1 with monotone value zero (mixing). The two isolated false positives are retained.

Status	Math-7B
Draws	256
Nominal positive	1
Gap within radius	2
Rival support too small	7
Unseen rival possible	44
Mode not certified	65
No observed rival	1
No eligible draw	0
Problems	120
Estimated direction available	56
Audit mode correct	51

Law	n	U	M
Population A	10^3	0	0
Population A	10^4	0	0
Population A	10^5	140	140
Population A	10^6	200	200
Population B	10^4	0	0
Population B	10^6	1	0
Separated calibrated	250	186	186
Separated calibrated	10^3	200	200
Separated calibrated	4,000	200	200
Separated calibrated	16,000	200	200
Calibrated counterexample	10^4	200	0
Calibrated counterexample	10^5	200	0
Calibrated counterexample	10^6	200	0
Calibrated counterexample	10^7	200	0
Three-bin mixing	10^4	200	1
Three-bin mixing	10^5	200	0
Three-bin mixing	10^6	200	0

A fourth law is calibrated but requires a nonmonotone bonus for an exponent gain. Its joint masses are

	$V = .2$	$V = .3$	$V = .4$
$a^\star$	.120	.060	.080
R	.024	0	.120
each of 20 other rivals	.0228	.007	0

Here $p = .26$, $q = .144$, $D = .52564$, and the monotone program value is zero: the leading rival's conditional score distribution dominates the correct answer's in upper tails. Hence no bounded positive nondecreasing weight improves the exponent. The unrestricted optimum bonuses the lower two bins, $b = (1, 1, 0)$; at $\eta = .25$ the exponent rises from .01716 to .02449. A middle-bin-only bonus also improves it, to .02046. Thus exact calibration and useful score information do not by themselves imply that larger scores should receive larger bonuses.

The fifth law illustrates why mixtures, rather than single rivals, matter. Its joint rows are $(.18, .04, .18)$ for the correct answer, $(.24, .06, 0)$ and $(0, .06, .24)$ for two leading rivals, on three ordered score bins. The correct conditional law is at TV distance .45 from either rival, but only .10 from their equal mixture. The unrestricted optimum is $b = (1, 0, 1)$ and the monotone value is zero. This law is not calibrated.

Certification frequencies. For A, none of the 200 samples certifies at 10^3 or 10^4 draws, 140 certify at 10^5, and all 200 at 10^6. The small frequency margin $.34 - .33$ prevents mode certification at the two smaller settings. In the separated law, 186/200 certify at 250 draws and 200/200 at 1,000. The calibrated nonmonotone law certifies unrestricted directions in all 200 samples at each tested size from 10^4 to 10^7, and monotone directions in none. The mixing law similarly separates the programs. Population B produces one false unrestricted certification among 200 samples at 10^6; the mixing law produces one false monotone certification among 200 at 10^4. These are observed simulation counts, not proofs of the procedure's error level. Table 31 gives all settings.

Computational checks. The records supply the joint laws, primal–dual programs, monotone counterpart, confidence-radius calculations and Chernoff optimization. A separate verification script reports 17 passing checks, covering the score-law and derivative identities, six exponent entries, the half-floor expansion, exact two- and three-vote probabilities, tie-independent losses at $N = 65$ and 129, 96 small program checks, and unseen-rival handling. These calculations support the stated examples; proofs establish the general results.

F.7 CERTIFICATES WITH DATA-SELECTED SCORE BINS

Can the separation criterion in Theorem 1 be certified when the score bins are chosen from the audit itself? The earlier nominal statuses in Section F.5 use such bins. Here a uniform conditional-CDF bound accounts for their selection. We compare fixed bins, bins learned from separate rounds, and bins learned from the certification audit; for each valid bound, the readout is the number of problems with a certified direction. Known-law simulations then check recovery and false positive outputs against known truth. This retrospective analysis uses the same stored audit records and keeps the earlier outputs unchanged.

A bound uniform over bin boundaries. For an eligible answer label a, write F_a for its conditional score CDF, $\widehat{F}_a$ for the empirical CDF from n_a occurrences, and Q_a for its score law. Conditional on the label sequence, iid scores satisfy the Dvoretzky–Kiefer–Wolfowitz bound (Massart, 1990). Allocate $\delta/2$ to the frequency intervals and the other half across the audit mode c and the possibly-leading observed rivals J_{obs}. With

$$\delta_t = \frac{\delta/2}{1 + |J_{\mathrm{obs}}|}, \qquad \epsilon_a = \min\left\{1, \sqrt{\frac{\log(2/\delta_t)}{2n_a}}\right\},$$

the event $\|\widehat{F}_a - F_a\|_\infty \le \epsilon_a$ holds simultaneously for these labels with conditional probability at least $1 - \delta/2$. The mode and rival set depend only on the label counts, so this conditional allocation remains valid after their selection.

For a step bonus with ordered-bin values $b_1, \dots, b_k \in [0, 1]$, write $V(b) = \sum_{s=1}^{k-1} |b_{s+1} - b_s|$ for its variation. Expressing the step function as a constant plus its successive jumps gives, on that same event,

$$|(\widehat{Q}_a - Q_a)b| \le V(b)\epsilon_a.$$

This bound holds for every boundary choice and every bonus, including those chosen from the audit scores. The certificate maximizes

$$\max_{0 \le b_s \le 1} \min_{a \in J_{\mathrm{obs}}} \left\{(\widehat{Q}_c - \widehat{Q}_a)b - V(b)(\epsilon_c + \epsilon_a)\right\}.$$

Auxiliary variables for the absolute successive differences make this a linear program, solved with the HiGHS solver (Huangfu & Hall, 2018) through SciPy. The monotone restriction has $V(b) \le 1$; an unrestricted four-bin bonus can have variation up to three. A positive value, together with the simultaneous mode-frequency gate and exclusion of unobserved leading rivals, certifies an exponent-improving direction at sufficiently weak bonus intensities relative to the reference mode under the iid premises. Correctness of that reference answer must be checked separately. At $\delta = .05$, coverage is per problem, not simultaneous across problems, regimes or sample sizes. The certificate does not select a finite-budget accuracy guarantee.

