A Posterior Predictive for Pass@k

Established Empirical Bayes, Measured on Released Rollouts

Peter Cotton

First version: September 2, 2026. This version: September 2, 2026.

Abstract

A prompt that fails a handful of scored attempts is rated impossible by the extrapolation $1-(1-\hat{p})^k$, and rated so at every k. The ingredients of the repair are established: the bias was flagged by Chen et al. (2021), the population-level fix by a fitted difficulty prior is published (Kazdan et al., 2025), and the per-prompt posterior predictive measured here is precisely the estimator of Vérine et al. (2026), who condition on each task's count and average the posterior predictions. This note is an independent evaluation of that estimator, on released rollout records with a genuinely extrapolative design: two training attempts per prompt, six held out, nine thousand prompts. The plug-in understates the aggregate pass@6 by seventeen points and, used as a reachability screen at the threshold a published diagnostic uses, declares 5,511 prompts unreachable of which a third succeed. The posterior predictive removes most of the damage, and the study separates how: shrinking the point estimate accounts for the bulk of the log-loss improvement, posterior integration for a further measurable step, and a fixed Jeffreys prior captures nearly all of the fitted families' gain. Limits are measured rather than assumed, including per-count calibration and the insensitivity of prompt rankings.

1 Introduction

Language models are improved after pretraining by having them attempt problems and scoring the attempts automatically: a unit test passes, a final answer matches, a verifier accepts. The evaluations that steer this process sample each prompt several times, because a model that solves a problem once in sixteen tries can often be trained into solving it reliably. Pass@k, the probability that at least one of k sampled attempts succeeds, averaged over prompts, is the working currency of post-training evaluation, and decisions of real consequence ride on it: which prompts to train on, whether a fine-tuning recipe helped, when to stop sampling.

When $n \geq k$ scored attempts per prompt are available, the combinatorial estimator of Chen et al. (2021), $1 - \binom{n-c}{k}/\binom{n}{k}$ with c the correct count, is unbiased for the per-prompt coverage, and unbiasedness for $k \leq n$ is as much as frequentist estimation can ask. It is not the same as optimal prediction: a decision scored by log loss or Brier score is a different objective, and the distinction matters below. The device this note measures arises when a coverage number is wanted at depth k from an estimate $\hat{p} = s/m$: evaluate the formula at the point estimate and report $1 - (1 - \hat{p})^k$: the plug-in, in the terminology of Vérine et al. (2026), who call it the naive plug-in estimator. Malladi et al. (2026) use it inside a coverage diagnostic, writing "we convert pass@1 to an estimated pass@16 value using $f_{16}(p) := 1 - (1 - p)^{16}$"; their pass@1 estimates rest on at least sixteen samples per problem per evaluation seed across three seeds (their Appendix D.3), so theirs is an instance of the plug-in transformation at substantial sampling depth rather

than extrapolation beyond it, a distinction this note keeps. The structural point transfers to their setting; the magnitudes measured here do not.

Chen et al. (2021) already cautioned that the plug-in is biased, and their combinatorial estimator addresses $k \leq n$. The population-level repair by a fitted difficulty prior has been published: Kazdan et al. (2025) fit a beta prior by marginal maximum likelihood and report the prior expectation of coverage, with zero-success problems handled by prior mass rather than declared impossible. The per-prompt posterior predictive is also published: Vérine et al. (2026) write "we therefore use the posterior predictive pass@k for each task and average across tasks," derive the beta-function-ratio form of the estimator, and additionally supply zero-one inflated priors and a theorem ordering population pass@k at fixed pass@1. Public evaluation code implements the beta-binomial version as well. This note claims no new estimator. What it adds is an independent evaluation on new data, under a design that actually extrapolates and scores per prompt: how large the plug-in's defects are against held-out batches, how much of the repair is plain shrinkage versus posterior integration, whether a fixed prior suffices, where the fitted models are and are not calibrated, and what changes in a threshold decision of the kind diagnostics actually take. A probit-normal prior is run alongside the beta and performs comparably; no advantage is claimed for it.

