Bernoulli Transformer LPE Report

Automated run

February 17, 2026

Artifacts Used

- Checkpoint: `checkpoints/bernoulli_transformer_L1_D16_seq1024.pt`
- Training/diagnostics log: `logs/bernoulli_diagnostics_seq1024_M200_L1000.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

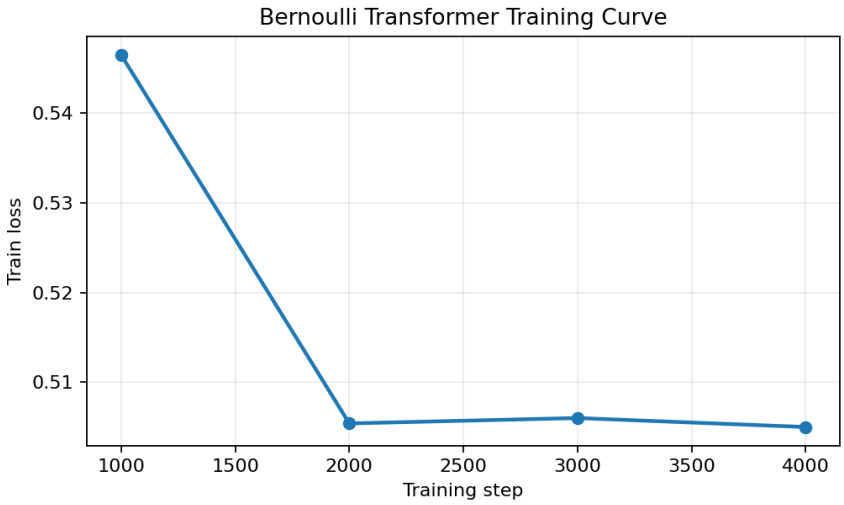

Figure 1: Logged training loss over optimization steps.

4. Final Loss vs Bayes Predictive

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

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.

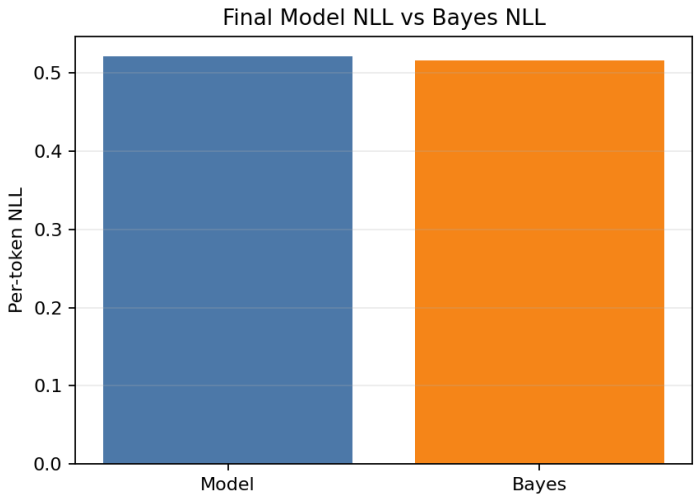

Figure 2: Per-token NLL: trained model vs Bayes optimal predictor.

5. Posterior-Sample Quality

Diagnostics used 30 trials, posterior samples per trial=200, rollout length=1000.

Metric	Value	Interpretation
$\lvert\log_{10}(\hat{p}/p)\rvert$ p50	0.069	Multiplicative error in log space
$\lvert\log_{10}(\hat{p}/p)\rvert$ p90	0.184	Tail log-error
$\lvert\log_{10}(\hat{p}/p)\rvert$ p95	0.204	Tail log-error
Posterior CV mean	1.066	Variability of Rao-Blackwell terms
Posterior CV median	0.812	Typical variability
Posterior-mean MSE	6.798e-05	Rollout mean vs exact Bayes mean
Posterior-mean correlation	0.9995	Rollout mean vs exact Bayes mean
$\log_{10}$ true-vs-est corr	0.9996	Alignment in log-probability space

Table 4: Posterior-sampling diagnostics.

