# Markov-$k$ Transformer LPE Report (k=2)

Automated run

February 20, 2026

## Artifacts Used

- Transformer artifacts: `artifacts/markov_k2_overnight_20260220_034625/attempt_01_a01_l8d256_seed1`
- Report figures and diagnostics: `artifacts/markov_k2_gaplt1_report/figures` and `artifacts/markov_k2_`

## 1. Model Architecture

| Field | Value |
|---|---|
| Transformer type | Decoder-only, causal self-attention |
| Positional encoding | Learned absolute positional embeddings |
| Markov order | $k = 2$ |
| Layers ($L$) | 8 |
| Model width ($d_{\text{model}}$) | 256 |
| Heads ($H$) | 8 |
| MLP width ($d_{\text{mlp}}$) | 1024 |
| Pre-norm | True |
| Trainable parameters | 7,368,194 |

Table 1: k=2 transformer architecture used for this report.

## 2. Training Details

| Field | Value |
|---|---|
| Training objective | Autoregressive cross-entropy on next-token prediction |
| Sequence length | 4000 |
| Batch size | 8 |
| Gradient accumulation | 8 |
| Optimizer | AdamW |
| Learning rate | $2.50 \times 10^{-4}$ (cosine decay to $2.00 \times 10^{-6}$) |
| Weight decay | 0.001 |
| Warmup schedule | Linear warmup for 500 steps |
| Gradient clipping | 0.8 |
| Total steps run | 2300 (early-stop target reached) |
| Early-stop target | Step-1 gap $\leq 0.95\%$ |

Table 2: Training setup for the high-capacity k=2 run.

Data generation process (explicit Markov parameterization):

- General $k$-Markov parameterization:

$$\theta_s \equiv P(x_t = 1 \mid x_{t-k:t-1} = s), \qquad s \in \{0,1\}^k.$$

- Prior over parameters: independent Beta per state

$$\theta_s \sim \mathrm{Beta}(1,1) \ \text{ independently for all } s \in \{0,1\}^k.$$

- Sequence generation for one training example of length $T = 4000$:

$$x_1, \ldots, x_k \overset{\text{i.i.d.}}{\sim} \mathrm{Bernoulli}(0.5),$$

$$x_t \mid x_{t-k:t-1} \sim \mathrm{Bernoulli}\left(\theta_{x_{t-k:t-1}}\right), \qquad t \geq k+1.$$

- For this report ($k = 2$), the latent parameters are

$$\theta_{00}, \theta_{01}, \theta_{10}, \theta_{11} \overset{\text{i.i.d.}}{\sim} \mathrm{Beta}(1,1),$$

with

$$P(x_t = 1 \mid x_{t-2:t-1} = ab) = \theta_{ab}, \qquad ab \in \{00, 01, 10, 11\}.$$

## 3. Training Curve

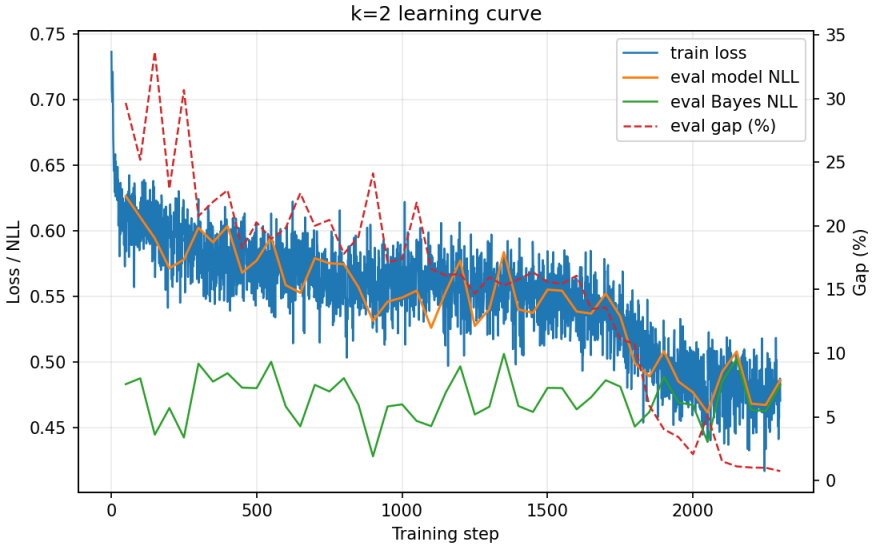


Figure 1: Training curve for k=2. Training loss is logged every gradient step. Evaluation curves are drawn only at true eval checkpoints and linearly connected between checkpoints.

