A BAYESIAN PERSPECTIVE ON TASK RETRIEVAL AND TASK LEARNING IN IN-CONTEXT LEARNING ON MARKOVIAN DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

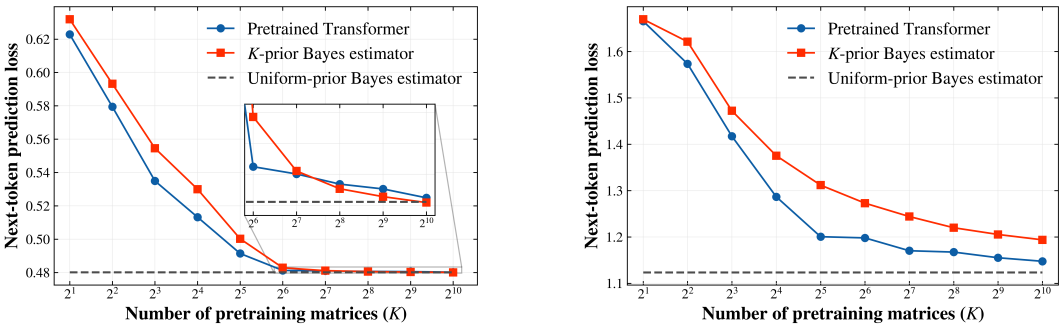
In-context learning enables pretrained transformers to adapt from a prompt without parameter updates, but it remains unclear when this behavior arises from retrieving memorized pretraining tasks (i.e., *task retrieval*) versus learning a rule from in-context data (i.e., *task learning*). We study this question through a finite-mixture Markov data model, where each latent Markov chain represents a pretraining task. Our analysis identifies the ratio of prompt length to the number of pretraining tasks as a key factor. We first establish quantitative guarantees showing that the finite-prior Bayesian predictor associated with the pretraining tasks favors retrieval at large ratios and learning from the prompt at sufficiently small ratios. We further show that, at small ratios, well-trained transformers make predictions close to those of the finite-prior Bayes predictor. At larger ratios, our experiments show increasing deviations from Bayes, with transformers exhibiting a stronger relative preference for learning. Our findings clarify how prompt length and pretraining diversity jointly shape in-context learning and when Bayesian prediction explains transformer behavior.

1 INTRODUCTION

Over the past few years, transformers (Vaswani et al., 2017), powered by the attention mechanism, have become one of the most widely used architectures in modern deep learning and achieved remarkable empirical success across a wide range of domains, including computer vision (Bao et al., 2023; Chen et al., 2020), natural language processing (Brown et al., 2020b; Devlin et al., 2019), and scientific machine learning (Geneva & Zabaras, 2022). A hallmark emergent capability of transformers is *in-context learning* (ICL) (Brown et al., 2020a; Olsson et al., 2022), through which pretrained transformers (PTs) adapt to new tasks without parameter updates by leveraging information provided in the input context, for example in translation, question answering, and arithmetic or logical reasoning. In particular, prior works (Lin & Lee, 2024; Pan et al., 2023) have identified two distinct modes of ICL behavior: *task retrieval*, where the model identifies and invokes a relevant pretrained task, and *task learning*, where the model infers an underlying rule from an in-context prompt and applies it to new queries.

A central question is what governs the competition between task retrieval and task learning. Empirical studies identify pretraining diversity as a key factor (Nguyen & Reddy, 2025; Chan et al., 2022). In regression experiments, Raventós et al. (2023); Lu et al. (2025) identify a diversity threshold below which transformers resemble Bayesian predictors restricted to the pretraining task family and above which they learn unseen tasks from context. In Markov data models, greater task diversity can shift transformers from relying on memorized transition matrices to estimating transition probabilities from the prompt (Park et al., 2025; Lepage et al., 2025) (see Figure 1). In addition, the competition between task retrieval and task learning also depends on the length of the in-context prompt. In classification experiments, additional demonstrations improve task learning but yield limited gains in task recognition (Pan et al., 2023). From a Bayesian perspective, Lin & Lee (2024) show that retrieval dominates with few examples, while learning becomes more prominent as more examples are provided. Context-dependent shifts between retrieval and inference strategies also occur in Markov data models (Park et al., 2025). Together, these studies identify pretraining task

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Number of states $S = 2$ (b) Number of states $S = 4$

Figure 1: **Transition from task retrieval to task learning in ICL on Markov data.** We report the next-token prediction loss on new test sequences for the finite-prior Bayes predictor (see Lemma 3.2(i)), the uniform-prior Bayes predictor (see Lemma 3.2(ii)), and the pretrained transformer (PT). When the number of pretraining matrices is small, the PT performs similarly to the K-prior Bayes predictor, suggesting that it relies on memorized pretraining tasks. In contrast, when the number of pretraining matrices is large, its performance approaches that of the uniform-prior Bayes predictor, indicating that it infers the transition rule from the in-context prompt. Please refer to Section E.1 for more details.

diversity and prompt length as key factors shaping retrieval–learning transitions. However, quantitative guarantees characterizing their joint effects in finite mixtures of Markov chains remain limited.

A growing body of work uses Markov data models as a simple yet expressive framework for studying transformers' ICL behavior and its underlying mechanisms (Edelman et al., 2024; Ekbote et al., 2025b; Lepage et al., 2025; Makkuva et al., 2025; Rajaraman et al., 2024b). Each transition rule defines a task, allowing the number of pretraining tasks and the prompt length to be varied independently. These models also make the distinction between retrieval and learning concrete: retrieval relies on transition rules learned during pretraining, whereas learning estimates a transition rule from the prompt. Please refer to Appendix A for more related work.

Our contributions. We develop a framework for understanding how pretraining task diversity and prompt length jointly shape task retrieval and task learning in Markov data models. Our main contributions are summarized as follows:

- **Quantitative characterization of retrieval and learning.** For a mixture of K binary Markov chains, we establish task retrieval and learning bounds on the finite-prior Bayes predictor for regular prompts of length n (see Theorem 3.3). When n/K is large, its prediction approaches that of the most likely pretraining task, with an error bound that decays exponentially in n/K. When n/K is sufficiently small, it approaches the uniform-prior Bayes predictor with an $O(K^{-1/2})$ error bound at a fixed confidence level.
- **Conditions for pretrained transformers to follow Bayes.** We prove that pretrained transformers closely approximate the finite-prior Bayes predictor on regular unseen-task prompts when both n/K and the achieved pretraining excess risk are sufficiently small (Theorem 3.4). We further bound their prediction error relative to the uniform-prior Bayes predictor, separating the effects of finite task diversity and imperfect pretraining. This transfers the Bayesian task-learning guarantee to pretrained transformers.
- **Empirical characterization of deviations from Bayes.** With architecture and training protocol fixed, our experiments show that transformers generally depart further from finite-prior Bayes as n/K increases. These deviations are accompanied by a growing preference shift toward task learning relative to the Bayes predictor, helping explain why their retrieval–learning phase diagrams differ (see Figures 3 and 4).

Notation. Given a matrix $\boldsymbol{A}$, we denote its (i, j)-th entry by a_{ij}. Given a vector $\boldsymbol{a}$, we use $\|\boldsymbol{a}\|$ to denote its Euclidean norm and a_i to denote its i-th entry. We use $[n]$ to denote the set $\{1, \ldots, n\}$ for a positive integer n, and $\boldsymbol{e}_i \in \mathbb{R}^d$ to denote the i-th standard basis vector in $\mathbb{R}^d$. We use $\mathrm{Unif}(0, 1)$ to denote the uniform distribution on the interval $(0, 1)$, and $\mathbb{I}\{A\}$ to denote the indicator function of an event A, which equals 1 if A occurs and 0 otherwise. Moreover, δ_z denotes the Dirac measure at z, and $|\Omega|$ denotes the cardinality of a finite set Ω.

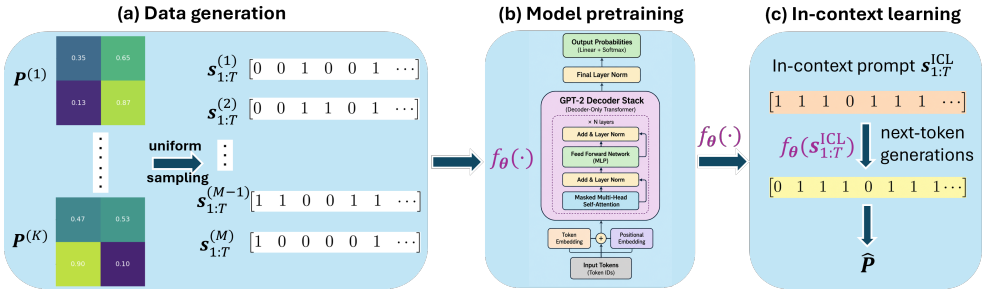


Figure 2: **Finite-prior Markov data generation, model pretraining, and in-context evaluation.** We first generate K latent Markov chains according to Definition 2.1. At each training step, we select one of these chains uniformly at random and generate a fresh training sequence $s_{1:T}^{(i)}$ from it. Then, we train a transformer on the resulting sequences by minimizing the objective in (2). During evaluation, we provide an in-context prompt $s_{1:n}^{\mathrm{ICL}}$ and evaluate the model's next-token prediction.

2 Problem Setup

In this section, we introduce a first-order Markov data model, define the transformer pretraining objective, and describe the in-context prediction setting. Figure 2 illustrates the overall workflow, from data generation and pretraining to in-context evaluation.

Markov chain model for token sequences. Let $\{0,1\}$ denote the binary alphabet. We consider token sequences $s_{1:T} := (s_1, \ldots, s_T) \in \{0,1\}^T$ generated by a binary first-order Markov chain. Specifically, each next token depends only on the current token, i.e., $\mathbb{P}(s_{n+1} = j \mid s_{1:n}) = \mathbb{P}(s_{n+1} = j \mid s_n)$ for every $n \geq 1$ and $j \in \{0,1\}$. We denote the Markov transition matrix by $\boldsymbol{P} = (\boldsymbol{P}_{i,j})_{i,j\in\{0,1\}}$, where $\boldsymbol{P}_{i,j} := \mathbb{P}(s_{n+1} = j \mid s_n = i)$. Here, we parameterize this matrix as

$$\boldsymbol{P}(p,q) = \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix}, \quad p, q \in (0,1). \tag{1}$$

We assume that the initial token s_1 is drawn from a fixed distribution $\boldsymbol{\mu} = (\mu_0, \mu_1)$, independent of (p,q), where $\mu_0, \mu_1 \in (0,1)$ and $\mu_0 + \mu_1 = 1$. For simplicity, we write $s_{1:T} \sim (\boldsymbol{\mu}, \boldsymbol{P}(p,q))$ to indicate that a sequence $s_{1:T}$ is generated by a binary Markov chain with initial distribution $\boldsymbol{\mu}$ and transition matrix $\boldsymbol{P}(p,q)$.

