Bernoulli Transformer LPE Report

Automated run

February 18, 2026

Artifacts Used

- Checkpoint: `checkpoints/bernoulli_transformer_L1_D16_seq1024.pt`
- Training/diagnostics log: `logs/bernoulli_diagnostics_beta11_fixed_target.log`
- Diagnostics CSV: `plots/bernoulli_posterior_sampling_diagnostics.csv`

1. Model Architecture

Field	Value
Transformer type	Decoder-only, causal self-attention
Positional encoding	none
Layers (L)	1
Model width (d_{model})	16
Heads (H)	1
MLP width (d_{mlp})	16
Pre-norm	True
Trainable parameters	1,794

Table 1: Bernoulli-transformer architecture reconstructed from checkpoint and script defaults.

2. Training Details

Field	Value
Data generation	$p \sim \text{Beta}(1,1)$, then sequence $y_t \sim \text{Bernoulli}(p)$
Training objective	Autoregressive cross-entropy on next-token prediction
Sequence length	1024
Batch size	64
Optimizer	AdamW (weight decay 0.01)
Learning rate	0.000300
Warmup steps	400
Gradient clipping	1.00
Total steps	4000

Table 2: Training setup from the logged run.

3. Training Curve

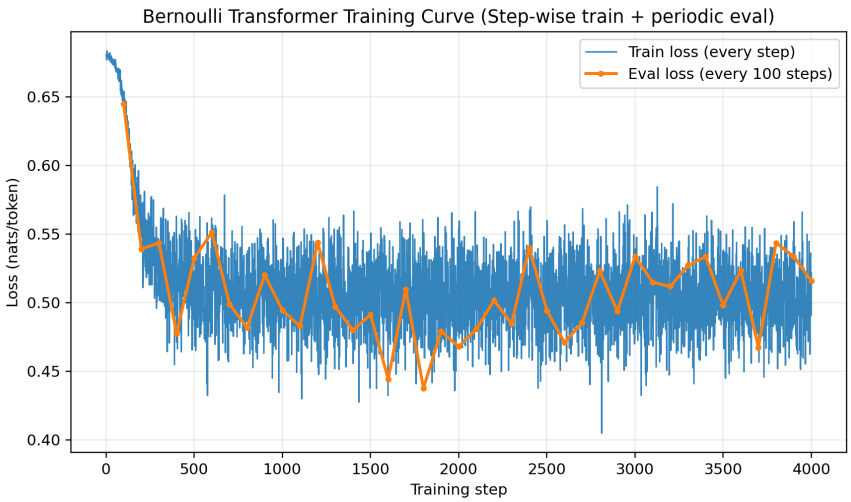

Figure 1: Training curve with training loss logged at every gradient step (1 to 4000) and evaluation loss logged every 100 steps.

4. Final Loss vs Bayes Predictive

Held-out evaluation used 130,560 prediction tokens drawn from the same Beta-Bernoulli process.

For the Bayes baseline, the prior is $p \sim \mathrm{Beta}(1,1)$. For a sequence $y_1, \ldots, y_T$, the Bayes predictive probability of the next bit at step $t \geq 2$ is

$$q_t \equiv P(y_t = 1 \mid y_{1:t-1}) = \frac{1 + \sum_{i=1}^{t-1} y_i}{2 + (t-1)}.$$

The per-token Bayes negative log-likelihood is

$$\ell_t^{\mathrm{Bayes}} = -\Big[y_t \log q_t + (1 - y_t) \log(1 - q_t) \Big].$$

So the reported Bayes NLL is the average over all predicted tokens in the held-out set:

$$\mathrm{NLL}_{\mathrm{Bayes}} = \frac{1}{N_{\mathrm{tok}}} \sum_{b=1}^{B} \sum_{t=2}^{T} \ell_{b,t}^{\mathrm{Bayes}},$$

with $B = 512$, $T = 256$, and $N_{\mathrm{tok}} = B(T-1) = 130{,}560$.