Three binning regimes. The deterministic equal-width edges are $.25, .5, .75$. Split-estimated edges are pooled score quartiles from round r0, with certification using only later rounds: Math-7B uses 128 audit candidates, versus 256 for the other regimes. Finally, each problem's own audit quartiles are used with the variation bound. The first two regimes also permit the fixed-partition multinomial bound. All counts use the same 1,001-class frequency allocation and unseen-rival check as Section F.5; the observed eligible labels in these records are canonical integers.

Table 32 gives every valid combination. On both certified Math-7B problems, the most frequent answer in the audit data was the correct answer. For problem 2006-II-3, the variation-adjusted margin is $.406348$ with equal-width bins and $.362870$ with same-audit quartiles. The latter regime also certifies 2002-II-1 with the much smaller margin $.001481$. All these witnesses are monotone zero–one bonuses of variation one. They provide concrete score directions linked to the sufficient condition in Theorem 1, with bin selection included in the uncertainty bound. Split-estimated bins certify no Math-7B problem; their smaller audit samples prevent attributing this difference to binning alone. A zero yield means that the audit does not clear the certificate's uncertainty gates, rather than identifying the population separation as zero. Full failure statuses, bin edges, directions and radii are retained per problem.

Table 32: Positive outputs in the selection-aware certificate analysis. Counts agree for unrestricted and monotone programs. The first two regimes use fixed partitions relative to their certification sample; same-audit quartiles require the variation bound. These are yields, not measurements of real-data coverage or finite-budget accuracy.

Bin regime	Bound	Math-7B (120)
Equal width	Fixed-bin	1
Equal width	Variation	1
Split-estimated quartiles	Fixed-bin	0
Split-estimated quartiles	Variation	0
Same-audit quartiles	Variation	2

Worked example: a certified bonus on an AIME problem. We expand the already-reported certificate for AIME 2006-II-3 to show how Theorem 1 applies to actual model outputs. The problem asks for the largest k such that 3^k divides $1 \cdot 3 \cdot 5 \cdots 199$. Counting factors of 3, 9, 27 and 81 gives $k = 33 + 11 + 4 + 1 = 49$. This answer is checked independently of the score certificate.

The audit uses the 256 Math-7B candidates from development rounds r0 and r1; all have eligible answers. Table 33 shows the observed answer–score distribution. The correct answer 49 appears 188 times, the leading wrong answer 48 appears 46 times, and each other answer appears at most four times. These are sample counts; the frequency bounds below connect them to the population premises of the theorem.

Answer	$V < 1/2$	$V \geq 1/2$	Total	Mean bonus
49 (correct)	17	171	188	$171/188 \approx .9096$
48 (leading wrong answer)	39	7	46	$7/46 \approx .1522$
Other 13 answers, combined	10	12	22	—
All answers	66	190	256	—

Table 33: A real-problem application of the positive-bonus criterion: Math-7B on AIME 2006-II-3. Entries count audit candidates; the bonus is $b(v) = \mathbf{1}\{v \geq 1/2\}$. The other answers are grouped only for display; the certificate treats each wrong label separately.

Step 1: identify the relevant rival. Rounding outward, the simultaneous frequency intervals give $p_{49} \geq .606832$ and $.093181 \leq p_{48} \leq .296734$. Every other observed label has upper bound at most .077091, and an unseen label has upper bound .043147. Thus, on the frequency confidence event, 49 is the correct population mode and $J = \{48\}$: only 48 can be a leading rival.

Step 2: certify a separating bonus. The equal-width-bin witness is $b(v) = \mathbf{1}\{v \geq 1/2\}$, with total variation one. Its observed mean gap is

$$\widehat{\mu}_{49} - \widehat{\mu}_{48} = \frac{171}{188} - \frac{7}{46} \approx .757401.$$

The selection-aware score bounds have radii .116180 and .234872. On their confidence event, the population gap therefore satisfies

$$\mu_{49} - \mu_{48} \geq \frac{171}{188} - \frac{7}{46} - \sqrt{\frac{\log 160}{376}} - \sqrt{\frac{\log 160}{92}} \approx .406348 > 0.$$

Together with Step 1, this is exactly the specified-bonus condition in Theorem 1.

Step 3: translate separation into an improvement. For all sufficiently small fixed $\eta > 0$, the rule $h_\eta(v) = 1 + \eta \mathbf{1}\{v \geq 1/2\}$ has a strictly larger error exponent than plurality. In plain terms, giving every candidate one vote and adding a small bonus for a score of at least one half makes errors decrease faster as the number of independent samples grows. The theorem requires separation from the leading rival, so a few high-scoring candidates with rare wrong answers do not invalidate this conclusion at sufficiently weak intensity.

For a numerical illustration, define a finite population law by the audit frequencies: each answer and score region receives its observed count divided by 256. Keeping all 14 wrong labels distinct and

taking the minimum of their Chernoff exponents gives

$$I(0) = .207691, \qquad I(.1) = .230635, \qquad I(.25) = .261745,$$

with 48 the limiting rival at each intensity. Here $I(\eta)$ is the error exponent of $1 + \eta b(V)$ for this empirical law treated as the population; these numerical rates are not certified rates for the generator. The statistical certificate above establishes the improving direction for the underlying law at sufficiently weak intensity.

