# TWO CALLS, TWO MOMENTS, AND THE VOTE-ACCURACY CURVE OF REPEATED LLM INFERENCE

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Repeated sampling can improve LLM accuracy, and quantifying the gains from additional calls is essential for allocating test-time compute. We study binary decisions, a fundamental setting where repeated answers to the same question are aggregated by majority vote. We show that two independently sampled responses per example in a validation set with known answers constrain the latent distribution of example-level success probabilities to a class consistent with the paired outcomes. Optimizing over this class yields sharp accuracy and gain bounds at every finite voting budget. A shared large-sample confidence region accounts for validation uncertainty. We also obtain sharp infinite-vote bounds and moment-matched forecasts of the vote-accuracy curve. On QNLI and QQP, our method successfully distinguishes settings with voting gains above a target margin from those with little room for improvement.

## 1 INTRODUCTION

Repeated sampling has become a standard way to spend test-time compute. Self-consistency samples multiple reasoning paths before aggregating final answers (Wang et al., 2023); pass@$k$ evaluation measures success across repeated code-generation attempts (Chen et al., 2021); and inference-scaling work studies repeated candidate generation as a performance axis (Brown et al., 2024). A parallel line uses consistency and entropy of repeated generations as uncertainty evidence (Manakul et al., 2023; Kuhn et al., 2023; Farquhar et al., 2024). These methods raise a statistical question: for a fixed model, prompt, and sampler, how much extra accuracy can majority voting extract from more calls?

The difficulty is that consistency is not correctness. Repeated calls may agree because the system reliably solves an example, or because it reliably makes the same mistake. Conversely, a model with lower single-call accuracy may retain randomness that voting can average away. LLM outputs can vary even under nominally deterministic serving settings (Atıl et al., 2025). With black-box access, evaluating voting on a new task requires repeated responses and additional inference compute.

We study binary decisions, the simplest setting for this question, with applications including LLM safety review (Han et al., 2024). Each call answers the same question, and voting succeeds when a majority of calls are correct. Our analysis applies to any stochastic binary predictor whose calls on each example are independent with a common success probability, which can vary across examples. Averaging majority-vote success over examples gives the vote-accuracy curve: accuracy as the number of calls increases. This curve says whether three, five, or more calls improve on one.

Estimating the full distribution of success probabilities would require many repeated responses per example. Two calls already constrain the vote-accuracy curve: on a validation set with known answers, we record the correctness of two responses per example. The single-call success rate identifies the mean success probability; the rate of two correct responses identifies its second moment. These rates also determine the correlation of correctness across calls on the same example. The paired outcomes turn a potentially many-call validation problem into a two-moment problem: distributions with the same moments can yield different vote-accuracy curves. Their smallest and largest accuracies at each budget give the sharp bounds.

The central reduction makes this problem tractable. For every finite voting budget, optimizing over all success-probability distributions with these two moments reduces to distributions supported on three points and matching quadratic upper and lower bounds. Each endpoint is attained by



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.

a compatible distribution, making the bounds sharp. The same reduction bounds the gain from increasing an existing voting budget. A shared large-sample confidence region for the moments supports simultaneous budget comparisons while accounting for validation uncertainty. We derive closed-form three-vote bounds and sharp infinite-vote bounds. Two moments also support useful point completions when a single curve is desired. Maximum-entropy and Gaussian-probit models fitted to these moments provide full-curve forecasts.

Experiments on QNLI and QQP show why single-call accuracy does not order the benefits of repeated inference. Highly correlated calls can leave little room for improvement, while other settings yield positive lower bounds on voting gains. In held-out evaluations, the two-call analysis establishes gains exceeding half a percentage point. An observed ranking reversal further shows that a model with lower single-call accuracy can overtake another after voting. These results guide decisions about starting voting and increasing the number of calls.

## 2 Related work

Repeated-sampling methods differ in their aggregation targets. Self-consistency aggregates sampled reasoning paths (Wang et al., 2023); adaptive-consistency uses agreement-based stopping (Aggarwal et al., 2023), while confidence-informed self-consistency weights answers by model confidence (Taubenfeld et al., 2025). Code-generation and inference-scaling studies evaluate repeated attempts through pass@$k$ and larger sampling budgets (Chen et al., 2021; Brown et al., 2024). Pass@$k$ measures the chance of at least one successful attempt. Plurality selects the most frequent answer, so its success also depends on how incorrect answers split.

Agreement-based uncertainty methods use repetitions as evidence about reliability. SelfCheckGPT compares sampled responses for hallucination detection (Manakul et al., 2023); semantic uncertainty and entropy group linguistically different but equivalent answers (Kuhn et al., 2023; Farquhar et al., 2024). Empirical studies examine how uncertainty proxies discriminate errors and calibrate confidence (Savage et al., 2025; Vashurin et al., 2025; Xiao et al., 2025), while ConU calibrates prediction sets for correctness coverage (Wang et al., 2024). Repeated-prompt studies emphasize the dependence of measurements on the same example (Gallo et al., 2025), and Nowak (2025) analyzes the allocation between distinct prompts and repeats when estimating self-consistency.

The C-bound controls weighted-majority risk through the first two margin moments, with PAC-Bayesian estimation guarantees (Germain et al., 2015). Liu (2026) characterizes the complete odd-budget voting curve through a signed signature. Moment-constrained optimization uses small-support extreme points of moment sets (Winkler, 1988; Pinelis, 2016) and polynomial majorant/minorant duals (Karlin & Studden, 1966; Bertsimas & Popescu, 2005). Maximum entropy (Jaynes, 1957) and normal-ogive item-response models (Albert, 1992; Lord, 1980) select latent laws.

## 3 Repeated calls and two-moment information

To formalize the validation probe, fix a model, inference protocol, target population, and correctness evaluator $e : \mathcal{Y} \times \mathcal{Z} \to \{0, 1\}$. For ground truth $Y_i$ and possibly structured call output $Z_{ij}$, the evaluator includes parsing, normalization, and an invalid-output rule. The score $B_{ij}$ equals one for a correct response and zero otherwise. Let $q_i$ be the success probability on example $i$, let $\pi$ be its law across target examples, and write $Q \sim \pi$ for a generic draw:

$$B_{ij} = e(Y_i, Z_{ij}), \qquad q_i \overset{\text{i.i.d.}}{\sim} \pi, \qquad B_{ij} \mid q_i \overset{\text{ind}}{\sim} \text{Bernoulli}(q_i). \tag{1}$$

Repeated calls are independent conditional on $q_i$, although marginally dependent because they share the example. Independent sampling units are example draws, not individual calls. A fixed distribution over prompts is also a fixed inference protocol when every call samples its prompt afresh and normalizes the answer to the same task label. Model or checkpoint changes define different protocols. Hidden serving-state reuse, feedback between calls, or correlated decoding streams need not satisfy (1) (see Appendix D.3).

For $K = 2n + 1$ calls, let $P_n(q)$ be the probability that a majority is correct on an example with success probability $q$. Averaging over examples gives the population accuracy $V_n(\pi)$:

$$P_n(q) = \sum_{\ell=n+1}^{2n+1} \binom{2n+1}{\ell} q^\ell (1-q)^{2n+1-\ell}, \qquad V_n(\pi) = \int P_n(q)\ \mathrm{d}\pi(q).$$

Thus $V_0 = \mu = \mathbb{E}Q$ is single-call accuracy and $V_1$ is *three*-vote accuracy. Fair binary tie-breaking makes an even budget equal in accuracy to the preceding odd budget (Appendix A.1). Selecting the first answer on a binary tie is also conditionally fair; counting ties as incorrect changes the objective.

The binary scores need not come from a binary output vocabulary. When outputs are normalized to one of two task labels, ordinary label majority has accuracy $V_n$ and needs no ground truth at deployment. For structured outputs, $V_n$ instead denotes the event that a strict majority passes the evaluator. With a unique correct answer this event guarantees plurality success, so its lower bound transfers to plurality, but its upper bound does not. The two-sided decisions below apply only to a deployed rule whose accuracy is $V_n$.

**Proposition 3.1.** *Under* (1)*,*

$$\mathbb{E}[B_{i1}] = \mathbb{E}[B_{i2}] = \mu := \mathbb{E}[Q], \qquad \mathbb{E}[B_{i1}B_{i2}] = \nu := \mathbb{E}[Q^2].$$

*For* $0 < \mu < 1$*,*

$$\rho = \mathrm{Corr}(B_{i1}, B_{i2}) = \frac{\nu - \mu^2}{\mu(1-\mu)} \in [0, 1]. \tag{2}$$

*The feasible set is exactly*

$$0 \le \mu \le 1, \qquad \mu^2 \le \nu \le \mu.$$

Appendix A.2 proves the moment identities and constructs a latent law for every feasible pair.

The pair distribution gives the operational meaning of the moments:

$$\mathbb{P}(B_{i1} = 1, B_{i2} = 1) = \nu, \quad \mathbb{P}(B_{i1} \neq B_{i2}) = 2(\mu - \nu), \quad \mathbb{P}(B_{i1} = 0, B_{i2} = 0) = 1 - 2\mu + \nu. \tag{3}$$

The disagreement probability $2(\mu - \nu) = 2\mathbb{E}[Q(1-Q)]$ measures within-example randomness; the covariance $\nu - \mu^2 = \mathrm{Var}(Q)$ measures between-example heterogeneity. Together they decompose the variability of a single correctness score:

$$\mu(1-\mu) = \mathrm{Var}(Q) + \mathbb{E}[Q(1-Q)].$$

At a fixed mean, large $\rho$ means that more variability is tied to the example and less remains across calls on that example. Successes and failures are then more reproducible. Section 4 quantifies the disagreement needed to guarantee voting gains.

Writing $\mathcal{P}([0,1])$ for all probability distributions on $[0,1]$, the latent laws compatible with the two-call distribution form the moment class

$$\mathcal{M}(\mu, \nu) = \left\{ \pi \in \mathcal{P}([0,1]) : \int q\ \mathrm{d}\pi(q) = \mu, \quad \int q^2\ \mathrm{d}\pi(q) = \nu \right\}. \tag{4}$$

The population bounds below hold uniformly over $\pi \in \mathcal{M}(\mu, \nu)$. For a budget of $2n + 1$ calls, the sharp accuracy interval is

$$L_n(\mu, \nu) = \inf_{\pi \in \mathcal{M}(\mu,\nu)} \int P_n(q)\ \mathrm{d}\pi(q), \qquad U_n(\mu, \nu) = \sup_{\pi \in \mathcal{M}(\mu,\nu)} \int P_n(q)\ \mathrm{d}\pi(q). \tag{5}$$

Sharpness means that no smaller interval covers all compatible latent laws. With exact moments, compatible laws can still give different accuracies; finite validation adds sampling uncertainty.

## 4 Sharp vote-accuracy bounds

With only a mean $\mu$, the law $(1-\mu)\delta_0 + \mu\delta_1$ (where $\delta_x$ is a point mass at $x$) remains compatible with the data and gives $V_n = \mu$ at every budget. Thus one call cannot guarantee a strictly positive gain. Appendix A.3 gives the exact mean-only three-vote envelope.

To see what the second moment adds, compare this population with one in which every example has success probability $\mu$. Both have single-call accuracy $\mu$. In the first, every example is either always correct or always wrong, so voting leaves accuracy unchanged. In the second, $Q = \mu$ and three-vote gain is $P_1(\mu) - \mu = \mu(1-\mu)(2\mu-1)$, which is positive for $1/2 < \mu < 1$. Their second moments are $\nu = \mu$ and $\nu = \mu^2$, respectively. Two calls distinguish these explanations of the same average accuracy; the sharp bounds extend this comparison to every intermediate feasible moment pair.

### 4.1 Three-support reduction and quadratic dual

The class (4) includes arbitrary discrete, continuous, and mixed latent laws, so optimizing over it is initially an infinite-dimensional problem. The next theorem shows that each best- or worst-case vote accuracy is attained by a population with at most three latent success probabilities. Quadratic envelopes establish the corresponding endpoints over the entire class.

**Theorem 4.1.** *For every feasible $(\mu, \nu)$ and every fixed $n \geq 1$, both extrema in* (5) *are attained by laws with at most three support points in $[0, 1]$. The quadratic dual extrema are also attained, with*

$$L_n(\mu,\nu) = \max\{a + b\mu + c\nu : a + bq + cq^2 \leq P_n(q)\ \forall q \in [0,1]\},$$
$$U_n(\mu,\nu) = \min\{a + b\mu + c\nu : a + bq + cq^2 \geq P_n(q)\ \forall q \in [0,1]\}.$$

The proof is in Appendix A.4.

Compactness of $\mathcal{M}(\mu, \nu)$ and continuity of $P_n$ give an optimizer, which linearity allows us to choose at an extreme point. The three constraints fix total mass, mean, and second moment. If a finite-support law has positive weights on four distinct points $q_j$, the four vectors $(1, q_j, q_j^2)$ are linearly dependent. One can perturb the weights in either direction while preserving every constraint, so that law is not extreme. The finite-moment extreme-point theorem extends this argument to general measures (Winkler, 1988; Pinelis, 2016). These three-point laws attain the bounds.