In our setting, each task is specified by a latent Markov transition matrix. As illustrated in Figure 2, we generate the pretraining data sequences using the finite-mixture Markov model as follows.

Definition 2.1. *Draw $(p^{(k)}, q^{(k)}) \overset{i.i.d.}{\sim} \mathrm{Unif}(0,1)^2$ for each $k \in [K]$. The pretraining task family consists of K latent Markov chains with the k-th task specified by the transition matrix $\boldsymbol{P}(p^{(k)}, q^{(k)})$ for each $k \in [K]$. Conditioned on the sampled task family, the training prior is $\Pi_{\mathrm{train}} := \sum_{k=1}^{K} \delta_{(p^{(k)}, q^{(k)})}/K$.*

According to this model, conditioned on the task family $\{(p^{(k)}, q^{(k)})\}_{k=1}^{K} \overset{i.i.d.}{\sim} \mathrm{Unif}(0,1)^2$, a pretraining sequence is generated by first drawing a latent task label $J \sim \mathrm{Unif}([K])$ and then drawing $s_{1:T} \mid J = k \sim \big(\boldsymbol{\mu}, \boldsymbol{P}(p^{(k)}, q^{(k)})\big)$.

Pretraining. We consider an L-layer causal transformer parameterized by $\boldsymbol{\theta} \in \Theta$, where $L \geq 2$ and $\boldsymbol{\theta}$ collects all model parameters, including the token embeddings, attention and feedforward weights, relative positional encodings, and readout parameters, and Θ denotes the corresponding parameter space. Given a binary sequence $s_{1:T} \in \{0,1\}^T$, the transformer produces layer-wise hidden representations $\{\boldsymbol{x}_n^{(\ell)}\}_{n=1}^{T}$ for $\ell = 0, \ldots, L$. At each position n, the final representation $\boldsymbol{x}_n^{(L)}$ is passed through a readout map to produce $f_{\boldsymbol{\theta}}(s_{1:n}) \in (0,1)$, interpreted as the predicted probability that the next token is 1. We denote the resulting transformer hypothesis class by $\mathcal{F} := \{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$. The formal specification of the architecture and its parameterization is deferred to Appendix B.

At the population level, we model pretraining using the standard next-token prediction objective for autoregressive language modeling (Brown et al., 2020a). Specifically, for any measurable next-token

predictor f, we define the expected next-token cross-entropy (CE) loss as

$$\mathcal{L}_K(f) := \frac{1}{T-\hat{T}} \mathbb{E}_{(p,q)\sim\Pi_{\text{train}},\ s_{1:T}\sim(\boldsymbol{\mu},\boldsymbol{P}(p,q))} \left[\sum_{n=\hat{T}}^{T-1} \ell\left(s_{n+1}, f(s_{1:n})\right) \right], \tag{2}$$

where $\hat{T} \in [T-1]$ is a context warm-up length and $\ell(\cdot,\cdot)$ denotes the binary CE loss defined by

$$\ell(x,y) := -x\log(y) - (1-x)\log(1-y). \tag{3}$$

Notably, we study a trained predictor $f_{\widehat{\boldsymbol{\theta}}} \in \mathcal{F}$ through its excess population risk relative to the finite-prior Bayes predictor. Our analysis takes the achieved excess risk as given, without analyzing how training attains it.

In-context learning. At inference time, given an in-context prompt $s_{1:n} \in \{0,1\}^n$, we study the next-token prediction $f(s_{1:n})$. Throughout the paper, we use *task retrieval* and *task learning* as operational descriptions of this prediction-level behavior. Specifically, we say that f exhibits task-retrieval behavior when its prediction is close to the next-token transition probability of the pretraining Markov chain most compatible with the prompt. In contrast, we say that f exhibits task-learning behavior when its prediction is close to the empirical Markov predictor computed from the transition statistics of the prompt.

3 Main Results

In this section, we first introduce notation and definitions in Section 3.1. We then characterize task retrieval and task learning for the finite-prior Bayes predictor in Section 3.2, and study how these behaviors are inherited by pretrained transformers in Section 3.3.

3.1 Preliminary Setup

Given a prefix $s_{1:n} \in \{0,1\}^n$ with $n \geq 2$, we define the transition counts

$$N_{ij}(s_{1:n}) := \textstyle\sum_{t=1}^{n-1} \mathbb{I}\{s_t = i,\ s_{t+1} = j\}, \quad \forall i,j \in \{0,1\}. \tag{4}$$

Thus, $N_{ij}(s_{1:n})$ records the number of occurrences of the transition $i \to j$ in the prefix. When the underlying prefix is clear from context, we write N_{ij} in place of $N_{ij}(s_{1:n})$. Moreover, define

$$N_0 := N_{00} + N_{01}, \quad N_1 := N_{10} + N_{11}, \quad \hat{c}_0 := \frac{N_0}{n-1}, \quad \hat{c}_1 := \frac{N_1}{n-1}. \tag{5}$$

N_0 and N_1 are the numbers of observed transitions originating from states 0 and 1, respectively, while $\hat{c}_0$ and $\hat{c}_1$ are their empirical frequencies. The corresponding empirical transition matrix is

$$\hat{\boldsymbol{P}} := \boldsymbol{P}(\hat{p},\hat{q}) = \begin{bmatrix} 1-\hat{p} & \hat{p} \\ \hat{q} & 1-\hat{q} \end{bmatrix}, \quad \hat{p} := \frac{N_{01}}{N_0}, \quad \hat{q} := \frac{N_{10}}{N_1}. \tag{6}$$

Throughout the main analysis, whenever $\hat{p}$ and $\hat{q}$ are used, we restrict attention to prefixes for which $N_0, N_1 > 0$ so that these quantities are well defined.

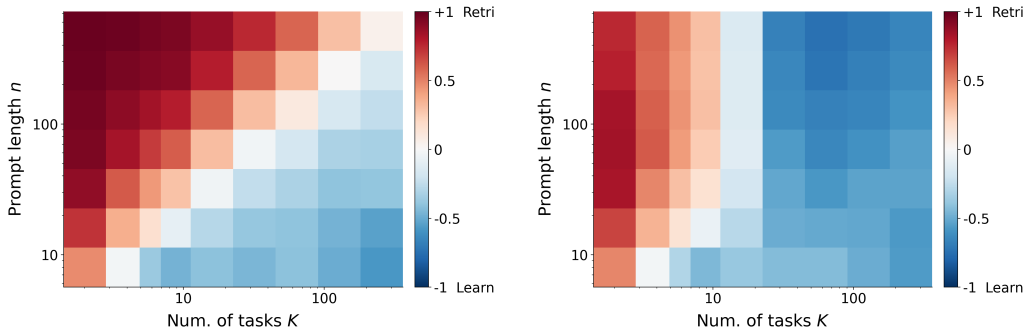
Definition 3.1. *Fix constants $\eta, \gamma \in (0, 1/2)$, independent of the prefix length n and the number of pretraining tasks K. We say that a prefix is (η,γ)-regular if $\hat{p}, \hat{q} \in [\eta, 1-\eta]$ and $\hat{c}_0 \in [\gamma, 1-\gamma]$.*

Bayes predictors. Let $\mathcal{G}$ denote the class of all measurable next-token predictors, and $\mathcal{L}_{\text{unif}}$ denote the population loss defined as in (2) with the finite-prior prior Π_{train} replaced by the continuous uniform prior $\text{Unif}((0,1)^2)$. We define the finite-prior and population-prior Bayes predictors as global minimizers of their respective population losses:

$$f_K^* \in \arg\min_{f\in\mathcal{G}} \mathcal{L}_K(f), \quad f_{\text{unif}}^* \in \arg\min_{f\in\mathcal{G}} \mathcal{L}_{\text{unif}}(f).$$

Lemma 3.2. *For every $n \in [T-1]$ and every prefix $s_{1:n} \in \{0,1\}^n$, the following statements hold: (i) The Bayes predictor under the finite-prior prior Π_{train} is*

$$f_K^*(s_{1:n}) = \sum_{k=1}^{K} \alpha_k \boldsymbol{P}^{(k)}_{s_n,1}, \quad \alpha_k := \frac{\mathbb{P}\left(s_{1:n} \mid p^{(k)}, q^{(k)}\right)}{\sum_{j=1}^{K} \mathbb{P}\left(s_{1:n} \mid p^{(j)}, q^{(j)}\right)},\ \forall k \in [K], \tag{7}$$



(a) finite-prior Bayes predictor (b) Pretrained transformer

Figure 3: **Retrieval-learning phase diagrams.** The x-axis is the number of latent Markov matrices K and y-axis is the prompt length n. Each cell is colored according to the relative alignment score $s(K, n)$ defined in (16). Deeper red indicates stronger relative alignment with the most likely pretraining task (task retrieval), whereas deeper blue indicates stronger relative alignment with the uniform-prior Bayes predictor (task learning). Pretrained transformers exhibit behavior similar to that of the finite-prior Bayes predictor when the prompt length n is small. However, as n increases, their retrieval–learning patterns diverge.

where $\boldsymbol{P}^{(k)}_{s_n,1}$ *denotes the transition probability from state* s_n *to* 1 *and* $\mathbb{P}\left(s_{1:n} \mid p^{(k)}, q^{(k)}\right)$ *is the prefix likelihood under the* k*-th Markov chain.*
(ii) The Bayes estimator under the continuous uniform prior $\mathrm{Unif}(0,1)^2$ *is*

$$f^*_{\mathrm{unif}}(s_{1:n}) = \frac{N_{s_n,1} + 1}{N_{s_n} + 2}. \tag{8}$$

The proof is deferred to Section C.2. Here, (7) expresses f^*_K as a posterior-weighted average of the next-token transition probabilities of the pretraining tasks, while (8) identifies f^*_{unif} as the Laplace-smoothed empirical next-token predictor $\hat{\boldsymbol{P}}_{s_n,1}$ (Demidenko, 2025), where $\hat{\boldsymbol{P}}$ is defined in (6).

3.2 Analysis of the finite-prior Bayes Predictor

We now characterize the retrieval and learning behavior of the finite-prior Bayes predictor f^*_K. The following theorem establishes prediction bounds that reveal the role of the ratio of the prefix length n to the number of pretraining tasks K in governing the transition between these regimes.