This is a retrospectively selected illustration of the existing audit. The certificate uses $\delta = .05$ per problem under iid sampling; coverage is not simultaneous across all 120 audited problems. Its conclusion concerns the error exponent, with the usable intensity range depending on the population law. In the separate evaluation round r2, plurality, raw weighting and count-first each return the correct answer on all 16 eight-candidate pools for this problem.

A complementary example: one problem, three generators. AIME 2011-I-15 asks: if $x^3 - 2011x + m$ has three integer roots a, b, c, what is $|a| + |b| + |c|$? The correct answer is 98. This problem provides a concrete view of how scores change finite decisions across generators. It was selected retrospectively for exposition: among the 252 shared problems, it is the only one with a positive net raw-weighting gain for all three generators. It is an illustration of the observed voting rules; the statistical certificate for the population criterion is the separate 2006-II-3 example above.

Table 34 shows the three pools in which raw weighting corrects an error. Math-7B has three eligible answers, each appearing once: 90, 98 and 102. First-occurrence plurality chooses 90, whereas the score .9609 selects 98. Phi-4 instead has a wrong count leader: 100 appears twice and 98 once. The single correct answer has more score mass (.9297) than the two wrong-leader candidates together (.3533). Math-1.5B supplies another count override: three occurrences of 90 have total score .4727, while two occurrences of 98 have total 1.6484. Thus the same final correction can come from resolving a count tie or from overcoming a one-vote lead. The aggregate decomposition in Table 25 distinguishes these mechanisms.

Answer	Count	Raw-score sum	Half-floor sum
Math-7B, r1: 3 eligible of 8 generated candidates			
90	1	.2812	.5000
98	1	.9609	.9609
102	1	.1797	.5000
Phi-4, r0: 7 eligible of 8 generated candidates			
90	1	.0559	.5000
98	1	.9297	.9297
100	2	.3533	1.0000
112	1	.0972	.5000
102	1	.2490	.5000
118	1	.1138	.5000
Math-1.5B, r0: 8 eligible of 8 generated candidates			
92	1	.4883	.5000
42	1	.0610	.5000
90	3	.4727	1.5000
98	2	1.6484	1.6484
100	1	.0588	.5000

Table 34: Answer groups in the three correcting pools for AIME 2011-I-15. The correct answer is bold. Every eligible answer group is shown, in first-occurrence order; remaining candidates have no eligible boxed answer. Scores are rounded for display. The raw half-floor uses $h(v) = \max\{v, 1/2\}$. Each pool has eight generated candidates.

Table 35 gives all nine runs for this problem. Raw weighting adds one correct decision per generator. The half-floor also corrects Math-7B's tie and Math-1.5B's one-vote lead, but retains Phi-4's wrong leader: its two half-votes sum to 1.0000, above the correct answer's .9297. This exhibits the tradeoff between limiting count overrides and permitting a sufficiently strong score signal to change a vote. The table retains the unchanged errors and the Math-1.5B wrong-to-wrong change. None of these individual pools establishes calibration, a population mode or an error exponent.

Generator	Run	Plurality	Raw V	Half-floor	Raw change
Math-7B	r0	314	314	314	Unchanged
	r1	90	**98**	**98**	Tied maximum
	r2	**98**	**98**	**98**	Unchanged
Phi-4	r0	100	**98**	100	Count override
	r1	**98**	**98**	**98**	Unchanged
	r2	**98**	**98**	**98**	Unchanged
Math-1.5B	r0	90	**98**	**98**	Count override
	r1	92	90	92	Count override
	r2	92	92	92	Unchanged

Table 35: All three runs for each generator on the retrospectively selected AIME 2011-I-15 example. Correct decisions are bold. The full-cohort results, rather than this selected problem, measure efficacy.

Table 36: New known-law runs: positive certificates out of 200 samples per size. Counts coincide for the unrestricted and monotone programs and for both valid bounds where applicable. Discrete laws use their fixed score bins; the continuous null uses equal-width bins and same-audit quartiles with the variation bound. Earlier runs in Table 31 retain their original counts.

Law	Audit sizes	Positive certificates
Population B (null)	$10^4, 10^6$	$0, 0$
Three-bin hull null	$10^4, 10^5$	$0, 0$
Continuous null	$10^3, 10^4, 10^5$	$0, 0, 0$
Population A	$10^5, 10^6$	$130, 200$
Separated calibrated law	$250, 1000$	$181, 200$

Known-law checks. The population criterion in Theorem 1 supplies the truth for these checks: null laws have no separating direction, and positive laws have one. Repeating the audit measures how often the certificate fires incorrectly or recovers a direction at each sample size. The continuous null also exercises bins chosen from the same scores. Each setting uses 200 independent audit samples. The null laws are Population B, a three-bin hull-null law with joint rows $(.16, .08, .16)$, $(.21, .06, .03)$ and $(.03, .06, .21)$ on scores $.2, .4, .6$, and a continuous law with answer probabilities $.55, .45$ whose two conditional score laws are both $\mathrm{Beta}(2, 2)$. The positive laws are Population A and the separated calibrated law in Section F.6. The simulations use each law's known finite answer alphabet.

Every tested null setting has zero positive outputs (Table 36). As an explicitly invalid comparison, the fixed-bin bound was also applied to same-audit quartiles: it yields one positive Math-7B output and none on the continuous null samples. The observed zero on the continuous null does not validate that procedure; uniformity over boundary selection comes from the CDF argument. The known-law frequencies illustrate power at the listed sample sizes, without providing a universal audit-size rate or a stopping rule.