For a true expectation sanity check, let $n = t - 1$, and define $S_n = \sum_{i=1}^{n} y_i$. Condition on $S_n = s$. Then

$$p \mid y_{1:n} \sim \mathrm{Beta}(s+1, n-s+1), \quad q_s \equiv \mathrm{P}(y_{n+1} = 1 \mid y_{1:n}) = \frac{s+1}{n+2}.$$

Given this prefix, the Bayes one-step loss at $n+1$ is random because y_{n+1} is random:

$$\ell_{n+1}^{\mathrm{Bayes}} = \begin{cases} -\log q_s, & y_{n+1} = 1, \\ -\log(1 - q_s), & y_{n+1} = 0. \end{cases}$$

Also $y_{n+1} \mid y_{1:n} \sim \text{Bernoulli}(q_s)$, so

$$\mathrm{E}\Big[\ell_{n+1}^{\text{Bayes}} \mid S_n = s\Big] = q_s(-\log q_s) + (1-q_s)(-\log(1-q_s)) = h(q_s),$$

where $h(q) = -q\log q - (1-q)\log(1-q)$. Therefore

$$\mathrm{E}\Big[\ell_{n+1}^{\text{Bayes}} \mid S_n = s\Big] = h\left(\frac{s+1}{n+2}\right).$$

Now average over S_n:

$$\mathrm{E}[\ell_{n+1}^{\text{Bayes}}] = \sum_{s=0}^{n} \mathrm{P}(S_n = s)\, h\left(\frac{s+1}{n+2}\right).$$

Under the Beta(1,1) prior/data model, $S_n \sim \text{BetaBinomial}(n,1,1)$, which is uniform on $\{0,\dots,n\}$, so

$$\mathrm{E}[\ell_{n+1}^{\text{Bayes}}] = \frac{1}{n+1}\sum_{s=0}^{n} h\left(\frac{s+1}{n+2}\right).$$

So the exact expected per-token Bayes NLL for $T = 256$ is

$$\mathrm{E}[\text{NLL}_{\text{Bayes}}] = \frac{1}{255}\sum_{n=1}^{255} \mathrm{E}[\ell_{n+1}^{\text{Bayes}}] \approx 0.50850 \text{ nats/token}.$$

Back-of-the-envelope: for large n, the points $(s+1)/(n+2)$ form an almost-uniform grid on $[0,1]$, so

$$\mathrm{E}[\ell_{n+1}^{\text{Bayes}}] \approx \int_0^1 h(q)\,dq = \int_0^1 [-q\log q - (1-q)\log(1-q)]\,dq = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}.$$

So we should expect the Bayes NLL to be near 0.5 nats/token. The finite-T value 0.50850 is slightly above 0.5 because early prediction steps are less certain and contribute larger loss.

Why is the reported value 0.516166 still reasonable? The table reports one finite held-out draw ($B = 512$ sequences), not the exact expectation. Repeating this finite evaluation across random seeds gives noticeable spread (empirically, mean ≈ 0.5088, std ≈ 0.0085 nats/token for this setup), so a single run near 0.516 is within typical sampling fluctuation.

Metric	Value	Notes
Model NLL	0.520968	Per-token negative log-likelihood
Bayes NLL	0.516166	Exact Beta-Bernoulli predictive
Gap	0.93%	(model − Bayes)/Bayes

Table 3: Model predictive quality against the Bayes-optimal baseline.

To compare next-token probabilities directly, we generated an additional evaluation set of mixed-length contexts. Specifically, we sampled 3,000 contexts independently as follows:

- Sample context length $n \sim \text{Uniform}\{1,\dots,256\}$.
- Sample latent $p \sim \text{Beta}(1,1)$.
- Sample context bits $x_{1:n}$ i.i.d. from Bernoulli(p).

For each context, we computed:

- Bayes predictive $P(y_{n+1} = 1 \mid x_{1:n}) = (1 + \sum_{t=1}^{n} x_t)/(n + 2)$.
- Model predictive $P_\theta(y_{n+1} = 1 \mid x_{1:n})$ from the softmax of the model's next-token logits.

The Pearson correlation between these two quantities is $r = 0.9958$ (MAE $= 0.00954$).