Theorem 3.3. *Suppose that we draw the pretraining task family* $\{(p^{(k)}, q^{(k)})\}_{k=1}^K$ *according to Definition 2.1. Fix* $\eta, \gamma \in (0, 1/2)$*. The following holds for every* (η, γ)*-regular prefix* $s_{1:n}$*, every integer* $K \geq 1$*, and* $\delta \in (0, 1]$*:*
(i) ***Task retrieval:*** *It holds with probability at least* $1 - \delta$ *that*

$$\left| f^*_K(s_{1:n}) - \boldsymbol{P}^{(k^\star)}_{s_n,1} \right| \leq 2e \exp\left(-\frac{\delta n}{\kappa_1 K}\right), \tag{9}$$

where $k^\star \in \arg\max_{k \in [K]} \mathbb{P}\left(s_{1:n} \mid p^{(k)}, q^{(k)}\right)$ *with ties broken by choosing the smallest index, and* $\kappa_1 > 0$ *is a constant only depending on* η *and* γ*.*
(ii) ***Task learning:*** *If* $n/K \leq c/\log(3/\delta)$*, it holds with probability at least* $1 - \delta$ *that*

$$|f^*_K(s_{1:n}) - f^*_{\mathrm{unif}}(s_{1:n})| \leq \kappa_2 \sqrt{\frac{\log(3/\delta)}{K}}, \tag{10}$$

where $c, \kappa_2 > 0$ *are constants only depending on* η *and* γ*.*

The proof of this theorem is deferred to Section C.3. Note that the probability bounds in this theorem are taken over the sampled pretraining task family, with the regular prefix held fixed. Before we proceed, let us make some remarks on the significance of this theorem.

- **Dependence of retrieval and learning on** n/K**.** At a fixed confidence level, Theorem 3.3 identifies sufficient conditions for task retrieval and task learning, as illustrated in Figure 3(a). When

n/K is large, the prediction error relative to the most likely pretraining task admits an upper bound that decays exponentially in n/K. When n/K is sufficiently small, the discrepancy from the uniform-prior Bayes predictor admits an upper bound of order $K^{-1/2}$. Moreover, we empirically evaluate the scaling behavior predicted by the bounds (9) and (10) in Figure 5.

- **From Bayes predictor to transformers.** Recent empirical studies show that increasing task diversity K can shift trained transformers away from finite-prior Bayesian prediction and toward the continuous-prior benchmark (Raventós et al., 2023; Park et al., 2025). Our experiments further reveal that the retrieval–learning phase diagrams of pretrained transformers and the finite-prior Bayes predictor do not fully coincide, with more pronounced discrepancies at longer contexts; see Figure 3. These findings suggest that Theorem 3.3 provides the most reliable guidance for transformer behavior in the shorter-context regime. At longer contexts, transferring the Bayesian guarantees to pretrained transformers requires additional control of the discrepancy between the trained model and the finite-prior Bayes predictor.

Takeaway: The finite-prior Bayes predictor f_K^* in (7) favors retrieving the most likely pretraining task when n/K is large and learning from the prompt when n/K is sufficiently small, with respective error bounds that decay exponentially in n/K and scale as $K^{-1/2}$.

3.3 When Do Pretrained Transformers Behave Like the Bayes Predictors?

The preceding result characterizes retrieval and learning for the finite-prior Bayes predictor f_K^*, which minimizes the population pretraining objective in (2) over all measurable predictors. However, pretraining optimizes this objective within the transformer function class $\mathcal{F}$, and thus the trained predictor $f_{\widehat{\boldsymbol{\theta}}}$ may deviate from f_K^*, as illustrated in Figure 3b. We study when pretrained transformers follow finite-prior Bayes closely on unseen tasks and how their deviations relate to retrieval and learning.

Prediction error and pretraining excess risk. Fix $\eta, \gamma \in (0, 1/2)$ and a context length $n \in \{\hat{T}, \dots, T-1\}$ with $n \geq 2$. For each prefix $s_{1:n} \in \{0,1\}^n$, we define the finite-prior and uniform-prior marginal likelihoods, respectively, by

$$\rho_K(s_{1:n}) := \mathbb{E}_{(p,q)\sim\Pi_{\text{train}}}\left[\mathbb{P}(s_{1:n} \mid p, q)\right], \ \rho_{\text{unif}}(s_{1:n}) := \mathbb{E}_{(p,q)\sim\text{Unif}(0,1)^2}\left[\mathbb{P}(s_{1:n} \mid p, q)\right]. \quad (11)$$

Here, $\rho_K(s_{1:n})$ is the probability of observing $s_{1:n}$ when the generating task is chosen uniformly from the K pretraining tasks, whereas $\rho_{\text{unif}}(s_{1:n})$ is the corresponding probability when (p, q) is sampled uniformly from $(0,1)^2$. Let $\mathcal{E}_n$ denote the event that the random prefix $s_{1:n}$ is (η, γ)-regular. For any two predictors f and g, define their root mean squared prediction error (RMSE) on regular unseen-task prompts by

$$\text{RMSE}_n(f, g) := \left(\mathbb{E}_{s_{1:n}\sim\rho_{\text{unif}}}\left[|f(s_{1:n}) - g(s_{1:n})|^2 \,\middle|\, \mathcal{E}_n\right]\right)^{1/2}.$$

In particular, $\text{RMSE}_n(f_{\widehat{\boldsymbol{\theta}}}, f_K^*)$ and $\text{RMSE}_n(f_{\widehat{\boldsymbol{\theta}}}, f_{\text{unif}}^*)$ measure how closely the transformer follows the finite-prior and uniform-prior Bayes predictors, respectively.

For the trained predictor $f_{\widehat{\boldsymbol{\theta}}}$, we define its excess population cross-entropy at context length n relative to f_K^* by

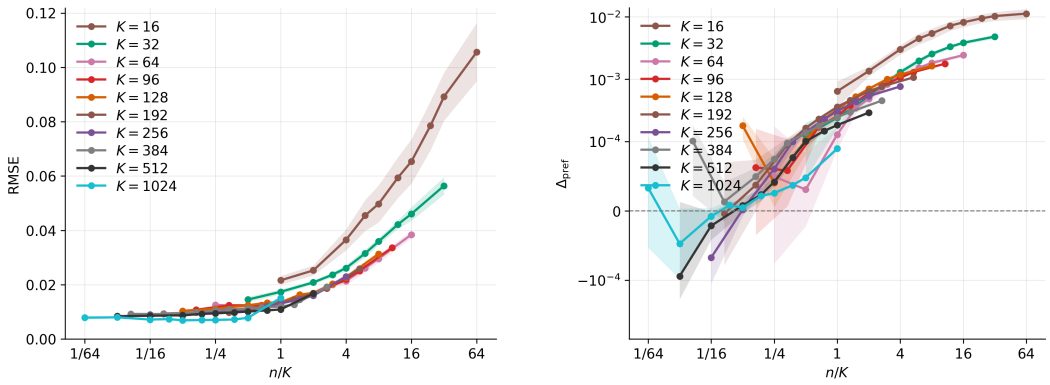
$$\epsilon_{K,n} := \mathbb{E}_{(p,q)\sim\Pi_{\text{train}}, s_{1:n+1}\sim(\boldsymbol{\mu}, \boldsymbol{P}(p,q))}\left[\ell\big(s_{n+1}, f_{\widehat{\boldsymbol{\theta}}}(s_{1:n})\big) - \ell\big(s_{n+1}, f_K^*(s_{1:n})\big)\right] \quad (12)$$

This excess risk averages over length-n prefixes under the pretraining distribution without conditioning on regularity, whereas the RMSE is evaluated on regular prompts from unseen tasks.

Pretrained transformer behavior for small n/K. We first establish the following theorem that bounds the trained transformer's prediction errors relative to both the finite-prior and uniform-prior Bayes predictors when n/K is sufficiently small.

Theorem 3.4. *Suppose that the pretraining task family $\{(p^{(k)}, q^{(k)})\}_{k=1}^K$ is sampled according to Definition 2.1. Fix $\eta, \gamma \in (0, 1/2)$. There exist constants $c_1, c_2 > 0$, depending only on (η, γ), such that the following holds. For every $n \in \{\hat{T}, \dots, T-1\}$ with $n \geq 2$ and $\rho_{\text{unif}}(\mathcal{E}_n) > 0$, and every $\delta \in (0,1)$, if*

$$\frac{n}{K} \leq \frac{1}{c_1 \log(12n^2/\delta)},$$



(a) $\text{RMSE}_n(f_{\widehat{\boldsymbol{\theta}}}, f_K^*)$ (b) Learning preference Δ_{pref}

Figure 4: **Transformer-to-Bayes errors and learning preference as n/K increases. (a)** Prediction error relative to finite-prior Bayes, $\text{RMSE}_n(f_{\widehat{\boldsymbol{\theta}}}, f_K^*)$ (see (13)). **(b)** Relative preference shift Δ_{pref} (see (14)), where positive values indicate a shift toward learning compared with finite-prior Bayes. Curves show means over seven task libraries per K; shaded bands indicate ± 1 standard error across libraries. Experimental details are provided in Section 4.2.

with probability at least $1-\delta$ *over the sampled pretraining task family,*

$$\text{RMSE}_n(f_{\widehat{\boldsymbol{\theta}}}, f_K^*) \le c_2\sqrt{\epsilon_{K,n}}, \quad \text{RMSE}_n(f_{\widehat{\boldsymbol{\theta}}}, f_{\text{unif}}^*) \le c_2\left(\sqrt{\frac{\log(12n^2/\delta)}{K}} + \sqrt{\epsilon_{K,n}}\right). \tag{13}$$

- The proof is deferred to Section D.3. Under the stated condition, the first inequality (13) shows that the transformer makes predictions close to f_K^* on regular unseen-task prompts when $\epsilon_{K,n}$ is small. The second inequality in (13) compares the transformer with f_{unif}^*, which contains two terms: the error bound for approximating f_{unif}^* by f_K^*, established in Theorem 3.3, and the transformer-to-Bayes error bound from the first inequality.
- This theorem helps interpret the small-n/K region of the phase diagrams in Figure 3. At fixed n, a sufficiently large K makes both n/K and the finite-library approximation term small. If $\epsilon_{K,n}$ is also small, the theorem guarantees that the pretrained transformer $f_{\widehat{\boldsymbol{\theta}}}$ is close to the uniform-prior Bayes predictor f_{unif}^* and thus exhibits task-learning behavior. At fixed K, shorter contexts can satisfy the same condition on n/K. Provided that $\epsilon_{K,n}$ is small, the first inequality then guarantees agreement with the finite-prior Bayes predictor f_K^*.

Pretrained transformer behavior for large n/K. The preceding theorem shows that pretrained transformers approximate the finite-prior Bayes predictor when both n/K and the pretraining excess risk $\epsilon_{K,n}$ are sufficiently small. We now examine empirically how their predictions diverge as n/K increases. Under the experimental setup in Section 4.2, with the architecture and training protocol fixed across K, we plot $\text{RMSE}_n(f_{\widehat{\boldsymbol{\theta}}}, f_K^*)$ against n/K in Figure 4a. It is observed that $\text{RMSE}_n(f_{\widehat{\boldsymbol{\theta}}}, f_K^*)$, computed on regular unseen-task prefixes and averaged over seven task libraries, generally increases with n/K, while its magnitude also varies with K. This trend indicates that larger ratios are associated with weaker agreement between the transformer and finite-prior Bayes.