2 Related work

Brown et al. (2024) established repeated sampling as an inference-time scaling axis and model aggregate coverage curves with exponentiated power laws, computing per-problem pass rates by the combinatorial estimator. Schaeffer et al. (2025) explain those power laws through the heavy-tailed distribution of per-problem success probabilities, fitting a three-parameter Kumaraswamy law to per-problem point estimates and forecasting coverage through the integral that defines aggregate pass@k. Kazdan et al. (2025) fit a beta difficulty prior by marginal maximum likelihood and report aggregate coverage as the prior expectation of $1 - (1 - p)^k$, with a budget-allocation application. Vérine et al. (2026) supply the estimator this note evaluates: binomial counts under a shared beta difficulty prior fitted by marginal maximum likelihood, the posterior updated with each task's count, the posterior expectation of future pass@k taken per task and averaged. Their beta-function ratio is algebraically the rising-factorial ratio below, and they distinguish the prior expectation of coverage from the average of posterior predictions, the same distinction Proposition 2's cautions require. They go further than this note in two respects, zero-one inflated priors and an ordering theorem for population pass@k at fixed pass@1. Elsewhere in the evaluation literature, Hariri et al. (2025) replace pass@k reporting with Bayesian credible intervals at the model level, and Miller (2024) bring frequentist standard errors to eval reporting. Malladi et al. (2026) supply the motivating diagnostic; nothing here disputes their method, and the question is confined to the extrapolation step inside it.

3 The estimation problem

The practitioner's situation is this. There are N prompts and a compute budget that allows m scored attempts at each, so the data are success counts $s_1, \ldots, s_N$ with $s_i \in \{0, \ldots, m\}$. Decisions are then taken about a larger budget $k > m$ that will not be spent yet, or ever: whether a prompt is worth training on depends on its chance of being solved within k attempts, and whether a fine-tuning method helped depends on how many prompts cross that bar. The problem is to

choose the function of s_i that stands in for the unavailable deeper measurement, and the quality of a choice is measured the way predictions are: per prompt by log loss and Brier score against later attempts, in aggregate by the error of the reported curve.

The model under which the problem has an exact answer is minimal. Each prompt carries a latent per-attempt success probability p_i, drawn independently from a difficulty distribution G on $[0,1]$; conditional on p_i, attempts are independent Bernoulli(p_i) trials, so $s_i \sim \mathrm{Bin}(m, p_i)$. Two estimands follow. Per prompt, the probability that a fresh batch of k attempts contains a success,

$$q_k(p_i) \;=\; 1-(1-p_i)^k, \tag{1}$$

to be predicted from s_i alone; and in aggregate, the pass@k the whole population would exhibit,

$$\pi_k \;=\; \int_0^1 \big(1-(1-p)^k\big)\, dG(p). \tag{2}$$

The conditional independence of attempts given p_i is not the assumption under attack: attempts are drawn with independent seeds, and the model grants that. What varies across the field is p_i itself, and the whole question is how to respect that variation when s_i is a noisy glimpse of it.

4 The plug-in and its failure mode

At $s_i = 0$ the plug-in predicts zero at every k, however small m is, so each eventual success on such a prompt costs unbounded log loss; the mirror failure occurs at $s_i = m$ for prompts that then fail everything. The aggregate defect was flagged by Chen et al. (2021); the following restates it with strictness and the $k = 1$ boundary.

Proposition 1 (The plug-in is biased downward for every difficulty). Fix $m \geq 1$ and $k \geq 2$, and let $s \sim \mathrm{Bin}(m,p)$. Then

$$\mathbb{E}\Big[\,1-\big(1-\tfrac{s}{m}\big)^k\Big] \;\leq\; 1-(1-p)^k, \tag{3}$$

with equality only at $p \in \{0,1\}$; at $k = 1$ equality holds for all p. Consequently the expected aggregate plug-in curve lies below π_k for every G not concentrated on $\{0,1\}$.