A quadratic upper envelope $r(q) = a + bq + cq^2 \geq P_n(q)$ has known expectation $a + b\mu + c\nu$. A supporting hyperplane of the convex hull of $(q, q^2, P_n(q))$ gives a matching envelope at an interior moment pair. The nonnegative gap $r(q) - P_n(q)$ has zero expectation at an optimizer, so every positive-weight support point satisfies $r(q) = P_n(q)$. Reversing the inequality gives the lower bound. The global envelope and an attaining distribution together establish sharpness.

This dual representation also gives a direct computation from the moments and the majority polynomial. For an upper bound, minimize $a + b\mu + c\nu$ over the three coefficients of $r(q) = a + bq + cq^2$, subject to $r(q) \geq P_n(q)$ on $[0, 1]$. An exchange method starts with constraints at a finite set of points and solves the resulting linear program. It then checks the polynomial slack $r - P_n$ at both endpoints and every real stationary point in the interval, where its minimum must occur. Any violating point is added to the constraints before solving again. Applying the same procedure to $-P_n$ gives the lower bound. The implementation checks residual slack and adjusts the constant term outward to account for numerical error; Appendix C gives the computation details. Increasing the budget raises the polynomial degree; the dual retains three coefficients.

The same moment class also gives sharp bounds on gains between budgets: replacing $P_n$ by $P_n - P_m$ in Theorem 4.1 optimizes the increment under one latent law. Extremizing laws can change with the budget, so direct increment bounds can be tighter than differences of separately optimized endpoints. For example, the gain from three to five calls is $V_2 - V_1 = \mathbb{E}[P_2(Q) - P_1(Q)]$. A positive lower bound establishes improvement for every compatible population; a negative upper bound establishes a loss. This uses the same two moments, with the polynomial-objective extension proved in Appendix A.4.

### 4.2 Closed-form three-vote bounds

For three votes, the bounds give an explicit answer to the first useful budget increase beyond one call. Since $P_1(q) = 3q^2 - 2q^3$, the only unobserved quantity is the third moment. Its exact range under $(\mu, \nu)$ yields the following closed form.

**Theorem 4.2.** *Let $\mu \in (0,1)$ and $\rho \in [0,1]$, with $\nu = \mu^2 + \rho\mu(1-\mu)$. Put $s = \mu(1-\mu)(1-\rho)$. Then, with the usual endpoint conventions at $\mu = 0$ and $\mu = 1$,*

$$L_1(\mu,\nu) = \nu + 2\frac{(\mu-\nu)^2}{1-\mu} = \mu + s\{2\mu(1-\rho) - 1\}, \tag{6}$$

$$U_1(\mu,\nu) = 3\nu - 2\frac{\nu^2}{\mu} = \mu + s\{2\mu + 2\rho(1-\mu) - 1\}. \tag{7}$$

*Moreover $U_1(\mu,\nu) - L_1(\mu,\nu) = 2\mu(1-\mu)\rho(1-\rho) \le 1/8$, and if $\mu > 1/2$, then $V_1 > \mu$ for every law in $\mathcal{M}(\mu,\nu)$ exactly when $\rho < 1 - 1/(2\mu)$.*

**Proof idea: bound the missing third moment.** Put $m_3 = \mathbb{E}Q^3$, $a = \nu/\mu$, and $b = (\mu-\nu)/(1-\mu)$; feasibility places $a, b$ in $[0,1]$. Integrating $q(q-a)^2 \ge 0$ and $(1-q)(q-b)^2 \ge 0$ gives

$$\frac{\nu^2}{\mu} \;\le\; m_3 \;\le\; \nu - \frac{(\mu-\nu)^2}{1-\mu}.$$

Substituting into $V_1 = 3\nu - 2m_3$ gives the two accuracy bounds. Equality is realized on the zero sets of these polynomials: support $\{0, a\}$ attains the upper accuracy, and $\{b, 1\}$ attains the lower accuracy, with weights fixed by the mean. Thus the same elementary inequalities both bound the missing moment and locate the populations proving sharpness.

The improvement condition has a direct interpretation in the observable pair table. For $0 < \mu < 1$,

$$L_1(\mu,\nu) > \mu \quad\Longleftrightarrow\quad \mathbb{P}(B_1 \ne B_2) > 2\,\mathbb{P}(B_1 = B_2 = 0).$$

Exactly-one-correct pairs must be more than twice as likely as both-wrong pairs. This balance guarantees improvement for every latent law compatible with the two moments. The condition concerns population probabilities; Section 4.4 accounts for their estimation from a finite pair table. The proof and extremizers are in Appendix A.5.

### 4.3 Sharp infinite-vote endpoint

The infinite-vote endpoint describes where repetition leads under a fixed latent law. On an example with $q > 1/2$, an increasingly large majority becomes correct with probability tending to one; for $q < 1/2$, it becomes incorrect with probability tending to one. For each fixed latent law, bounded convergence therefore gives the infinite-vote endpoint

$$V_\infty(\pi) := \lim_{n\to\infty} V_n(\pi) = \pi((1/2, 1]) + \frac{1}{2}\pi(\{1/2\}). \tag{8}$$

The half-weight at $1/2$ follows from odd-vote symmetry. Unlike a finite majority polynomial, this objective depends discontinuously on the position of latent mass relative to the voting threshold. Nevertheless, the same two moments admit an exact endpoint characterization.

**Theorem 4.3.** *Let $U_\infty(\mu,\nu) = \sup_{\pi\in\mathcal{M}(\mu,\nu)} V_\infty(\pi)$ and $L_\infty(\mu,\nu) = \inf_{\pi\in\mathcal{M}(\mu,\nu)} V_\infty(\pi)$. Set $G(1/2, 1/4) = 1/2$, and for every other feasible $(\mu,\nu)$ define*

$$G(\mu,\nu) = \begin{cases} \dfrac{\nu - \mu^2}{\nu + 1/4 - \mu}, & \mu \le 1/2,\ \nu \le \mu/2, \\ \min\{1,\ 3\mu - 2\nu\}, & \textit{otherwise}. \end{cases}$$

*Then $U_\infty(\mu,\nu) = G(\mu,\nu)$ and $L_\infty(\mu,\nu) = 1 - G(1-\mu,\ 1-2\mu+\nu)$. These are sharp infima and suprema; one-sided moment-matching sequences may be needed in place of attaining laws.*

The two branches describe how much population mass can lie on the winning side of $q = 1/2$. Writing $\sigma^2 = \nu - \mu^2$, the fractional branch is $\sigma^2/\{\sigma^2 + (1/2-\mu)^2\}$. When the mean is below the threshold, placing mass just above it requires compensating mass below the mean. The ratio compares the available latent variance with the squared distance that must be crossed. The condition $\nu \le \mu/2$ is precisely what keeps that compensating support point nonnegative.

Outside this regime, the support constraint matters differently. A compensating point cannot move below zero; in the nonsaturated branch, extremal sequences instead use the endpoints 0 and 1 together

with a point just above $1/2$. The quadratic majorant $3q - 2q^2$ matches this geometry and gives $3\mu - 2\nu$. When the moments permit all mass to lie above the voting threshold, or to approach such a configuration, the upper endpoint reaches one, explaining the cap. Reflection $Q \mapsto 1 - Q$ gives the lower endpoint. Appendix A.6 gives the constructions and one-sided limits.

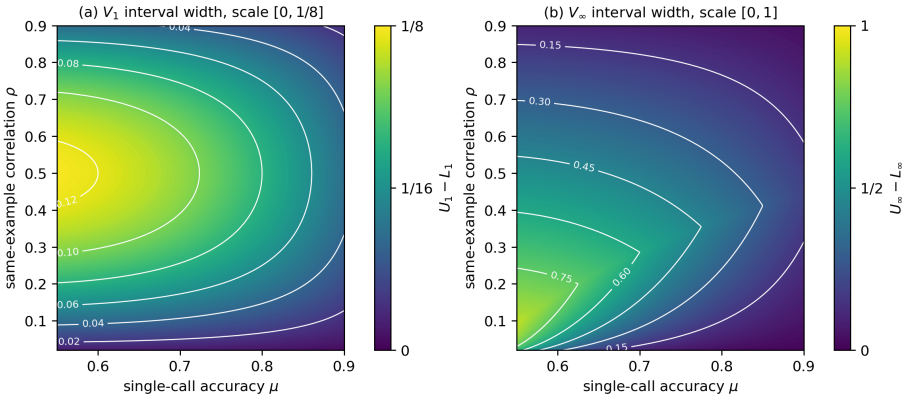


Figure 1: Population identification widths from two moments. (a) Exact three-vote width $U_1 - L_1$, on the scale $[0, 1/8]$. (b) Exact infinite-vote width $U_\infty - L_\infty$, on the different scale $[0, 1]$. Both panels use horizontal $\mu$ and vertical $\rho$; each panel has its own color scale. Widths describe variation over latent laws with known moments.

Figure 1 shows the contrast with three votes. At $\mu = 1/2$ and $0 < \rho < 1$, the exact widths are

$$U_1 - L_1 = \tfrac{1}{2}\rho(1 - \rho), \qquad U_\infty - L_\infty = 1 - \rho.$$

As positive $\rho$ tends to zero, the finite width vanishes while the endpoint width approaches one. At $\rho = 0$, $Q = 1/2$ is forced and both widths are zero. Thus tight finite-budget bounds do not imply a tight infinite-vote extrapolation: the endpoint has its own sharp geometry.

### 4.4 Statistical inference from two calls

The population bounds treat $(\mu, \nu)$ as known. The two-call moment estimates are subject to sampling uncertainty. From $N$ labeled example draws, form

$$X_i = \big((B_{i1} + B_{i2})/2,\ B_{i1}B_{i2}\big), \qquad \hat{\theta} = (\hat{\mu}, \hat{\nu}) = N^{-1}\sum_i X_i,$$

The coordinates record average correctness and joint success. Let $\hat{\Sigma}$ be their sample covariance.

For iid draws with nonsingular covariance, use the feasible Wald region

$$\mathcal{C}^{\mathrm{W}}_\alpha = \{\theta \in \mathcal{F} : N(\theta - \hat{\theta})^\top \hat{\Sigma}^{-1}(\theta - \hat{\theta}) \le \chi^2_{2,1-\alpha}\}, \tag{9}$$

where $\mathcal{F} = \{(\mu, \nu) : 0 \le \mu \le 1,\ \mu^2 \le \nu \le \mu\}$. It uses the covariance between the estimated moments and has nominal confidence $1 - \alpha$ under the usual large-sample approximation.

For each moment pair in this region, compute its sharp interval. Taking the union gives confidence sets for accuracy and gain over one call:

$$\mathcal{S}^{\mathrm{W}}_{n,\alpha} = \bigcup_{(\mu,\nu)\in\mathcal{C}^{\mathrm{W}}_\alpha} [L_n(\mu, \nu), U_n(\mu, \nu)],$$
$$\mathcal{G}^{\mathrm{W}}_{n,\alpha} = \bigcup_{(\mu,\nu)\in\mathcal{C}^{\mathrm{W}}_\alpha} [L_n(\mu, \nu) - \mu, U_n(\mu, \nu) - \mu].$$

We report the smallest interval containing each projected set. Whenever the moment region contains the true pair, these sets cover all corresponding vote accuracies and gains simultaneously for that protocol. The same construction applies to the infinite-vote endpoints in Section 4.3. The statistical coverage is asymptotic; sharpness refers to the population moment bounds.

Projecting the direct incremental-gain bounds over the same moment region gives simultaneous comparisons between budgets, including comparisons selected after inspecting the intervals. A positive lower gain bound establishes improvement at the nominal confidence level. Appendix B.1 details gain projection, protocol adjustments, and budget selection.

## 5 PARAMETRIC CURVE COMPLETIONS

A curve completion selects one success-probability distribution matching the moments and uses it to forecast accuracy across budgets. Maximum entropy and a latent-difficulty Gaussian-probit model each give a unique fit for every interior moment pair, with predictions inside the sharp bounds.

### 5.1 MAXIMUM-ENTROPY COMPLETION

Maximum entropy (ME) selects the density with the largest differential entropy subject to the observed moments (Jaynes, 1957). Fixing a mean and variance on the real line gives a Gaussian; on $[0, 1]$, the corresponding density is quadratic-exponential:

$$\psi(\lambda, \kappa) = \log \int_0^1 \exp\{\lambda(q - 1/2) + \kappa(q - 1/2)^2\} \, \mathrm{d}q,$$
$$f_{\lambda,\kappa}(q) = \exp\{\lambda(q - 1/2) + \kappa(q - 1/2)^2 - \psi(\lambda, \kappa)\}\mathbf{1}_{[0,1]}(q).$$

For interior $(\mu, \nu)$, the unique parameters minimize the strictly convex objective

$$J_{\mu,\nu}(\lambda, \kappa) = \psi(\lambda, \kappa) - \lambda(\mu - 1/2) - \kappa(\nu - \mu + 1/4).$$

The parameter $\lambda$ tilts the density across the voting threshold, with $\mathrm{sign}(\lambda) = \mathrm{sign}(\mu - 1/2)$; at fixed $\lambda$, $\kappa$ shifts mass toward the center or endpoints. The completed curve and its limit are

$$V_n^{\mathrm{ME}} = \int_0^1 P_n(q) f_{\lambda,\kappa}(q) \, \mathrm{d}q, \qquad V_\infty^{\mathrm{ME}} = \int_{1/2}^1 f_{\lambda,\kappa}(q) \, \mathrm{d}q.$$

Proposition G.1 gives the entropy characterization and uniqueness; the proof is in Appendix A.