The RMSE quantifies the magnitude of the transformer's deviation from finite-prior Bayes. To measure the associated change in retrieval–learning preference, we compare the two predictors' retrieval-minus-learning mean squared errors:

$$\Delta_{\text{pref}} := \left(\text{RMSE}_n^2(f_{\widehat{\boldsymbol{\theta}}}, h) - \text{RMSE}_n^2(f_{\widehat{\boldsymbol{\theta}}}, f_{\text{unif}}^*)\right) - \left(\text{RMSE}_n^2(f_K^*, h) - \text{RMSE}_n^2(f_K^*, f_{\text{unif}}^*)\right), \tag{14}$$

where $h(s_{1:n}) := \boldsymbol{P}_{s_n,1}^{(k^\star)}$ is the retrieval benchmark defined in Theorem 3.3(i). Note that positive Δ_{pref} indicates a relative preference shift toward learning compared with finite-prior Bayes.

As shown in Figure 4b, Δ_{pref} remains near zero at small n/K and becomes predominantly positive at larger ratios. Together with the RMSE results, this suggests that larger n/K is associated with both greater deviations from finite-prior Bayes and a relative preference shift toward learning. These observations are consistent with the differences between the transformer and finite-prior Bayes alignment patterns in Figure 3.

Theoretical intuition. A mismatch between the pretraining and test prompt distributions may help explain the increase in $\mathrm{RMSE}_n(f_{\widehat{\theta}}, f_K^*)$ with n/K in Figure 4a. The pretraining excess risk is averaged under ρ_K, whereas the RMSE is evaluated under ρ_{unif} conditional on regularity. As n/K grows, some sets of regular prompts may receive little probability under ρ_K relative to ρ_{unif}, weakening the control of test errors by pretraining excess risk. Proposition D.2 and its proof make this dependence on prompt coverage explicit. Thus, small pretraining excess risk need not translate into comparably small test RMSE.

Takeaway: When n/K and the pretraining excess risk $\epsilon_{K,n}$ are sufficiently small, pretrained transformers closely follow finite-prior Bayes f_K^* on regular unseen-task prompts. With sufficiently large K, they also approximate uniform-prior Bayes f_{unif}^* and thus exhibit task-learning behavior. Empirically, larger n/K is associated with greater deviations from f_K^* and a relative preference shift toward learning under the tested training protocol.

4 Experimental Results

In this section, we empirically examine our theoretical predictions for retrieval and learning on Markov data and evaluate how closely trained transformers follow the Bayes predictors. Motivated by Theorem 3.3, we define the retrieval and learning errors of a next-token predictor f, respectively, as

$$|f(s_{1:n}) - \boldsymbol{P}_{s_n,1}^{(k^\star)}| \quad \text{and} \quad |f(s_{1:n}) - f_{\mathrm{unif}}^*(s_{1:n})|, \tag{15}$$

where $k^\star$ is defined in Theorem 3.3(i). The former measures agreement with the most likely pretraining task, whereas the latter measures agreement with the uniform-prior Bayes predictor.

4.1 Empirical verification of the Theory on the Bayes Predictor

In this subsection, we first compare the retrieval–learning patterns of finite-prior Bayes and trained transformers, then examine the empirical scaling of the prediction errors bounded in Theorem 3.3.

Data generation. According to Definition 2.1, we sample a task library that contains 256 parameter pairs drawn i.i.d. from $\mathrm{Unif}(0,1)^2$. For each K, the first K pairs define a task family that remains fixed during pretraining. Each sequence is generated by sampling $J \sim \mathrm{Unif}([K])$, drawing s_1 from $\boldsymbol{\mu} = (1/2, 1/2)$, and sampling subsequent tokens according to $\boldsymbol{P}(p^{(J)}, q^{(J)})$. Each training run uses $64{,}000$ freshly sampled sequences of length $T = 513$, with the total training data budget held fixed across K.

Training setup. We use a causal transformer with 4 layers, 2 attention heads per layer, and model and feed-forward dimensions both equal to 128. The model uses learned relative positional embeddings for keys and values, ReLU activations, and post-layer normalization, with a binary vocabulary and a maximum context length of 512. For each task family and each K, we train a separate model using AdamW for solving Problem (2), with a peak learning rate of 10^{-3} and weight decay of 0.01. Training consists of $2{,}000$ optimizer updates with an effective batch size of 32, using a 10% linear warmup followed by cosine learning-rate decay.

Transition between retrieval and learning. We first examine how the number of tasks $K \in \{2, 4, 8, 16, 32, 64, 128, 256\}$ and prompt length $n \in \{8, 16, 32, 64, 128, 256, 512\}$ jointly shape retrieval and learning. For a predictor f, let $d_{\mathrm{retri}}(K, n)$ and $d_{\mathrm{learn}}(K, n)$ denote the empirical means of the retrieval and learning errors defined in (15), averaged over sampled task families and regular in-context prefixes. We define the relative alignment score as

$$s(K, n) := \frac{d_{\mathrm{learn}}(K, n) - d_{\mathrm{retri}}(K, n)}{d_{\mathrm{learn}}(K, n) + d_{\mathrm{retri}}(K, n)} \in [-1, 1]. \tag{16}$$

Here, positive values indicate closer alignment with retrieval, and negative values indicate closer alignment with learning.

For $f = f_K^*$, we visualize the relative alignment score for the finite-prior Bayes predictor in Figure 3a. We observe that the finite-prior Bayes predictor becomes more learning-like as K increases

and more retrieval-like as n increases. The approximately diagonal boundary is consistent with the role of n/K identified by Theorem 3.3. Then, we pre-train transformers on the generated sequences and visualize their score in Figure 3b. We observe that it exhibits a similar pattern at short contexts, but differs at longer contexts: for some library sizes, its alignment shifts back toward learning while the Bayes predictor becomes increasingly retrieval-like.

Scaling of retrieval and learning with n and K. We evaluate the finite-prior Bayes predictor f_K^* along four slices of the (K, n) plane. We fix 64 test sequences, independently of the task families, whose prefixes are $(0.1, 0.1)$-regular at every evaluated prompt length. For each (K, n), we sample $2{,}048$ independent task libraries and reuse them across prompt lengths. For each fixed prefix, we compute the empirical 0.99-quantile of each error in (15) over task libraries, matching the probability space of Theorem 3.3. In Figure 5, solid curves show the median of these quantiles across prefixes, and shaded regions indicate their full range. Dashed curves are empirically calibrated reference curves of the forms $c_1 \exp(-c_2 n/K)$ and $c_3 K^{-1/2}$.

The retrieval error decreases approximately exponentially with n at $K = 4$ and increases with K at $n = 10{,}000$ (see Figures 5a and 5b), consistent with the retrieval bound (9). The finite-to-uniform Bayes error remains approximately constant across the tested prompt lengths at $K = 2{,}000$ and decreases approximately as $K^{-1/2}$ at $n = 64$ (see Figures 5c and 5d). These observations are consistent with the form of the learning bound (10), which has no explicit dependence on n within its stated regime.

4.2 Alignment Between Pretrained Transformers and the Bayes Predictor

In this subsection, we examine how transformer agreement with finite-prior Bayes and the associated preference shift between retrieval and learning vary with n/K. We evaluate all combinations of $K, n \in \{16, 32, 64, 96, 128, 192, 256, 384, 512, 1024\}$, training a separate model on each of seven independently sampled task libraries per K.

Training setup. We use the same depth, number of attention heads, and model dimension as in Section 4.1, but increase the feedforward dimension to 256 and the maximum context length to 1024. The model uses learned absolute positional embeddings, GELU activations, and pre-layer normalization with a final layer normalization. Each model is trained for $30{,}000$ AdamW updates with an effective batch size of 128, a peak learning rate of 6×10^{-4}, and weight decay of 10^{-4}. The architecture and training protocol are fixed across K.

Evaluation on unseen tasks. We sample $1{,}024$ test tasks independently from $\mathrm{Unif}(0,1)^2$, independently of the training libraries, and generate 16 sequences of length 1024 per task. All models use this shared test set, with prefixes of the same sequences evaluated across context lengths. At each n, we compute $\mathrm{RMSE}_n(f_{\widehat{\boldsymbol{\theta}}}, f_K^*)$ and the preference shift Δ_{pref} on $(0.1, 0.1)$-regular prefixes.

We estimate $\epsilon_{K,n}$ by averaging the value in (12) over the length-n prefixes of $4{,}096$ fresh sequences drawn from each training library. These sequences are independent of the training and validation sets. All metrics, including RMSE, are computed separately for each library and then averaged equally across the seven libraries. Results are plotted against n/K, with a separate curve for each K. Shaded bands indicate ± 1 standard error across libraries, conditional on the shared test set and fixed model initialization. We plot $\mathrm{RMSE}_n(f_{\widehat{\boldsymbol{\theta}}}, f_K^*)$ and Δ_{pref} against n/K in Figures 4a and 4b, respectively.

5 Conclusion

In this work, we studied how pretraining task diversity and prompt length jointly shape retrieval and learning in finite mixtures of binary Markov chains. Our finite-sample bounds characterize when the finite-prior Bayes predictor approximates the most likely pretraining task or the uniform-prior Bayes predictor, highlighting the role of n/K. We further showed that, under the stated small-n/K condition, low pretraining excess risk guarantees agreement with finite-prior Bayes on regular unseen-task prompts. Empirically, larger n/K is associated with greater deviations from finite-prior Bayes and a relative preference shift toward learning under the tested training protocol. Future work could extend the analysis to higher-dimensional task families and investigate how training dynamics shape transformer deviations from finite-prior Bayes.

AI USE STATEMENT

In this work, we used generative AI tools to assist with polishing the writing of the manuscript, including improving phrasing and transitions in both the main text and the proofs. All AI-assisted revisions were carefully reviewed and edited by the authors to ensure their correctness and consistency with the intended content. We also used generative AI tools to assist with writing code for experimental verification and plotting, and all resulting code was checked by the authors for correctness. We have not used generative AI tools for helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, clean and reformat dataset, support qualitative and thematic data analysis, and generate synthetic data sets is not applicable to this work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A VIT backbone for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22669–22679, 2023.

Alberto Bietti, Vivien Cabannes, Diane Bouchacourt, Herve Jegou, and Leon Bottou. Birth of a transformer: A memory viewpoint. *Advances in Neural Information Processing Systems*, 36: 1560–1588, 2023.

Tom Brown, Benjamin Mann, Nick Ryder, and others. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020a.

Tom Brown, Benjamin Mann, Nick Ryder, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020b.

