Extrapolation Is an Assumption: Sharp Limits for Pass@k Forecasting

Pahan Dewasurendra
Johns Hopkins University

Abstract

Repeated sampling can reveal capabilities and risks that are invisible in one language-model attempt, motivating forecasts of pass@k from much smaller response banks. We ask what such data support without a parametric law for task difficulty. With b attempts per task, the count distribution identifies exactly the first b moments of the latent success-probability distribution. We prove that the worst-case diameter of pass@k's identified set is exactly twice the best degree-b uniform approximation error for x^k. A theorem of Newman and Rivlin then yields explicit binomial-tail bounds and a sharp extrapolation transition at $k \asymp b^2$. We extend the limit to finite task sets through a Hellinger modulus and to adaptive policies with a per-task cap. We also give two finite-sample confidence intervals: a global polynomial interval that attains the limiting worst-case diameter and a data-adaptive projection interval. At 128 tasks and 78 attempts per task, any honest 95% interval for population pass has expected-length lower bounds 0.276 and 0.686 for pass@1000 and pass@10000. In synthetic audits, narrow taskwise Beta intervals have zero coverage under seven nonparametric alternatives. Across 22 public response banks, model-free intervals include every 10,000-response reference but are necessarily wide at distant horizons. Pass@k extrapolation can be useful, but its uncertainty must disclose the structural assumptions that make it possible.

1 Introduction

Repeated inference is now a practical scaling axis for language models. Large response banks show substantial gains from trying a task many times (Brown et al., 2024). The aggregate pass@k curve is also used to forecast capability and safety at budgets far beyond those directly sampled. Recent methods fit heavy-tailed task-difficulty laws (Schaeffer et al., 2025), use a population Beta-binomial model and adaptive allocation (Kazdan et al., 2025), or place independent Beta posteriors on tasks (Anonymous, 2026; Hariri et al., 2026). These approaches answer an important conditional question: what does pass@k look like if the chosen structural model extrapolates correctly? Related scaling theories derive power-law curves from Beta or regularly varying latent difficulty models (Levi, 2025, 2026).

We study the complementary question: what is learned before that assumption? For a task with one-attempt success probability P, the target is

$$\theta_k(F) = \mathbb{E}_F[1 - (1-P)^k], \tag{1}$$

where F is the population distribution of P. We observe b independent attempts per task. This is a classical binomial-mixture model. It is also the incidence model underlying species-accumulation curves (Colwell et al., 2004). Our contribution is not a new mixture representation. It is a sharp account of the information boundary for the pass@k functional, together with finite-task inference and an audit of current forecasting practice.

The boundary is governed by polynomial approximation. The complete count histogram identifies all polynomials of degree at most b. Consequently, two task populations are observationally equivalent exactly when they agree on the first b moments. We show that the largest possible pass@k gap between such populations is

$$D_{b,k} = 2 \inf_{q \in \mathcal{P}_b} \|x^k - q(x)\|_{\infty,[0,1]}.$$

This equality turns a 1976 approximation theorem (Newman and Rivlin, 1976) into explicit limits for pass@k. If $H \sim \text{Binomial}(2k, 1/2)$, then

$$\frac{\Pr(|H-k| > b)}{2e} \le D_{b,k} \le 2\Pr(|H-k| > b).$$

Thus model-free extrapolation is accurate uniformly over task populations when $k \ll b^2$, and can be highly ambiguous when k reaches or exceeds that scale.

Our main contributions are:

- We characterize the sharp identified set for each observable count law by moment duality, and prove the exact worst-case diameter above.
- We derive confidence-set lower bounds that persist with infinitely many tasks, a finite-task Hellinger modulus, and an indistinguishability extension to adaptive sampling with a deterministic per-task cap.
- We construct global and data-adaptive finite-sample confidence intervals. The global interval is honest for heterogeneous fixed tasks and is asymptotically sharp in worst-case width.
- We test the limits against current methods. A numerically moment-matched pair has count probabilities equal within 1.7×10^{-12} at $b = 8$, but pass@1000 values 0.167 and 0.950. In 3,200 synthetic datasets, taskwise Beta intervals are well calibrated only under their model-correct case. A 59,400-forecast audit on 22 public response banks finds increasing point error and intervals whose narrowness does not reflect nonparametric ambiguity.

The result is not that all extrapolation should stop at b. Structure can be scientifically justified and empirically useful. Rather, forecasts beyond the identified regime should separate sampling uncertainty from assumptions about the unobserved lower tail of task success probabilities.

2 Model and Identification

Observation model. For tasks $i = 1, \dots, M$, let P_i be iid from an unknown probability law F on $[0, 1]$. Conditional on P_i,

$$S_i \mid P_i \sim \text{Binomial}(b, P_i).$$

The population count probabilities are

$$\pi_s(F) = \binom{b}{s} \int p^s (1-p)^{b-s} \, dF(p), \qquad s = 0, \dots, b.$$

We first study population identification, which corresponds to knowing $\pi(F)$ exactly. Section 3 returns to finite M and also covers fixed heterogeneous task probabilities.

Proposition 1 (Observable information)**.** *The count law $\pi(F)$ identifies exactly $\int q(p)dF(p)$ for every $q \in \mathcal{P}_b$, equivalently the first b moments of F. Hence θ_k is identified for $k \le b$, with the usual combinatorial pass@k estimator, but not generally for $k > b$.*

This follows because the $b + 1$ binomial probabilities are expectations of the Bernstein basis, which spans $\mathcal{P}_b$. Finite-moment identification is well known in binomial mixtures. Basu et al. (2026) recently used the same boundary for exact inference on mixing-distribution CDFs and quantiles.

2.1 Exact identified sets

Let $\mu_j = \int p^j dF(p)$, including $\mu_0 = 1$, be the moments recovered from $\pi(F)$. Generalized moment duality gives the exact endpoints

$$L_k(\mu) = \sup_{\substack{q \in \mathcal{P}_b \\ q \le h_k}} \sum_{j=0}^{b} q_j \mu_j, \quad U_k(\mu) = \inf_{\substack{q \in \mathcal{P}_b \\ q \ge h_k}} \sum_{j=0}^{b} q_j \mu_j, \quad (2)$$

where $h_k(p) = 1 - (1-p)^k$ and $q(p) = \sum_j q_j p^j$. The primal problems optimize $\int h_k dF$ over all laws with moments μ. Compact generalized moment problems have no duality gap (de Klerk and Laurent, 2019; Halaseh et al., 2026). A dual optimizer need not exist for boundary moment vectors, so (2) is stated using a supremum and an infimum. Convexity shows that the identified set is the full interval $[L_k(\mu), U_k(\mu)]$.

The width depends on the observed count law. The following theorem identifies its largest possible value.

Theorem 1 (Sharp worst-case diameter)**.** *Let*

$$E_{b,k} = \inf_{q \in \mathcal{P}_b} \|x^k - q(x)\|_{\infty,[0,1]}.$$

Then

$$\sup_{\mu} \{U_k(\mu) - L_k(\mu)\} = D_{b,k} = 2E_{b,k}.$$

Moreover, there are two probability laws attaining this gap while inducing the same complete count distribution.

Proof sketch. If F and G match through degree b, every $q \in \mathcal{P}_b$ cancels. The target difference is therefore at most $2\|x^k - q\|_\infty$. For the reverse inequality, quotient-space duality for $C[0, 1]/\mathcal{P}_b$ and the Riesz representation theorem give a signed measure ν of total variation one that annihilates $\mathcal{P}_b$ and integrates x^k to $E_{b,k}$. Since ν annihilates constants, its positive and negative parts each have mass $1/2$. Normalizing them yields the required probability laws and gap $2E_{b,k}$. Full proofs are in the supplement.