### 5.2 LATENT-DIFFICULTY GAUSSIAN-PROBIT COMPLETION

The latent-difficulty Gaussian-probit (LDGP) completion models variation in success probabilities through example difficulty, following normal-ogive item-response models (Albert, 1992; Lord, 1980):

$$Q = \Phi(\eta - \gamma Z), \qquad Z \sim N(0, 1), \qquad \gamma \geq 0.$$

Here $Z$ represents example difficulty, $\eta$ locates performance relative to this scale, $\gamma$ controls the effect of difficulty on success probability, and $\Phi$ is the standard normal cdf. Writing $r = \gamma^2/(1+\gamma^2)$, moment matching gives

$$t = \Phi^{-1}(\mu), \quad \nu = \Phi_2(t, t; r), \quad \gamma = \sqrt{\frac{r}{1-r}}, \quad \eta = \frac{t}{\sqrt{1-r}},$$

where $\Phi_2(\cdot, \cdot; r)$ is the bivariate standard normal cdf; $r$ is distinct from the correctness correlation $\rho$. For fixed $t$, the second-moment map increases strictly from $\mu^2$ to $\mu$ as $r$ runs from zero to one, giving a unique fit for every interior moment pair. Its vote curve is

$$V_n^{\mathrm{LDGP}} = \mathbb{E}\big[P_n\{\Phi(\eta - \gamma Z)\}\big], \qquad V_\infty^{\mathrm{LDGP}} = \Phi(\eta/\gamma).$$

Proposition G.2 includes the boundary case $\gamma = 0$. Thus two calls determine a full curve under a familiar difficulty model. Table 11 compares the two completions at matched moments; their finite-budget forecasts are close in the examples shown, while the limiting forecasts reflect different allocations of mass around $q = 1/2$.

## 6 TWO-CALL DECISIONS ON LLM VALIDATION DATA

We evaluate binary judgments on QQP (are two questions duplicates?) and QNLI (does a sentence answer a question?) (Wang et al., 2019) with Llama 3.1-8B, Phi-4-mini, and Qwen2.5-7B served through Ollama; each call asks a Yes/No question that is scored against the task label. Figure 2a shows ten settings, with Q denoting QQP and N denoting QNLI. Q01–Q03 and N01–N03 use temperature 1; Q05 and N04 use Qwen at $0.2$, all on $8{,}192$ examples per task.

Two additional protocols, fixed on development data and evaluated on separate held-out examples, assess the gain bound: Q04 asks Llama at temperature 2 about longer QQP pairs, and N05 asks Phi at temperature 2 QNLI questions whose wording is resampled on every call, each with $6{,}144$ draws and a prespecified $1{,}024$-draw subset that receives a third call. Appendix E lists all settings and Appendix B.3 the prompts and sampling rules. A gain is the difference of percentage accuracies $100(V_n - V_0)$ within a setting.

The full direct-prompt grid contains 18 task–model–temperature settings, with five responses per example. We estimate moments from the first two responses. Empirical three-vote accuracy averages over the ten possible triples of recorded responses; five-vote accuracy uses all five. Each gain uses the single-call baseline averaged over the same recorded responses (Appendix C). All empirical three- and five-vote accuracies fall within the projected $95\%$ regions for their settings (Table 7). These are compatibility checks using shared responses.

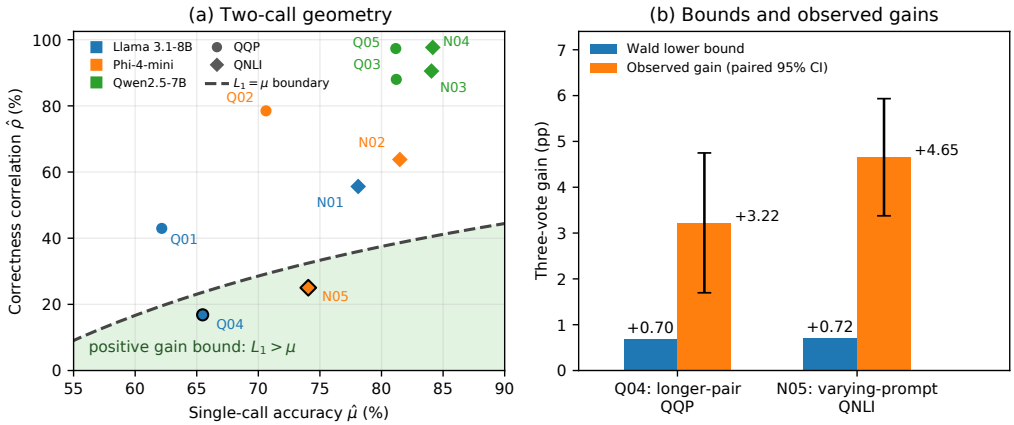


Figure 2: Two-call geometry and three-vote gains. (a) Ten settings by single-call accuracy and correctness correlation; colors give the model, markers the task, and outlined markers the held-out settings Q04 and N05. The gray dashed curve bounds the green region where the lower gain bound is positive. (b) Lower confidence bounds on the three-vote gain for Q04 and N05 (Wald projection, nominal joint confidence $95\%$) and the gains observed with a third call on $1{,}024$ prespecified draws (paired $95\%$ intervals), in percentage points.

The vertical coordinate in Figure 2a, the correlation between two calls on the same example, is what single-call accuracy omits: a population of always-correct and always-wrong examples matches any mean and gains nothing from voting. At a fixed single-call accuracy, higher correctness correlation means fewer disagreements, limiting how much three-vote accuracy can change (Section 6.1). The held-out settings lie inside the shaded region, where every population compatible with the estimated moments gains from three votes. The projected bounds below account for estimation uncertainty. Several direct-prompt settings outside this region show realized gains. Their estimated moments are compatible with both positive and negative population gains. On the same QQP examples, raising Llama's temperature from $0.2$ to $2.0$ lowers single-call accuracy and correctness correlation while increasing the five-vote gain from $-0.06$ to $+2.57$ percentage points (Table 7, Q06 and Q07).

For Q04 and N05, projecting the Wald region of the two-call statistics through the sharp lower bound establishes three-vote gains of at least $+0.70$ and $+0.72$ percentage points at nominal joint confidence $95\%$ (Figure 2b; Appendix B). The third calls then show observed gains of $+3.22$ and $+4.65$ points, with paired $95\%$ intervals $[1.70, 4.75]$ and $[3.38, 5.93]$ (Table 3). The projected minima bound population gains; the third-call estimates measure observed gains.

Both gain bounds exceed half a percentage point at the reported nominal joint confidence level. Minimization over the moment region and all compatible populations accounts for sampling uncertainty and the unknown success-probability distribution (Appendix B.1).

### 6.1 BUDGET COMPARISONS

Two-call upper bounds also quantify how much improvement remains possible at a specified budget. Since $P_1(q) - q = q(1-q)(2q-1)$, averaging over the population gives

$$|V_1 - \mu| \le \mathbb{E}[Q(1-Q)] = \mu - \nu = \tfrac{1}{2}\mathbb{P}(B_1 \neq B_2),$$

so three-vote accuracy moves by at most half the disagreement rate. In N04, Qwen at temperature $0.2$, the two calls disagree on about $0.6\%$ of examples; a Wald upper bound on this rate, applied to the inequality and its five-vote analogue, limits the gain to $+0.46$ and $+0.47$ points at familywise $95\%$ confidence across the direct-prompt settings (Appendix D). These bounds quantify the limited benefit of three or five calls without evaluating either budget directly.

The broader direct-prompt grid shows the same pattern: Qwen has the highest single-call accuracy in both tasks, yet its observed five-vote gain is at most $0.17$ percentage points on QNLI and $0.04$ on QQP across the three temperatures. Together with the Llama temperature comparison, this shows why baseline accuracy and repeatability provide different information about the value of extra calls.

For the same half-percentage-point target, the lower bounds for Q04 and N05 support additional calls, whereas the three- and five-vote upper bounds for N04 exclude that improvement at their reported confidence levels. These comparisons concern gains over each setting's own baseline; absolute accuracy and call cost also matter for deployment.

The choice of call budget also matters because single-call rankings can change after voting. Among the additional QNLI temperature settings, N09 (Phi-4-mini at $T = 2.0$) trails N06 (Llama 3.1-8B at $T = 0.2$) by $0.86$ with one call but leads by $1.04$ after five votes, measured as differences of percentage accuracies. This voting-induced rank reversal uses recorded-repeat comparisons on the same examples (Table 9). It shows why protocols should be compared at the intended vote budget rather than ranked by single-call accuracy alone.

The point forecasts also connect the two-call moments to empirical voting accuracy. Across the 18 direct-prompt settings, LDGP has mean absolute errors of $0.21$ and $0.22$ percentage points at three and five votes, compared with $0.36$ and $0.35$ for the interval midpoint (Table 10). Both model completions overpredict on average. Appendix F.1 also reports randomized-mixture results based on fitted per-example probabilities, illustrating how independent component selection can change repeatability and voting gains.

The supplementary nine-call GSM8K evaluation records positive strict-majority correctness gains but leaves their sign unresolved by the two-call Wald intervals. Appendix D.2 reports the full curve and separate plurality outcomes; Appendix D.3 examines the role of conditional independence.

## 7 CONCLUSION

Two labeled calls per example turn repeatability into evidence for allocating test-time compute. Their joint correctness table identifies the moments needed for sharp, computable accuracy bounds at every finite budget and for gains between budgets. The three-vote case yields a direct improvement criterion. Projecting a confidence region for the two moments through these bounds supports budget comparisons with validation uncertainty included. In the binary-judgment evaluations, recorded gains vary across settings; the two-call bounds establish positive gains in some settings and limit the possible improvement in another.

The infinite-vote bounds extend this account to limiting accuracy, while the moment-matched completions turn the same information into full-curve forecasts. The observed rank reversal further shows why model choice and call budget should be considered together. A contextual extension could use example features to estimate conditional moments and guide individual call budgets, with regression uncertainty propagated through the bounds. The same two-call validation sample supports accuracy and gain bounds across finite deployment budgets.

## STATEMENT ON THE USE OF AI

Generative AI tools were used to revise wording and improve the clarity and readability of the manuscript; to identify potentially relevant literature, references, and search directions, and to assist with research execution, including drafting, debugging, and refining code. For mathematical proofs, AI was used primarily to review and check the authors' arguments for potential errors, inconsistencies, or gaps. AI-assisted manuscript revisions, literature-search suggestions, feedback on research ideas and proofs, and generated or modified code were reviewed and verified by the authors before incorporation into the final work. The authors independently assessed the accuracy and appropriateness of these outputs and retain full responsibility for the methodology, proofs, implementation, analysis, and conclusions presented in the paper.

## REPRODUCIBILITY STATEMENT

The supplementary materials provide the experimental protocols, scored responses, and code for reproducing the tables, figures, and confidence bounds. The appendices contain the mathematical proofs, confidence derivations, and essential evaluation settings.

## REFERENCES

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

James H. Albert. Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics*, 17(3):251–269, 1992. doi: 10.3102/10769986017003251. URL https://doi.org/10.3102/10769986017003251.

Berk Atıl, Sarp Aykent, Alexa Chittams, Lisheng Fu, Rebecca J. Passonneau, Evan Radcliffe, Guru Rajan Rajagopal, Adam Sloan, Tomasz Tudrej, Ferhan Ture, Zhe Wu, Lixinyu Xu, and Breck Baldwin. Non-determinism of "deterministic" LLM system settings in hosted environments. In *Proceedings of the 5th Workshop on Evaluation and Comparison of NLP Systems*, pp. 135–148, Mumbai, India, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.eval4nlp-1.12. URL https://aclanthology.org/2025.eval4nlp-1.12/.

Dimitris Bertsimas and Ioana Popescu. Optimal inequalities in probability theory: A convex optimization approach. *SIAM Journal on Optimization*, 15(3):780–804, 2005. doi: 10.1137/S1052623401399903. URL https://doi.org/10.1137/S1052623401399903.

Bradley Brown, Jordan Juravsky, Ryan Saul 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. doi: 10.48550/arXiv.2407.21787. URL https://arxiv.org/abs/2407.21787.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021. doi: 10.48550/arXiv.2107.03374. URL https://arxiv.org/abs/2107.03374.

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.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024. doi: 10.1038/s41586-024-07421-0. URL https://doi.org/10.1038/s41586-024-07421-0.