## 4. Final Loss vs Bayes Predictive

Held-out evaluation uses sequences from the same Beta-Bernoulli Markov-2 process.

For a prefix $x_{1:t-1}$ and current state $s_t \in \{00, 01, 10, 11\}$, with transition counts $n_{s,0}, n_{s,1}$, the Bayes predictive is

$$P(x_t = 1 \mid x_{1:t-1}) = \frac{1 + n_{s_t,1}}{2 + n_{s_t,0} + n_{s_t,1}}.$$

The per-token Bayes NLL is

$$\ell_t^{\text{Bayes}} = -\big[x_t \log q_t + (1 - x_t) \log(1 - q_t)\big],$$

where $q_t = P(x_t = 1 \mid x_{1:t-1})$.

| Metric | Value | Notes |
|---|---|---|
| Model NLL | 0.505525 | Per-token negative log-likelihood |
| Bayes NLL | 0.501614 | Exact Bayes predictive on held-out samples |
| Gap | 0.78% | (model − Bayes)/Bayes |

Table 3: Model predictive quality against Bayes baseline (k=2).

To compare next-token probabilities directly, we generated 200 mixed-length contexts:

- Sample context length uniformly from $[64, 4095]$.
- Sample one latent Markov-2 process from the priors above and draw context bits.

- Compute Bayes and model $P(x_{n+1} = 1 \mid x_{1:n})$.

Summary: Pearson $r = 0.9933$, MAE $= 0.02438$.

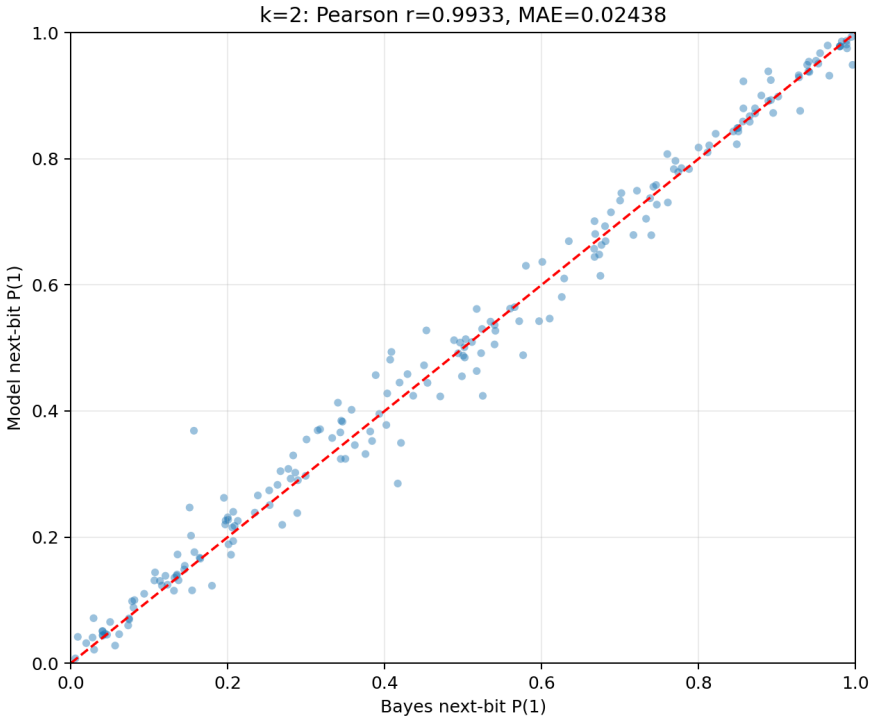


Figure 2: Model vs Bayes next-token probability on mixed-length contexts for k=2.