F.8 Learning a weight with separate problem roles

Can a weight chosen using some problems improve accuracy on different problems? Theorem 1 characterizes whether exponent-improving weights exist for a known population law. This experiment evaluates the finite-budget accuracy of a restricted score-bin family when the weight must be learned. Training problems nominate weights, development problems select a rule or retain plurality, and test-role problems measure the selected rule's paired accuracy difference from plurality. This separates fitting, selection and evaluation within the stored experiment.

Learning aggregation is established: PAC-Bayesian bounds support learning classifier-vote weights (Masegosa et al., 2020), and RISC learns answer rankings from features relative to the candidate set (Marina et al., 2026). Our comparison concerns the scalar score-bin family and retrospective problem roles described below.

We retrospectively divide the stored problems into training, development and test roles: 60/30/30 Math-7B problems, using a seeded permutation within the existing strata. Fitting and selection

use disjoint problems; the test-role evaluation uses sixteen eight-candidate pools per Math-7B problem. These records and earlier whole-cohort outcomes had already been inspected. Preprocessing is also shared: the inherited bin edges, $(.08, .35, .89)$, came from pooled audit-score quartiles across problems now assigned to all three roles. The analysis therefore separates roles within this retrospective procedure, rather than providing a new blinded test.

Candidate weights and selection. The frozen lattice contains weights $1 + \eta b(V)$ for four-bin bonuses $b_s \in \{0, 1/2, 1\}$, excluding constant bonuses, and $\eta \in \{1/2, 1, 2, 4\}$. Its 312 candidates partition into 48 nondecreasing and 264 nonmonotone weights. Training ranks them by accuracy against the answer key, averaged over 500 audit-resampled pools of eight per problem. The five best from each group, plus count-first, raw weighting and the standard half-floor, give 13 development nominees. Plurality is the constant control. Development gains $\widehat{g}_\ell$ use 500 audit-resampled pools per development problem.

The declared primary screen requires

$$\widehat{g}_\ell - r > 0, \qquad r = \sqrt{2\log(13/.05)/J_{\mathrm{dev}}},$$

with gains measured as proportions. It chooses the passing rule with largest development gain, or retains plurality if none passes. The radius is 60.89 percentage points; none of the 13 Math-7B nominees passes. Retaining plurality is the primary selection outcome, not an abstention from answering the problem. These screens concern resampled audit gains; shared preprocessing and finite audit laws preclude reading them as certificates of gains on fresh generator draws.

A secondary screen, included in the original declaration, instead requires the $.05/13$ quantile of a 10,000-draw development problem-bootstrap distribution to exceed zero. This is a nominal bootstrap screen, not a finite-sample simultaneous guarantee. Its adjustment covers each development nominee set, not all objectives, configurations and test comparisons. Six Math-7B nominees pass (five under exact training-gain ranking with stable candidate-order ties; the plan specified no nomination tie convention, and the sealed run resolved one exactly tied training score by floating-point summation order, as its filed addendum records). Maximizing development gain among them selects $b = (0, 0, 1/2, 1)$ at $\eta = 2$, or bin weights $(1, 1, 2, 3)$.

Test-role results. Plurality accuracy is 38.125% on 480 Math-7B pools. The secondary-selected Math weight reaches 41.04%, a gain of 2.92 points $[.21, 6.25]$ under the nominal 95% test problem-cluster bootstrap. The primary remains plurality under the primary screen. Table 37 reports every nominee for the correct-answer objective; the fixed-rule contrasts are inconclusive, and no direct comparison establishes that the selected weight outperforms those rules. The positive result is a secondary finding on 30 test-role problems with the retrospective provenance above.

This finding shows that a weight selected on different problem roles can retain an observed accuracy gain in this finite-budget evaluation. It complements Theorem 1: the theorem identifies when a known per-problem law admits an exponent-improving weight, while this test measures one common learned weight across problems. The primary screen's retention of plurality follows from its recorded uncertainty radii; it does not establish that useful weights are absent.

Agreement objective and oracle comparison. Two additional comparisons clarify what the learner is optimizing and how much room remains within its nominees. The agreement objective uses the most frequent audit answer as its target; the oracle uses correctness outcomes on the test problems to select the best nominee separately for each problem. Comparing these with the rule trained on correct-answer labels distinguishes a label-free target from correctness and from hindsight selection. Replacing accuracy against the answer key with agreement with each problem's audit mode again makes the primary screen retain plurality. Its bootstrap alternative selects count-first on Math-7B. This is a different objective from correctness. An oracle choosing separately for each test problem, using that problem's correctness outcomes, attains 43.96% on Math-7B (44.17% under exact training-gain ranking with stable candidate-order ties). This maximizes only over the 13 nominees for the correct-answer objective plus plurality, not the full lattice, and is not a deployable rule.

Execution record. The sealed declaration includes both selection screens. The executed subsampling resets the generator for each problem with an added 100000 times its ordinal, differing from the declared role-level stream. A sentinel collision caused the first execution's validation to fail after

Table 37: Every nominee for the correct-answer objective evaluated on test-role problems. Entries give accuracy changes from plurality in percentage points with nominal 95% problem-bootstrap intervals. For lattice weights, b lists the four bonus values and η is the intensity in $1 + \eta b$. A star marks the development bootstrap screen's selected weight; the primary screen retains plurality. All 13 Math-7B nominees are reported.