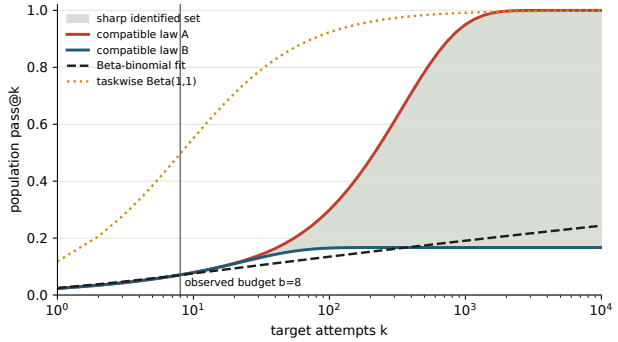

Figure 1: Two task populations induce the same count law for $b = 8$ but diverge after the observed budget. The shaded region is the sharp identified set for their shared moments. Parametric fits select narrow trajectories inside it.

Corollary 1 (Extrapolation horizon)**.** *For* $H \sim \mathrm{Binomial}(2k, 1/2)$*, let* $T_{b,k} = \Pr(|H - k| > b)$*. Then*

$$\frac{T_{b,k}}{2e} \le D_{b,k} \le 2T_{b,k}.$$

In particular, the worst-case identification transition is $b \asymp \sqrt{k}$*, or* $k \asymp b^2$*.*

To obtain the corollary, map approximation of x^k on $[0, 1]$ by degree b polynomials to approximation of x^{2k} on $[-1, 1]$ by degree $2b$ polynomials. Newman and Rivlin (1976) bound that error between $T_{b,k}/(4e)$ and $T_{b,k}$. The b^2 approximation transition is theirs. Its interpretation as the exact pass@k ambiguity is new here.

Corollary 2 (What a probability floor buys)**.** *Restrict both task populations to* $P \in [p_0, 1]$*, for a declared* $p_0 \in [0, 1]$*. The sharp worst-case diameter becomes*

$$D_{b,k}(p_0) = (1 - p_0)^k D_{b,k}.$$

Set $x = 1 - p$ and rescale $x = (1 - p_0)y$. Degree-b moment matching and polynomial approximation are preserved, while the target monomial is multiplied by $(1 - p_0)^k$. This quantifies the value and the risk of a lower tail assumption. At $b = 78, k = 10000$, the unrestricted diameter is 0.285. A known floor $p_0 = 10^{-4}$ reduces it to 0.105, while $p_0 = 10^{-3}$ reduces it below 1.3×10^{-5}. Such a floor can make distant forecasts precise, but it rules out impossible and extremely hard tasks. With finitely many attempts, its validity comes from domain knowledge, not from seeing no success on a task.

A sharp ambiguity example. Figure 1 numerically constructs the dual extremizers at $b = 8$. Each is supported on five probabilities. Their induced count probabilities differ by at most 1.7×10^{-12}, while pass@1000 is 0.16689 under one and 0.95019 under the other. The resulting diameter 0.78330 agrees with $2E_{8,1000}$ to 10^{-11}. The same calculation gives approximate diameters 0.509 at $b = 16, k = 1000$, 0.707 at $b = 32, k = 10000$, and 0.389 at $b = 64, k = 10000$.

3 Honest Inference

Identification ambiguity does not vanish as the number of tasks grows. It is also not the only obstacle at finite M.

Theorem 2 (Confidence-set limits)**.** *Let* C *have coverage at least* $1-\alpha$ *under two task-population laws* F, G*. If they induce identical count laws and have target gap* Δ*, then*

$$\Pr\{\mathrm{diam}(C) \ge \Delta\} \ge 1 - 2\alpha,$$
$$\mathbb{E}[\mathrm{diam}(C)] \ge (1 - 2\alpha)\Delta.$$

More generally, write q_F, q_G *for their one-task count laws and* $\rho(q_F, q_G) = \sum_s \sqrt{q_F(s)q_G(s)}$*. Then under* F*,*

$$\Pr_F\{\mathrm{diam}(C) \ge \Delta\} \ge 1 - 2\alpha - \sqrt{1 - \rho(q_F, q_G)^{2M}}. \quad (3)$$

The first result is a union bound under the shared data law. For (3), change measure for one coverage event and use $\mathrm{TV}(q_F^M, q_G^M) \le \sqrt{1 - \rho(q_F, q_G)^{2M}}$. This motivates the finite-task modulus

$$\omega_{M,b,k}(v) = \sup\{|\theta_k(F) - \theta_k(G)| : \rho(q_F, q_G) \ge (1 - v^2)^{1/(2M)}\}.$$

Every honest interval spans $\omega_{M,b,k}(v)$ with probability at least $1 - 2\alpha - v$. A grid-restricted convex program gives valid witnesses, not upper bounds, for this modulus.

Table 1 reports repaired feasible witnesses for the common public-data scale $M = 128, b = 78$. At pass@1000, the infinite-task identification diameter is only 0.00049, yet finite task sampling forces a much larger expected-length bound 0.276. At pass@10000, both sources of ambiguity are substantial.

Proposition 2 (Adaptive policies)**.** *Suppose a possibly randomized finite-budget adaptive policy samples every task at most* b *times. If two mixing laws agree through moment* b*, they induce the same distribution over the policy's full transcript. Theorem 2 therefore applies unchanged.*

Every binary history of length at most b has probability equal to an integral of a degree-b polynomial in p. Induction over the adaptive transcript proves

Table 1: Model-free 95% interval limits at $M = 128, b = 78$. The lower bound is on expected length for population $\theta_k(F)$. "Global" is the worst-case width of the construction in (4).

target k	$D_{b,k}$	length lower	global width
100	$< 3.9\times10^{-32}$	.022	.335
1,000	.00049	.276	.982
10,000	.28495	.686	1.000

the claim. The deterministic cap matters. This result does not directly cover policies that can sample one task without limit.

3.1 Two honest constructions

Global polynomial interval. Write a degree-b polynomial in the Bernstein basis,

$$q(p) = \sum_{s=0}^{b} a_s \binom{b}{s} p^s (1-p)^{b-s} = \mathbb{E}_p[a_S].$$

If $\|q - h_k\|_\infty \le \epsilon$ and $R = \max_s a_s - \min_s a_s$, Hoeffding's inequality (Hoeffding, 1963) gives the interval

$$C_{\text{lin}} = [\bar{a} - \epsilon - Rc_M,\ \bar{a} + \epsilon + Rc_M] \cap [0,1],$$
$$c_M = \sqrt{\frac{\log(2/\alpha)}{2M}}, \tag{4}$$

where $\bar{a} = M^{-1}\sum_i a_{S_i}$. This is valid conditionally for any fixed heterogeneous probabilities $p_1, \dots, p_M$, with target $M^{-1}\sum_i h_k(p_i)$. For iid tasks, it also covers population $\theta_k(F)$ by applying the same argument to the mixture counts. We choose q by a linear program minimizing $\epsilon + Rc_M$. Continuous-domain bias is certified between grid points using a derivative bound. As $M \to \infty$, its optimized worst-case width tends to $2E_{b,k} = D_{b,k}$.