## 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 context. Contexts were generated as:

- Sample latent transitions independently: $\theta_{00}, \theta_{01}, \theta_{10}, \theta_{11} \sim \text{Beta}(1, 1)$.
- Sample one context of length 1000 from the Markov-2 process.
- Compute exact Bayes posterior marginals from context transition counts:

$$\theta_s \mid x_{1:n} \sim \text{Beta}(n_{s,1} + 1, n_{s,0} + 1), \quad s \in \{00, 01, 10, 11\}.$$

- Estimate model-implied posterior by 200 rollouts of length 1000 per context; each rollout yields one $\hat{\theta} \in [0, 1]^4$ from continuation transition frequencies.

| Metric | Mean | Median | p90 | p95 | Max | Perfect-match expected mean |
|---|---|---|---|---|---|---|
| Wasserstein-1 | 0.0386 | 0.0326 | 0.0639 | 0.0718 | 0.1129 | 0.0028 |
| KS(CDF) | 0.4128 | 0.3932 | 0.5563 | 0.5845 | 0.6509 | 0.0579 |
| CvM-int | 0.0146 | 0.0112 | 0.0254 | 0.0369 | 0.0510 | 0.0001 |
| PIT-KS | 0.4148 | 0.3967 | 0.5579 | 0.5870 | 0.6525 | 0.0605 |
| PIT-CvM | 18.8527 | 17.1613 | 30.6424 | 33.0650 | 40.0807 | 0.1651 |
| Quantile RMSE | 0.0412 | 0.0351 | 0.0657 | 0.0783 | 0.1178 | 0.0034 |
| Coverage MAE | 0.2877 | 0.2750 | 0.4060 | 0.4496 | 0.4850 | 0.0199 |

Table 4: Summary of posterior-distribution similarity metrics across 30 Markov-2 contexts, with finite-sample perfect-match baselines.