Bonus or rule	Intensity	Test effect [95% interval]
Math-7B: 30 problems, 480 pools; plurality 38.125%		
$(0, 0, 1/2, 1)$	4	$+3.12$ $[-1.04, 8.33]$
$(0, 0, 1/2, 1)$	2 ($\star$)	$+2.92$ $[.21, 6.25]$
$(0, 1/2, 1/2, 1)$	4	$+1.88$ $[-.62, 4.58]$
$(0, 0, 1, 1)$	4	$+2.50$ $[-1.67, 7.30]$
$(0, 0, 1/2, 1)$	1	$+1.88$ $[-.21, 4.58]$
$(1/2, 0, 1, 1)$	4	$+1.67$ $[-1.67, 5.83]$
$(1/2, 0, 1, 1)$	2	$+1.67$ $[-.62, 4.79]$
$(1/2, 0, 1/2, 1)$	4	$+2.92$ $[-.21, 7.71]$
$(1/2, 0, 1, 1)$	1	$+1.46$ $[-.21, 3.54]$
$(1/2, 0, 1, 1)$	$1/2$	$+.83$ $[-.42, 2.29]$
Count-first	—	$+1.25$ $[-.62, 3.54]$
Raw weighting	—	$+2.71$ $[-2.50, 8.75]$
Half-floor	—	$+1.88$ $[-.21, 4.58]$

results were displayed; its log is retained, followed by the recorded repair and reseal. These implementation reconciliations accompany the unchanged declaration and do not turn the retrospective analysis into a prospective study.

G POST HOC COUNT-THEN-SCORE READOUT AT EIGHT CANDIDATES

Question and status. Does using scores only to break count ties improve the Math-7B replication, and how much of raw weighting's gain does it recover? Section 5 and Proposition 1 motivates this policy: keep every unique count winner and use summed scores among tied leaders. We apply the complete rule to every saved pool from Section F, retaining the same 252 problems, three runs per problem, and 756 decisions. Paired comparisons with plurality measure the value of tie refinement; comparisons with raw weighting measure the effect of also allowing lower-count answers to win. A count-pattern decomposition then locates the decisions responsible for each difference.

The rule was proposed after all original prospective outcomes had been inspected. This is a *post hoc* readout and does not inherit the prospective status of the original rules. Its implementation ran before the requested review and had no pre-execution seal; an independent reconstruction from the raw pools subsequently reproduced all decisions, both original comparators, and the reported intervals. It evaluates count-first score tie-breaking in this cohort; the proposition's conditional bonus-dominance premise is a separate population condition.

Complete policy and numerical conventions. Use the original candidate order, boxed-integer eligibility, canonical answers, and raw scores. For each eligible answer a, let C_a be its count and S_a the sum of its carriers' raw scores. Retain $T = \arg\max_a C_a$, select a member of $\arg\max_{a \in T} S_a$, and resolve any remaining tie by first occurrence in the original candidate order. Scores are reconstructed from the finite stored log score by exponentiation, with the original positive-score fallback and zero otherwise. Sums start at zero and use sequential floating-point addition in original candidate order; the policy compares these unnormalized computed sums and treats only exact equality as a score tie. In contrast, the original raw-V comparator retains its original normalization and 10^{-12} tolerance for maximum answer mass. This distinction is part of the evaluated specification, not a change to the original comparator.

A nonempty eligible pool with all score sums zero still chooses within T by first occurrence. An empty eligible pool uses the original standardized first-extractable fallback, or abstains if no such answer exists. On these data, all seven empty pools, spanning five problems, abstain under every rule and remain incorrect in every denominator; no candidate has a zero raw score. The policy preserves all 508 unique plurality winners. Its other branches are score selection on 240 tied pools, first-occurrence

resolution of equal sums on one tied pool, and the seven empty-pool abstentions. Here $N = 8$ counts generated candidates before eligibility; the proposition instead fixes the number of eligible iid draws. Matching its selection rule does not establish its sampling or conditional bonus-dominance premise.

Complete comparisons and uncertainty. Table 38 reports both fixed contrasts on the whole cohort. The policy is correct on 292/756 decisions (38.62%), compared with 274/756 (36.24%) for first-occurrence plurality and 295/756 (39.02%) for raw V. Its difference from plurality is $+2.38$ percentage points, supported by 19 gains and one loss on 16 distinct problems. Its difference from raw V is -0.40 percentage points, with six gains and nine losses on 14 distinct problems. The latter intervals include both negative and positive effects; they do not establish equivalence or noninferiority to raw V.

Each estimand averages the three paired run differences within each problem and weights the 252 problems equally. We reuse the original verified bootstrap indices: 10,000 replicates, NumPy `default_rng(28391)`, lexicographically sorted problem IDs, 140 stratum-A and 112 stratum-B problem draws per replicate, with all three runs held together and A drawn before B. Percentile endpoints use linear quantiles. We report central 95% intervals and the fixed local Bonferroni-two 97.5% intervals for these two contrasts. This local adjustment does not account for outcome-informed policy selection. The intervals are descriptive and rely on comparable independent problem clusters conditional on the observed runs; the 756 decisions are not independent inferential units. They are not distribution-free guarantees for regenerating answers to these fixed questions.

Table 38: Complete post hoc count-then-score policy on the saved $N = 8$ cohort: 252 problems, three runs each, and 756 decisions per rule. The policy is correct on 292/756 decisions (38.62%). Differences and intervals are percentage points, policy minus comparator. Both fixed contrasts are retained. Local Bonferroni-two intervals do not correct outcome-informed policy selection.