Robert J. Gallo, Michael Baiocchi, Thomas R. Savage, and Jonathan H. Chen. Establishing best practices in large language model research: An application to repeat prompting. *Journal of the American Medical Informatics Association*, 32(2):386–390, 2025. doi: 10.1093/jamia/ocae294. URL https://doi.org/10.1093/jamia/ocae294.

Pascal Germain, Alexandre Lacasse, François Laviolette, Mario Marchand, and Jean-Francis Roy. Risk bounds for the majority vote: From a PAC-Bayesian analysis to a learning algorithm. *Journal of Machine Learning Research*, 16:787–860, 2015. URL https://www.jmlr.org/papers/v16/germain15a.html.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. WildGuard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of LLMs. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/0f69b4b96a46f284b726fbd70f74fb3b-Abstract-Datasets_and_Benchmarks_Track.html.

E. T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106(4):620–630, 1957. doi: 10.1103/PhysRev.106.620. URL https://doi.org/10.1103/PhysRev.106.620.

Samuel Karlin and William J. Studden. *Tchebycheff Systems: With Applications in Analysis and Statistics*. Interscience Publishers, New York, 1966. URL https://books.google.com/books?id=oRluQgAACAAJ.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=VD-AYtP0dve.

Yi Liu. When can voting help, hurt, or change course? exact structure of binary test-time aggregation, 2026. URL https://arxiv.org/abs/2605.05592.

Frederic M. Lord. *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1980.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.557. URL https://aclanthology.org/2023.emnlp-main.557/.

Robert Nowak. Estimating the self-consistency of LLMs. *arXiv preprint arXiv:2509.19489*, 2025. doi: 10.48550/arXiv.2509.19489. URL https://arxiv.org/abs/2509.19489.

Ollama. Generate a completion. Documentation, 2026a. URL https://docs.ollama.com/api/generate. Accessed 2026-05-03.

Ollama. llama3.1:8b model page. Model documentation, 2026b. URL https://ollama.com/library/llama3.1:8b. Accessed 2026-05-03.

Ollama. phi4-mini model page. Model documentation, 2026c. URL https://ollama.com/library/phi4-mini. Accessed 2026-05-03.

Ollama. qwen2.5:7b model page. Model documentation, 2026d. URL https://ollama.com/library/qwen2.5:7b. Accessed 2026-05-03.

Iosif Pinelis. On the extreme points of moments sets. *Mathematical Methods of Operations Research*, 83(3):325–349, 2016. doi: 10.1007/s00186-015-0530-0. URL https://doi.org/10.1007/s00186-015-0530-0.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Thomas Savage, John Wang, Robert Gallo, Abdessalem Boukil, Vishwesh Patel, Seyed Amir Ahmad Safavi-Naini, Ali Soroush, and Jonathan H. Chen. Large language model uncertainty proxies: Discrimination and calibration for medical diagnosis and treatment. *Journal of the American Medical Informatics Association*, 32(1):139–149, 2025. doi: 10.1093/jamia/ocae254. URL https://doi.org/10.1093/jamia/ocae254.

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, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.1030. URL https://aclanthology.org/2025.findings-acl.1030/.

Roman Vashurin, Ekaterina Fadeeva, Artem Vazhentsev, Lyudmila Rvanova, Daniil Vasilev, Akim Tsvigun, Sergey Petrakov, Rui Xing, Abdelrahman Sadallah, Kirill Grishchenkov, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, Maxim Panov, and Artem Shelmanov. Benchmarking uncertainty quantification methods for large language models with LM-polygraph. *Transactions of the Association for Computational Linguistics*, 13:220–248, 2025. doi: 10.1162/tacl_a_00737. URL https://aclanthology.org/2025.tacl-1.11/.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJ4km2R5t7.

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://openreview.net/forum?id=1PL1NIMMrw.

Zhiyuan Wang, Jinhao Duan, Lu Cheng, Yue Zhang, Qingni Wang, Xiaoshuang Shi, Kaidi Xu, Heng Tao Shen, and Xiaofeng Zhu. ConU: Conformal uncertainty in large language models with correctness coverage guarantees. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 6886–6898, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.404. URL https://aclanthology.org/2024.findings-emnlp.404/.

Gerhard Winkler. Extreme points of moment sets. *Mathematics of Operations Research*, 13(4): 581–587, 1988. doi: 10.1287/moor.13.4.581. URL https://doi.org/10.1287/moor.13.4.581.

Quan Xiao, Debarun Bhattacharjya, Balaji Ganesan, Radu Marinescu, Katya Mirylenka, Nhan H. Pham, Michael Glass, and Junkyu Lee. The consistency hypothesis in uncertainty quantification for large language models. In *Proceedings of the Forty-First Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pp. 4636–4651. PMLR, 2025. URL https://proceedings.mlr.press/v286/xiao25a.html.

# A PROOFS OF MAIN RESULTS

This appendix gives formal proofs for the statements whose main-text presentations are abbreviated.

## A.1 EVEN-BUDGET COLLAPSE

Let $X \sim \mathrm{Bin}(2n-1, q)$ be the number of correct responses among the first $2n-1$ calls and let $Y \sim \mathrm{Bernoulli}(q)$ be the last call, independent of $X$. For $2n$ binary votes with uniform random tie-breaking, the success probability is

$$\mathbb{P}(X + Y > n) + \frac{1}{2}\mathbb{P}(X + Y = n).$$

Compared with the preceding odd-budget accuracy $\mathbb{P}(X \geq n)$, only the boundary events $X = n-1$ and $X = n$ matter. Since

$$q\mathbb{P}(X = n-1) = (1-q)\mathbb{P}(X = n),$$

which follows from $\binom{2n-1}{n-1} = \binom{2n-1}{n}$, the two boundary contributions cancel and this $2n$-call rule has the same success probability as the $(2n-1)$-call strict-majority rule.

## A.2 PROOF OF PROPOSITION 3.1

Conditional on $q_i$, the two correctness indicators are independent Bernoulli variables with success probability $q_i$. Therefore

$$\mathbb{E}[B_{i1} \mid q_i] = q_i, \qquad \mathbb{E}[B_{i1}B_{i2} \mid q_i] = \mathbb{E}[B_{i1} \mid q_i]\mathbb{E}[B_{i2} \mid q_i] = q_i^2.$$

Averaging over examples gives $\mathbb{E}B_{i1} = \mu$ and $\mathbb{E}(B_{i1}B_{i2}) = \nu$. Since $B_{ij}$ is binary,

$$\mathrm{Var}(B_{ij}) = \mu(1-\mu), \qquad \mathrm{Cov}(B_{i1}, B_{i2}) = \nu - \mu^2,$$

which yields (2) whenever $0 < \mu < 1$.

The feasibility inequalities follow from $0 \leq Q \leq 1$: Jensen's inequality gives $\nu = \mathbb{E}Q^2 \geq (\mathbb{E}Q)^2 = \mu^2$, and $Q^2 \leq Q$ gives $\nu \leq \mu$. Conversely, every pair satisfying $\mu^2 \leq \nu \leq \mu$ is attainable. If $\mu = 0$ or $\mu = 1$, the only possible laws are $\delta_0$ or $\delta_1$. For $0 < \mu < 1$, the two-point law

$$\pi = (1-p)\delta_0 + p\delta_a, \qquad a = \frac{\nu}{\mu}, \qquad p = \frac{\mu^2}{\nu},$$

with the limiting interpretation when $\nu = \mu^2$, has mean $\mu$ and second moment $\nu$ because $a \in [\mu, 1]$ and $p \in [0, 1]$. This proves exact feasibility. Finally, (3) follows from

$$\mathbb{P}(B_{i1} = 1, B_{i2} = 1) = \mathbb{E}Q^2, \qquad \mathbb{P}(B_{i1} = 1, B_{i2} = 0) = \mathbb{P}(B_{i1} = 0, B_{i2} = 1) = \mathbb{E}[Q(1-Q)].$$

## A.3 MEAN-ONLY EXAMPLES AND THE ONE-CALL ENVELOPE

For $\pi_\mu^{\mathrm{det}} = (1-\mu)\delta_0 + \mu\delta_1$, the majority polynomial satisfies $P_n(0) = 0$ and $P_n(1) = 1$, hence $V_n(\pi_\mu^{\mathrm{det}}) = \mu$ for every odd budget. For $\pi_\mu^{\mathrm{hom}} = \delta_\mu$, $V_n(\pi_\mu^{\mathrm{hom}}) = P_n(\mu)$ by definition. The endpoint formula follows by evaluating (8) for these two laws.

For three votes, the mean-only lower and upper ranges are

$$L_1^{\mathrm{one}}(\mu) = \begin{cases} P_1(\mu), & \mu \leq 1/4, \\ (9\mu - 1)/8, & \mu \geq 1/4, \end{cases} \qquad U_1^{\mathrm{one}}(\mu) = \begin{cases} 9\mu/8, & \mu \leq 3/4, \\ P_1(\mu), & \mu \geq 3/4. \end{cases}$$

Indeed,

$$P_1(q) = 3q^2 - 2q^3, \qquad P_1'(q) = 6q(1-q),$$

so the lower convex envelope of $P_1$ on $[0, 1]$ is $P_1(q)$ on $[0, 1/4]$ and the tangent chord

$$\ell_-(q) = \frac{9q - 1}{8}$$

on $[1/4, 1]$. The two pieces match in value and slope at $q = 1/4$, and $\ell_-(1) = P_1(1) = 1$. Jensen's inequality applied to the lower convex envelope gives the lower bound; the linear branch is attained by laws supported on $\{1/4, 1\}$. By the symmetry $P_1(q) = 1 - P_1(1-q)$, the upper concave envelope is $\ell_+(q) = 9q/8$ on $[0, 3/4]$ and $P_1(q)$ on $[3/4, 1]$; the linear branch is attained by laws supported on $\{0, 3/4\}$. This proves the upper formula as well. Since $L_1^{\mathrm{one}}(\mu) \leq \mu$ for every interior $\mu$, one labeled call alone cannot guarantee three-vote improvement.

### A.4 Proof of Theorem 4.1

The moment class (4) is tight and weakly closed on the compact space $[0, 1]$, hence compact. Because $P_n$ is continuous, $\pi \mapsto \int P_n \, d\pi$ is continuous and linear, so both primal extrema are attained.

An extremizer may be chosen to have at most three support points. The feasible set is a two-moment slice of the probability simplex, and a linear functional attains its maximum and minimum at extreme points of this compact convex slice. Under the three linear constraints $\int 1 \, d\pi = 1$, $\int q \, d\pi = \mu$, and $\int q^2 \, d\pi = \nu$, every extreme point has support size at most three. In the atomic case this is immediate: if four distinct support points $q_1, \ldots, q_4$ carry positive mass, then the four vectors $(1, q_j, q_j^2) \in \mathbb{R}^3$ are linearly dependent, so a nonzero signed perturbation preserves all three moments; small positive and negative perturbation sizes keep the measure nonnegative, contradicting extremality. The general case follows from the finite-moment extreme-point theorem of Winkler–Caratheodory.

For the quadratic dual, first consider the upper endpoint at an interior moment pair $\mu^2 < \nu < \mu$. Any quadratic $r(q) = a + bq + cq^2$ satisfying $r(q) \geq P_n(q)$ on $[0, 1]$ gives

$$\int P_n(q) \, d\pi(q) \leq \int r(q) \, d\pi(q) = a + b\mu + c\nu$$

for every $\pi \in \mathcal{M}(\mu, \nu)$, so the dual majorant value is an upper bound. To see exactness and attainment, consider the compact convex set

$$\mathcal{A}_n = \operatorname{conv}\{(1, q, q^2, P_n(q)) : q \in [0, 1]\} \subset \mathbb{R}^4.$$

Its vertical section at $(1, \mu, \nu)$ is the closed interval $\{(1, \mu, \nu, v) : L_n(\mu, \nu) \leq v \leq U_n(\mu, \nu)\}$. At an interior moment pair, a supporting hyperplane at the upper boundary has a nonzero last coefficient: otherwise it would support the projected moment region at an interior point. It can therefore be normalized to have last coefficient $-1$, giving coefficients $(a, b, c)$ such that $a + bq + cq^2 \geq P_n(q)$ for every $q$ and $a + b\mu + c\nu = U_n(\mu, \nu)$. Thus the dual minimum is attained and equals the primal upper endpoint. Applying the same argument to the lower boundary, or equivalently to $-P_n$, gives an attained quadratic minorant and the formula for $L_n$.

At a degenerate moment pair, the finite-vote polynomial's bounded second derivative gives matching quadratic envelopes for the unique feasible law. The general continuous-objective extension is restricted to interior moment pairs, omitting degenerate cases such as $(0, 0)$; there the same proof gives attained quadratic majorants and minorants and three-atom primal extremizers. In particular, any polynomial objective, or any linear combination of finite-vote accuracies, is optimized in the same three-atom moment space. Quantities such as the largest possible marginal gain $V_3 - V_1$ are therefore obtained by running the same program with objective $P_3 - P_1$.