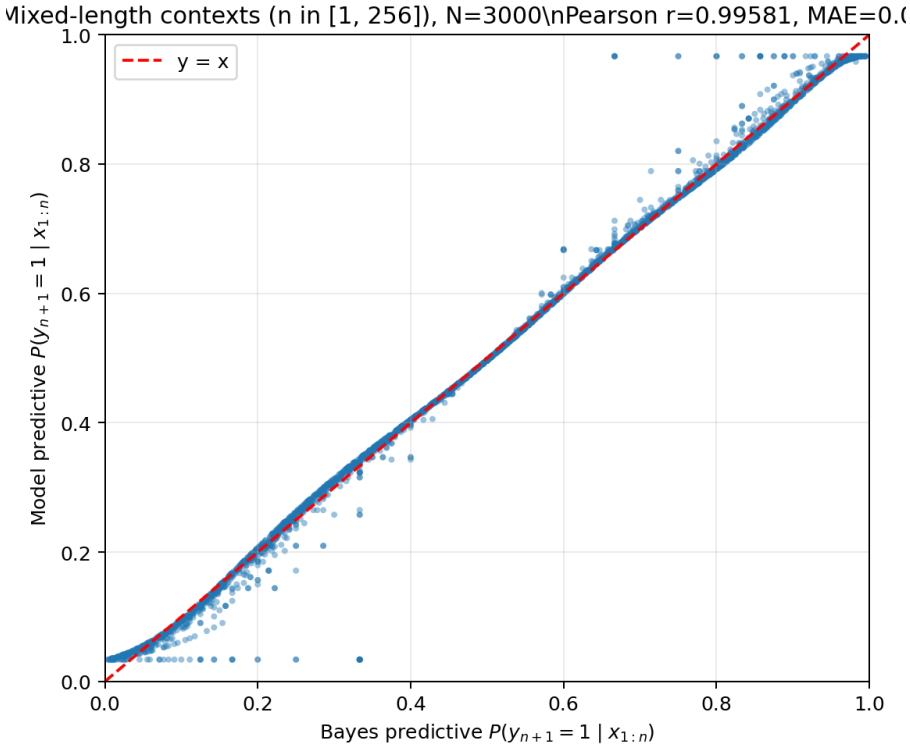

Figure 2: Model vs Bayes next-token $P(y_{n+1} = 1)$ on 3,000 mixed-length contexts ($n \in [1, 256]$).

The two horizontal bands are caused by *pure* contexts (all zeros or all ones). In this evaluation set, 113/3000 contexts (3.77%) are pure: 51 all-zero and 62 all-one contexts. On these, the model prediction is almost length-invariant, concentrating near 0.0339 (all-zero) or 0.9667 (all-one), which creates the bottom/top lines. By contrast, the Bayes predictor for pure contexts still depends on length:

$$P_{\text{Bayes}}(y_{n+1} = 1 \mid x_{1:n} \equiv 0) = \frac{1}{n+2}, \qquad P_{\text{Bayes}}(y_{n+1} = 1 \mid x_{1:n} \equiv 1) = \frac{n+1}{n+2}.$$

So short pure contexts can have non-extreme Bayes values (e.g., 1/3 or 2/3 at $n = 1$) while the model remains near its saturated values. Quantitatively, 45 of the 113 pure contexts have $n \leq 8$, and 42 of the 113 pure contexts have Bayes predictive in $(0.1, 0.9)$.