Quantity	Versus plurality	Versus raw V
Comparator correct / total	274/756	295/756
Comparator accuracy (%)	36.24	39.02
Difference (pp)	$+2.381$	-0.397
Pointwise 95% interval	$[1.190, 3.836]$	$[-1.455, 0.661]$
Local 97.5% interval	$[0.926, 3.968]$	$[-1.720, 0.794]$
Changed predictions	174	73
Problems with any changed prediction	100	60
Gains / losses	19/1	6/9
Wrong-to-wrong changes	154	58
Changed predictions, both correct	0	0
Distinct discordant problems	16	14
Net-positive / net-negative problems	15/1	6/8
Problems without correctness discordance	236	238
Problems with both gain and loss	0	0
Discordant problems with zero net change	0	0

A gain is a decision on which the policy is correct and its comparator is wrong; a loss reverses these roles. A discordant problem contains at least one gain or loss across its three runs. A wrong-to-wrong change alters the predicted answer without altering correctness. The seven empty-pool abstentions are retained as incorrect. Intervals use the original 10,000 A/B-stratified problem-cluster bootstrap replicates, keeping all three runs together.

Which count patterns contribute? This decomposition asks whether the observed gain comes from selecting among answers seen once or from combining scores for repeated answers. Both are count ties in Proposition 1, but only the latter sums multiple scores for each tied leader. Table 39 partitions all 756 pools using eligible canonical-answer counts: seven empty pools, 508 unique maxima, 203 singleton maximum ties (largest count one), and 38 repeated-answer maximum ties (largest count at least two). These categories are exhaustive and disjoint at the pool level. Their represented-problem counts overlap, because a problem's three runs can fall in different categories. All paired correctness cells and wrong-to-wrong prediction changes are retained. Contributions use the same denominator of 756, so they sum to the complete-policy effect; they are not within-category accuracy differences.

Of the 18 net gains over plurality, 13 occur in singleton ties (14 gains, one loss; $+1.720$ percentage points of the whole cohort) and five in repeated-answer ties (five gains, no losses; $+0.661$ percentage points). The larger contribution is therefore score selection among answers observed once. In those 203 pools the policy coincides with raw V on every decision; this agreement was checked under the two numerical tie conventions. In repeated-answer ties it compares sums over multiple carriers. Both contributions are descriptive: no subgroup intervals, hypothesis tests, or causal mechanism claims are attached to them.

Where can count-tie refinement change correctness? A further post hoc check flags a pool when at least two answers attain the maximum eligible count and the correct answer, in its canonical form, belongs to that set. This is a pathwise *opportunity*: correctness could differ between rules that select within the maximum-count set. Empty and unique-maximum pools are false by definition. There are 48 such pools out of 756, comprising 34 singleton ties and 14 repeated-answer ties, and spanning 37 distinct problems. All 19 gains and the one loss of the policy versus plurality occur inside these 48 pools; the 141 prediction changes outside them are all wrong-to-wrong. Thus the observed $+18$ net gains lie within the part of these saved data where the theory's tie mechanism can change correctness. The flag uses the answer key retrospectively. It estimates neither achievable gain nor a conditional success probability or the theorem's $1 - 1/|T|$ conditional gain cap, and it supplies no evidence for conditional bonus dominance. It does not restrict correctness differences against raw V, which can select outside the maximum-count set. Whole-policy intervals are unchanged; no conditional rate or additional subgroup inference is reported.

Reconciliation with raw weighting. Which decisions change when scores may override answer counts? Section 5 separates count-tie refinement from changing a count winner. We cross-classify each pool's count pattern with whether raw weighting selects a maximum-count answer, then record paired gains and losses. This locates the empirical consequence of that distinction for the two complete rules. Table 40 retains the complete joint partition by count status and the raw-V choice. Raw V chooses within the maximum-count set on 676 pools, and the policy agrees with it on every one; this is an observed numerical check, not an automatic consequence of their different tie arithmetic. Raw V chooses below the maximum count on 73 pools: 65 with a unique maximum and eight with tied maxima. All eight tied-pool overrides are repeated-answer ties. On those eight pools, the policy differs from plurality twice, both wrong-to-wrong, and differs from raw V eight times (one gain, one loss, six wrong-to-wrong changes). Consequently the entire net deficit against raw V arises among the 65 unique-maximum overrides: five gains and eight losses for the policy. Empty and unique-maximum pools make no contribution to its plurality contrast.

Scope. This readout establishes the specified policy's accuracy and the exact count patterns responsible for its gain in this Math-7B cohort at $N = 8$. Its preservation of unique count winners and concentration of changes in tied leaders connect directly to Section 5. The theorem's conditional bonus-dominance premise remains a population condition; this readout evaluates the policy on observed pools. The separate generation-budget extension compares raw weighting with nine- and ten-candidate plurality (Section F.2).

Post hoc controls: selection and random tie-breaking. These controls ask whether the score-based gains depend on plurality's first-occurrence tie rule, and whether aggregating scores improves on simply selecting the highest-scored candidate. Uniform-random ties provide the conditional baseline in Proposition 1; highest-score selection uses score ranking without pooling answer support. Comparing all rules on the same candidates isolates these readout choices. On the same 756 pools we compared, after all prospective and tie-policy outcomes were known, the single highest-scored eligible candidate (selection key: stored log score, then score, then smallest candidate index, so exact score ties resolve to the first occurrence), uniform-random-tie plurality (expected correctness $\mathbf{1}\{\text{correct answer} \in T\}/|T|$ over the tied maximum-count set T) and the raw half-floor against plurality, raw weighting and the count-then-score policy. Table 41 reports all eight comparisons in three separately adjusted families (selection, three comparisons; random tie, three; half-floor, two), each with its pointwise interval, its family level and the signed support. Selecting the highest-scored candidate answers 295 of 756 pools correctly, the same number as raw weighting, while the two rules differ on 61 pools (11 gains and 11 losses). Raw weighting, count-then-score and the raw half-floor each exceed expected uniform-random-tie plurality under their respective family adjustments, and no

Table 39: Exhaustive post hoc count-pattern partition of both complete policy contrasts. Correctness cells are ordered (policy, comparator): 10 is a gain and 01 a loss. W counts different predictions within the 00 cell; Disc. counts distinct problems with correctness discordance (output field `discordant_problems`). Opp. counts pools with at least two maximum-count answers, including the correct answer in its canonical form. The last column is $100(10-01)/756$, the contribution in percentage points to the whole-cohort difference.