Proof. Write $\varphi(x) = (1-x)^k$ on $[0,1]$. Its second derivative is $\varphi''(x) = k(k-1)(1-x)^{k-2} \geq 0$, strictly positive on $[0,1)$ when $k \geq 2$, so φ is convex, and strictly convex there. The estimate $\hat{p} = s/m$ satisfies $\mathbb{E}\,\hat{p} = p$. Jensen's inequality gives $\mathbb{E}\,\varphi(\hat{p}) \geq \varphi(\mathbb{E}\,\hat{p}) = (1-p)^k$, and the inequality is strict whenever $\hat{p}$ is nondegenerate, which is exactly $p \in (0,1)$. Negating and adding one reverses the inequality into the display. At $k = 1$, φ is affine and Jensen holds with equality. The aggregate statement follows by integrating the pointwise inequality against G and taking expectations over the counts. □

The sign is universal; the magnitude is not monotone in k. A direct calculation at $m = 2$, $p = 0.9$ gives bias 0.0315 at $k = 3$ and 0.0212 at $k = 4$: extrapolating further can cost less, because the bias also vanishes as $q_k(p)$ saturates toward one. The experiments below happen to show bias growing over the displayed range of k; that is a fact about this data's difficulty distribution, not a consequence of convexity.

5 The posterior predictive

Give p a prior with two parameters fitted by marginal maximum likelihood across prompts, and predict with the posterior expectation

$$\widehat{\text{pass}}_k^{\text{post}} \;=\; \mathbb{E}\big[\,1-(1-p)^k \;\big|\; s\,\big]. \tag{4}$$

Under a Beta(a,b) prior the posterior is Beta$(a+s,\; b+m-s)$ and

$$\mathbb{E}\big[(1-p)^k \mid s\big] = \prod_{j=0}^{k-1} \frac{b+m-s+j}{a+b+m+j}, \tag{5}$$

a ratio of rising factorials, so Equation (4) is closed-form. Under a probit-normal prior, $p=\Phi(\theta)$ with $\theta \sim N(\mu,\tau^2)$, the marginal likelihood and the posterior moment are one-dimensional Gauss–Hermite quadratures over θ; this is the family that composes with correlated extensions, since a factor-correlated portfolio of successes conditions to exactly this form (Cotton, 2026).

Because q_k is concave, the posterior predictive sits *below* the plug-in evaluated at the posterior mean: $\mathbb{E}[q_k(p) \mid s] \le q_k(\mathbb{E}[p \mid s])$. Comparing those two predictors therefore separates two effects that are easy to conflate: shrinking the noisy point estimate toward the population, and integrating the curvature of q_k over the posterior. The study below prices them separately.

Proposition 2 (What the posterior predictive is). Let Y indicate a success among k fresh attempts at the same prompt, drawn independently of the training attempts given p. Under the model, with prior equal to the true difficulty law G: (i) $\mathbb{E}[Y \mid s] = \mathbb{E}[\,1-(1-p)^k \mid s\,]$, the exact conditional success probability; (ii) it minimizes expected Brier score and expected log loss among all predictors that depend on the data only through s; (iii) its population expectation equals π_k: $\mathbb{E}_s\big[\mathbb{E}[q_k(p) \mid s]\big] = \pi_k$ when the outer expectation is taken under the model's marginal law of s.