Data-adaptive projection. The global interval protects its width before observing the histogram. We also project a simultaneous confidence band for the count CDF. Let $\widehat{G}_s = M^{-1}\sum_i \mathbf{1}\{S_i \le s\}$. For iid tasks, the DKW radius is $\sqrt{\log(2/\alpha)/(2M)}$ (Massart, 1990). For fixed, independent heterogeneous tasks, a union bound over $s = 0, \dots, b-1$ gives

$$\delta_{\text{fix}} = \sqrt{\frac{\log(2b/\alpha)}{2M}}.$$

Partition $[0,1]$ at $e_0 < \cdots < e_J$, and let w_j be the mixing mass in bin j. The binomial CDF $g_s(p)$ decreases in p, while $h_k(p)$ increases. Every compatible law therefore satisfies

$$\sum_j w_j g_s(e_{j+1}) \le \widehat{G}_s + \delta, \qquad \sum_j w_j g_s(e_j) \ge \widehat{G}_s - \delta.$$

Minimizing $\sum_j w_j h_k(e_j)$ and maximizing $\sum_j w_j h_k(e_{j+1})$ over these constraints gives an outer confidence interval. Endpoint monotonicity makes it rigorous at every finite partition size. This is a specialized finite-sample projection method. Generic projection inference for partially identified moment models is developed by Kaido et al. (2019).

4 Experiments

We ask three questions. Do the lower bounds become material at current response-bank sizes? Do model-based intervals retain coverage off model? Do public forecasts expose the same horizon effect? Code, fixed seeds, all atomic witness laws, and row-level results accompany the supplement.

Numerical programs. Best approximation and moment-pair linear programs use shifted Chebyshev bases. The finite-task modulus is solved as a convex Hellinger-affinity program on 1,001 support points. We repair each numerical solution into an explicit feasible pair before reporting it. The global interval uses 8,001 constraint points plus a continuous derivative certificate. The projection interval uses 1,001 probability bins. Independent tests check primal-dual agreement, count-law matching, modulus feasibility, and interval containment.

Synthetic coverage. We simulate 400 datasets in each of eight settings. These include a model-correct $F = \text{Beta}(1,1)$, a rare-success $\text{Beta}(0.35, 5)$, both sides of the exact $b = 8, k = 1000$ pair, and both sides of finite-task modulus pairs at $b = 78$ for $k = 1000, 10000$. Each dataset has 128 tasks. The target is the finite-task mean of $h_k(P_i)$. We compare the two honest intervals with independent taskwise $\text{Beta}(1,1)$ posterior intervals, computed from 2,000 shared posterior draws.

Figure 2(a,b) shows a sharp split. The taskwise Beta interval covers 96.25% under the exact Uniform model, with mean width 0.0059. It has zero coverage under every other law. Its mean widths remain between 0.0198 and 0.0550 on the adversarial pairs. Both honest procedures cover all 3,200 datasets. Projection materially adapts on benign histograms: its mean width is 0.190 rather than 0.752 under Uniform, and 0.547 rather than 0.806 under the rare Beta law. It correctly approaches width one on the indistinguishable pairs.

Public response-bank audit. We use all 22 panels released by Brown et al. (2024), covering four benchmarks and model sizes from 70M to 70B. Each panel contains 127–140 tasks and 10,000 responses per task. For each of 50 shuffles, a total budget of

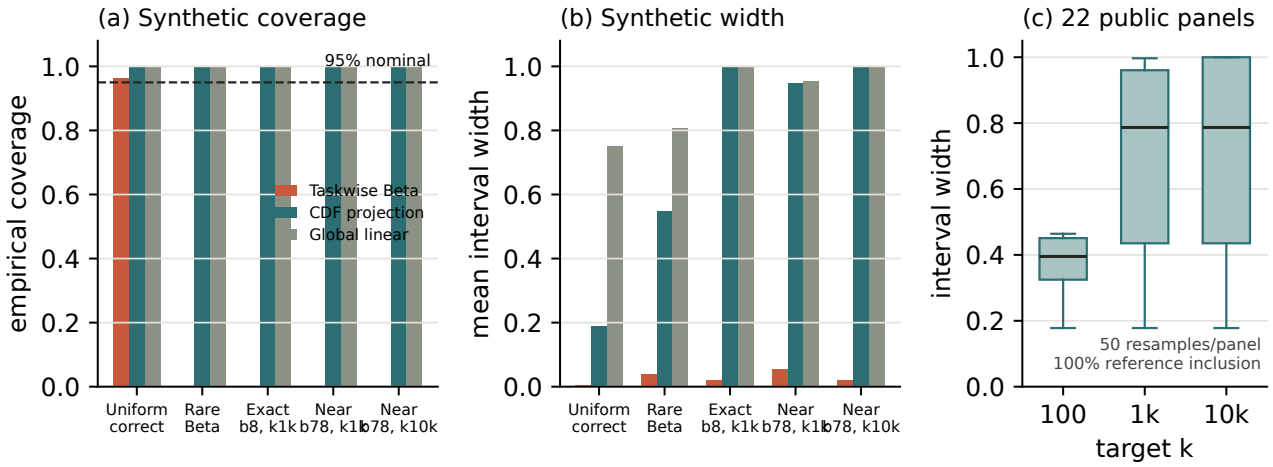

Figure 2: Honest uncertainty adapts when the histogram is favorable and stays wide when the data cannot distinguish the target. Panels (a) and (b) average the two sides of each exact or finite-task pair. Panel (c) shows fixed-task projection widths over 22 public panels and 50 resamples per panel. Reference inclusion is descriptive because the 10,000-response bank is finite.

Table 2: Public audit at total budget 10,000. Point columns report MAE over 1,100 panel-resamples. The last two columns report the taskwise Beta interval coverage and projection mean width.

k	Pop. B.	SGT	Task B.	TB cov.	Proj. width
100	.018	.016	.290	.090	.371
1,000	.041	.200	.384	.089	.714
10,000	.070	.353	.333	.091	.721

10,000 is allocated equally, giving 71–78 observed attempts per task. The remaining full bank defines empirical pass@k references. We evaluate population Beta-binomial MLE (Kazdan et al., 2025), direct plug-in, binomial smoothed Good–Toulmin (Orlitsky et al., 2016), taskwise Beta, and our fixed-task projection interval.

Table 2 shows that the population Beta model is a useful point forecast on average. Its MAE nevertheless grows fourfold from pass@100 to pass@10000, and 17% of its far-horizon forecasts err by more than 0.1. Smoothed Good–Toulmin is competitive at pass@100 but deteriorates beyond its logarithmic extrapolation regime. Taskwise Beta intervals are extremely narrow at the far target, with mean width 0.0118, and include the finite-bank reference only 9.1% of the time. This is not a formal coverage evaluation because the reference is noisy and reuses the bank. It does show that posterior concentration within a taskwise model is not a substitute for extrapolation robustness.

The projection interval includes every reference. Its mean width rises from 0.371 at pass@100 to 0.714 at pass@1000 and 0.721 at pass@10000. At the two farther targets, 40.5% of intervals are wider than 0.9. The method is informative for favorable panels and candidly uninformative where the available attempts do not resolve rare successes.

Figure 3 separates allocation from extrapolation. Dynamic allocation reduces mean population-Beta MAE from 0.0697 to 0.0667, only 4.3%, and does not uniformly dominate equal allocation across panels. Better allocation can reduce sampling error, but Proposition 2 shows that a per-task cap cannot resolve the underlying moment ambiguity.

5 Related Work and Limitations

Species extrapolation. The functional in (1) is the normalized incidence-based species accumulation curve of Colwell et al. (2004). That work gives unbiased interpolation and parametric finite-mixture extrapolation. Smoothed Good–Toulmin estimators attain logarithmic extrapolation factors in abundance and Bernoulli-product incidence models (Orlitsky et al., 2016). Rational continuations and parametric accumulation-rate models can work much farther (Daley and Smith, 2013; Shestopaloff, 2025), but their bootstrap intervals inherit structural bias. For unbounded total richness, Mao and Lindsay (2007) prove a more severe discontinuity that permits only one-sided inference. Our bounded finite-horizon target instead has a nontrivial, exactly quantified diameter.