Stephanie Chan, Adam Santoro, Andrew Lampinen, Jane Wang, Aaditya Singh, Pierre Richemond, James McClelland, and Felix Hill. Data distributional properties drive emergent in-context learning in transformers. In *Advances in Neural Information Processing Systems*, volume 35, pp. 18878–18891, 2022.

Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International Conference on Machine Learning*, pp. 1691–1703. PMLR, 2020.

Siyu Chen, Heejune Sheen, Tianhao Wang, and Zhuoran Yang. Unveiling induction heads: Provable training dynamics and feature learning in transformers. In *Advances in Neural Information Processing Systems*, volume 37, pp. 66479–66567, 2024.

Eugene Demidenko. Laplace's law of succession estimator and m-statistics. *The American Statistician*, 79(3):311–319, 2025.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pp. 4171–4186, 2019.

Ezra Edelman, Nikolaos Tsilivis, Benjamin Edelman, Eran Malach, and Surbhi Goel. The evolution of statistical induction heads: In-context learning markov chains. *Advances in Neural Information Processing Systems*, 37:64273–64311, 2024.

Chanakya Ekbote, Marco Bondaschi, Nived Rajaraman, Jason D Lee, Michael Gastpar, Ashok Vardhan Makkuva, and Paul Pu Liang. What one cannot, two can: Two-layer transformers provably represent induction heads on any-order markov chains. *arXiv preprint arXiv:2508.07208*, 2025a.

Chanakya Ekbote, Ashok Vardhan Makkuva, Marco Bondaschi, Nived Rajaraman, Michael Gastpar, Jason D. Lee, and Paul Pu Liang. What one cannot, two can: Two-layer transformers provably represent induction heads on any-order markov chains. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b.

Nicholas Geneva and Nicholas Zabaras. Transformers for modeling physical systems. *Neural Networks*, 146:272–289, 2022.

Yifan Hao, Chenlu Ye, Chi Han, and Tong Zhang. Transformers as multi-task learners: Decoupling features in hidden markov models. *arXiv preprint arXiv:2506.01919*, 2025.

Simon Lepage, Jeremie Mary, and David Picard. Markov chain estimation with in-context learning. *arXiv preprint arXiv:2508.03934*, 2025.

Ziqian Lin and Kangwook Lee. Dual operating modes of in-context learning. In *International Conference on Machine Learning*, 2024.

Yue M Lu, Mary Letey, Jacob A Zavatone-Veth, Anindita Maiti, and Cengiz Pehlevan. Asymptotic theory of in-context learning by linear attention. *Proceedings of the National Academy of Sciences*, 122(28):e2502599122, 2025.

Ashok Vardhan Makkuva, Marco Bondaschi, Adway Girish, Alliot Nagle, Martin Jaggi, Hyeji Kim, and Michael Gastpar. Attention with markov: A curious case of single-layer transformers. In *International Conference on Learning Representations*, 2025.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11048–11064, 2022.

Aliakbar Nafar, Kristen Brent Venable, and Parisa Kordjamshidi. Learning vs retrieval: The role of in-context examples in regression with LLMs. *arXiv preprint arXiv:2409.04318*, 2024.

Alex Nguyen and Gautam Reddy. Differential learning kinetics govern the transition from memorization to generalization during in-context learning. In *International Conference on Learning Representations*, 2025.

Eshaan Nichani, Alex Damian, and Jason D Lee. How transformers learn causal structure with gradient descent. In *International Conference on Machine Learning*, pp. 38018–38070, 2024.

Catherine Olsson, Nelson Elhage, Neel Nanda, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Jane Pan, Tianyu Gao, Howard Chen, and Danqi Chen. What in-context learning "learns" in-context: Disentangling task recognition and task learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8298–8319, 2023.

Core Francisco Park, Ekdeep Singh Lubana, and Hidenori Tanaka. Competition dynamics shape algorithmic phases of in-context learning. In *International Conference on Learning Representations*, 2025.

Nived Rajaraman, Marco Bondaschi, Ashok Vardhan Makkuva, Kannan Ramchandran, and Michael Gastpar. Transformers on markov data: Constant depth suffices. *Advances in Neural Information Processing Systems*, 37:137521–137556, 2024a.

Nived Rajaraman, Jiantao Jiao, and Kannan Ramchandran. An analysis of tokenization: Transformers under markov data. *Advances in Neural Information Processing Systems*, 37:62503–62556, 2024b.

Allan Raventós, Mansheej Paul, Feng Chen, and Surya Ganguli. Pretraining task diversity and the emergence of non-bayesian in-context learning for regression. In *Advances in Neural Information Processing Systems*, volume 36, pp. 14228–14246, 2023.

Aditya Varre, Gizem Yüce, and Nicolas Flammarion. Learning in-context n-grams with transformers: Sub-n-grams are near-stationary points. In *International Conference on Machine Learning*, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*, 2022.

Haolin Yang, Hakaze Cho, and Naoya Inoue. Localizing task recognition and task learning in in-context learning via attention head analysis. In *International Conference on Learning Representations*, 2026.

Oğuz Kaan Yüksel and Nicolas Flammarion. On the sample complexity of next-token prediction. In *International Conference on Artificial Intelligence and Statistics*, volume 258, pp. 694–702. PMLR, 2025.

Oussama Zekri, Ambroise Odonnat, Abdelhakim Benechehab, Linus Bleistein, Nicolas Boullé, and Ievgen Redko. Large language models as markov chains. *arXiv preprint arXiv:2410.02724*, 2024.

Ruida Zhou, Chao Tian, and Suhas Diggavi. An information-theoretic approach to understanding transformers' in-context learning of variable-order markov chains. In *International Conference on Artificial Intelligence and Statistics*, 2026.

Appendices

To simplify our development, we introduce some additional assumptions. Let $\mathrm{Bern}(a)$ denote the Bernoulli distribution with mean a and $\mathrm{KL}(\cdot,\cdot)$ denote the Kullback–Leibler divergence. For non-negative quantities a and b, we write $a \lesssim_\xi b$ if there exists a constant $C_\xi > 0$, depending only on ξ, such that $a \le C_\xi b$. We write $a \asymp_\xi b$ if both $a \lesssim_\xi b$ and $b \lesssim_\xi a$ hold. Unless otherwise specified, these constants are independent of all other problem parameters and may differ from one occurrence to another.

A Related Work

Studies of transformers on Markov data. Extensive works have studied transformers on Markov data models as tractable testbeds for understanding ICL. One line of work (Bietti et al., 2023; Chen et al., 2024; Ekbote et al., 2025a; Nichani et al., 2024; Rajaraman et al., 2024a) shows that two-layer single-head transformers can learn Markov chains under suitable architectural constructions. They also provide a training-dynamics analysis, showing that gradient descent can find such solutions under appropriate conditions. Another line of work (Edelman et al., 2024; Makkuva et al., 2025; Varre et al., 2025) characterizes the loss landscape of transformers, showing how data distribution and architecture influence global optima, bad local minima, and transitions from simple unigram predictors to context-dependent bigram predictors. In addition, some works have studied other aspects of transformers on Markov data, including generalization bounds (Yüksel & Flammarion, 2025; Zekri et al., 2024), induction heads (Chen et al., 2024; Edelman et al., 2024), and Bayesian inference (Hao et al., 2025; Xie et al., 2022; Zhou et al., 2026).

Task retrieval and learning in ICL. Min et al. (2022); Pan et al. (2023) demonstrate that the effectiveness of ICL depends on two distinct mechanisms through which transformers exploit in-context information, namely task retrieval and task learning. Later, Lin & Lee (2024) provides the first explanation of task retrieval and task learning in ICL through a linear regression framework. Recently, Nafar et al. (2024) studies ICL in real-world regression settings and argues that its behavior ranges from retrieving pretrained task information to learning from in-context examples, depending on task familiarity and the richness of the provided examples. More recently, Yang et al. (2026) provides a unified mechanistic framework for understanding how LLMs perform ICL by identifying two specialized types of attention heads, namely task-recognition heads and task-learning heads. Moreover, our work is closely related to Lepage et al. (2025); Park et al. (2025), which study ICL in a finite mixture of Markov chains and demonstrate that transformers exhibit retrieval and learning behaviors, with transitions governed by data diversity and prompt length.

B Transformer Architecture

We consider an L-layer causal transformer that maps a binary sequence $s_{1:T} \in \{0,1\}^T$ to layer-wise hidden representations $\{\boldsymbol{x}_n^{(\ell)}\}_{n=1}^T$ for $\ell = 0,\dots,L$. At each position n, the final representation $\boldsymbol{x}_n^{(L)}$ is used to predict the next token s_{n+1}. The L-layer transformer is defined as follows:

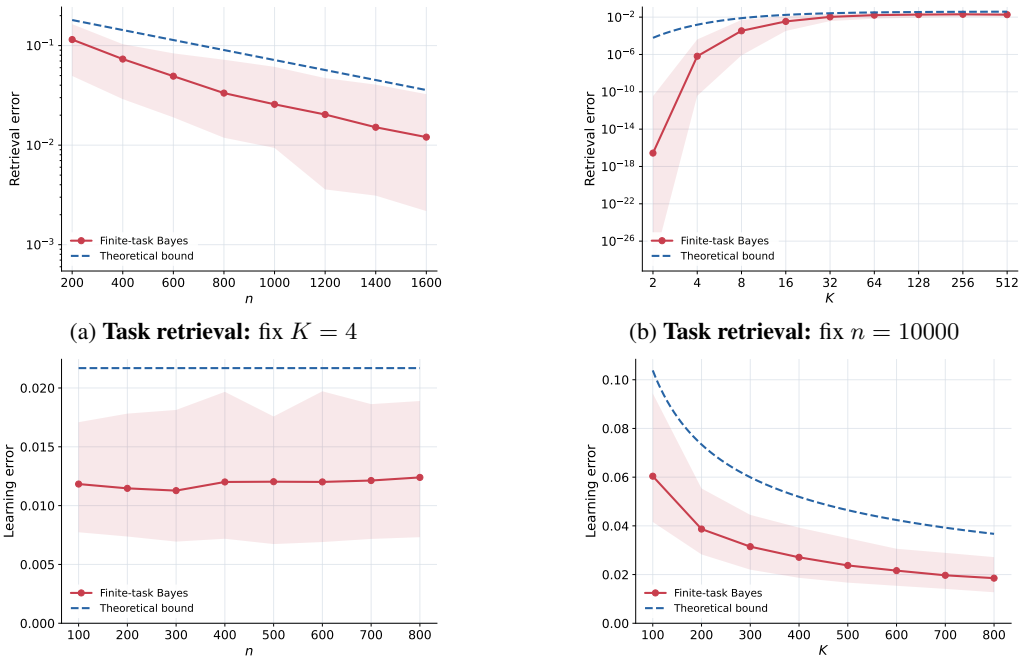
- *Embedding layer.* For each $n \in [T]$, we define

$$\boldsymbol{x}_n^{(0)} = \boldsymbol{\Phi} \boldsymbol{e}_{s_n+1} \in \mathbb{R}^d, \tag{17}$$

  where $\boldsymbol{\Phi} \in \mathbb{R}^{d\times S}$ is the token embedding matrix, $S = 2$, and $\boldsymbol{e}_{s_n+1} \in \mathbb{R}^S$ is the standard basis vector corresponding to token $s_n \in \{0,1\}$.