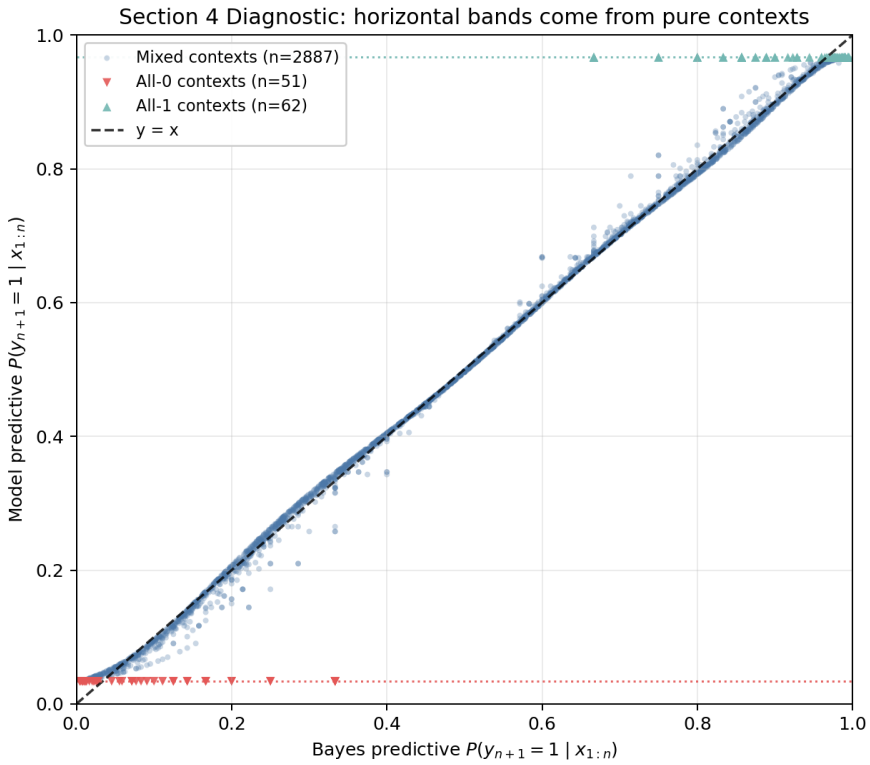

Figure 3: Section 4 diagnostic scatter with contexts colored by purity (mixed, all-zero, all-one). The horizontal bands correspond to pure contexts.

5. Posterior-Sample Quality

Diagnostics used 30 trials, posterior samples per trial=200, rollout length=1000. For this section, each trial corresponds to one generated input string (context). The 30 contexts were generated as follows:

- For each trial, sample a latent Bernoulli parameter $p_{\text{true}} \sim \text{Beta}(1, 1)$.
- Generate one input context string of length 50 i.i.d. from Bernoulli(p_{true}).
- From that context, compute the exact Bayes posterior (Beta(1,1) prior) and its posterior mean.
- Estimate the model-implied posterior from rollouts (200 rollouts of length 1000). For each context, we plot a histogram of these 200 rollout-derived p-samples and overlay the true Bayes posterior density.