Functional estimation and partial identification. Polynomial approximation is a standard tool for mini-

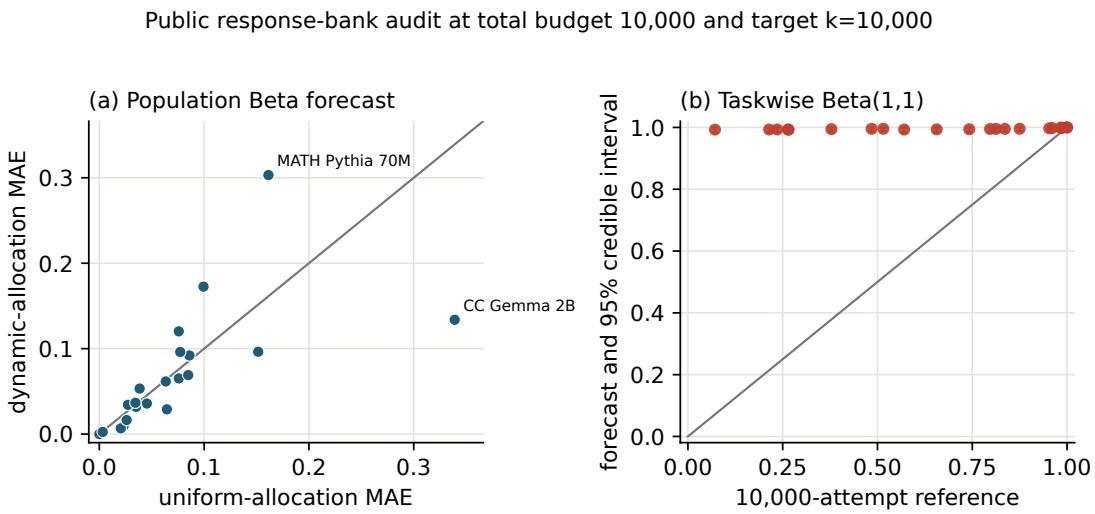

Figure 3: Public response-bank forecasts at total budget 10,000 and target $k = 10{,}000$. Left: dynamic allocation usually tracks uniform allocation rather than removing model error. Right: taskwise Beta credible intervals are much narrower than their discrepancies from the finite-bank reference.

max functional estimation (Jiao et al., 2015). Moment projection is also classical, and our confidence-band projection is related to generic partially identified inference (Kaido et al., 2019). A concurrent paper studies partial identification of latent truth from dependent LLM reports (Chen et al., 2026). Its target and observation model differ from repeated-attempt success. Our contribution is the pass@k-specific diameter and horizon, the finite-task modulus, and honest constructions tied to current evaluation practice.

Limitations. The model assumes conditionally independent attempts with a stable task-specific probability. Prompt variation, correlated decoding, verifier errors, and nonstationary systems require a different observation model. Zhou et al. (2026) show that deterministic activation interventions can reach some tasks missed by ordinary sampling, reinforcing that P is specific to a declared decoding regime rather than an intrinsic measure of task hardness. The Hellinger modulus values are feasible lower-bound witnesses, not certified global optima. Our public references are finite response banks, not latent truth, and the projection intervals intentionally trade width for assumption-robust coverage. Finally, favorable scientific restrictions on the lower tail of F can yield much tighter valid inference. Developing sensitivity analyses indexed by such restrictions is an important next step.

6 Conclusion

With b attempts per task, pass@k is a finite-moment extrapolation problem. Its sharp nonparametric ambiguity is $2E_{b,k}$, with a transition near $k = b^2$. Finite task sampling can make honest intervals wide even before that population boundary is reached. Parametric forecasts remain valuable when their assumptions are defensible, but narrow uncertainty beyond the observed budget is evidence about a model as well as evidence from data. Reporting that distinction makes repeated-inference scaling claims more auditable.

Reproducibility Statement

The supplement provides complete proofs, data provenance, optimization details, and commands. The artifact contains deterministic tests, all witness laws, raw and summarized rows, and scripts that regenerate every table and figure. Public data are retrieved from the authors' released repository.

A Identification Proofs

A.1 Observable information

Let

$$B_{s,b}(p) = \binom{b}{s} p^s (1-p)^{b-s}, \qquad s = 0, \dots, b,$$

be the degree-b Bernstein basis. If $S \mid P = p$ is binomial, then $\Pr(S = s) = \int B_{s,b}(p) dF(p)$.

Proof of Proposition 1. The $b+1$ Bernstein polynomials form a basis for $\mathcal{P}_b$. Hence the count probabilities determine the integral of every polynomial in $\mathcal{P}_b$. Conversely, every linear functional of the count probabilities is the expectation of a polynomial in their span, so no other linear functional of F is directly determined.

For $k \le b$, conditional on $S = s$, the probability that a uniformly chosen subset of k observed attempts contains at least one success is

$$\widehat{h}_k(s) = 1 - \frac{\binom{b-s}{k}}{\binom{b}{k}},$$

with the ratio defined as zero when $b - s < k$. Conditional exchangeability gives $\mathbb{E}_p[\widehat{h}_k(S)] = 1 - (1-p)^k$. Averaging over tasks and F proves unbiasedness.

For $k > b$, $h_k(p) = 1 - (1-p)^k$ has degree k and is not in $\mathcal{P}_b$. Theorem 1 constructs two laws that agree on every element of $\mathcal{P}_b$ but disagree on h_k, proving nonidentification in general. □

A.2 The identified interval for fixed moments

For a feasible moment vector $\mu = (\mu_0, \dots, \mu_b)$, define

$$\mathcal{F}(\mu) = \left\{ F : \int p^j \, dF(p) = \mu_j, \ j = 0, \dots, b \right\}.$$

The set is weakly compact because probability measures on $[0,1]$ are weakly compact and the moment constraints are continuous. The functional $F \mapsto \int h_k dF$ is continuous, so its minimum and maximum over $\mathcal{F}(\mu)$ exist. Convexity implies that every value between the two endpoints is attained.

The lower primal problem and its generalized moment dual are

$$\inf_{F \in \mathcal{F}(\mu)} \int h_k \, dF \quad \text{and} \quad \sup_{q \in \mathcal{P}_b} \left\{ \sum_{j=0}^{b} q_j \mu_j : q \le h_k \right\}.$$

The upper pair reverses the optimization and the envelope. Feasibility and compact support give equality of primal and dual values (de Klerk and Laurent, 2019; Halaseh et al., 2026). A dual envelope need not attain its optimum when μ lies on the boundary of the moment cone. This is why the main paper uses a supremum and an infimum.

Because $h_k - q$ and $q - h_k$ are univariate polynomials, their nonnegativity on $[0,1]$ can also be represented exactly by the Markov–Lukacs certificate. If a polynomial r has even degree $2d$,

$$r(p) = \sigma_0(p) + p(1-p)\sigma_1(p),$$

where σ_0, σ_1 are sums of squares of degrees at most $2d$ and $2d-2$. If its degree is odd $2d+1$,

$$r(p) = p\sigma_0(p) + (1-p)\sigma_1(p),$$

where both sums of squares have degree at most $2d$. Gram-matrix representations turn the endpoint problems into finite semidefinite programs. Their size grows with k, so our largest-horizon computations instead use linear programs on Chebyshev grids and monotone outer projection.

A.3 Sharp worst-case diameter

Proof of Theorem 1. Set $X = 1 - P$. Matching the first b moments of P is equivalent to matching those of X. Let F, G be any two such laws and let $q \in \mathcal{P}_b$. Since $\int q \, d(F - G) = 0$,

$$\begin{aligned} |\theta_k(F) - \theta_k(G)| &= \left| \int x^k \, d(F-G)(x) \right| \\ &= \left| \int (x^k - q(x)) \, d(F-G)(x) \right| \\ &\le 2 \|x^k - q\|_\infty. \end{aligned}$$

Taking the infimum over q proves $D_{b,k} \le 2E_{b,k}$.