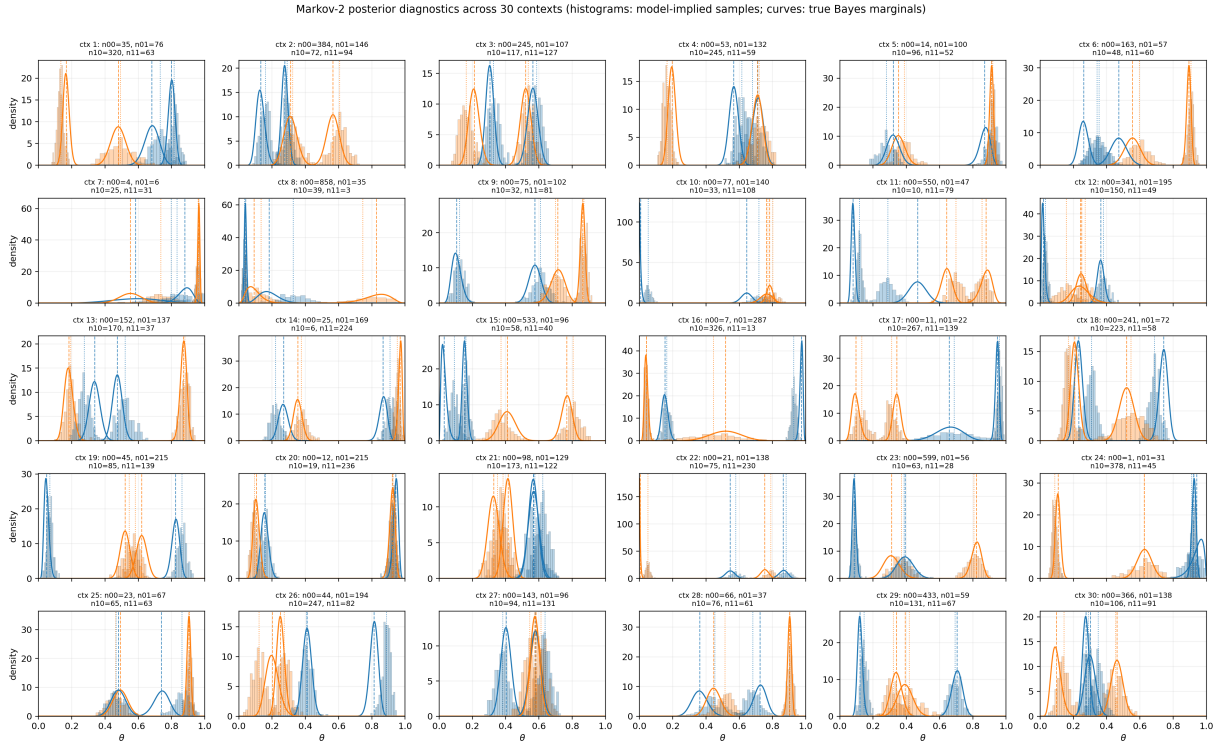


Figure 3: Posterior-shape diagnostics for 30 contexts (5x6 grid). Histograms are model-implied samples by state; curves are true Bayes posterior marginals.

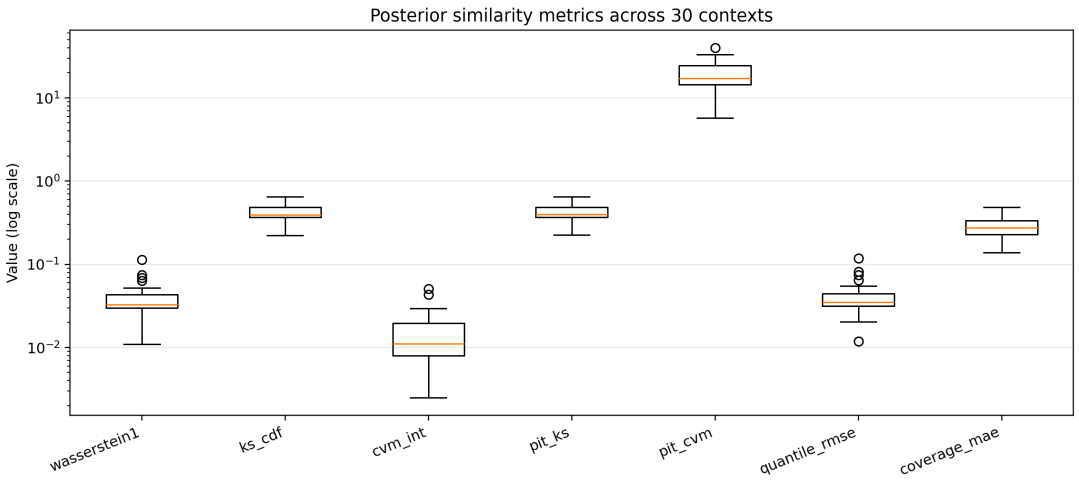


Figure 4: Boxplots of per-context posterior similarity metrics (30 values per metric, log scale).

For each metric, the reported context-level value is computed per state and then averaged, preserving the same metric family and table structure as the Bernoulli and k=1 reports while adapting to the 4-parameter posterior.

## 6. LPE Estimation Quality and Naive MC Baseline

For this k=2 LPE rerun (Step 3 only; same checkpoint as Sections 1–5):

- Number of contexts: 8 (all length 1000).
- Target string length: 100.
- Target mode: `balanced` (de Bruijn-based construction).
- The same fixed target string is used for all 8 contexts in this run.
- Posterior samples per context: $M = 200$; rollout length $L = 1000$.

A binary de Bruijn cycle $B(2, n)$ is a cyclic bit string of length $2^n$ such that, with wrap-around, every length-$n$ binary substring appears exactly once. For k=2 we use order $n = k + 1 = 3$, so the base cycle length is $2^3 = 8$.