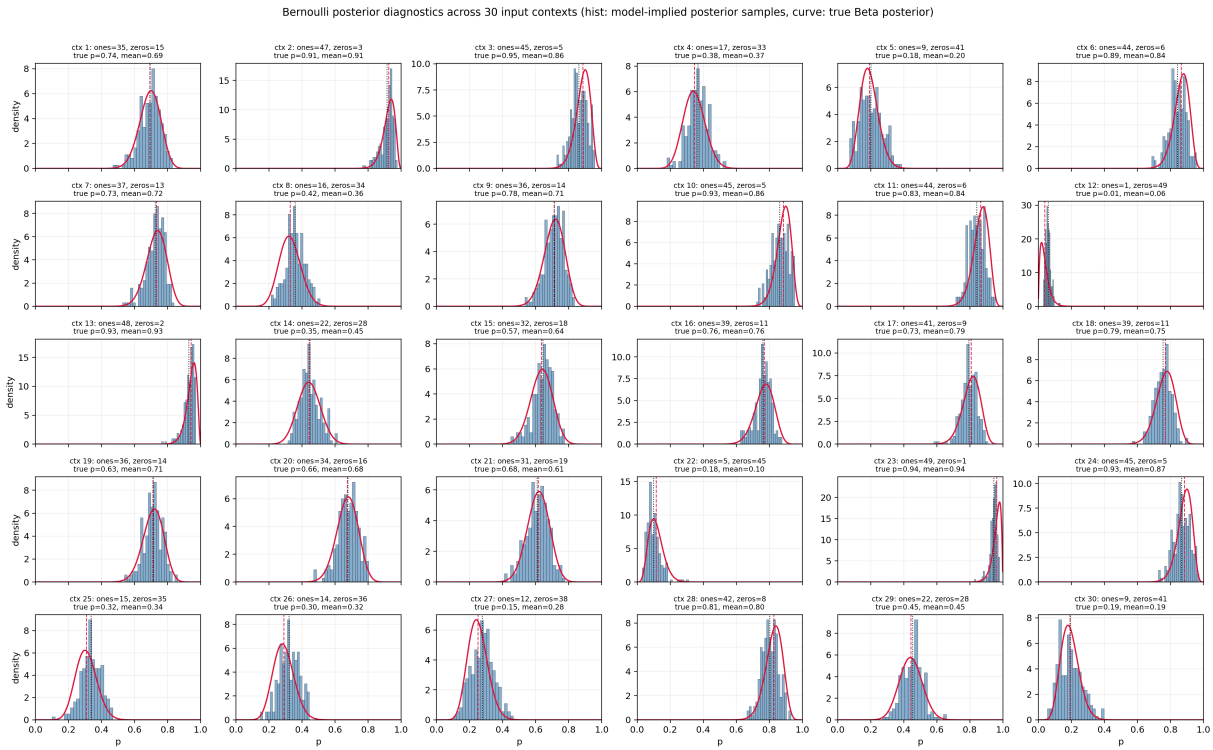

Figure 4: Posterior-shape diagnostics for 30 generated input contexts (5x6 grid). In each panel, blue histogram: model-implied posterior samples from 200 rollouts of length 1000; red curve: true Bayes posterior $p \mid$ context under Beta(1,1).

To quantify posterior-distribution similarity for each context, we computed:

- W_1: Wasserstein-1 distance between empirical and true posterior CDFs.
- KS(CDF): Kolmogorov-Smirnov distance between empirical and true posterior CDFs.
- CvM-int: integrated squared CDF error $\int_0^1 (F_n(p) - F(p))^2 \, dp$.
- PIT-KS and PIT-CvM: goodness-of-fit metrics after probability-integral transform $u_i = F_{\text{true}}(p_i)$ against Uniform(0,1). PIT-CvM is the Cramér-von Mises statistic W^2, which measures global (integrated) mismatch between the empirical CDF of $\{u_i\}$ and the uniform CDF.
- Quantile RMSE: RMSE of $\{5, 25, 50, 75, 95\}\%$ sample quantiles vs true posterior quantiles.
- Coverage MAE: mean absolute error of empirical central interval coverage at 50/80/90/95%.

For finite-sample interpretation, we also report a "perfect-match expected mean" baseline: the expected value of each metric when samples are drawn directly from the true posterior with the same sample size ($n = 200$).

Metric	Mean	Median	p90	p95	Max	Perfect-match expected mean
Wasserstein-1	0.0164	0.0148	0.0270	0.0297	0.0331	0.0048
KS(CDF)	0.1764	0.1645	0.2428	0.3545	0.5334	0.0590
CvM-int	0.00221	0.00161	0.00464	0.00524	0.00812	0.00015
PIT-KS	0.1794	0.1677	0.2459	0.3602	0.5403	0.0604
PIT-CvM	3.4308	2.2047	5.4743	10.9525	22.5864	0.1657
Quantile RMSE	0.0176	0.0166	0.0279	0.0293	0.0304	0.0060
Coverage MAE	0.0445	0.0412	0.0841	0.1012	0.1188	0.0199

Table 4: Summary of posterior-distribution similarity metrics across the 30 contexts, with finite-sample perfect-match baselines.

Overall, the posterior match is reasonably good in the typical case but not uniformly strong across all contexts. Relative to finite-sample perfect-match baselines, median errors are often a few times larger (e.g., W_1 median 0.0148 vs baseline mean 0.0048, quantile RMSE median 0.0166 vs 0.0060, coverage MAE median 0.0412 vs 0.0199). The tail metrics show difficult contexts: KS(CDF) reaches 0.5334 (baseline mean 0.0590), and PIT-CvM is heavy-tailed (p95 = 10.95, max = 22.59, baseline mean 0.1657). So the method is generally aligned but still exhibits substantial context-dependent mismatch in the harder cases.