Count pattern	Pools	Problems	Opp.	00	01	10	11	W	Disc.	pp
Policy minus first-occurrence plurality										
Empty	7	5	0	7	0	0	0	0	0	0
Unique maximum	508	216	0	244	0	0	264	0	0	0
Singleton tie	203	103	34	184	1	14	4	146	12	+1.720
Repeated-answer tie	38	34	14	28	0	5	5	8	4	+0.661
Whole cohort	**756**	**252**	**48**	**463**	**1**	**19**	**273**	**154**	**16**	**+2.381**
Policy minus raw V										
Empty	7	5	0	7	0	0	0	0	0	0
Unique maximum	508	216	0	236	8	5	259	52	12	−0.397
Singleton tie	203	103	34	185	0	0	18	0	0	0
Repeated-answer tie	38	34	14	27	1	1	9	6	2	0
Whole cohort	**756**	**252**	**48**	**455**	**9**	**6**	**286**	**58**	**14**	**−0.397**

Singleton ties have maximum eligible-answer count one and at least two labels; repeated-answer ties have maximum count at least two and at least two maximizing labels. Represented-problem counts overlap across rows and must not be added; the whole-cohort entry is their union. The discordant-problem sets happen to be disjoint across categories in these data. The 48 opportunity pools span 37 distinct problems: 27 represented among singleton ties and 12 among repeated-answer ties, with overlapping membership. Opportunity counts locate where a rule restricted to maximum-count answers could change correctness; they are not achievable-gain estimates or conditional success rates. All seven empty pools abstain under all rules. No subgroup intervals or tests are used.

aggregation rule establishes an advantage over highest-score selection. The analysis was bound and reviewed before the final reviewed execution and independently reconstructed from the candidate features afterwards. The half-floor family was finalized after selection and random-tie outcomes had been inspected by members of the team; the three local adjustments do not account for outcome-informed nomination. The separate generation-budget extension evaluates raw weighting against nine- and ten-candidate plurality (Section F.2).

H Relation to Classical Aggregation and Calibration

From juror competences to sampled scores. In classical binary aggregation, each juror has a competence parameter p_i; under the independent symmetric competence model and equal prior odds, the optimal decision uses $\log[p_i/(1-p_i)]$, which is negative for an anti-expert (Nitzan & Paroush, 1982; Shapley & Grofman, 1984). Berend & Kontorovich (2014) analyze finite-sample error, consistency, and estimated competences in this setting. Conditionally on its score V_i, a calibrated generated answer has correctness probability V_i, suggesting the juror analogy. But the sampled score is a random attribute of a generated answer, rather than a fixed juror parameter; our fixed-problem condition does not specify likelihoods under alternative true answers. A Bayes-optimal log-odds interpretation needs that additional channel model. Our positive-weight classes concern uniform preservation over all admissible laws, rather than optimality within a known channel.

Where the off-diagonal enters. Optimal polychotomous committee decisions were already studied by Ben-Yashar & Paroush (2001). More directly, Kuncheva & Rodríguez (2014) distinguish uniform incorrect-label allocation from a full confusion matrix: under class-conditional independence and equal priors, uniform allocation gives the effective vote weight $\log[p_i(m-1)/(1-p_i)]$, where m is the number of answer classes, whereas a general known confusion matrix gives answer-dependent log-likelihood contributions. The Dawid–Skene model estimates such observer-specific matrices across items (Dawid & Skene, 1979). Gao et al. (2016) establish optimal exponents for Dawid–Skene aggregation and give conditions for attaining them with estimated confusion matrices. Our score-law criterion (Theorem 1) characterizes existence of an improving bounded

Table 40: Complete joint partition by eligible-answer count status and the raw-V selection. "Within" and "below" refer to the maximum-count set; "undefined" denotes no defined selection before fallback. Correct counts are policy/comparator, G/L are decision-level gains/losses, W is wrong-to-wrong changes, and Disc. is distinct discordant problems (output field `discordant_problems`). All zero-effect and empty cells are retained.

Count status	Raw-V status	Pools	Problems	Correct	Changed	G/L	W	Disc.
Policy minus first-occurrence plurality								
Unique	Within	443	192	259/259	0	0/0	0	0
Unique	Below	65	55	5/5	0	0/0	0	0
Unique	Undefined	0	0	0/0	0	0/0	0	0
Tied	Within	233	124	27/9	172	19/1	152	16
Tied	Below	8	8	1/1	2	0/0	2	0
Tied	Undefined	0	0	0/0	0	0/0	0	0
Empty	Within	0	0	0/0	0	0/0	0	0
Empty	Below	0	0	0/0	0	0/0	0	0
Empty	Undefined	7	5	0/0	0	0/0	0	0
Whole cohort		**756**	**252**	**292/274**	**174**	**19/1**	**154**	**16**
Policy minus raw V								
Unique	Within	443	192	259/259	0	0/0	0	0
Unique	Below	65	55	5/8	65	5/8	52	12
Unique	Undefined	0	0	0/0	0	0/0	0	0
Tied	Within	233	124	27/27	0	0/0	0	0
Tied	Below	8	8	1/1	8	1/1	6	2
Tied	Undefined	0	0	0/0	0	0/0	0	0
Empty	Within	0	0	0/0	0	0/0	0	0
Empty	Below	0	0	0/0	0	0/0	0	0
Empty	Undefined	7	5	0/0	0	0/0	0	0
Whole cohort		**756**	**252**	**292/295**	**73**	**6/9**	**58**	**14**

Pools form a disjoint partition; represented problems can occur in multiple rows and their counts must not be summed. All eight tied-pool lower-count raw-V selections occur in repeated-answer ties. The policy agrees with raw V on all 676 within-maximum pools, including the 233 tied-maximum pools. This table is descriptive and supplies no subgroup inference.