For the reverse inequality, let $C = C[0,1]$ with the uniform norm, and consider the quotient Banach space $C/\mathcal{P}_b$. The norm of the coset $[x^k]$ is its distance to $\mathcal{P}_b$, namely $E_{b,k}$. By Hahn–Banach, there is a continuous linear functional Λ on C with operator norm one such that

$$\Lambda(q) = 0 \quad (q \in \mathcal{P}_b), \qquad \Lambda(x^k) = E_{b,k}.$$

The Riesz representation theorem gives a finite signed measure ν with total variation one and $\Lambda(f) = \int f \, d\nu$. Since the constant function lies in $\mathcal{P}_b$, $\nu([0,1]) = 0$. In the Jordan decomposition $\nu = \nu^+ - \nu^-$, both parts therefore have mass $1/2$.

Define probability measures $F^\star = 2\nu^+$ and $G^\star = 2\nu^-$. They match every polynomial of degree at most b, and

$$\int x^k \, d(F^\star - G^\star) = 2E_{b,k}.$$

Their count laws match by the Bernstein representation. Their pass@k values have the same absolute gap because the constant part of $1 - x^k$ cancels. Thus $D_{b,k} \ge 2E_{b,k}$. □

A.4 Newman–Rivlin mapping

Proof of Corollary 1. Let $E_n^{[-1,1]}(f)$ denote best degree-n uniform approximation on $[-1,1]$. If $q \in \mathcal{P}_b$, then $q(z^2)$ is an even degree-$2b$ polynomial and

$$\sup_{z\in[-1,1]} |z^{2k} - q(z^2)| = \sup_{x\in[0,1]} |x^k - q(x)|.$$

Conversely, symmetrizing any degree-$2b$ approximant to the even function z^{2k} cannot increase its error. The resulting even polynomial is $q(z^2)$ for a $q \in \mathcal{P}_b$. Hence

$$E_{b,k} = E_{2b}^{[-1,1]}(z^{2k}).$$

Newman and Rivlin (1976) prove that

$$\frac{1}{4e} P_{n,m} \le E_m^{[-1,1]}(z^n) \le P_{n,m},$$

where $P_{n,m}$ is the probability that the absolute difference between heads and tails in n fair tosses exceeds m. With $n = 2k, m = 2b$, that event is $|H - k| > b$ for $H \sim \text{Binomial}(2k, 1/2)$. Multiplication by two and Theorem 1 give the displayed bounds.

For large k, the standardized threshold is $b/\sqrt{k/2} = \sqrt{2}b/\sqrt{k}$. The central limit theorem therefore places the transition at $b/\sqrt{k}$ of constant order. If $b/\sqrt{k} \to \infty$, the tail and diameter vanish. If that ratio remains bounded, the lower bound stays away from zero along nondegenerate limits. □

A.5 Declared support restrictions

Proof of Corollary 2. Under $P \in [p_0, 1]$, set $X = 1 - P \in [0, a]$ with $a = 1 - p_0$. If $a = 0$, the target is constant and the claim is immediate. Otherwise write $X = aY$. The map from laws of X to laws of $Y \in [0,1]$ is bijective. Matching degree-b moments is preserved under this scaling, and

$$\left|\mathbb{E}_F X^k - \mathbb{E}_G X^k\right| = a^k \left|\mathbb{E}_F Y^k - \mathbb{E}_G Y^k\right|.$$

Taking the supremum over matching laws and applying Theorem 1 gives $D_{b,k}(p_0) = a^k D_{b,k}$. □

B Confidence-Set Lower Bounds

B.1 Exactly matching count laws

First part of Theorem 2. Let $A_F = \{\theta_k(F) \in C\}$ and $A_G = \{\theta_k(G) \in C\}$. Under a shared data law,

$$\Pr(A_F \cap A_G) \ge 1 - \Pr(A_F^c) - \Pr(A_G^c) \ge 1 - 2\alpha.$$

On the intersection, C contains two points separated by Δ, so its diameter is at least Δ. The expected-length claim follows from $\mathbb{E}[\text{diam}(C)] \ge \Delta \Pr\{\text{diam}(C) \ge \Delta\}$. □

B.2 Nearby finite-task experiments

Lemma 1 (Affinity bound)**.** *For probability mass functions q, r, define $\rho(q,r) = \sum_s \sqrt{q_s r_s}$. Then*

$$\text{TV}(q,r) \le \sqrt{1 - \rho(q,r)^2}.$$

For M iid observations, $\rho(q^M, r^M) = \rho(q,r)^M$.

Proof. Cauchy–Schwarz gives

$$\begin{aligned}\text{TV}(q,r) &= \frac{1}{2}\sum_s |\sqrt{q_s} - \sqrt{r_s}|(\sqrt{q_s} + \sqrt{r_s}) \\ &\le \frac{1}{2}\sqrt{\sum_s (\sqrt{q_s} - \sqrt{r_s})^2}\sqrt{\sum_s (\sqrt{q_s} + \sqrt{r_s})^2} \\ &= \sqrt{(1 - \rho(q,r))(1 + \rho(q,r))}.\end{aligned}$$

The product identity follows by factorizing the sum over product outcomes. □

Finite-task part of Theorem 2. Let P_F, P_G denote the two full data laws. Coverage and total variation give

$$\begin{aligned}P_F(A_G) &\ge P_G(A_G) - \text{TV}(P_F, P_G) \\ &\ge 1 - \alpha - \text{TV}(P_F, P_G).\end{aligned}$$

Combining this with $P_F(A_F) \ge 1 - \alpha$ yields

$$P_F(A_F \cap A_G) \ge 1 - 2\alpha - \text{TV}(P_F, P_G).$$

The preceding lemma bounds the final term by $\sqrt{1 - \rho(q_F, q_G)^{2M}}$. On the intersection the interval spans the target gap. □

B.3 Convex modulus program and repair

On support points $p_1, \ldots, p_J$, let w, z be two vectors in the probability simplex. Define the $(b+1) \times J$ matrix

$$A_{sj} = \binom{b}{s} p_j^s (1 - p_j)^{b-s}$$

and $t_j = h_k(p_j)$. For a desired product-TV budget v, the grid-restricted program is

$$\begin{aligned}\text{maximize} \quad & t^\top (w - z) \\ \text{subject to} \quad & w, z \ge 0, \quad \mathbf{1}^\top w = \mathbf{1}^\top z = 1, \\ & \sum_{s=0}^{b} \sqrt{(Aw)_s (Az)_s} \ge (1 - v^2)^{1/(2M)}.\end{aligned}$$

The affinity is concave in (w, z), so its superlevel set is convex. The objective is linear. Any feasible solution

is an actual pair of mixing laws, and therefore a valid lower-bound witness for the unrestricted modulus.

Numerical solvers can return weights with tiny negative entries or affinity slightly below the requested value. We first clip negative weights and renormalize. Let $m = (w+z)/2$ and $d = (w-z)/2$. We then replace the pair by

$$w_\lambda = m + \lambda d, \qquad z_\lambda = m - \lambda d, \qquad 0 \le \lambda \le 1.$$

At $\lambda = 0$, the two laws coincide and have affinity one. We use bisection to select the largest feasible λ, followed by a relative 10^{-10} shrinkage. Every reported target gap is recomputed from this repaired pair. Thus solver optimality labels do not affect validity of the lower bounds.