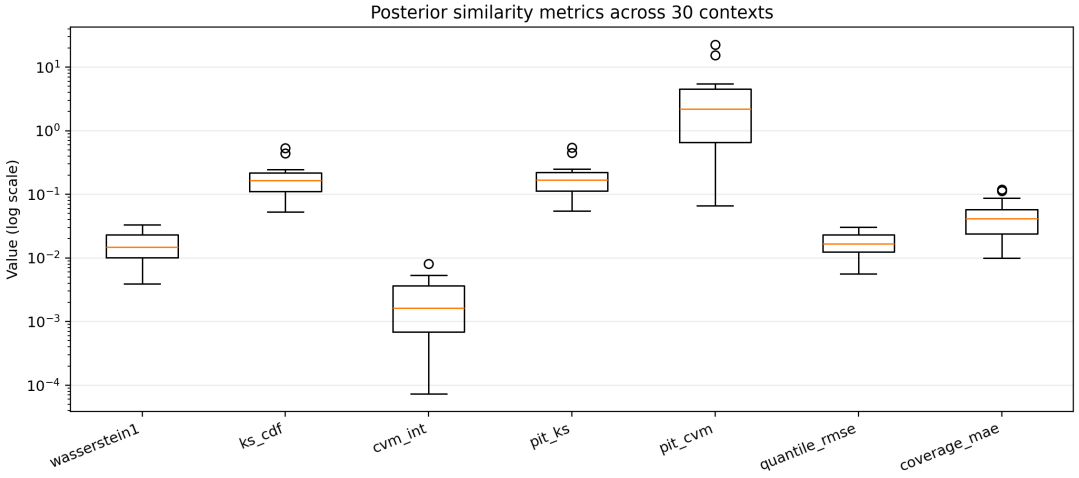

Figure 5: Boxplots of per-context posterior similarity metrics (30 values per metric, log scale).

Full per-context and summary metrics are available in: `artifacts/bernoulli_transformer_report/figures/` and `artifacts/bernoulli_transformer_report/figures/posterior_context_metrics_summary_5x6.csv`.

6. LPE Estimation Quality and Naive MC Baseline

The LPE numbers in this section come from the diagnostics run in `logs/bernoulli_diagnostics_beta11_fixed_ta` and `plots/bernoulli_posterior_sampling_diagnostics.csv`. For each of 30 trials:

- Sample latent Bernoulli parameter $p_{\text{true}} \sim \text{Beta}(1, 1)$.
- Generate one context $x_{1:n}$ of length $n = 50$ i.i.d. from Bernoulli(p_{true}).

- Use one fixed target string $y_{1:m}$ of length $m = 50$, shared across all contexts, with exactly 25 ones and 25 zeros (`target_mode=alternating`).

For each trial, the true event probability $P(y_{1:m} \mid x_{1:n})$ is computed analytically under the Beta-Bernoulli model with prior $\mathrm{Beta}(1, 1)$, and compared to the rollout-posterior estimator:

$$\text{If } s = \sum_{t=1}^{n} x_t, \quad p \mid x_{1:n} \sim \mathrm{Beta}(s+1,\ n-s+1).$$

For a fixed target string with $k = \sum_{t=1}^{m} y_t$ ones,

$$P(y_{1:m} \mid x_{1:n}) = \int_0^1 p^k (1-p)^{m-k} \, \frac{p^s (1-p)^{n-s}}{B(s+1, n-s+1)} \, dp = \frac{B(s+k+1,\ n-s+m-k+1)}{B(s+1,\ n-s+1)}.$$

In our run ($n = m = 50$, fixed target has $k = 25$), this becomes

$$P(y_{1:50} \mid x_{1:50}) = \frac{B(s+26,\ 76-s)}{B(s+1,\ 51-s)}.$$

The estimator compared against this exact value is:

$$\widehat{P}_{\mathrm{post}}(y_{1:m} \mid x_{1:n}) = \frac{1}{M} \sum_{j=1}^{M} f(\hat{p}_j), \quad f(p) = p^k (1-p)^{m-k},$$

where $k = \sum_{t=1}^{m} y_t$, $M = 200$, and each $\hat{p}_j$ is the fraction of ones in one model rollout of length $L = 1000$ conditioned on $x_{1:n}$. Reported errors are per-trial relative errors $|\widehat{P}_{\mathrm{post}} - P|/P$, summarized by percentiles.

For the "equal compute budget" comparison, we match total generated tokens per trial. The posterior method generates M rollouts of length L, so its generation cost is $M \cdot L$ tokens. A direct naive Monte Carlo estimator would instead generate R rollouts of the target length m, so its cost is $R \cdot m$ tokens. Setting these equal gives $R \approx (ML)/m$, hence:

$$R_{\mathrm{eq}} = \left\lfloor \frac{M \cdot L}{m} \right\rfloor = \left\lfloor \frac{200 \times 1000}{50} \right\rfloor = 4000.$$

This is why the naive baseline in the table uses $R_{\mathrm{eq}} = 4000$: it spends approximately the same generation compute as the posterior estimator.