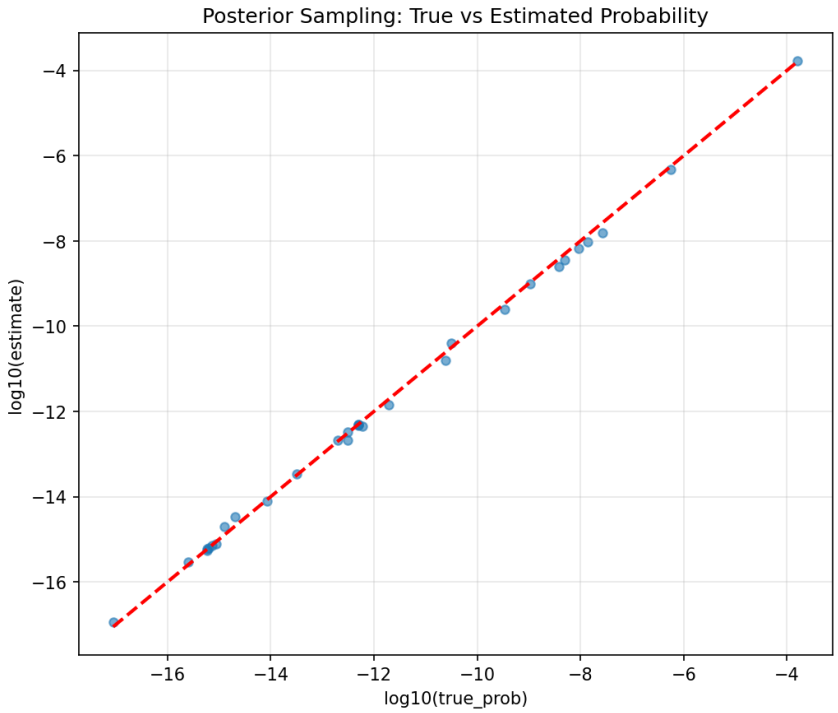

Figure 3: Diagnostics: $\log_{10}$(true prob) vs $\log_{10}$(estimated prob).

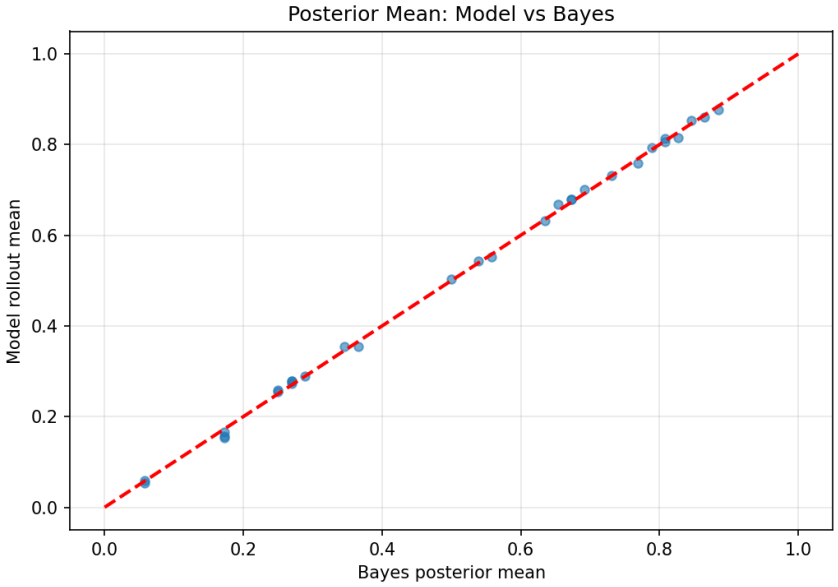

Figure 4: Diagnostics: rollout-derived posterior mean vs exact Bayes posterior mean.

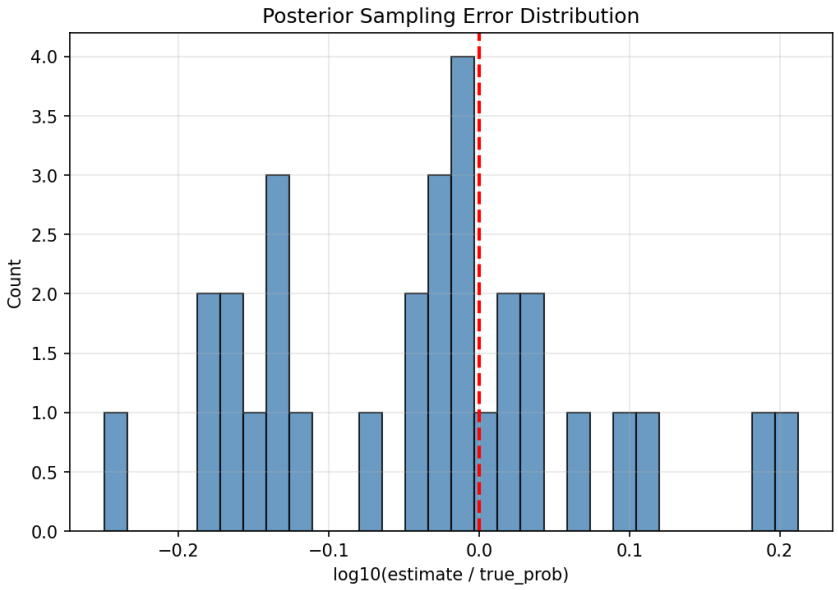

Figure 5: Diagnostics: histogram of $\log_{10}(\hat{p}/p)$ estimation error.