Target generation matches `make_target_bits`:

1. Build $B(2, 3)$.
2. Sample a random cyclic offset.
3. Rotate, repeat, and truncate to length $m = 100$.

True event probability is computed by teacher forcing under the trained transformer:

$$P_{\mathrm{TF}}(y_{1:m} \mid x_{1:n}) = \prod_{t=1}^{m} P_{\mathrm{model}}(y_t \mid x_{1:n}, y_{1:t-1}) \,.$$

Here $x_{1:n}$ is the observed context (length $n$) and $y_{1:m}$ is the fixed target (length $m$). We use this teacher-forced model probability as the "true" probability for LPE relative-error evaluation.

The rollout-posterior estimator is

$$\widehat{P}_{\text{post}}(y_{1:m} \mid x_{1:n}) = \frac{1}{M} \sum_{j=1}^{M} f(\hat{\theta}^{(j)}),$$

where

$$f(\theta) \equiv P_\theta(y_{1:m} \mid x_{1:n}) = \prod_{t=1}^{m} \theta_{s_t}^{y_t} \big(1 - \theta_{s_t}\big)^{1-y_t}.$$

Equal-compute naive MC uses

$$R_{\text{eq}} = \left\lfloor \frac{ML}{m} \right\rfloor = \left\lfloor \frac{200 \times 1000}{100} \right\rfloor = 2000,$$

and

$$\text{relative SE}_{\text{naive}} = \sqrt{\frac{1-P}{R_{\text{eq}}P}}, \qquad P(\text{zero hits}) = (1-P)^{R_{\text{eq}}}.$$

| Statistic | Posterior method \|rel err\| | Naive MC expected rel SE | Naive MC $P$(zero hits) |
|---|---|---|---|
| p50 | $2.72 \times 10^{2}$ | $2.12 \times 10^{18}$ | 1.000000 |
| p90 | $8.39 \times 10^{4}$ | $4.49 \times 10^{25}$ | 1.000000 |
| p95 | $1.56 \times 10^{5}$ | $9.67 \times 10^{25}$ | 1.000000 |

Table 5: k=2 posterior estimator error vs expected naive-MC error under matched token budget.

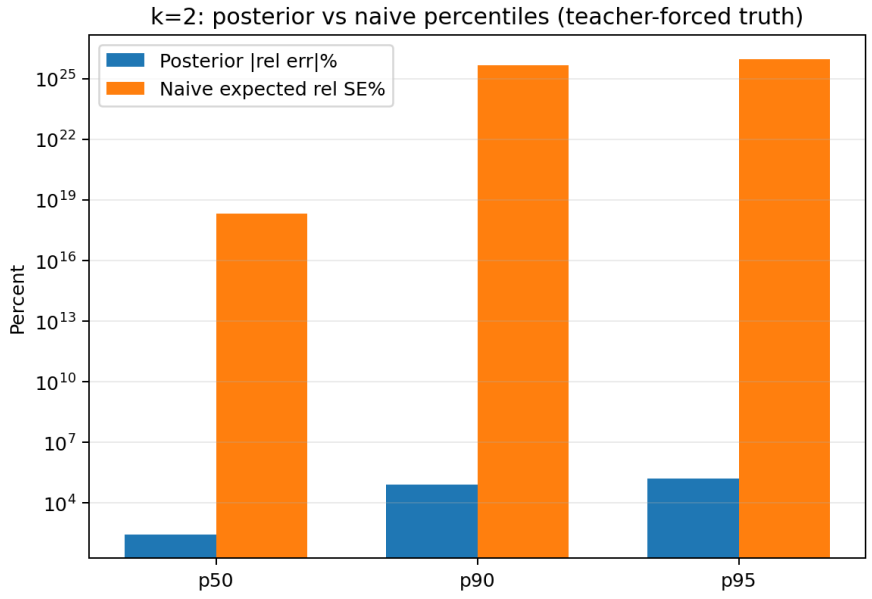


Figure 5: Percentile-level comparison of posterior method error and naive-MC expected error (k=2, teacher-forced truth, log scale).