- *Causal self-attention and feedforward layers.* For each layer $\ell = 0,\dots,L-1$ and position $n \in [T]$, we first compute the causal self-attention update

$$\widetilde{\boldsymbol{x}}_n^{(\ell+1)} = \boldsymbol{x}_n^{(\ell)} + \boldsymbol{W}_O^{(\ell)} \sum_{i=1}^{n} \mathrm{att}_{n,i}^{(\ell)} \left(\boldsymbol{W}_V^{(\ell)} \boldsymbol{x}_i^{(\ell)} + \boldsymbol{p}_{n-i}^{(\ell),V} \right) \in \mathbb{R}^d, \tag{18}$$



(a) **Task retrieval:** fix $K = 4$

(b) **Task retrieval:** fix $n = 10000$

(c) **Task learning:** fix $K = 2000$

(d) **Task learning:** fix $n = 64$

Figure 5: **Task retrieval and learning errors of the finite-prior Bayes predictor.** For each fixed regular prefix, we compute the empirical 0.99-quantile of the error over independently sampled task families. Solid curves show the median of these quantiles across prefixes, and shaded regions show their full range. Dashed curves are separately calibrated upper guides with the theory-predicted forms $c_1 \exp(-c_2 n/K)$ for retrieval and $c_3 K^{-1/2}$ for learning. Please refer to Section 4.1 for experimental details.

where

$$\mathrm{att}_{n,i}^{(\ell)} = \frac{\exp\Big(\Big\langle \boldsymbol{W}_K^{(\ell)} \boldsymbol{x}_i^{(\ell)} + \boldsymbol{p}_{n-i}^{(\ell),K}, \boldsymbol{W}_Q^{(\ell)} \boldsymbol{x}_n^{(\ell)} \Big\rangle\Big)}{\sum_{j=1}^{n} \exp\Big(\Big\langle \boldsymbol{W}_K^{(\ell)} \boldsymbol{x}_j^{(\ell)} + \boldsymbol{p}_{n-j}^{(\ell),K}, \boldsymbol{W}_Q^{(\ell)} \boldsymbol{x}_n^{(\ell)} \Big\rangle\Big)}, \qquad i \in [n]. \tag{19}$$

Here, $\boldsymbol{W}_Q^{(\ell)}, \boldsymbol{W}_K^{(\ell)}, \boldsymbol{W}_V^{(\ell)} \in \mathbb{R}^{d_h^{(\ell)} \times d}$ are the query, key, and value projection matrices, respectively; $\boldsymbol{W}_O^{(\ell)} \in \mathbb{R}^{d \times d_h^{(\ell)}}$ is the output projection matrix; and $\boldsymbol{p}_{n-i}^{(\ell),K}, \boldsymbol{p}_{n-i}^{(\ell),V} \in \mathbb{R}^{d_h^{(\ell)}}$ are the relative positional encodings for the key and value representations.

We then apply a position-wise feedforward update:

$$\boldsymbol{x}_n^{(\ell+1)} = \widetilde{\boldsymbol{x}}_n^{(\ell+1)} + \boldsymbol{W}_2^{(\ell)} \mathrm{ReLU}\Big(\boldsymbol{W}_1^{(\ell)} \widetilde{\boldsymbol{x}}_n^{(\ell+1)}\Big), \tag{20}$$

where $\boldsymbol{W}_1^{(\ell)} \in \mathbb{R}^{d_{\mathrm{ff}} \times d}$ and $\boldsymbol{W}_2^{(\ell)} \in \mathbb{R}^{d \times d_{\mathrm{ff}}}$ are learnable weight matrices, and d_{ff} is the hidden dimension of the feedforward network.

- *Output layer.* At each position n, the final representation is mapped to a scalar prediction through a parameterized readout map:

$$f_{\boldsymbol{\theta}}(s_{1:n}) := \mathbb{P}\left(s_{n+1} = 1 \mid s_{1:n}\right) = \Psi_{\boldsymbol{\theta}}\Big(\boldsymbol{x}_n^{(L)}\Big) \in (0,1), \tag{21}$$

where $\Psi_{\boldsymbol{\theta}} : \mathbb{R}^d \to (0,1)$. The output $f_{\boldsymbol{\theta}}(s_{1:n})$ is interpreted as the model's predicted probability that $s_{n+1} = 1$, and $\boldsymbol{\theta}$ denotes the collection of all model parameters.

C Characterization of the Bayes Predictor

Throughout this section, we consider the binary first-order Markov model and assume that

$$(p, q) \sim \Pi, \quad s_{1:T} \mid (p, q) \sim (\boldsymbol{\mu}, \boldsymbol{P}(p, q)),$$

where Π is a prior distribution on $(p, q) \in (0, 1)^2$, $\boldsymbol{\mu}$ is an initial distribution independent of (p, q), and

$$\boldsymbol{P}(p, q) := \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix}.$$

Condition on the pretraining task family $\{(p^{(k)}, q^{(k)})\}_{k=1}^{K}$ defined in Definition 2.1. For any prefix $s_{1:n} \in \{0, 1\}^n$, let $\mathbb{P}\left(s_{1:n} \mid p^{(k)}, q^{(k)}\right)$ denote the likelihood assigned to the observed prefix $s_{1:n}$ by the k-th Markov chain with a transition matrix $\boldsymbol{P}(p^{(k)}, q^{(k)})$ in the pretraining task family. We define the finite-prior and uniform-prior marginal likelihoods by

$$\rho_K(s_{1:n}) := \mathbb{E}_{(p,q)\sim\Pi_{\text{train}}}\left[\mathbb{P}(s_{1:n} \mid p, q)\right], \ \rho_{\text{unif}}(s_{1:n}) := \mathbb{E}_{(p,q)\sim\text{Unif}(0,1)^2}\left[\mathbb{P}(s_{1:n} \mid p, q)\right].$$

For an (η, γ)-regular prefix with positive marginal probability, the likelihood is maximized at the empirical transition parameters $(\hat{p}, \hat{q})$. We therefore define the maximum-normalized likelihood by

$$w_n(p, q) := \frac{\mathbb{P}(s_{1:n} \mid p, q)}{\mathbb{P}(s_{1:n} \mid \hat{p}, \hat{q})}, \quad Z_n(s_{1:n}) := \mathbb{E}_{(p,q)\sim\text{Unif}(0,1)^2}[w_n(p, q)]. \tag{22}$$

Here, $Z_n(s_{1:n})$ is the data-dependent uniform-prior normalizing constant of the relative likelihood.

C.1 Bayes Predictor with a General Prior

We first establish the following characterization of the Bayes predictor under a general prior distribution Π.

Lemma C.1. *Let Π be a prior distribution on $(p, q) \in (0, 1)^2$. For any $n \in \{1, \ldots, T-1\}$ and any prefix $s_{1:n} \in \{0, 1\}^n$ with positive marginal probability, consider the one-step log-loss*

$$\ell\big(s_{n+1}, f(s_{1:n})\big) = -s_{n+1} \log f(s_{1:n}) - (1 - s_{n+1}) \log\big(1 - f(s_{1:n})\big),$$

where $f(s_{1:n}) \in (0, 1)$.
(i) The unique Bayes estimator of minimizing the posterior expected log-loss, i.e.,

$$f^*(s_{1:n}) \in \arg\min_{x\in(0,1)} \mathbb{E}\left[\ell\big(s_{n+1}, x\big) \mid s_{1:n}\right], \tag{23}$$

is given by

$$f^*(s_{1:n}) = \mathbb{P}(s_{n+1} = 1 \mid s_{1:n}) = \begin{cases} \mathbb{E}[p \mid s_{1:n}], & s_n = 0, \\ \mathbb{E}[1 - q \mid s_{1:n}], & s_n = 1. \end{cases} \tag{24}$$

(ii) It holds that

$$f^*(s_{1:n}) = \frac{\displaystyle\int \boldsymbol{P}_{s_n,1}(p, q)\, \mathbb{P}(s_{1:n} \mid p, q)\, \Pi(dp\, dq)}{\displaystyle\int \mathbb{P}(s_{1:n} \mid p, q)\, \Pi(dp\, dq)}, \tag{25}$$

where $\mathbb{P}(s_{1:n} \mid p, q)$ denotes the likelihood of the observed prefix $s_{1:n}$ under the Markov chain with transition parameters (p, q) and $\Pi(dp\, dq)$ denotes integration with respect to the prior distribution of (p, q).

Proof. (i) For ease of exposition, we define

$$\beta(s_{1:n}) := \mathbb{P}(s_{n+1} = 1 \mid s_{1:n}) = \mathbb{E}[s_{n+1} \mid s_{1:n}].$$

By the tower property and the conditional Markov property, we have

$$\beta(s_{1:n}) = \mathbb{E}[\mathbb{P}(s_{n+1} = 1 \mid s_{1:n}, p, q) \mid s_{1:n}] = \mathbb{E}\Big[[\boldsymbol{P}(p, q)]_{s_n,1} \,\Big|\, s_{1:n}\Big]. \tag{26}$$

Since $p, q \in (0,1)$ almost surely under Π, we have $[\boldsymbol{P}(p,q)]_{s_n,1} \in (0,1)$ almost surely. Hence, $\beta(s_{1:n}) \in (0,1)$. For any $u \in (0,1)$, the definition of the log-loss and the fact that $s_{n+1} \in \{0,1\}$ give

$$\begin{aligned}\mathbb{E}[\ell(s_{n+1}, u) \mid s_{1:n}] &= \mathbb{E}[-s_{n+1} \log u - (1 - s_{n+1}) \log(1-u) \mid s_{1:n}] \\ &= -\beta(s_{1:n}) \log u - \big(1 - \beta(s_{1:n})\big) \log(1-u).\end{aligned}$$

For notational simplicity, we write $\beta := \beta(s_{1:n})$ and define $\varphi(u) := -\beta \log u - (1-\beta)\log(1-u)$ for all $u \in (0,1)$. Differentiating gives

$$\varphi'(u) = -\frac{\beta}{u} + \frac{1-\beta}{1-u} = \frac{u-\beta}{u(1-u)}, \quad \varphi''(u) = \frac{\beta}{u^2} + \frac{1-\beta}{(1-u)^2} > 0, \quad \forall u \in (0,1).$$

Therefore, φ is strictly convex on $(0,1)$, and its unique minimizer is $u = \beta$. It follows that $f^*(s_{1:n}) = \beta(s_{1:n}) = \mathbb{P}(s_{n+1} = 1 \mid s_{1:n})$. Finally, since

$$[\boldsymbol{P}(p,q)]_{s_n,1} = \begin{cases} p, & s_n = 0, \\ 1-q, & s_n = 1, \end{cases}$$

(26) yields (24).