B.4 Adaptive policies

Proof of Proposition 2. Condition on the policy's external random seed, so its decisions are deterministic functions of the observed transcript. Consider any finite transcript compatible with the policy. For task i, suppose it contains $t_i \le b$ observations with s_i successes. Integrating its latent probability gives the history probability

$$\int p^{s_i}(1-p)^{t_i - s_i}\, dF(p).$$

The integrand has degree $t_i \le b$, so this value is unchanged when F is replaced by a law G with matching first b moments. Latent task probabilities are independent, hence the probability of the collection of task histories is the product of these integrals. Policy decisions add only indicators that the transcript follows the deterministic rule. Therefore every complete transcript has the same probability under F and G. Averaging over the external random seed proves the claim for a randomized policy. □

C Honest Interval Constructions

C.1 Global Bernstein interval

Theorem 3. *Fix arbitrary $p_1, \dots, p_M \in [0,1]$, and independently draw $S_i \sim \mathrm{Binomial}(b, p_i)$. If $\|q - h_k\|_\infty \le \epsilon$, where*

$$q(p) = \sum_{s=0}^{b} a_s B_{s,b}(p),$$

and $R = \max_s a_s - \min_s a_s$, then the global polynomial interval in the main paper covers

$$\theta_{M,k} = \frac{1}{M}\sum_{i=1}^{M} h_k(p_i)$$

with probability at least $1-\alpha$. If instead P_i are iid from F, the same interval also covers population $\theta_k(F)$ with that probability.

Proof. Let $\bar{a} = M^{-1}\sum_i a_{S_i}$. The Bernstein identity gives

$$\mathbb{E}[\bar{a}] = \frac{1}{M}\sum_i q(p_i).$$

The deterministic approximation error therefore satisfies

$$|\mathbb{E}[\bar{a}] - \theta_{M,k}| \le \epsilon.$$

Each a_{S_i} lies in an interval of length R. Hoeffding's inequality for independent, not necessarily identically distributed variables yields

$$\Pr\{|\bar{a} - \mathbb{E}\bar{a}| > Rc_M\} \le 2\exp(-2Mc_M^2) = \alpha.$$

The triangle inequality gives coverage before clipping, and intersecting with $[0,1]$ cannot remove the target.

For iid tasks, the mixture counts S_i are iid and $\mathbb{E}[a_{S_i}] = \int q(p)dF(p)$. Repeating the same approximation and concentration argument gives population coverage. □

Design program. On evaluation points u_ℓ, the implemented linear program minimizes

$$\epsilon + c_M(a_{\max} - a_{\min})$$

over $a_0, \dots, a_b, \epsilon, a_{\min}, a_{\max}$, subject to

$$|q(u_\ell) - h_k(u_\ell)| \le \epsilon, \qquad a_{\min} \le a_s \le a_{\max}.$$

After solving, we certify the continuous error. The Bernstein derivative identity gives

$$q'(p) = b\sum_{s=0}^{b-1}(a_{s+1} - a_s)B_{s,b-1}(p),$$

so $|q'(p)| \le b\Delta_a$, where $\Delta_a = \max_s |a_{s+1} - a_s|$. Also $h_k'(p) = k(1-p)^{k-1}$, which decreases in p. On adjacent grid points $u_\ell, u_{\ell+1}$, every p is within half the interval width of an endpoint. Hence the continuous error on that interval is at most

$$\max\{|q(u_\ell) - h_k(u_\ell)|,\, |q(u_{\ell+1}) - h_k(u_{\ell+1})|\} + \frac{u_{\ell+1} - u_\ell}{2}\left[k(1-u_\ell)^{k-1} + b\Delta_a\right].$$

We use the maximum of these bounds plus 10^{-10} as ϵ.

Asymptotic sharpness. In the ideal continuous design problem, every feasible polynomial has $\epsilon \ge E_{b,k}$. Conversely, for any $\eta > 0$, choose a polynomial with error at most $E_{b,k} + \eta$. Its finite Bernstein coefficient range is multiplied by $c_M \to 0$. Therefore the optimized half-width converges to $E_{b,k}$, and the full worst-case width converges to $2E_{b,k} = D_{b,k}$.

Table 3: Finite-task modulus witnesses for $M = 128, b = 78, \alpha = 0.05$. All displayed product-TV bounds are at most the requested v.

k	v	target gap	span probability	expected-length lower	solver label
100	.05	.00531	.850	.00451	optimal-inaccurate
100	.10	.01063	.800	.00850	optimal
100	.20	.02139	.700	.01497	optimal
100	.30	.03246	.600	.01948	optimal
100	.40	.04407	.500	.02203	optimal
1,000	.05	.28233	.850	.23998	optimal-inaccurate
1,000	.10	.33187	.800	.26549	optimal-inaccurate
1,000	.20	.39457	.700	.27620	optimal-inaccurate
1,000	.30	.44060	.600	.26436	optimal-inaccurate
1,000	.40	.47946	.500	.23973	optimal-inaccurate
10,000	.05	.80759	.850	.68645	optimal-inaccurate
10,000	.10	.83681	.800	.66945	optimal-inaccurate
10,000	.20	.86959	.700	.60871	optimal-inaccurate
10,000	.30	.89134	.600	.53480	optimal-inaccurate
10,000	.40	.90786	.500	.45393	optimal-inaccurate

C.2 Count-CDF projection interval

For thresholds $s = 0, \dots, b-1$, write

$$g_s(p) = \Pr\{\text{Binomial}(b, p) \le s\}.$$

This function is decreasing in p. The pass target $h_k(p)$ is increasing.

Theorem 4. *For iid tasks from F, use radius*

$$\delta_{\text{iid}} = \sqrt{\frac{\log(2/\alpha)}{2M}}.$$

For fixed $p_1, \dots, p_M$ with independent counts, use

$$\delta_{\text{fix}} = \sqrt{\frac{\log(2b/\alpha)}{2M}}.$$

The endpoint-envelope linear program in the main paper gives a $(1-\alpha)$ outer confidence interval for $\theta_k(F)$ in the iid case, and for $\theta_{M,k}$ in the fixed case.

Proof. In the iid case, the counts S_i are iid draws from the mixture count CDF $G_s = \int g_s(p) dF(p)$. The DKW inequality (Massart, 1990) gives

$$\Pr\left\{\max_s |\widehat{G}_s - G_s| > \delta_{\text{iid}}\right\} \le \alpha.$$

In the fixed case, $\mathbf{1}\{S_i \le s\}$ are independent variables in $[0, 1]$, with average expectation

$$G_{M,s} = \frac{1}{M} \sum_i g_s(p_i).$$

For each threshold, Hoeffding's inequality gives a two-sided error probability at most $2\exp(-2M\delta^2)$. A union bound over the b nontrivial thresholds gives the stated radius.

It remains to prove projection validity on either simultaneous-band event. Let $F_\star$ be F in the iid setting and the empirical law $M^{-1}\sum_i \delta_{p_i}$ in the fixed setting. On a partition $[e_j, e_{j+1}]$, set $w_j = F_\star([e_j, e_{j+1}])$, with a fixed convention at shared endpoints. Monotonicity gives

$$\sum_j w_j g_s(e_{j+1}) \le \int g_s \, dF_\star \le \sum_j w_j g_s(e_j).$$

Combining these inequalities with the confidence band shows that the true bin masses are feasible for the linear constraints. A second use of monotonicity gives

$$\sum_j w_j h_k(e_j) \le \int h_k \, dF_\star \le \sum_j w_j h_k(e_{j+1}).$$

The minimized lower envelope can only be smaller than the left side, and the maximized upper envelope can only be larger than the right side. The projected interval therefore contains the target on the simultaneous-band event. □

Numerical conditioning. To compute the upper endpoint, maximizing h_k is implemented as minimizing the failure probability $(1-p)^k$, then subtracting the optimum from one. This avoids an objective numerically indistinguishable from the constant one at large k. Each endpoint first uses HiGHS, then retries its dual simplex and interior-point methods at a relaxed 10^{-8} feasibility tolerance. Every reported run completed successfully after this reparameterization.