### A.5 Proof of Theorem 4.2

For three votes,

$$P_1(q) = 3q^2 - 2q^3, \qquad V_1 = 3\nu - 2m_3, \quad m_3 := \mathbb{E}Q^3.$$

Thus the upper bound on $V_1$ is obtained by minimizing $m_3$, and the lower bound by maximizing $m_3$, under the first two moments.

For any $a \in [0, 1]$,

$$q(q - a)^2 \geq 0.$$

Expanding and integrating gives

$$m_3 - 2a\nu + a^2\mu \geq 0.$$

Choosing $a = \nu/\mu$ for $0 < \mu < 1$ yields

$$m_3 \geq \frac{\nu^2}{\mu}.$$

Equality is attained by a law supported on $\{0, \nu/\mu\}$ with the appropriate weights, so

$$U_1(\mu, \nu) = 3\nu - 2\frac{\nu^2}{\mu}.$$

Similarly,

$$(1-q)(q-b)^2 \geq 0$$

implies

$$\nu - 2b\mu + b^2 - m_3 + 2b\nu - b^2\mu \geq 0.$$

Choosing $b = (\mu - \nu)/(1-\mu)$ gives

$$m_3 \leq \nu - \frac{(\mu-\nu)^2}{1-\mu}.$$

Equality is attained by a law supported on $\{(\mu-\nu)/(1-\mu), 1\}$, hence

$$L_1(\mu,\nu) = 3\nu - 2\left\{\nu - \frac{(\mu-\nu)^2}{1-\mu}\right\} = \nu + 2\frac{(\mu-\nu)^2}{1-\mu}.$$

The endpoint cases $\mu = 0$ and $\mu = 1$ are forced by feasibility.

Substituting $\nu = \mu^2 + \rho\mu(1-\mu)$ gives the equivalent $\rho$-form in (6)–(7). Direct subtraction yields

$$U_1 - L_1 = 2\mu(1-\mu)\rho(1-\rho) \leq 2 \cdot \frac{1}{4} \cdot \frac{1}{4} = \frac{1}{8}.$$

Finally, for $\mu > 1/2$, uniform improvement over one call is equivalent to $L_1(\mu,\nu) > \mu$. By (6),

$$L_1 - \mu = \mu(1-\mu)(1-\rho)\{2\mu(1-\rho) - 1\}.$$

The prefactor is positive except in the degenerate deterministic case, so strict improvement holds exactly when $2\mu(1-\rho) > 1$, i.e.,

$$\rho < 1 - \frac{1}{2\mu}.$$

Figure 3 visualizes the three-vote gain lower bound $L_1 - \mu$ over the feasible two-call moment region. The zero contour is exactly the theorem's improvement boundary $\rho = 1 - 1/(2\mu)$ for $\mu > 1/2$.

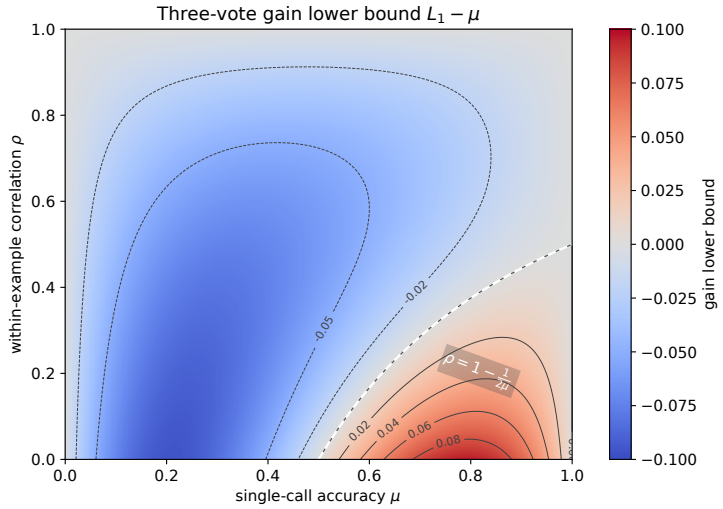


Figure 3: Three-vote gain lower bound $L_1 - \mu$ implied by Theorem 4.2. Warm regions guarantee a uniform gain from three votes over one call; cool regions indicate that no such uniform gain is guaranteed from the two-call moments alone. The dashed white contour is the zero-gain boundary $\rho = 1 - 1/(2\mu)$.

### A.6 Proof of Theorem 4.3

Let

$$h(q) = \mathbf{1}\{q > 1/2\} + \frac{1}{2}\mathbf{1}\{q = 1/2\}.$$

The upper endpoint is the supremum of $\int h\, d\pi$ over $\mathcal{M}(\mu,\nu)$. The deterministic cases $\mu = 0, 1$ are immediate. The exceptional point $(\mu,\nu) = (1/2, 1/4)$ forces $Q \equiv 1/2$, so the value is $1/2$.

For the first nontrivial upper branch, assume $\mu \le 1/2$ and $\nu \le \mu/2$. For $a < 1/2$ define

$$\psi_a(q) = \left(\frac{q-a}{1/2-a}\right)^2.$$

Then $\psi_a(q) \ge h(q)$ on $[0,1]$: it is nonnegative below the threshold, equals one at the threshold, and is at least one above it. Therefore

$$\int h\, d\pi \le \frac{\nu - 2a\mu + a^2}{(1/2-a)^2}.$$

Minimizing the right side over $a < 1/2$ gives

$$a^\star = \frac{\mu/2 - \nu}{1/2 - \mu}$$

and the value

$$\frac{\nu - \mu^2}{\nu + 1/4 - \mu}.$$

Here the denominator is positive away from the exceptional point because

$$\nu + \frac{1}{4} - \mu = (\nu - \mu^2) + (\mu - 1/2)^2.$$

On this branch, feasibility gives $a^\star \le \mu < 1/2$, while $\nu \le \mu/2$ is equivalent to $a^\star \ge 0$. Thus the optimizing quadratic is admissible. If $\nu = \mu^2$, the degenerate law $\delta_\mu$ is sharp. Otherwise, put $t = 1/2 + \varepsilon$ and

$$x_\varepsilon = \frac{\mu t - \nu}{t - \mu}, \qquad p_\varepsilon = \frac{\mu - x_\varepsilon}{t - x_\varepsilon}.$$

For all sufficiently small $\varepsilon > 0$, the branch inequalities give $0 \le x_\varepsilon < 1/2 < t$ and $0 \le p_\varepsilon \le 1$. The two-point law $(1-p_\varepsilon)\delta_{x_\varepsilon} + p_\varepsilon \delta_t$ matches $(\mu, \nu)$, and its endpoint value is $p_\varepsilon$, which tends to $(\nu - \mu^2)/(\nu + 1/4 - \mu)$ as $\varepsilon \downarrow 0$.

For the remaining upper branch, the quadratic

$$\psi_0(q) = 3q - 2q^2$$

satisfies $\psi_0(q) \ge h(q)$ on $[0,1]$: it is nonnegative on $[0, 1/2)$, equals one at $1/2$, and is at least one on $(1/2, 1]$. Hence

$$\int h\, d\pi \le 3\mu - 2\nu,$$

with the trivial additional bound $\int h\, d\pi \le 1$. If $3\mu - 2\nu > 1$, a two-point law supported on $\{x, 1\}$ with $x = (\mu - \nu)/(1-\mu) > 1/2$ attains endpoint value one whenever this branch is active. If $3\mu - 2\nu \le 1$, set $s = 1/2 + \varepsilon$ and use support $\{0, s, 1\}$ with weights

$$w_s = \frac{\mu - \nu}{s(1-s)}, \qquad w_1 = \frac{\nu - s\mu}{1-s}, \qquad w_0 = 1 - w_s - w_1.$$

They match the two moments. The weights are nonnegative for all sufficiently small $\varepsilon > 0$. Feasibility gives $w_s \ge 0$. On the present branch, either $\mu > 1/2$ or $\nu > \mu/2$, and feasibility gives $\nu/\mu > 1/2$ when $\mu > 0$; hence $w_1 \ge 0$ after taking $s \in (1/2, \nu/\mu)$. Finally,

$$w_0 = \frac{\nu - \mu + s(1-\mu)}{s}.$$

If $1 - 3\mu + 2\nu > 0$, then $w_0 > 0$ for $s$ close to $1/2$. On the boundary $1 - 3\mu + 2\nu = 0$, this formula gives $w_0 = (1-\mu)(s - 1/2)/s \ge 0$. Since $s > 1/2$, the endpoint value is $w_s + w_1 = 1 - w_0$, which tends to $3\mu - 2\nu$ as $\varepsilon \downarrow 0$. This proves the stated $U_\infty$ formula.

For the lower endpoint, use the exact symmetry

$$h(q) + h(1-q) = 1,$$

including at $q = 1/2$, where both terms equal $1/2$. If $\widetilde{Q} = 1 - Q$, then

$$\mathbb{E}\widetilde{Q} = 1 - \mu, \qquad \mathbb{E}\widetilde{Q}^2 = 1 - 2\mu + \nu.$$

Therefore

$$L_\infty(\mu, \nu) = 1 - U_\infty(1-\mu,\, 1 - 2\mu + \nu),$$

and substituting the upper formula with these reflected moments gives the displayed reflection formula. The same limiting constructions prove sharpness of the lower bound.

### A.7 Proof of Proposition G.1

Let $X = Q - 1/2$. The log-partition function

$$\psi(\lambda, \kappa) = \log \int_{-1/2}^{1/2} e^{\lambda x + \kappa x^2}\, dx$$

is finite and smooth for all $(\lambda, \kappa) \in \mathbb{R}^2$. Differentiating under the integral sign gives

$$\nabla\psi(\lambda, \kappa) = \big(\mathbb{E}_{\lambda,\kappa} X, \mathbb{E}_{\lambda,\kappa} X^2\big),$$

and its Hessian is the covariance matrix of $(X, X^2)$ under the corresponding density. This covariance matrix is positive definite because the curve $x \mapsto (x, x^2)$ is not contained in any affine line on an interval of positive length. Hence $\psi$ is strictly convex.

The convex support of $(X, X^2)$ has interior

$$\{(m, s) : -1/2 < m < 1/2,\ m^2 < s < 1/4\}.$$

Let the target be $(m, s) = (\mu - 1/2, \nu - \mu + 1/4)$ in this interior. The dual objective $J_{\mu,\nu}(\lambda, \kappa) = \psi(\lambda, \kappa) - \lambda m - \kappa s$ is strictly convex. It is also coercive: along any sequence $(\lambda_j, \kappa_j)$ with norm tending to infinity, a subsequential direction $u$ exposes a strictly positive gap between the support function $\sup_x u \cdot (x, x^2)$ and $u \cdot (m, s)$, because $(m, s)$ is an interior point of the convex support. The compact-support Laplace bound for $\psi$ then makes $J_{\mu,\nu}(\lambda_j, \kappa_j) \to \infty$. Hence $J_{\mu,\nu}$ has a unique minimizer, and its first-order conditions give

$$\mathbb{E}Q = \mu, \qquad \mathbb{E}Q^2 = \nu.$$

For entropy optimality, let $g$ be any absolutely continuous density on $[0, 1]$ with the same two moments as $f_{\lambda,\kappa}$. Since $\log f_{\lambda,\kappa}$ is affine in $1$, $q - 1/2$, and $(q - 1/2)^2$, matching moments gives

$$H(g) = H(f_{\lambda,\kappa}) - D_{\mathrm{KL}}(g \,\|\, f_{\lambda,\kappa}).$$

Nonnegative KL divergence, zero exactly when $g = f_{\lambda,\kappa}$ almost everywhere, proves uniqueness.

Finally, for $q \in [0, 1]$,

$$\frac{f_{\lambda,\kappa}(q)}{f_{\lambda,\kappa}(1 - q)} = \exp\{2\lambda(q - 1/2)\}.$$

The map $\lambda \mapsto \mathbb{E}_{\lambda,\kappa} Q$ is strictly increasing for fixed $\kappa$ because its derivative is $\mathrm{Var}_{\lambda,\kappa}(Q) > 0$, and at $\lambda = 0$ the density is symmetric around $1/2$. Hence $\mathrm{sign}(\lambda) = \mathrm{sign}(\mu - 1/2)$.