(ii) Conditional on the parameter pair (p,q), the chain is first-order Markov. Therefore, we have

$$\mathbb{P}(s_{n+1} = 1 \mid s_{1:n}, p, q) = \mathbb{P}(s_{n+1} = 1 \mid s_n, p, q) = \boldsymbol{P}_{s_n,1}(p,q).$$

Hence, the tower property implies $f^*(s_{1:n}) = \mathbb{P}(s_{n+1} = 1 \mid s_{1:n}) = \mathbb{E}[\boldsymbol{P}_{s_n,1}(p,q) \mid s_{1:n}]$. Applying Bayes' rule, the posterior distribution of (p,q) given $s_{1:n}$ is

$$\Pi(dp\, dq \mid s_{1:n}) = \frac{\mathbb{P}(s_{1:n} \mid p,q)\, \Pi(dp\, dq)}{\displaystyle\int \mathbb{P}(s_{1:n} \mid p,q)\, \Pi(dp\, dq)}.$$

Substituting this posterior representation into the conditional expectation above yields (25). □

C.2 Proof of Lemma 3.2

Proof of Lemma 3.2. (i) Throughout the proof, we condition on the realized pretraining task family $\{(p^{(k)}, q^{(k)})\}_{k=1}^K$. Let $Z \sim \mathrm{Unif}([K])$ denote the latent task index. Thus, for every $k \in [K]$, we have $\mathbb{P}(Z = k) = 1/K$ and

$$\mathbb{P}(s_{1:n} \mid Z = k) = \mathbb{P}(s_{1:n} \mid p^{(k)}, q^{(k)}). \tag{27}$$

By Bayes' rule, we have

$$\mathbb{P}(Z = k \mid s_{1:n}) = \frac{\mathbb{P}(s_{1:n} \mid Z = k)\mathbb{P}(Z = k)}{\displaystyle\sum_{r=1}^K \mathbb{P}(s_{1:n} \mid Z = r)\mathbb{P}(Z = r)} = \alpha_k(s_{1:n}), \tag{28}$$

where the last equality follows from $\mathbb{P}(Z = k) = 1/K$ for each $k \in [K]$ and the definition of α_k in (7).

Conditioning on $Z = k$, the sequence follows the k-th Markov chain. Therefore, by the conditional Markov property, we have $\mathbb{P}(s_{n+1} = 1 \mid s_{1:n}, Z = k) = \boldsymbol{P}^{(k)}_{s_n,1}$. Using Lemma C.1(i) and the law of total probability, we obtain

$$\begin{aligned} f_K^\star(s_{1:n}) = \mathbb{P}(s_{n+1} = 1 \mid s_{1:n}) &= \sum_{k=1}^K \mathbb{P}(Z = k \mid s_{1:n}) \mathbb{P}(s_{n+1} = 1 \mid s_{1:n}, Z = k) \\ &\overset{(27),(28)}{=} \sum_{k=1}^K \alpha_k(s_{1:n}) \boldsymbol{P}^{(k)}_{s_n,1}. \end{aligned}$$

This proves (7).

(ii) Because the initial distribution $\boldsymbol{\mu}$ is fixed and does not depend on (p, q), the likelihood of the observed prefix $s_{1:n}$ is

$$\mathbb{P}(s_{1:n} \mid p, q) = \mu_{s_1}(1-p)^{N_{00}} p^{N_{01}} q^{N_{10}} (1-q)^{N_{11}}.$$

Under the product-uniform prior $\Pi = \mathrm{Unif}\big((0,1)^2\big)$, the prior density with respect to Lebesgue measure is constant. Hence, Bayes' rule gives

$$\Pi(dp, dq \mid s_{1:n}) \propto (1-p)^{N_{00}} p^{N_{01}} q^{N_{10}} (1-q)^{N_{11}},\ \forall (p,q) \in (0,1)^2.$$

Therefore, conditioned on $s_{1:n}$, the random variables p and q remain independent, with

$$p \mid s_{1:n} \sim \mathrm{Beta}(N_{01}+1, N_{00}+1), \quad q \mid s_{1:n} \sim \mathrm{Beta}(N_{10}+1, N_{11}+1). \tag{29}$$

According to (24) in Lemma C.1, we have

$$f^*_{\mathrm{unif}}(s_{1:n}) = \mathbb{P}(s_{n+1}=1 \mid s_{1:n}) = \begin{cases} \mathbb{E}[p \mid s_{1:n}], & s_n = 0, \\ \mathbb{E}[1-q \mid s_{1:n}], & s_n = 1. \end{cases}$$

A $\mathrm{Beta}(a,b)$ random variable has mean $a/(a+b)$. Therefore, (29) gives

$$\mathbb{E}[p \mid s_{1:n}] = \frac{N_{01}+1}{N_{01}+N_{00}+2}, \quad \mathbb{E}[q \mid s_{1:n}] = \frac{N_{10}+1}{N_{10}+N_{11}+2}.$$

Hence, we have

$$\mathbb{E}[1-q \mid s_{1:n}] = 1 - \mathbb{E}[q \mid s_{1:n}] = \frac{N_{11}+1}{N_{10}+N_{11}+2}.$$

Combining the two cases, together with (5), proves (8). □

Remark C.2. For any prefix $s_{1:n} \in \{0,1\}^n$ with $n \geq 2$, the likelihood of observing $s_{1:n}$ under the k-th Markov chain is

$$\mathbb{P}(s_{1:n} \mid p^{(k)}, q^{(k)}) = \mu_{s_1}(1-p^{(k)})^{N_{00}} (p^{(k)})^{N_{01}} (q^{(k)})^{N_{10}} (1-q^{(k)})^{N_{11}},$$

where $N_{ij} := N_{ij}(s_{1:n})$ is defined in (4). Then, we have

$$\begin{aligned} \log\left(\frac{\mathbb{P}(s_{1:n} \mid p^{(k)}, q^{(k)})}{\mu_{s_1}}\right) &= N_{00}\log(1-p^{(k)}) + N_{01}\log p^{(k)} + N_{10}\log q^{(k)} + N_{11}\log(1-q^{(k)}) \\ &= -N_0 \ell(\hat{p}, p^{(k)}) - N_1 \ell(\hat{q}, q^{(k)}) = -(n-1) D(\hat{\boldsymbol{P}}, \boldsymbol{P}^{(k)}), \end{aligned}$$

where the second equality follows from (5) and (6). This implies

$$\mathbb{P}(s_{1:n} \mid p^{(k)}, q^{(k)}) = \mu_{s_1} \exp\Big(-(n-1) D(\hat{\boldsymbol{P}}, \boldsymbol{P}^{(k)})\Big). \tag{30}$$

We next quantify how well a candidate transition matrix explains the observed prefix $s_{1:n}$. For any candidate transition matrix $\boldsymbol{P} = \boldsymbol{P}(p,q)$, we define its weighted empirical transition CE relative to $\hat{\boldsymbol{P}}$ by

$$D\big(\hat{\boldsymbol{P}}, \boldsymbol{P}\big) := \hat{c}_0\, \ell(\hat{p}, p) + \hat{c}_1\, \ell(\hat{q}, q) = -\frac{1}{n-1} \sum_{t=1}^{n-1} \log \boldsymbol{P}_{s_t, s_{t+1}}, \tag{31}$$

where ℓ is defined in (3). Notably, $D(\hat{\boldsymbol{P}}, \boldsymbol{P})$ is the normalized negative log-likelihood of the observed transitions under $\boldsymbol{P}$, so a smaller value indicates greater compatibility with the observed prefix.

C.3 Proof of Theorem 3.3

To prove this theorem, we first establish some preliminary results. Recall that $Z_n(\cdot)$ is defined in (22).

Lemma C.3. *Fix $\eta, \gamma \in (0, 1/2)$. For every (η,γ)-regular prefix $s_{1:n}$, it holds that*

$$Z_n(s_{1:n}) \asymp_{\eta,\gamma} \frac{1}{n}. \tag{32}$$

Proof. Since the initial-distribution term cancels in the likelihood ratio, it follows from (4) and (22) that

$$w_n(p,q) = \frac{p^{N_{01}}(1-p)^{N_{00}}}{\hat{p}^{N_{01}}(1-\hat{p})^{N_{00}}} \frac{q^{N_{10}}(1-q)^{N_{11}}}{\hat{q}^{N_{10}}(1-\hat{q})^{N_{11}}}.$$

Consequently, we have

$$Z_n(s_{1:n}) = \frac{\mathrm{Beta}(N_{01}+1, N_{00}+1)}{\hat{p}^{N_{01}}(1-\hat{p})^{N_{00}}} \frac{\mathrm{Beta}(N_{10}+1, N_{11}+1)}{\hat{q}^{N_{10}}(1-\hat{q})^{N_{11}}}. \tag{33}$$

Since $\hat{p}, \hat{q} \in [\eta, 1-\eta]$, the standard Stirling bounds give, uniformly over all such prefixes,

$$\begin{aligned}\frac{\mathrm{Beta}(N_{01}+1, N_{00}+1)}{\hat{p}^{N_{01}}(1-\hat{p})^{N_{00}}} &= \frac{N_{01}!\, N_{00}!}{(N_0+1)!} \frac{N_0^{N_0}}{N_{01}^{N_{01}} N_{00}^{N_{00}}} \asymp \frac{1}{N_0+1}\sqrt{\frac{N_{01}N_{00}}{N_0}} \\ &= \frac{\sqrt{N_0\hat{p}(1-\hat{p})}}{N_0+1} \asymp_\eta \frac{1}{\sqrt{N_0}}.\end{aligned}$$

Similarly, we have

$$\frac{\mathrm{Beta}(N_{10}+1, N_{11}+1)}{\hat{q}^{N_{10}}(1-\hat{q})^{N_{11}}} \asymp_\eta \frac{1}{\sqrt{N_1}}.$$

Therefore, $Z_n(s_{1:n}) \asymp_\eta 1/\sqrt{N_0 N_1}$. By Definition 3.1, we have $N_0 = \hat{c}_0(n-1)$ and $N_1 = (1-\hat{c}_0)(n-1)$, where $\hat{c}_0 \in [\gamma, 1-\gamma]$. It follows that

$$\frac{1}{\sqrt{N_0 N_1}} = \frac{1}{(n-1)\sqrt{\hat{c}_0(1-\hat{c}_0)}} \asymp_\gamma \frac{1}{n},$$

which proves (32). □

The estimate in Lemma C.3 controls the total normalized likelihood mass. To analyze whether the finite-prior posterior is dominated by a single library task, we additionally need to understand how the uniform prior distributes mass across small likelihood-energy shells.