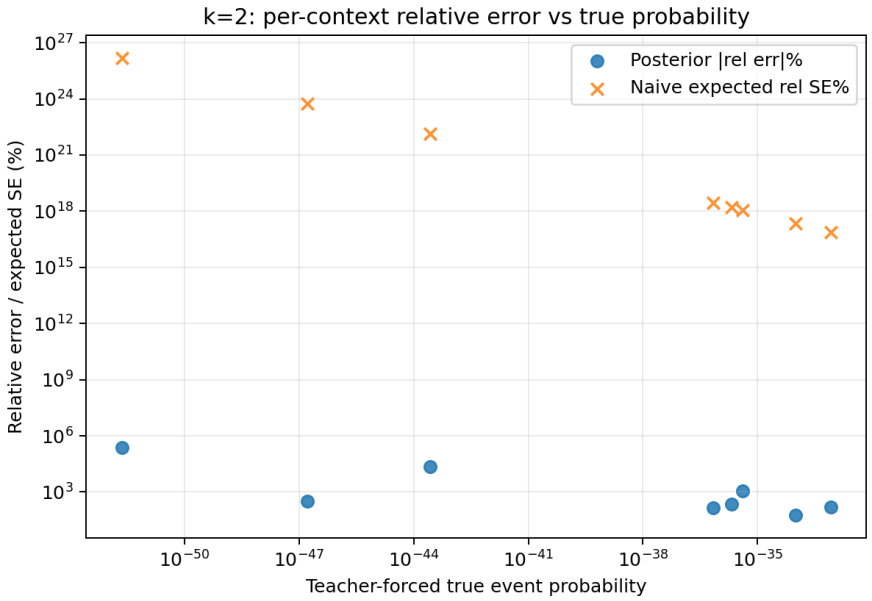


Figure 6: Per-context relative error versus teacher-forced true event probability (k=2, log-log).

| Context | True prob (teacher-forced) | Predicted prob | True posterior mean | True posterior sd | Sampled posterior mean | Sample |
|---|---|---|---|---|---|---|
| 0 | 2.680e-44 | 5.955e-42 | (0.492,0.044,0.658,0.059) | (0.031,0.011,0.025,0.055) | (0.573,0.048,0.618,0.081) | (0.045,0.0 |
| 1 | 6.898e-37 | 1.641e-36 | (0.477,0.113,0.320,0.530) | (0.025,0.019,0.028,0.061) | (0.534,0.114,0.272,0.565) | (0.037,0.0 |
| 2 | 9.999e-35 | 4.118e-35 | (0.589,0.128,0.494,0.549) | (0.030,0.019,0.028,0.052) | (0.616,0.142,0.512,0.476) | (0.045,0.0 |
| 3 | 2.061e-36 | 6.468e-36 | (0.529,0.255,0.655,0.843) | (0.040,0.028,0.031,0.019) | (0.537,0.283,0.671,0.840) | (0.061,0.0 |
| 4 | 2.266e-52 | 5.168e-49 | (0.034,0.098,0.220,0.765) | (0.006,0.046,0.064,0.100) | (0.022,0.063,0.501,0.521) | (0.007,0.0 |
| 5 | 4.105e-36 | 5.237e-35 | (0.623,0.503,0.413,0.860) | (0.040,0.040,0.039,0.015) | (0.685,0.546,0.393,0.852) | (0.054,0.0 |
| 6 | 1.618e-47 | 6.972e-47 | (0.667,0.545,0.455,0.994) | (0.149,0.144,0.144,0.003) | (0.719,0.505,0.696,0.992) | (0.217,0.1 |
| 7 | 8.236e-34 | 2.087e-33 | (0.397,0.553,0.668,0.724) | (0.037,0.034,0.033,0.022) | (0.444,0.491,0.695,0.748) | (0.057,0.0 |

Table 6: Per-context posterior diagnostics for the Step-3 rerun. Entries are vectors over states $(00, 01, 10, 11)$.

With teacher-forced model probability as truth, the posterior method remains far better than equal-compute naive MC across percentiles, but absolute errors are still large for difficult contexts.