### A.8 Proof of Proposition G.2

Let $Z, E_1, E_2$ be independent standard normal variables. Conditional on $Z$, two independent latent responses from the LDGP model satisfy

$$Q^2 = \mathbb{P}(E_1 \le \eta - \gamma Z,\ E_2 \le \eta - \gamma Z \mid Z).$$

After averaging over $Z$, this is the probability that the two normal variables $E_1 + \gamma Z$ and $E_2 + \gamma Z$ are both at most $\eta$. After standardization, they are bivariate standard normal with common threshold $t = \eta/\sqrt{1 + \gamma^2}$ and correlation $r = \gamma^2/(1 + \gamma^2)$. Hence

$$\mathbb{E}Q = \Phi(t), \qquad \mathbb{E}Q^2 = \Phi_2(t, t; r).$$

For a target $\mu$, $t = \Phi^{-1}(\mu)$ is fixed. As $r$ increases from 0 to 1, the bivariate normal cdf $\Phi_2(t, t; r)$ increases continuously from $\Phi(t)^2 = \mu^2$ to $\Phi(t) = \mu$. Plackett's formula gives strict monotonicity on $(0, 1)$

$$\frac{\partial}{\partial r}\Phi_2(t, t; r) = \phi_2(t, t; r) > 0.$$

For an interior feasible $\nu$, there is therefore a unique $r \in (0, 1)$ matching $\nu$, and then $\gamma = \sqrt{r/(1 - r)}$ and $\eta = t/\sqrt{1 - r}$ are unique. The boundary $\nu = \mu^2$ corresponds to $r = 0$ and $\gamma = 0$, while $\nu = \mu$ is the degenerate limit $r \uparrow 1$ and $\gamma \to \infty$ rather than a finite-parameter interior law. The formula for $V_n^{\mathrm{LDGP}}$ is the definition of the completed curve, and the endpoint follows from

$$\mathbb{P}\{\Phi(\eta - \gamma Z) > 1/2\} = \mathbb{P}(\eta - \gamma Z > 0) = \Phi(\eta/\gamma)$$

when $\gamma > 0$. When $\gamma = 0$, $Q \equiv \mu$; applying the definition of $V_\infty$ gives endpoint 1, $1/2$, or 0 according as $\mu > 1/2$, $\mu = 1/2$, or $\mu < 1/2$.

# B PAIR-TABLE CONFIDENCE AND HELD-OUT CONFIRMATION

## B.1 GAIN BOUNDS AND BUDGET COMPARISONS

For three votes, the lower endpoint of the projected gain is the convex minimization

$$\underline{\Delta}^{\mathrm{W}}_{1,\alpha} = \min_{(\mu,\nu)\in\mathcal{C}^{\mathrm{W}}_{\alpha}} \left\{ \nu - \mu + \frac{2(\mu-\nu)^2}{1-\mu} \right\}, \tag{10}$$

with the endpoint convention from Theorem 4.2. A positive value establishes a gain at the nominal Wald confidence level over the same protocol's single-call accuracy. The moment pair and single-call baseline are optimized together, rather than subtracting separately calibrated accuracy limits. The pair-table covariance and held-out calculations follow below.

For $J$ fixed protocols, use $\mathcal{C}^{\mathrm{W}}_{\alpha/J}$ for each to obtain nominal simultaneous confidence $1-\alpha$ by Bonferroni correction. Within a protocol, the same moment region covers all vote budgets. A positive lower gain bound establishes improvement; the upper bound describes how large that improvement could be. An interval containing zero leaves the sign unresolved.

One may therefore compare the gain intervals for several budgets using the same two-call validation data. Whenever the moment region contains the true pair, all projected comparisons hold together, including comparisons chosen after inspecting the intervals. This supports budget selection even when accuracy varies nonmonotonically with the number of calls.

For a required accuracy target, the projected accuracy intervals identify the smallest evaluated budget whose lower bound exceeds the target. A budget whose upper bound falls below the target can be excluded. When deciding whether to increase an existing budget, the corresponding gain interval instead answers whether the additional calls meet a required improvement. Both decisions use the same two-call validation sample. The budget comparison therefore separates the cost of learning about repetition from the number of calls eventually used at deployment.

## B.2 WALD GAIN PROJECTION FROM THE PAIR TABLE

Let $n_{ab}$ count example draws with first-call score $a$ and second-call score $b$, where $a, b \in \{0,1\}$. Put $N = \sum_{a,b} n_{ab}$ and $D = n_{01} + n_{10}$. The estimates and covariance are

$$\hat{\mu} = \frac{2n_{11}+D}{2N}, \qquad \hat{\nu} = \frac{n_{11}}{N}, \qquad \hat{\Sigma} = \frac{N}{N-1}\left\{ \begin{pmatrix} (\hat{\mu}+\hat{\nu})/2 & \hat{\nu} \\ \hat{\nu} & \hat{\nu} \end{pmatrix} - \hat{\theta}\hat{\theta}^{\top} \right\}. \tag{11}$$

This follows because $X_i$ takes only the values $(0,0)$, $(1/2,0)$, and $(1,1)$. Thus the pair counts suffice to reconstruct both the moment estimate and its Wald ellipse.

For each of Q04 and N05, we minimize the sharp lower gain in (10) over $\mathcal{C}^{\mathrm{W}}_{0.025}$. The objective is convex: its nonlinear term is a quadratic divided by the positive affine function $1-\mu$. Both fitted covariance matrices are nonsingular, and the minimizing moment pairs lie in the feasible interior. The two per-setting confidence levels of $97.5\%$ give nominal joint confidence $95\%$ by Bonferroni correction, conditional on the fixed development-selected protocols. The projection gives a lower confidence bound on the population gain.

## B.3 HELD-OUT POPULATIONS AND INFERENCE PROTOCOLS

Setting Q04 is QQP with Llama 3.1-8B; N05 is QNLI with Phi-4-mini. Both protocols were chosen using development data and fixed before confirmation. Each setting uses $6{,}144$ uniform independent draws with replacement from a source pool disjoint from development and the direct-prompt GLUE evaluation. The Q04 pool contains $45{,}406$ QQP pairs with more than 30 lowercased alphanumeric words across the two questions; the N05 pool contains $72{,}414$ QNLI examples. Repeated source examples receive fresh calls. A subset of $1{,}024$ draws, chosen before observing outcomes, receives a third call for the empirical check.

Both models run through Ollama at temperature $2.0$ and top-$p = 0.95$, without native thinking (Ollama, 2026a;b;c). Q04 uses a direct question about whether the two inputs ask the same thing. N05 uses eight fixed formulations: the direct entailment question has probability $9/16$, and each

alternative has probability $1/16$. Formulations are sampled independently on every call, and negative formulations are inverted to the original task label before voting. This defines one fixed randomized protocol; exact templates and remaining generation settings are supplied with the code.

Responses are parsed as binary Yes/No labels and scored against the task label. Invalid or truncated responses map to canonical No regardless of the formulation; all confirmation responses were valid. The sampling unit is the example draw, and the confidence calculation assumes fresh, conditionally independent calls under each fixed protocol.

### B.4 Pair counts and gain bounds

Table 1 reports the pair counts for Q04 and N05; Table 2 gives their moment estimates and three-vote gain bounds. $W$ and $C$ denote incorrect and correct scores. The plug-in endpoint uses estimated moments; the Wald lower bound also accounts for sampling uncertainty.

| Setting | $N$ | $WW$ | $WC$ | $CW$ | $CC$ |
|---|---|---|---|---|---|
| Q04: QQP, direct, length-defined | 6144 | 966 | 1156 | 1154 | 2868 |
| N05: QNLI, fixed prompt sampler | 6144 | 709 | 923 | 847 | 3665 |

Table 1: First-two-call counts for Q04 and N05, used for the Wald moment regions. The third calls are excluded.

| Setting | $\hat{\mu}$ (%) | $\hat{\rho}$ (%) | Plug-in lower gain | Wald lower gain |
|---|---|---|---|---|
| Q04 (QQP) | 65.48 | 16.83 | $+1.68$ | $+0.70$ |
| N05 (QNLI) | 74.06 | 25.03 | $+1.59$ | $+0.72$ |

Table 2: Three-vote gain lower bounds from the first two calls. Gains are differences of percentage accuracies: the plug-in column is $100\{L_1(\hat{\mu},\hat{\nu})-\hat{\mu}\}$ and the Wald column is $100\underline{\Delta}^{\mathrm{W}}_{1,0.025}$, at nominal joint confidence $95\%$ across Q04 and N05. Values are rounded after calculation. Table 1 reports the pair counts.

### B.5 Empirical three-call gains and a direct-prompt comparison

For each of Q04 and N05 let $m = 1024$, $S_i = B_{i1} + B_{i2} + B_{i3}$ on its three-call subset, and

$$Z_i = \mathbf{1}\{S_i \geq 2\} - S_i/3, \qquad \widehat{\Delta}_{\mathrm{check}} = m^{-1}\sum_i Z_i, \qquad \widehat{\mathrm{se}} = \mathrm{sd}(Z_1,\ldots,Z_m)/\sqrt{m}.$$

The paired Wald $95\%$ interval is $\widehat{\Delta}_{\mathrm{check}} \pm 1.96\,\widehat{\mathrm{se}}$, using the sample standard deviation. Its single-call baseline averages the three calls on the subset. These checks reuse the first two calls used to construct the gain bound and do not independently test its coverage.

| Setting | $m$ | Single (%) | Majority (%) | Gain | Paired Wald 95% interval |
|---|---|---|---|---|---|
| Q04 (QQP) | 1024 | 64.65 | 67.87 | $+3.22$ | $[+1.70, +4.75]$ |
| N05 (QNLI) | 1024 | 76.50 | 81.15 | $+4.65$ | $[+3.38, +5.93]$ |

Table 3: Predesignated three-call checks for Q04 and N05. Gains and their intervals are differences of percentage accuracies, $100\widehat{\Delta}_{\mathrm{check}}$. The paired Wald intervals summarize observed gains; Table 2 projects two-call moment ellipses through the sharp lower gain.

A matched direct-prompt QNLI control achieves $84.28\%$ single-call accuracy on the same $1{,}024$ N05 draws, compared with $81.15\%$ for N05's randomized protocol after three votes. This control is distinct from the direct-prompt QNLI settings, which use a different evaluation pool. The positive lower gain bound establishes improvement over N05's own single-call baseline; it does not establish superiority to direct prompting.

# C COMPUTATION DETAILS

**Finite-budget computation.** Finite-budget endpoints use the quadratic-envelope dual in Theorem 4.1. An exchange optimization constructs quadratic envelopes, checking their polynomial slack at both endpoints and at all real stationary points in $[0, 1]$. Adjusting the quadratic constant for numerical slack gives an outer endpoint. Wald projection optimizes over the moment ellipse as well as the latent law. Degenerate moment pairs are evaluated from their uniquely determined law.

**Wald regions and projection.** The sample covariance is $\hat{\Sigma} = (N-1)^{-1} \sum_i (X_i - \hat{\theta})(X_i - \hat{\theta})^\top$, equivalently (11). Every reported two-call confidence interval uses the feasible Wald region (9). The direct-prompt QQP and QNLI settings and the GSM8K curve use $\alpha = 0.05$ per setting; Q04 and N05 use $\alpha = 0.025$ each. The low-temperature gain comparison also reports an 18-condition Bonferroni adjustment. The without-replacement GLUE and GSM8K samples use the example-level covariance without a finite-population correction. All confidence levels have a large-sample interpretation.

| Quantity | Minimum SE | Median SE | Maximum SE |
|---|---|---|---|
| $\hat{\mu}$ | 0.38 | 0.42 | 0.51 |
| $\hat{\nu}$ | 0.41 | 0.48 | 0.55 |
| $\hat{\rho}$ | 0.33 | 0.74 | 1.19 |

Table 4: Example-level standard-error ranges for the direct-prompt QQP and QNLI settings, with every standard error multiplied by 100. The first two rows are obtained from the covariance matrix of $X_i = ((B_{i1} + B_{i2})/2, B_{i1}B_{i2})$; the $\hat{\rho}$ row uses the delta method. With $N = 8192$, both moments have standard errors below one unit on the displayed percentage scale.

**Mixture calculations.** With fixed mixture weights, a component policy is sampled independently on each call. The auxiliary mixture calculations align the recorded examples across component policies. For component $a$, let $S_{ia}$ count correct responses among its five calls on example $i$, and let $w_a$ be its selection probability. With $\hat{q}_{ia} = S_{ia}/5$, the mixture uses $\hat{q}_{i,\mathrm{mix}} = \sum_a w_a \hat{q}_{ia}$ and averages $P_n(\hat{q}_{i,\mathrm{mix}})$ for vote curves. These are finite-repeat plug-in summaries, not observations from executed mixture policies. For the second moment, within-component estimates use the proportion of response pairs that are both correct; cross-component estimates multiply the two success rates.

**Direct-prompt QQP and QNLI settings.** For each dataset, 8,192 examples are sampled uniformly without replacement from the QNLI or QQP training pool. Llama 3.1-8B, Phi-4-mini, and Qwen2.5-7B are evaluated at temperatures 0.2, 1.0, and 2.0, with five calls per example. The nine model–temperature combinations define the direct-prompt settings for each dataset; Table 7 gives the ID mapping. Fixed prompts request a Yes/No answer. The parser reads the JSON answer, falling back to a standalone Yes/No token; all recorded calls parse successfully, so label majority and correctness majority coincide on these records.

**Empirical vote estimator.** For GLUE and GSM8K, let $R$ be the number of calls and $S_i$ the number correct on example $i$. The empirical majority accuracy averages over all $K$-subsets:

$$T_{iK} = \frac{\sum_{\ell > K/2} \binom{S_i}{\ell}\binom{R-S_i}{K-\ell}}{\binom{R}{K}}, \qquad \widehat{V}_{(K-1)/2} = \frac{1}{N}\sum_i T_{iK},$$

Impossible binomial coefficients are zero, and $T_{i1} = S_i/R$. The resulting empirical accuracies are point summaries of the recorded calls. Confidence intervals obtained from two calls use Wald moment projection; the three-call checks in Appendix B.5 use paired scalar Wald intervals.