Table 4: Worst-case population identification. Grid-based approximation and moment-pair programs agree at the displayed precision.

b	k	approx. $D_{b,k}$	max moment resid.
4	128	.6677	2.1×10^{-15}
8	1,000	.7833	5.7×10^{-12}
16	1,000	.5086	5.1×10^{-15}
32	10,000	.7066	4.3×10^{-12}
64	10,000	.3886	1.6×10^{-14}
78	1,000	.000489	8.2×10^{-11}
78	10,000	.28495	7.3×10^{-13}

D Numerical Identification Results

The $b = 78, k = 100$ error is below double-precision resolution. We do not use the numerical zero as a theorem. Corollary 1 gives the rigorous bound

$$3.58 \times 10^{-33} \le D_{78,100} \le 3.89 \times 10^{-32}.$$

At full precision, the two induced count probability vectors differ by at most 1.7×10^{-12}. Their target values are 0.9501871171 and 0.1668880229. Their stored degree-b Chebyshev moment residual is 5.7×10^{-12}. For this ambiguity example, the numerical primal and dual diameters differ by less than 10^{-11}.

E Experiment Protocols and Additional Results

E.1 Synthetic interval audit

Each repetition first samples 128 latent task probabilities from the specified discrete or Beta law. It then samples one binomial count per task. Coverage is evaluated against the realized finite-task target

$$\theta_{M,k} = M^{-1} \sum_{i=1}^{M} [1 - (1 - p_i)^k].$$

This makes the comparison conditional on the sampled benchmark and matches the fixed-task guarantee of both honest intervals.

The exact pair is the law in Table 5. Finite-pair laws are the repaired $v = 0.2$ modulus witnesses for $b = 78, k \in \{1000, 10000\}$. The benign laws are $\mathrm{Beta}(1, 1)$ and $\mathrm{Beta}(0.35, 5)$, discretized by 20,001 equal-mass midpoint quantiles for simulation.

For each taskwise Bayesian interval, the posterior is

$$p_i \mid S_i \sim \mathrm{Beta}(S_i + 1, b - S_i + 1).$$

We use 2,000 independent joint posterior draws, compute the task-average pass@k for each draw, and take its 2.5% and 97.5% quantiles. The posterior mean is computed analytically. Population Beta point estimates use a ten-start Beta-binomial maximum-likelihood fit in log parameters. We also report the direct plug-in and the clipped binomial smoothed Good–Toulmin point estimates.

The honest intervals' empirical coverage of one is consistent with their guarantees but should not be interpreted as exact calibration. Their concentration radii are conservative at $M = 128$. The comparison of widths is the relevant positive result: the projection interval adapts substantially under Uniform and rare Beta laws, while retaining the required width on hard pairs.

E.2 Public response-bank audit

Data provenance. We use the correctness indicators from all 22 configurations of `ScalingIntelligence/monkey_business`, the response-bank release of Brown et al. (2024). The Hugging Face dataset revision observed during the audit was `a9f8f73bcd6948a57ed922cba4e48062ef95f553`. Its dataset card declares the MIT license. We fetch only the `is_corrects` and original task index columns from the hosted Parquet shards, sort by original task index, and store Boolean arrays plus source URLs. The 22 panels contain 127–140 tasks and 10,000 attempts per task.

References and resampling. For a response bank with $N = 10000$ attempts and c_i successful attempts on task i, the finite-bank pass reference is

$$\widehat{\theta}_k = \frac{1}{M} \sum_{i=1}^{M} \left[1 - \frac{\binom{N-c_i}{k}}{\binom{N}{k}} \right].$$

This is the expected without-replacement pass@k of a uniformly sampled subset from the bank. For each of 50 seeds, we independently shuffle each task's responses. Equal allocation uses the first $\lfloor B/M \rfloor$ responses per task. Dynamic allocation implements Algorithm 1 of Kazdan et al. (2025): it initializes each task, then allocates attempts according to the prescribed function of failures and successes. The nominal budgets are 1,000, 3,000, and 10,000. Targets are 100, 1,000, and 10,000.

Population Beta fit. Given possibly unequal b_i and counts s_i, the likelihood is

$$\prod_i \binom{b_i}{s_i} \frac{\mathrm{B}(s_i + \alpha, b_i - s_i + \beta)}{\mathrm{B}(\alpha, \beta)}.$$

The combinatorial factors do not affect optimization. We optimize $\log \alpha, \log \beta$ with analytic gradients from

Table 5: Atomic laws for the $b = 8, k = 1000$ ambiguity example. Weights shown here are rounded. The artifact stores full precision.

High pass@1000 law		Low pass@1000 law	
p	weight	p	weight
.002934	.940617	.000000	.833112
.149794	.038257	.041906	.134095
.501963	.011496	.311564	.018474
.854108	.006747	.692430	.008329
1.000000	.002882	.962090	.005991

Table 6: Synthetic 95% interval audit over 400 repetitions per scenario. "Task B." is taskwise Beta, "Projection" is fixed-task count-CDF projection, and "Global" is the Bernstein interval.

Scenario	Method	coverage	mean width	MAE	repetitions
Uniform, $b = 78, k = 1000$	Task B.	.9625	.0059	.0011	400
	Projection	1.0000	.1901	.0941	400
	Global	1.0000	.7523	.2606	400
Rare Beta, $b = 78, k = 1000$	Task B.	.0000	.0379	.1233	400
	Projection	1.0000	.5470	.1224	400
	Global	1.0000	.8056	.1637	400
Exact low, $b = 8, k = 1000$	Task B.	.0000	.0210	.8235	400
	Projection	1.0000	1.0000	.3318	400
	Global	1.0000	1.0000	.3318	400
Exact high, $b = 8, k = 1000$	Task B.	.0000	.0210	.0415	400
	Projection	1.0000	.9999	.4502	400
	Global	1.0000	1.0000	.4503	400
Finite first, $b = 78, k = 1000$	Task B.	.0000	.0549	.1384	400
	Projection	1.0000	.9472	.2792	400
	Global	1.0000	.9531	.2670	400
Finite second, $b = 78, k = 1000$	Task B.	.0000	.0551	.5350	400
	Projection	1.0000	.9490	.1165	400
	Global	1.0000	.9540	.1285	400
Finite first, $b = 78, k = 10000$	Task B.	.0000	.0198	.0204	400
	Projection	1.0000	1.0000	.4722	400
	Global	1.0000	1.0000	.4722	400
Finite second, $b = 78, k = 10000$	Task B.	.0000	.0198	.8900	400
	Projection	1.0000	1.0000	.3974	400
	Global	1.0000	1.0000	.3974	400

Table 7: Uniform-allocation public point forecasts at total budget 10,000. Each row contains 1,100 panel-resamples.

k	method	MAE	RMSE	fraction $> \|.1\|$
100	Population Beta	.0175	.0225	.000
100	SGT	.0161	.0223	.000
100	Plug-in	.0396	.0516	.055
100	Taskwise Beta	.2901	.3451	.729
1,000	Population Beta	.0414	.0558	.092
1,000	SGT	.2000	.2764	.605
1,000	Plug-in	.1481	.1845	.511
1,000	Taskwise Beta	.3843	.4846	.682
10,000	Population Beta	.0697	.1300	.170
10,000	SGT	.3529	.4688	.713
10,000	Plug-in	.2365	.2869	.737
10,000	Taskwise Beta	.3326	.4489	.682

Table 8: Fixed-task projection audit over 22 panels and 50 resamples.

k	ref. inclusion	mean width	median width	width $> .9$
100	1.000	.371	.395	.000
1,000	1.000	.714	.787	.405
10,000	1.000	.721	.787	.405

ten starting points, using bounds $[-14, 18]$ on both log parameters. Boundary fits are retained and flagged. Forecasts use

$$1 - \frac{\mathrm{B}(\alpha, \beta + k)}{\mathrm{B}(\alpha, \beta)}.$$

All 59,400 forecast rows used in the main point-estimate audit have finite objectives. The artifact preserves convergence, gradient, and boundary diagnostics.