For naive MC, let $I_r = \mathbf{1}\{Y_{1:m}^{(r)} = y_{1:m}\}$, where each rollout $Y_{1:m}^{(r)}$ has success probability $P \equiv P(y_{1:m} \mid x_{1:n})$. Then

$$\widehat{P}_{\mathrm{naive}} = \frac{1}{R} \sum_{r=1}^{R} I_r, \quad \mathrm{Var}(\widehat{P}_{\mathrm{naive}}) = \frac{P(1-P)}{R}.$$

So its relative standard error is

$$\frac{\sqrt{\mathrm{Var}(\widehat{P}_{\mathrm{naive}})}}{P} = \sqrt{\frac{1-P}{R\,P}}.$$

In the table, "Naive MC expected rel SE" is this quantity evaluated at $R = R_{\mathrm{eq}}$ and the exact per-trial true probability P, then summarized by p50/p90/p95 across the 30 trials. The "Naive MC P(zero hits)" column uses $P(\text{zero hits}) = (1-P)^R$ with the same R and P.

Statistic	Posterior method |rel err|	Naive MC expected rel SE	Naive MC P(zero hits)
p50	0.154	1307105.5	1.000000
p90	1.423	1055461232.9	1.000000
p95	2.050	2631950303.2	1.000000

Table 5: Posterior estimator error vs expected naive-MC error under matched token budget.

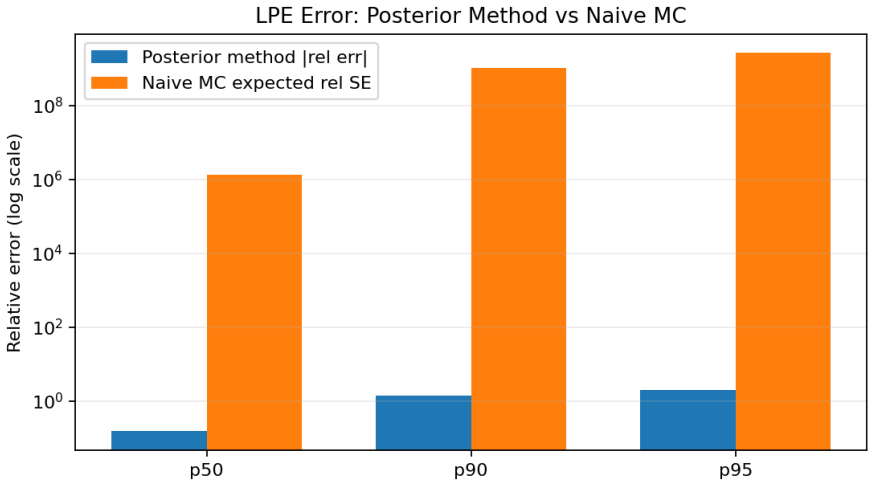

Figure 6: Percentile-level comparison of posterior method error and naive-MC expected error (log scale).

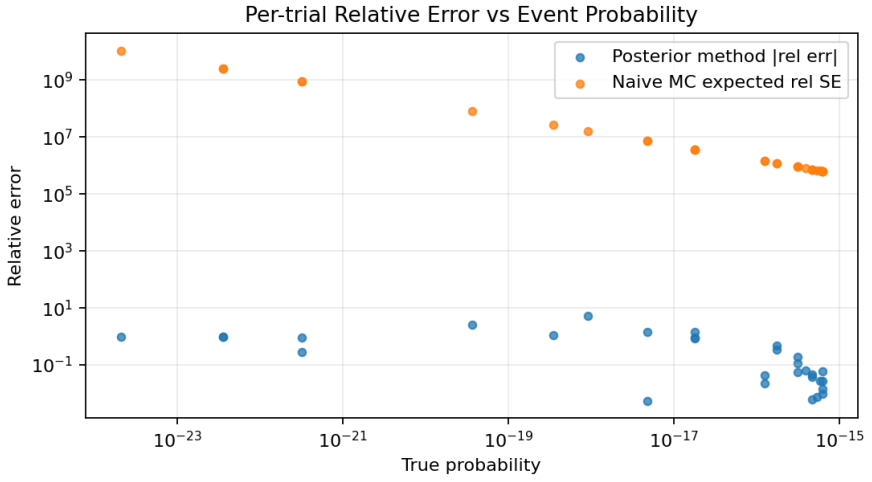

Figure 7: Per-trial relative error versus true event probability (log-log).