**Call accounting.** Two-call validation uses $2N$ model calls and at most $N$ ground-truth labels, with each label scoring both calls. Table 4 reports example-level standard errors for these coordinates.

# D ADDITIONAL GAIN BOUNDS AND HIGHER-BUDGET RESULTS

## D.1 SETTING N04: LOW-TEMPERATURE QWEN DECISION

For N04, the first two repeats give pair counts $(n_{00}, n_{01}, n_{10}, n_{11}) = (1272, 31, 20, 6869)$. Let $d = \mathbb{P}(B_1 \neq B_2) = 2(\mu - \nu)$. The polynomial inequalities

$$P_1(q) - q \leq q(1-q), \qquad P_2(q) - q \leq \frac{7\sqrt{7}}{18} q(1-q)$$

imply $V_1 - \mu \leq d/2$ and $V_2 - \mu \leq (7\sqrt{7}/18)d/2$. Indeed, the ratios to $q(1-q)$ are $2q-1$ and $(2q-1)\{1 + 3q(1-q)\}$; maximizing them on $[0,1]$ gives the constants above.

The linear projection of the Wald ellipse gives

$$d_U = \hat{d} + \sqrt{\chi^2_{2,1-\alpha/J}\, \hat{s}_d^2/N}, \qquad \hat{d} = \frac{51}{8192}, \qquad \hat{s}_d^2 = \frac{N}{N-1}\hat{d}(1-\hat{d}).$$

Here $J = 1$ gives a single-condition bound, while $J = 18$ accounts for choosing from the fixed 18-condition family. Combining this upper limit with the polynomial inequalities bounds both vote gains on the same moment-region event. The bounds have nominal large-sample confidence $1 - \alpha$.

| Calls | Point gain interval | Wald upper | 18-setting Wald upper |
|---|---|---|---|
| 3 | $[-0.2990, +0.3090]$ | 0.418 | 0.460 |
| 5 | $[-0.3203, +0.3203]$ | 0.430 | 0.474 |

Table 5: Setting N04: QNLI/Qwen2.5-7B at $T = 0.2$. All limits are on the absolute percentage-accuracy scale $100(V_n - V_0)$. The last two columns combine the Wald upper limit for disagreement with the polynomial gain bounds, at nominal level $1 - \alpha = 0.95$.

The full five-repeat record gives empirical gains of $+0.011$ and $-0.029$ on the percentage-accuracy scale. These point estimates use the subset estimator; two-call gain bounds use moment projection.

## D.2 GSM8K EVALUATION AND BUDGET CURVES

This supplementary higher-budget evaluation uses Qwen3-14B on GSM8K (Cobbe et al., 2021; Qwen Team, 2025). Table 6 reports the empirical accuracy curve and Wald-projected two-moment bounds.

The evaluation uses 496 problems from a uniform draw of the official GSM8K test set, after excluding the 16 problems used during development. Qwen3-14B generates nine fresh calls per problem at temperature 0.7 and top-$p = 0.95$, with native thinking disabled. The fixed evaluator removes commas and requires exact numeric agreement with the gold answer. All recorded responses were successfully generated and parsed.

The empirical curves use the subset estimator in Appendix C. The Wald intervals use only the first two calls, whose pooled pair counts are 32 both-wrong, 23 discordant, and 441 both-correct. All budgets use the same moment ellipse. The empirical three- and nine-vote gains over the single-call accuracy estimated from all nine calls are $+0.42$ and $+0.85$, percentage points, respectively.

| Calls $K$ | Empirical accuracy (nine-call record) | Wald-projected accuracy interval (first-two-call moments) |
|---|---|---|
| 1 | 91.89 | $[88.35, 94.11]$ |
| 3 | 92.31 | $[86.90, 96.13]$ |
| 5 | 92.50 | $[86.11, 96.32]$ |
| 7 | 92.69 | $[85.62, 96.46]$ |
| 9 | 92.74 | $[85.28, 96.58]$ |

Table 6: GSM8K strict-majority correctness accuracies and Wald-projected intervals (%). Empirical accuracies average all subsets of the nine-call record; the moment estimate uses the first two calls, with $\hat{\mu} = 91.23\%$. All intervals are projections of the same nominal 95% Wald region.

Strict-majority correctness guarantees plurality success for the unique correct numeric answer, while split wrong answers can make plurality more accurate. With numeric normalization and either uniform random or earliest-answer tie-breaking, the observed plurality curves are:

| $K$ | 1 | 3 | 5 | 7 | 9 |
|---|---|---|---|---|---|
| uniform-random-tie plurality (%) | 91.89 | 92.51 | 92.83 | 93.02 | 93.25 |
| first-answer-tie plurality (%) | 91.89 | 92.55 | 92.84 | 93.03 | 93.15 |

These plurality accuracies are secondary outcomes. The strict-majority lower bound also bounds plurality accuracy from below; the upper bound applies only to strict-majority accuracy.

### D.3 Matched dependence stress diagnostic

On a matched subset of 128 problems, a cache repeats each of three fresh generations in a block of three, intentionally violating conditional independence. The first two cached calls always agree, collapsing their plug-in interval to a point, yet the observed higher-budget accuracy can differ from that point (by up to 0.78 percentage points). This illustrates the role of conditional independence.

## E Detailed LLM geometry and results

Table 7 identifies the 18 direct-prompt settings; Tables 2 and 3 report Q04 and N05. Q01–Q05 and N01–N05 appear in Figure 2; the remaining settings extend the temperature comparisons. Q denotes QQP and N denotes QNLI. Q04 uses longer QQP pairs, while N05 varies the QNLI wording independently on each call. Both rules were developed and evaluated on separate pools.

Table 7 gives settings and results; Figure 4 shows the direct-prompt QQP and QNLI settings (Ollama, 2026b;c;d). For the 18 direct-prompt settings, the first two calls estimate the moment coordinates and confidence region; all five calls contribute to the subset-averaged empirical gains. Every observed three- and five-vote accuracy lies inside its projected row-wise Wald 95% region. Since the calculations share calls, this is a compatibility check rather than independent coverage validation.

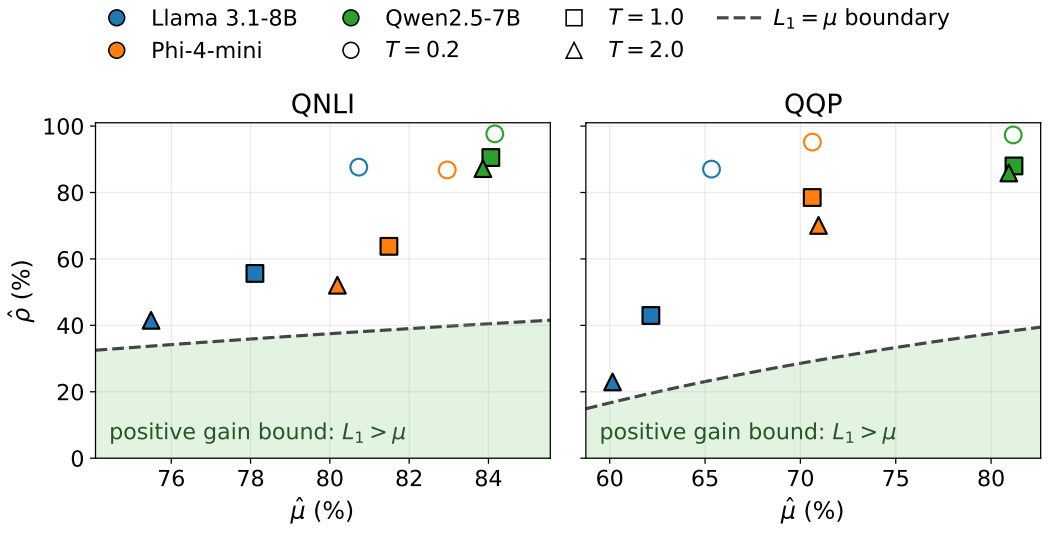


Figure 4: Direct-prompt QNLI settings (left) and QQP settings (right), by model (color) and temperature (shape), with horizontal $\hat{\mu}$ and vertical $\hat{\rho}$ in percent. Table 7 maps models and temperatures to IDs. The dashed curve is the population boundary $\rho = 1 - 1/(2\mu)$, and the shaded region is $L_1 > \mu$. Figure 2 shows the middle-temperature settings and the low-temperature Q05 and N04 alongside Q04 and N05, whose gains appear in its right panel; Table 7 gives the observed gains for the settings shown here.

| Setting | Policy | $\hat{\mu}$ (%) | $\hat{\rho}$ (%) | $100(V_1 - V_0)$ | $100(V_2 - V_0)$ | $[L_1, U_1]_{95}$ (%) | $[L_2, U_2]_{95}$ (%) |
|---|---|---|---|---|---|---|---|
| N06 | Llama 3.1-8B $T = 0.2$ | 80.73 | 87.64 | +0.11 | +0.08 | [78.10,83.59] | [77.67,83.73] |
| N01 | Llama 3.1-8B $T = 1.0$ | 78.11 | 55.63 | +1.39 | +2.03 | [74.61,85.17] | [72.92,86.78] |
| N07 | Llama 3.1-8B $T = 2.0$ | 75.49 | 41.51 | +1.62 | +2.47 | [72.98,84.17] | [71.80,86.99] |
| N08 | Phi-4-mini $T = 0.2$ | 82.96 | 86.79 | +0.11 | +0.08 | [80.46,85.71] | [80.01,85.85] |
| N02 | Phi-4-mini $T = 1.0$ | 81.49 | 63.79 | +0.94 | +1.21 | [78.16,87.13] | [76.70,88.02] |
| N09 | Phi-4-mini $T = 2.0$ | 80.19 | 52.08 | +1.34 | +1.98 | [77.29,87.25] | [75.90,88.84] |
| N04 | Qwen2.5-7B $T = 0.2$ | 84.16 | 97.66 | +0.01 | -0.03 | [82.87,85.45] | [82.85,85.46] |
| N03 | Qwen2.5-7B $T = 1.0$ | 84.06 | 90.53 | +0.08 | +0.17 | [81.98,86.25] | [81.73,86.32] |
| N10 | Qwen2.5-7B $T = 2.0$ | 83.86 | 87.19 | +0.05 | +0.14 | [81.48,86.47] | [81.07,86.60] |
| Q06 | Llama 3.1-8B $T = 0.2$ | 65.34 | 87.04 | -0.04 | -0.06 | [61.60,69.27] | [61.02,69.62] |
| Q01 | Llama 3.1-8B $T = 1.0$ | 62.16 | 42.97 | +1.47 | +1.92 | [56.90,70.98] | [54.07,74.85] |
| Q07 | Llama 3.1-8B $T = 2.0$ | 60.15 | 22.97 | +1.73 | +2.57 | [57.33,68.54] | [55.95,73.17] |
| Q08 | Phi-4-mini $T = 0.2$ | 70.63 | 95.17 | -0.10 | -0.10 | [68.46,72.82] | [68.36,72.88] |
| Q02 | Phi-4-mini $T = 1.0$ | 70.62 | 78.49 | +0.03 | -0.02 | [66.27,75.69] | [65.04,76.39] |
| Q09 | Phi-4-mini $T = 2.0$ | 70.95 | 70.09 | +0.12 | +0.32 | [66.15,77.17] | [64.33,78.39] |
| Q05 | Qwen2.5-7B $T = 0.2$ | 81.16 | 97.33 | +0.03 | +0.02 | [79.70,82.61] | [79.67,82.63] |
| Q03 | Qwen2.5-7B $T = 1.0$ | 81.20 | 88.00 | 0.00 | -0.01 | [78.65,83.96] | [78.24,84.10] |
| Q10 | Qwen2.5-7B $T = 2.0$ | 80.93 | 85.88 | +0.03 | +0.04 | [78.16,84.00] | [77.63,84.18] |

Table 7: Settings and results for the 18 direct-prompt QNLI (N) and QQP (Q) settings. Each row identifies the model, temperature, and setting ID. Each task uses $8192$ examples sampled without replacement, with five calls per example (Appendix C). $V_1$ and $V_2$ denote three- and five-vote accuracy; gains are in percentage points. The moment estimates use the first two calls; empirical gains use all five calls and their single-call baseline. The last two columns are row-wise nominal $95\%$ Wald accuracy regions.

# F Additional empirical summaries

The 18 direct-prompt conditions also support analyses of mixtures and point predictions.