Reference inclusion is not a frequentist coverage estimate. The full bank is finite, contains the sampled

prefix, and estimates rather than equals a latent target. We report it as a reproducible descriptive check. The synthetic study provides the controlled coverage experiment.

E.3 Compute and software

All optimization and data experiments ran in Python 3.10 on Argonne Swing. CPU computations were submitted to a GPU-only PBS queue but do not use the allocated accelerator. Core packages are NumPy, SciPy 1.15.3, pandas, Matplotlib, CVXPY, Clarabel, PyArrow, and fsspec. Four spawned workers run the coverage and projection audits with all BLAS thread counts set to one.

The full 400-repetition synthetic audit took 2 minutes 41 seconds on one node. The 3,300-row public projection audit took 2 minutes 16 seconds. The artifact uses fixed seeds, writes outputs only after all worker results return, and contains ten deterministic tests. Tests cover the combinatorial estimator, dynamic allocation, Beta-binomial recovery, taskwise posterior computation, smoothed Good–Toulmin interpolation, identification primal-dual agreement, continuous bias certification, finite-modulus feasibility, and projection containment.

F Additional Literature Boundary

The closest statistical model predates pass@k. Colwell et al. (2004) use the same incidence binomial mixture and accumulation functional. They recommend moment interpolation, fit finite mixtures for extrapolation, and caution that the asymptote is unreliable. Orlitsky et al. (2016) derive minimax smoothed Good–Toulmin estimators in both abundance and Bernoulli-product incidence models. Their estimable extrapolation factor grows logarithmically with sample size. Our work addresses a different quantity: the exact finite-b identified diameter, finite-task interval lower bounds, and honest pass@k projection.

Mao and Lindsay (2007) show that unseen-class odds in a zero-truncated Poisson mixture are discontinuous and that uniformly informative two-sided inference for total richness is impossible. That target is unbounded and asymptotic. Our finite-horizon target remains in $[0, 1]$, producing a bounded but nontrivial ambiguity that changes continuously with b and k. Daley and Smith (2013) use rational continuation for sequencing-library complexity, while Shestopaloff (2025) fits parametric accumulation-rate tails and uses a parametric bootstrap. Both demonstrate that structural continuation can be practically effective.

Basu et al. (2026) give exact pointwise CDF and quantile intervals for a binomial mixing distribution by test inversion. They explicitly discuss the finite-moment identification boundary. They do not study the pass@k functional, its sharp diameter, or the b^2 horizon. Kaido et al. (2019) provide generic asymptotically calibrated projection intervals for partially identified moment models. Our count-CDF interval is a specialized nonasymptotic outer projection whose validity follows from DKW or finite Hoeffding bands and monotone bin endpoints.

Recent pass@k papers impose useful structure rather than study nonparametric identification. Schaeffer et al. (2025) explain aggregate power laws through heavy lower tails. Levi (2025) derive the same curve under a Beta failure model, while Levi (2026) connect a regularly varying difficulty tail to training-dependent inference scaling. Kazdan et al. (2025) propose population Beta-binomial fitting and dynamic allocation. Anonymous (2026) propagate taskwise Beta or Dirichlet posteriors through richer repeated-sampling profiles. Hariri et al. (2026) use taskwise Bayesian inference to stabilize model evaluation. Our audits target the extrapolative assumptions in these population and taskwise models, not their within-model posterior algebra. A separate diagnostic by Zhou et al. (2026) shows that a deterministic intervention can reach some tasks missed by ordinary sampling. That result concerns the choice of decoding regime, not extrapolation from a fixed regime's finite count law.

References

Anonymous (2026). Success has a shape: TailPass@k for repeated-sampling agent evaluation. Project page. NeurIPS 2026 submission.

Basu, P., Brill, B., and Yekutieli, D. (2026). Exact confidence intervals for the mixing distribution from binomial mixture distribution samples. *Journal of Computational and Graphical Statistics*, 35(2):880–890.

Brown, B., Juravsky, J., Ehrlich, R., Clark, R., Le, Q. V., Ré, C., and Mirhoseini, A. (2024). Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*.

Chen, X., Rambachan, A., and Tamer, E. (2026). Partial identification from LLM prompts. *arXiv preprint arXiv:2606.15031*.

Colwell, R. K., Mao, C. X., and Chang, J. (2004). Interpolating, extrapolating, and comparing incidence-based species accumulation curves. *Ecology*, 85(10):2717–2727.

Daley, T. and Smith, A. D. (2013). Predicting the molecular complexity of sequencing libraries. *Nature Methods*, 10(4):325–327.

de Klerk, E. and Laurent, M. (2019). A survey of semidefinite programming approaches to the generalized problem of moments and their error analysis. In *World Women in Mathematics 2018*, pages 17–56. Springer.

Halaseh, S., Magron, V., and Skomra, M. (2026). Duality attainment and strict feasibility of the generalized moment problem and its relaxations. *arXiv preprint arXiv:2604.15072*.

Hariri, M., Samandar, A., Hinczewski, M., and Chaudhary, V. (2026). Don't pass@k: A Bayesian framework for large language model evaluation. In *International Conference on Learning Representations*.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30.

Jiao, J., Venkat, K., Han, Y., and Weissman, T. (2015). Minimax estimation of functionals of discrete distributions. *IEEE Transactions on Information Theory*, 61(5):2835–2885.

Kaido, H., Molinari, F., and Stoye, J. (2019). Confidence intervals for projections of partially identified parameters. *Econometrica*, 87(4):1397–1432.

Kazdan, J., Schaeffer, R., Allouah, Y., Sullivan, C., Yu, K., Levi, N., and Koyejo, S. (2025). Efficient prediction of pass@k scaling in large language models. *arXiv preprint arXiv:2510.05197*.

Levi, N. (2026). Learning shrinks the hard tail: Training-dependent inference scaling in a solvable linear model. *arXiv preprint arXiv:2601.03764*.

Levi, N. I. (2025). A simple model of inference scaling laws. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 33984–33998.

Mao, C. X. and Lindsay, B. G. (2007). Estimating the number of classes. *The Annals of Statistics*, 35(2):917–930.

Massart, P. (1990). The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283.

Newman, D. J. and Rivlin, T. J. (1976). Approximation of monomials by lower degree polynomials. *Aequationes Mathematicae*, 14(3):451–455.

Orlitsky, A., Suresh, A. T., and Wu, Y. (2016). Optimal prediction of the number of unseen species. *Proceedings of the National Academy of Sciences*, 113(47):13283–13288.

Schaeffer, R., Kazdan, J., Hughes, J., Juravsky, J., Price, S., Lynch, A., Jones, E., Kirk, R., Mirhoseini, A., and Koyejo, S. (2025). How do large language monkeys get their power (laws)? In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 53132–53176.

Shestopaloff, K. (2025). An accumulation rate curve estimator for total species. *Environmental and Ecological Statistics*, 32(1):311–345.

Zhou, L., Shah, S., Rodolà, E., and Dessì, R. (2026). Hard or just unreached? diagnosing the sampling blind spot in math-reasoning difficulty estimation. *arXiv preprint arXiv:2606.19636*.