6. LPE Estimation Quality and Naive MC Baseline

Equal-compute naive Monte Carlo budget was estimated as:

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

Statistic	Posterior method |rel err|	Naive MC expected rel SE	Naive MC P(zero hits)
p50	0.160	22232.2	1.000000
p90	0.352	654130.8	1.000000
p95	0.508	849890.3	1.000000

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

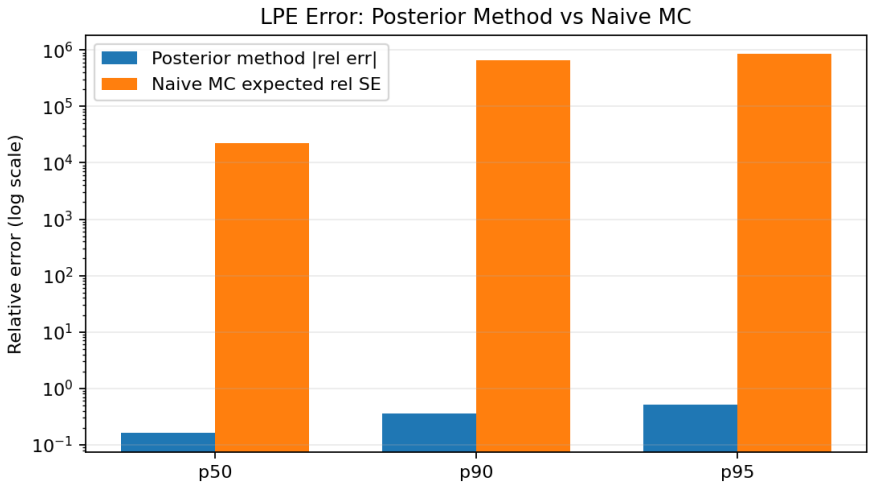

Figure 6: Percentile-level comparison of posterior method error and naive-MC expected error (log scale).

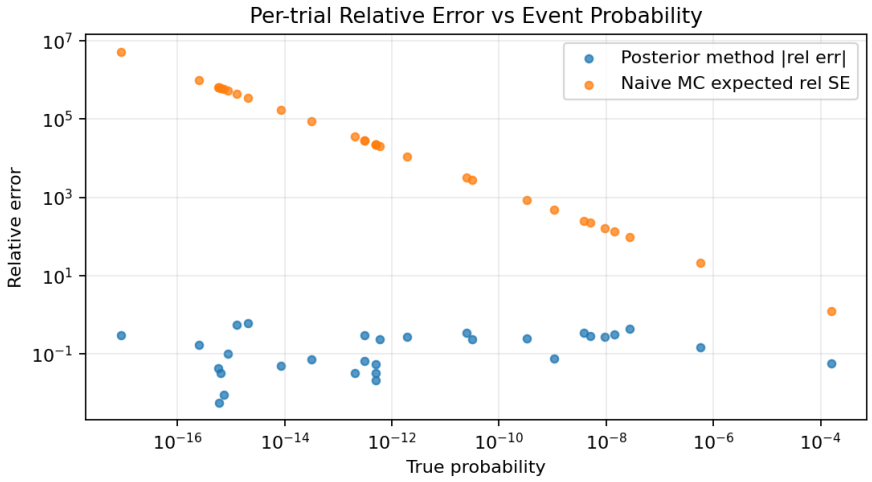

Figure 7: Per-trial relative error versus true event probability (log-log).

7. Additional Notes

- The Bayes gap is small (0.93%), indicating the transformer is close to Bayes-optimal one-step prediction on this task.
- LPE errors are still sensitive to event rarity: even with strong one-step calibration, tiny target probabilities produce heavy-tailed relative error.
- Under the same compute budget, naive direct-MC is effectively unusable on these trials (typical zero-hit probability near 1), while posterior sampling remains informative.