Proof. Given p, Y is Bernoulli with parameter $q_k(p)$ and is independent of s. The tower property gives $\mathbb{E}[Y \mid s] = \mathbb{E}\big[\mathbb{E}[Y \mid p, s] \;\big|\; s\big] = \mathbb{E}[q_k(p) \mid s]$, which is (i). For (ii), the conditional expectation minimizes mean squared error among s-measurable predictors, which is the Brier statement, and the true conditional probability minimizes expected log loss because the excess log loss of any other predictor is the expectation of a Kullback–Leibler divergence, which is nonnegative. For (iii), $\mathbb{E}\big[\mathbb{E}[q_k(p) \mid s]\big] = \mathbb{E}\, q_k(p) = \pi_k$, the tower property again. □

These are standard facts of Bayesian prediction, recorded to fix what is and is not claimed: they explain the method, and they are not a contribution. Two cautions attach to (iii). It is a statement about population expectations under the true model; the average of posterior predictions over a particular fitted sample need not equal the prior expectation of coverage under the fitted prior, so the population estimator of Kazdan et al. (2025) is computed as its own quantity in the tables below. In the present study the two happen to nearly coincide, because a two-parameter family fitted to counts on $m=2$ attempts nearly interpolates the three-cell count histogram; that coincidence is particular, not general.

6 Numerical study on released rollouts

The data are the publicly released Pass8 rollout records of the RLVE training suite: procedurally generated reasoning tasks, each with an automatic verifier, of the kind used as reinforcement-learning environments for language models. They carry nine thousand prompts across eighteen

environment families and eight scored samples per prompt from the Qwen3-4B-Thinking model.[1] RLVE shapes rewards, so a positive reward can denote partial credit: of 23,152 positive rewards, 2,988 lie strictly between zero and one. Success here means reward at least one; the released rows carry no separate accuracy field, and rerunning every table under the loose positive-reward definition changes the numbers but no conclusion.

The design extrapolates. Samples 0–1 are the training half ($m = 2$); samples 2–7 are held out, and the per-prompt target is a success among those six ($k = 6 > m$), a single Bernoulli outcome scored by log loss and Brier score. The aggregate reference for $k \leq 6$ is the combinatorial estimator on the six held-out samples only, disjoint from training, and it is a reference estimate, not truth. The fitted priors are Beta$(0.31, 0.79)$ and probit-normal $\mu = -1.06$, $\tau = 1.53$.

predictor of held-out pass@6	log loss	Brier
plug-in $1 - (1 - s/2)^6$	6.065	0.248
plug-in at the fitted posterior mean	0.570	0.185
Jeffreys Beta$(\frac{1}{2}, \frac{1}{2})$ posterior predictive	0.536	0.183
fitted beta posterior predictive	0.510	0.172
fitted probit-normal posterior predictive	0.509	0.172

Table 1: Per-prompt scores over nine thousand held-out Bernoulli outcomes, $m = 2$ training attempts, $k = 6$. The plug-in's log loss is computed with predictions clipped to $[10^{-12}, 1 - 10^{-12}]$; unclipped it is infinite, and the value is a clipping convention, not a stable measure: clipping at 10^{-6}, 10^{-3}, 10^{-2} gives 3.11, 1.63, 1.14. The decomposition is the point of the table: shrinkage alone (row two) removes most of the damage, posterior integration improves on row two by a further eleven percent of log loss and seven percent of Brier, and the fixed Jeffreys prior, with no fitted hyperparameters, captures nearly all of the fitted families' gain.

aggregate	$k = 1$	$k = 2$	$k = 4$	$k = 6$
reference estimate (six held-out samples)	0.280	0.386	0.498	0.558
plug-in from the training half	0.282	0.335	0.374	0.384
fitted beta, average of posteriors	0.282	0.388	0.491	0.546
fitted beta, prior expectation	0.282	0.388	0.491	0.546

Table 2: Aggregate pass@k from two training samples per prompt. The plug-in is 17.4 points under the reference at $k = 6$; the posterior rows are within 1.3 points. The prior-expectation row is the population estimator of Kazdan et al. (2025); it nearly coincides with the average of posterior predictions here because the fitted two-parameter family nearly interpolates the three-cell count histogram at $m = 2$, a coincidence of this design rather than an identity.