## F.1 Randomized mixture policies

A randomized policy would select a component independently on each call. Table 8 summarizes an auxiliary search for five-vote gain using the aligned plug-in estimates in Appendix C. These fitted, same-data selections require direct validation; they differ from the independently sampled prompt protocol used for the held-out QNLI gain evaluation.

| Dataset | Type | Mixture | $\hat{\mu}$ (%) | $\hat{\rho}$ (%) | $V_2$ (%) | $100(V_2 - \mu)$ |
|---|---|---|---|---|---|---|
| QNLI | pair | 0.45·Llama 3.1-8B (T=2.0) + 0.55·Phi-4-mini (T=2.0) | 77.91 | 37.85 | 81.01 | +3.10 |
| QNLI | triple | 0.40·Llama 3.1-8B (T=2.0) + 0.40·Phi-4-mini (T=2.0) + 0.20·Qwen2.5-7B (T=2.0) | 78.91 | 38.23 | 82.35 | +3.44 |
| QQP | pair | 0.40·Llama 3.1-8B (T=2.0) + 0.60·Qwen2.5-7B (T=0.2) | 72.72 | 42.43 | 77.89 | +5.17 |
| QQP | triple | 0.40·Llama 3.1-8B (T=2.0) + 0.10·Phi-4-mini (T=2.0) + 0.50·Qwen2.5-7B (T=0.2) | 71.73 | 39.28 | 76.76 | +5.04 |

Table 8: Mixtures selected for largest plug-in five-vote gain on the auxiliary grid. Values describe simulated independent component selection using fitted per-example success probabilities. The gain column is a difference of percentage accuracies, not a relative improvement.

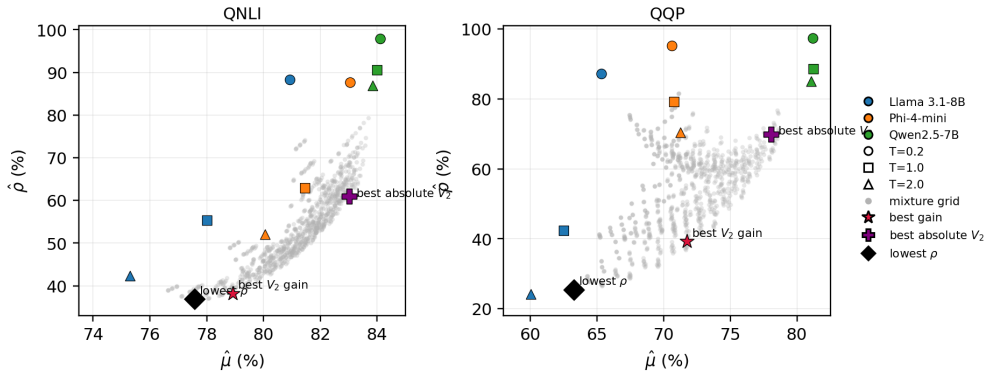


Figure 5: Plug-in mixture summaries in the two-call plane, estimated from recorded base-policy responses. Gray points are mixtures and colored markers are base policies. These fitted mixtures differ from the directly evaluated QNLI prompt sampler.

### F.2 Overtaking examples

Table 9 gives examples where policy $A$ has lower single-call accuracy than policy $B$ but higher accuracy after voting. The differences are paired on the same example set and reported as differences of percentage accuracies, $100\{V_n(A) - V_n(B)\}$.

| Dataset | Policy $A$ | Policy $B$ | $100\Delta_0$ | $100\Delta_1$ | $100\Delta_2$ |
|---|---|---|---|---|---|
| QNLI | N09: Phi-4-mini (T=2.0) | N06: Llama 3.1-8B (T=0.2) | -0.86 | +0.37 | +1.04 |
| QNLI | 0.30·Llama 3.1-8B (T=2.0) + 0.20·Phi-4-mini (T=1.0) + 0.50·Qwen2.5-7B (T=1.0) | Llama 3.1-8B (T=0.2) | -0.04 | +1.95 | +2.67 |
| QQP | 0.40·Llama 3.1-8B (T=2.0) + 0.20·Phi-4-mini (T=2.0) + 0.40·Qwen2.5-7B (T=0.2) | Phi-4-mini (T=1.0) | -0.05 | +3.13 | +4.56 |

Table 9: Policies with lower single-call accuracy can have higher accuracy after three or five votes. Here $\Delta_n = V_n(A) - V_n(B)$, so the displayed differences are on the percentage-accuracy scale. Mixture rows use the auxiliary plug-in calculation; base-policy comparisons use the recorded-repeat estimator.

### F.3 Point summaries of finite-vote accuracy

Table 10 compares three point predictions of finite-vote accuracy against the empirical three- and five-vote estimates in the 18 base-policy runs. The interval midpoint uses the sharp bounds evaluated at the first-two-call moment estimates, separately at each budget, whereas LDGP and maximum entropy select latent laws. LDGP has the lowest mean absolute error (MAE) and root mean squared error (RMSE) at both budgets. Both parametric completions have positive average bias.

| Target | Point summary | Bias | MAE | RMSE | Max abs. error |
|---|---|---|---|---|---|
| $V_1$ | Interval midpoint | +0.34 | 0.36 | 0.61 | 1.81 |
| $V_1$ | LDGP | +0.13 | 0.21 | 0.35 | 1.06 |
| $V_1$ | maximum entropy | +1.04 | 1.04 | 1.19 | 2.14 |
| $V_2$ | Interval midpoint | +0.22 | 0.35 | 0.71 | 2.20 |
| $V_2$ | LDGP | +0.15 | 0.22 | 0.38 | 1.08 |
| $V_2$ | maximum entropy | +1.22 | 1.22 | 1.38 | 2.52 |

Table 10: Point-prediction error across the direct-prompt QQP and QNLI settings. Errors are differences of percentage accuracies; positive bias denotes overprediction.

## G Point-completion constructions

The following statements give the two moment-matched laws used in Section 5.

### G.1 MAXIMUM-ENTROPY COMPLETION

On the real line, fixing mean and variance selects a Gaussian as the maximum-entropy law. On $[0,1]$, maximizing differential entropy relative to Lebesgue measure instead selects a quadratic-exponential density (Jaynes, 1957). For an interior feasible pair $(\mu,\nu)$, define

$$\psi(\lambda,\kappa)=\log\int_0^1 \exp\{\lambda(q-1/2)+\kappa(q-1/2)^2\}\,\mathrm{d}q,$$

with density

$$f_{\lambda,\kappa}(q)=\exp\{\lambda(q-1/2)+\kappa(q-1/2)^2-\psi(\lambda,\kappa)\}\mathbf{1}_{[0,1]}(q). \tag{12}$$

**Proposition G.1.** *For every interior feasible $(\mu,\nu)$ there is a unique $(\lambda,\kappa)$ such that* (12) *has first two moments $(\mu,\nu)$. This density uniquely maximizes differential entropy among absolutely continuous laws on $[0,1]$ with those moments. Equivalently, $(\lambda,\kappa)$ uniquely minimizes*

$$J_{\mu,\nu}(\lambda,\kappa)=\psi(\lambda,\kappa)-\lambda(\mu-1/2)-\kappa(\nu-\mu+1/4).$$

*Moreover,*

$$\operatorname{sign}(\lambda)=\operatorname{sign}(\mu-1/2),\qquad \frac{f_{\lambda,\kappa}(q)}{f_{\lambda,\kappa}(1-q)}=\exp\{2\lambda(q-1/2)\}.$$

The parameter $\lambda$ tilts the density across the decision boundary, while $\kappa$ controls concentration near the center versus the endpoints; the resulting point curve has coordinates $V_n^{\mathrm{ME}}=\int_0^1 P_n(q)f_{\lambda,\kappa}(q)\,\mathrm{d}q$ and $V_\infty^{\mathrm{ME}}=\int_{1/2}^1 f_{\lambda,\kappa}(q)\,\mathrm{d}q$.

### G.2 LATENT-DIFFICULTY GAUSSIAN-PROBIT COMPLETION

Latent-difficulty Gaussian-probit (LDGP) gives a complementary completion from the established normal-ogive/probit item-response model for binary correctness (Albert, 1992; Lord, 1980). Let $Z\sim N(0,1)$ be example difficulty and set

$$Q=\Phi(\eta-\gamma Z),\qquad \gamma\ge 0, \tag{13}$$

where $\Phi$ is the standard normal cdf. The ability parameter $\eta$ shifts the policy relative to the latent difficulty scale, and $\gamma$ controls how much example difficulty spreads the success probabilities.

**Proposition G.2.** *For every interior feasible pair $0<\mu<1$, $\mu^2<\nu<\mu$, there is a unique LDGP law of the form* (13) *matching $(\mu,\nu)$. Writing $t=\eta/\sqrt{1+\gamma^2}$ and $r=\gamma^2/(1+\gamma^2)$, the parameters are characterized by*

$$\mu=\Phi(t),\qquad \nu=\Phi_2(t,t;r),$$

*where $\Phi_2(\cdot,\cdot;r)$ is the bivariate standard normal cdf with correlation $r$. The completed curve is*

$$V_n^{\mathrm{LDGP}}=\mathbb{E}\big[P_n\{\Phi(\eta-\gamma Z)\}\big],\qquad V_\infty^{\mathrm{LDGP}}=\Phi(\eta/\gamma)$$

*for $\gamma>0$. When $\gamma=0$, the law is constant $Q\equiv\mu$, so the endpoint is $1$, $1/2$, or $0$ according as $\mu>1/2$, $\mu=1/2$, or $\mu<1/2$.*

After difficulty-scale normalization, the same two moments determine both LDGP parameters and hence every vote budget. Both completions produce coherent modeled curves inside the sharp bounds. Proofs are in Appendix A.

| $\rho$ | $V_1^{\mathrm{ME}}$ | $V_2^{\mathrm{ME}}$ | $V_\infty^{\mathrm{ME}}$ | $V_1^{\mathrm{LDGP}}$ | $V_2^{\mathrm{LDGP}}$ | $V_\infty^{\mathrm{LDGP}}$ |
|---|---|---|---|---|---|---|
| 0.05 | 77.15 | 81.25 | 97.38 | 77.23 | 81.34 | 96.33 |
| 0.20 | 74.57 | 76.68 | 82.64 | 74.58 | 76.63 | 81.99 |
| 0.50 | 72.64 | 73.52 | 74.95 | 71.56 | 72.08 | 73.13 |
| 0.80 | 71.35 | 71.61 | 71.78 | 70.23 | 70.30 | 70.43 |

Table 11: Moment-matched model curves at fixed $\mu=0.70$. Values are accuracies in percent; each completion produces its finite-vote coordinates and infinite endpoint from one selected latent law.

At fixed one-call accuracy in Table 11, lower correlation produces larger modeled voting gains in both completions. Their finite-budget predictions are relatively close, while their limiting predictions differ more: the two prescriptions allocate latent mass differently across the threshold $q = 1/2$. Appendix F.3 compares their finite-vote predictions with the LLM results. Completions give coherent forecasts across budgets; sharp bounds support decisions across all compatible laws.

## H EXISTING ASSETS, LICENSES, AND TERMS OF USE

Our experiments use existing benchmark datasets, model weights, and local inference software only for research evaluation. We do not redistribute third-party raw data, model weights, or software binaries in the anonymous supplementary materials.

**GLUE QNLI and QQP.** We use QNLI and QQP from GLUE (Wang et al., 2019). GLUE directs users to the original datasets and their terms. QNLI is derived from SQuAD, which is distributed under CC BY-SA 4.0. QQP is derived from the Quora Question Pairs release and is subject to the applicable Quora/Kaggle data-use terms. We use these data only for evaluation and do not redistribute raw examples. Relevant pages: `https://huggingface.co/datasets/nyu-mll/glue`, `https://rajpurkar.github.io/SQuAD-explorer/`, and `https://www.kaggle.com/competitions/quora-question-pairs/rules`.

**Models.** We use `llama3.1:8b`, `phi4-mini`, and `qwen2.5:7b` through Ollama for local inference. Llama 3.1 is governed by the Llama 3.1 Community License and Acceptable Use Policy; Phi-4-mini is released under the MIT License; Qwen2.5-7B-Instruct is released under Apache-2.0. We do not redistribute any model weights. Relevant pages: `https://www.llama.com/llama3_1/license/`, `https://huggingface.co/microsoft/Phi-4-mini-instruct`, and `https://huggingface.co/Qwen/Qwen2.5-7B-Instruct`.

**GSM8K.** We use the official GSM8K test data released with Cobbe et al. (2021). The OpenAI grade-school-math repository distributes the dataset and accompanying code under the MIT License. We do not redistribute the raw problems. Relevant pages: `https://github.com/openai/grade-school-math` and `https://github.com/openai/grade-school-math/blob/master/LICENSE`.

**Qwen3-14B.** We use the official `Qwen/Qwen3-14B` checkpoint (Qwen Team, 2025), released under Apache-2.0. We do not redistribute the model weights. Relevant page: `https://huggingface.co/Qwen/Qwen3-14B`.

**Ollama.** All generations are run locally through Ollama's `/api/generate` endpoint. Ollama is distributed under the MIT License, and its services are subject to its Terms of Service. Relevant pages: `https://github.com/ollama/ollama/blob/main/LICENSE`, `https://docs.ollama.com/api/generate`, and `https://ollama.com/terms`.

The anonymous supplementary materials provide manuscript and figure sources, scored responses, and reconstruction code, without redistributing third-party task text or model weights.