Comparison	Fam.	Δ (pp)	95%	Family	Support
Selection − plurality	A	+2.78	[+0.66, +5.03]	[+0.18, +5.56]	38/17/245
Raw weighting − selection	A	0.00	[−1.19, +1.19]	[−1.59, +1.46]	11/11/39
Count-then-score − selection	A	−0.40	[−1.98, +1.19]	[−2.38, +1.59]	16/19/95
Random-tie plur. − plurality	B	+0.58	[−0.33, +1.44]	[−0.55, +1.63]	48 (+38/−10)
Raw weighting − random-tie	B	+2.20	[+0.64, +3.83]	[+0.31, +4.26]	62 (+36/−26)
Count-then-score − random-tie	B	+1.80	[+0.74, +2.96]	[+0.52, +3.24]	48 (+28/−20)
Raw half-floor − selection	C	−0.53	[−2.12, +1.06]	[−2.38, +1.32]	16/20/197
Raw half-floor − random-tie	C	+1.67	[+0.59, +2.83]	[+0.46, +3.00]	49 (+27/−22)

Table 41: Post hoc controls on the prospective study's 756 saved pools (252 problems × 3 runs). Selection = the single highest-scored eligible candidate; random-tie = uniform-random-tie plurality (expected correctness). Correct pools: plurality 274, raw weighting 295, count-then-score 292, selection 295, raw half-floor 291; random-tie plurality expected 278.39. Families A (selection) and B (random tie) are each adjusted for three comparisons (98.33%), family C (half-floor) for two (97.5%); intervals use the A/B-stratified paired problem-cluster bootstrap keeping each problem's three runs together. Support: correctness gains/losses/changed wrong answers for the integer comparisons; for the random-tie comparisons, the number of pools whose expected correctness differs and the sign split. Declared after the prospective outcomes were known; independently reconstructed from the final reviewed execution.

positive score-only weight under a fixed answer–score law. Our conditional wrong-answer allocation $\pi_a(v) = P(A = a \mid V = v, R = 0)$ describes the corresponding allocation of error mass within one fixed problem and score stratum. It is allowed to vary with v and is not estimated across recurring workers or assumed uniform. The new question is what fixed positive score transforms guarantee without learning that allocation.

Scalar and multiclass calibration. For a probability vector Q, canonical calibration requires $P(Y = j \mid Q) = Q_j$ for every class; classwise calibration conditions only on Q_j, and confidence calibration conditions on the reported maximum and tests correctness of its maximizing label (Vaicenavicius et al., 2019; Kull et al., 2019). The latter two are weaker summaries of the vector condition. Our $P(R = 1 \mid V) = V$ is a binary correctness condition on a scalar verifier score. Its connection to confidence calibration is conceptual: a verifier score need not be the maximum coordinate of any predictive probability vector. Requiring the condition separately for every problem and candidate law is also stronger than pooling calibration across problems. The familiar loss of class information motivates our formulation; the additional result is the exact consequence for positive answer weighting, including the sharp coupling bounds and transformation classes.

Readout scope and signed weights. A fixed-temperature softmax vote is a fixed positive weight $h(v) = \exp(\beta v)$: its common sample normalizer cancels in the answer argmax. Nonzero β fails the universal preserving class, while remaining in the theorem's domain. A fixed filter $\mathbf{1}\{v > t\}$ with $t \geq 1/2$ belongs to the nonnegative branch when its target is defined. Best-of-N, adaptive temperatures and score-driven generation require separate reasoning. Kuang et al. (2026) derive MAP aggregation under a specified symmetric generation and score-likelihood model with conditional-independence assumptions; their weights can be signed. We instead ask when a bounded positive score-only weight improves plurality for one fixed joint answer–score law, without imposing a uniform allocation of incorrect answers.

Neighboring information and decision models. Surprisingly popular aggregation uses elicited predictions of others' answers (Prelec et al., 2017), an information model different from a scalar correctness-calibrated score. PRM continuation-probability calibration (Park et al., 2025) and generative step verification (Khalifa et al., 2026) address scorer estimation; exact terminal correctness calibration under the chosen scalarization and problem law remains a separate assumption here.

Best-of-N theory studies quality–compute trade-offs under imperfect rewards (Huang et al., 2025), while our decision rule sums fixed weights within answer classes. Huang et al. (2026, Theorems 3.1–3.2) compare iid self-consistency and best-of-N sample complexity when the correct answer is the

unique mode; their selection bound assumes its reward exceeds every incorrect answer's. Our results characterize fixed score-only voting weights without that ranking assumption. Chen et al. (2024) explain nonmonotone voting performance across mixtures of query difficulties; our comparisons hold the problem law fixed and compare weighting rules at a shared budget or through their limiting error rates.

Mode estimation and correctness. Sequential methods identify the unique mode of a finite categorical law from iid observations using confidence sequences (Anand Jain et al., 2022). Identifying a response mode does not by itself establish answer correctness. Feng et al. (2026) analyze margin-dependent plurality error and fixed or dynamic sample allocation; our fixed-law results classify score transforms and compare weighted correctness. Ota et al. (2026) certify prespecified targets as ordinary or confidence-weighted response modes with anytime-valid guarantees, distinct from answer correctness; our count-margin result bounds harmful reversals of correct plurality.