Calibration by count locates what the fitted model does and does not get right. For the 5,511 prompts with no training success the fitted beta predicts a held-out success rate of 0.318 against an observed 0.343, a modest undershoot; at counts one and two the predictions 0.843 and 0.982 sit close to the observed 0.834 and 0.977, where the plug-in's 0.984 and 1.000 overshoot. One structural limit is worth stating plainly: with a common prior and common m, every prompt with

[1] Dataset: `CL-From-Nothing/RLVE-Qwen3-4B-Thinking-2507-Pass8-Rollouts` on Hugging Face. Extraction and experiment scripts are committed under `research/cavity_calculus/exp2_passk/` in the `winning` repository.

the same count receives the same prediction, so the ranking of prompts is exactly the ranking by count under either method. The posterior changes thresholds and magnitudes, not order.

That is where the decision demonstration lives. The coverage diagnostic of Malladi et al. (2026) works on a base-reachable set, prompts whose estimated pass@16 lies in $[0.05, 0.95]$. Applying the lower edge as a reachability screen on predicted pass@6: the plug-in declares the 5,511 zero-count prompts unreachable, and 1,890 of them, a third, succeed within the six held-out attempts; the posterior predictive prices those same prompts at 0.318 and declares none unreachable, in line with their observed 0.343 rate. A screen built on the plug-in wrongly discards a third of what it discards at this sampling depth; a screen built on the posterior predictive makes no such error here, at the cost of declaring nothing unreachable, which is what a calibrated probability near one third requires.

7 Scope

The bias mechanism and the repair are established prior work; the measurements are the contribution, and they are bounded by their setting. The magnitudes scale with the noise in $\hat{p}$: two training attempts is an extreme of the small-sample regime, chosen to make the extrapolation genuine, and Malladi et al. (2026) operate at sixteen or more samples per problem per seed, where the noise-driven defects measured here are far smaller. The data cover one model on one environment suite, with success defined by full reward in the absence of a separate accuracy field; the loose definition changes every number and no conclusion. The two fitted families agree closely here, and agreement at $m = 2$ establishes little about the difficulty distribution: both families place no atom at $p = 0$, both therefore predict eventual success with probability approaching one for every prompt, and two attempts cannot identify the tail behavior that arbitrary extrapolation would need; the zero-one inflated priors of Vérine et al. (2026) address exactly this. What the study supports is narrower and useful: at small sampling depth, shrinkage of any reasonable kind repairs most of the plug-in's per-prompt damage, posterior integration adds a measured further step, and threshold decisions of the kind diagnostics take change materially. Two extensions are named and left undone: measuring the substitution inside a diagnostic at its native sampling depth on math and code benchmarks, and developing correlated, heterogeneous attempt portfolios, for which the probit-normal family composes with factor structure (Cotton, 2026) but for which nothing is demonstrated here.

References

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

M. Chen et al. Evaluating large language models trained on code. arXiv:2107.03374, 2021.

P. Cotton. Scalable inversion of contests with correlated performances, including softmax and multinomial probit. arXiv:2609.01133, 2026.

A. Hariri et al. Don't pass@k: a Bayesian framework for large language model evaluation. arXiv:2510.04265, 2025.

J. Kazdan, R. Schaeffer, et al. Efficient prediction of pass@k scaling in large language models. arXiv:2510.05197, 2025.

S. Malladi, S. Jelassi, D. Foster, J. T. Ash, and A. Krishnamurthy. TailSFT: filtered fine-tuning improves post-training performance. arXiv:2608.25756, 2026.

E. Miller. Adding error bars to evals: a statistical approach to language model evaluations. arXiv:2411.00640, 2024.

R. Schaeffer et al. How do large language monkeys get their power (laws)? arXiv:2502.17578, 2025.

A. Vérine, F. Le Bronnec, A. Allauzen, and B. Negrevergne. Estimating pass@k from fewer samples with hierarchical Bayesian priors. In *1st Workshop on Combining Theory and Benchmarks (CTB@ICML 2026)*, 2026.