Given a prefix $s_{1:n}$, we define the excess empirical transition cross-entropy as follows:

$$\begin{aligned}Y_n(p,q) &:= D\big(\hat{\boldsymbol{P}}, \boldsymbol{P}(p,q)\big) - D\big(\hat{\boldsymbol{P}}, \hat{\boldsymbol{P}}\big) \\ &= \hat{c}_0 \mathrm{KL}(\mathrm{Bern}(\hat{p}) \| \mathrm{Bern}(p)) + \hat{c}_1 \mathrm{KL}(\mathrm{Bern}(\hat{q}) \| \mathrm{Bern}(q)),\end{aligned} \tag{34}$$

where $D(\cdot,\cdot)$ is defined in (31) and $\hat{c}_i$ is defined in (5) for $i \in \{0,1\}$. According to (22), one can verify

$$w_n(p,q) = \exp\left(-(n-1)Y_n(p,q)\right). \tag{35}$$

Moreover, let

$$F_n(t) := \mathbb{P}_{(p,q)\sim\mathrm{Unif}(0,1)^2}\big(Y_n(p,q) \le t\big). \tag{36}$$

Note that Y_n and F_n depend on $s_{1:n}$, and we suppress this dependence for notational simplicity. For the fixed observed prefix $s_{1:n}$, $F_n(t)$ measures the uniform-prior mass of candidate tasks whose excess empirical transition cross-entropy on this prefix is at most t. Now, we prove a lemma that provides uniform control of the likelihood-energy distribution over regular prefixes.

Lemma C.4. *Fix $\eta, \gamma \in (0, 1/2)$ There exist constants $c_1, c_2 > 0$, depending only on (η, γ), such that for every (η, γ)-regular prefix $s_{1:n}$, the associated distribution function F_n satisfies*

$$c_1(v-u) \le F_n(v) - F_n(u) \asymp_{\eta,\gamma} v - u \le c_2(v-u), \quad \forall 0 \le u < v \le \frac{\gamma\eta^2}{2}. \tag{37}$$

Moreover, for every nonnegative measurable function $\phi : [0,\infty) \to [0,\infty]$,

$$\int_0^\infty \phi(t)\, dF_n(t) \le c_2 \int_0^\infty \phi(t)\, dt. \tag{38}$$

Proof. For $r \in [\eta, 1-\eta]$ and $c \in [\gamma, 1-\gamma]$, we define

$$d_r(x) := \mathrm{KL}(\mathrm{Bern}(r) \| \mathrm{Bern}(x)) = r \log\left(\frac{r}{x}\right) + (1-r)\log\left(\frac{1-r}{1-x}\right), \tag{39}$$

$$T_{c,r}(x) := \mathrm{sgn}(x-r)\sqrt{c\, d_r(x)}, \quad \forall x \in (0,1). \tag{40}$$

Since $d_r'(x) = (x-r)/(x(1-x))$ and $d_r(x) \to \infty$ as $x \to 0$ or $x \to 1$, $T_{c,r}$ is a strictly increasing bijection from $(0,1)$ onto $\mathbb{R}$. Write $x = T_{c,r}^{-1}(z)$. For $z \neq 0$, differentiating the both sides of $z^2 = (T_{c,r}(x))^2 = c\, d_r(x)$ w.r.t. z gives the inverse Jacobian

$$J_{c,r}(z) := \frac{dx}{dz} = \frac{2zx(1-x)}{c(x-r)} = \frac{2x(1-x)}{\sqrt{c}} \frac{\sqrt{d_r(x)}}{|x-r|}. \tag{41}$$

At $z = 0$, this extends continuously to

$$J_{c,r}(0) = \lim_{z \to 0} J_{c,r}(z) = \sqrt{\frac{2r(1-r)}{c}}.$$

Pinsker's inequality (see Lemma F.1) and $\log u \leq u - 1$ give, respectively,

$$2(x-r)^2 \leq d_r(x) \leq \frac{(x-r)^2}{x(1-x)}. \tag{42}$$

The upper bound in (42), together with $c \in [\gamma, 1-\gamma]$, implies

$$J_{c,r}(z) \leq \frac{2\sqrt{x(1-x)}}{\sqrt{c}} \leq \frac{1}{\sqrt{c}} \leq \frac{1}{\sqrt{\gamma}} =: C_J, \quad \forall z \in \mathbb{R}, \tag{43}$$

where the second inequality uses $x(1-x) \leq 1/4$ for all x.

Set $\kappa_1 := \gamma\eta^2/2$. If $z^2 = c\, d_r(x) \leq \kappa_1$, the lower bound in (42) yields

$$|x-r| \leq \sqrt{\frac{d_r(x)}{2}} \leq \sqrt{\frac{\kappa_1}{2c}} \leq \sqrt{\frac{\kappa_1}{2\gamma}} = \frac{\eta}{2}.$$

This, together with $r \in [\eta, 1-\eta]$, implies $x \in [\eta/2, 1-\eta/2]$. Using (41) and (42), we obtain

$$J_{c,r}(z) \geq \frac{2\sqrt{2}x(1-x)}{\sqrt{c}} \geq \frac{\sqrt{2}\,\eta(1-\eta/2)}{\sqrt{1-\gamma}} =: c_J, \quad |z| \leq \sqrt{\kappa_1}, \tag{44}$$

where the bound at $z = 0$ follows by continuity.

Apply the preceding transformation to the two transition parameters $z_0 := T_{\hat{c}_0,\hat{p}}(p)$ and $z_1 := T_{\hat{c}_1,\hat{q}}(q)$. According to (39) and (40), we have $z_0^2 = \hat{c}_0 d_{\hat{p}}(p) = \hat{c}_0 \mathrm{KL}(\mathrm{Bern}(\hat{p}) || \mathrm{Bern}(p))$ and $z_1^2 = \hat{c}_1 d_{\hat{q}}(q) = \hat{c}_1 \mathrm{KL}(\mathrm{Bern}(\hat{q}) || \mathrm{Bern}(q))$. By the definition of Y_n, we have $Y_n(p,q) = z_0^2 + z_1^2$. Since (p,q) is uniform on $(0,1)^2$, changing variables gives

$$F_n(v) - F_n(u) = \int_{\{(p,q)\in(0,1)^2 : u < Y_n(p,q) \leq v\}} dpdq = \int_{\{u < z_0^2 + z_1^2 \leq v\}} J_{\hat{c}_0,\hat{p}}(z_0) J_{\hat{c}_1,\hat{q}}(z_1)\, dz_0\, dz_1.$$

For $0 \leq u < v \leq \kappa_1$, both Jacobian factors lie in $[c_J, C_J]$. The integration region is an annulus of area $\pi(v-u)$, so

$$\pi c_J^2 (v-u) \leq F_n(v) - F_n(u) \leq \pi C_J^2 (v-u).$$

Therefore, (37) holds with $c_1 := \pi c_J^2$ and $c_2 := \pi C_J^2 = \pi/\gamma$.

Finally, for every nonnegative measurable function ϕ, the global Jacobian bound and polar coordinates give

$$\begin{aligned}\int_0^\infty \phi(t)\, dF_n(t) &= \int_{\mathbb{R}^2} \phi(z_0^2 + z_1^2) J_{\hat{c}_0,\hat{p}}(z_0) J_{\hat{c}_1,\hat{q}}(z_1)\, dz_0\, dz_1 \\ &\leq C_J^2 \int_{\mathbb{R}^2} \phi(z_0^2 + z_1^2)\, dz_0\, dz_1 = 2\pi C_J^2 \int_0^\infty \phi(\rho^2)\rho\, d\rho = c_2 \int_0^\infty \phi(t)\, dt,\end{aligned}$$

where the last equality uses $t = \rho^2$. This proves (38). All constants depend only on (η, γ), completing the proof. □

Based on the above lemma, we are ready to show the following lemma, which shows that with high probability over the sampled pretraining task family, the task with the largest weight carries nearly all posterior mass.

Lemma C.5. *Fix $\eta, \gamma \in (0, 1/2)$. Draw $(p^{(k)}, q^{(k)}) \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(0,1)^2$ for each $k \in [K]$ and write*

$$w_k := w_n(p^{(k)}, q^{(k)}), \quad \forall k \in [K]. \tag{45}$$

Let $k^\star \in \arg\max\{\mathbb{P}(s_{1:n} \mid p^{(k)}, q^{(k)}) : k \in [K]\}$ with ties broken by choosing the smallest index. For every (η, γ)-regular prefix $s_{1:n}$ and every integer $K \geq 1$, there exist constants $c_3 > 0$ depending only on (η, γ) such that for every $\epsilon \in (0, 1)$,

$$\mathbb{P}\left(1 - \alpha_{k^\star} > \epsilon\right) \leq c_3 K Z_n(s_{1:n}) \left(1 + \log \frac{2}{\epsilon}\right). \tag{46}$$

Proof. If $K = 1$, then $k^\star = 1$, $\alpha_1 = 1$, and $f_1^*(s_{1:n}) = \boldsymbol{P}^{(1)}_{s_n,1}$. Thus both claims hold deterministically. In the remainder of the proof, we assume $K \geq 2$.

Fix an (η, γ)-regular prefix $s_{1:n}$ and $\epsilon \in (0, 1)$. We define $Y_k := Y_n(p^{(k)}, q^{(k)})$ and $M := \min_{k \in [K]} Y_k$, where $Y_n(\cdot, \cdot)$ is defined in (34). The random variables $Y_1, \ldots, Y_K$ are independent with common distribution function F_n. By (38), their common distribution is absolutely continuous. Since $Y_k \geq 0$, this implies $F_n(0) = 0$. Moreover, the minimizing index is unique almost surely. Since $w_k = \exp\left(-(n-1)Y_k\right)$ by (35), this minimizing index is $k^\star$ due to $k^\star \in \arg\max\{\mathbb{P}(s_{1:n} \mid p^{(k)}, q^{(k)}) : k \in [K]\}$ and (22). This, together with (45) and $M := \min_{k \in [K]} Y_k$, yields

$$\alpha_{k^\star} = \frac{w_{k^\star}}{\sum_{r=1}^K w_r} = \frac{\exp(-(n-1)M)}{\exp(-(n-1)M) + \sum_{k \neq k^\star} \exp(-(n-1)Y_k)} = \frac{1}{1+\beta}, \tag{47}$$

where $\beta := \sum_{k \neq k^\star} \exp\left(-(n-1)(Y_k - M)\right)$. Hence, we have

$$1 - \alpha_{k^\star} = \beta/(1+\beta) \leq \beta. \tag{48}$$

It therefore suffices to bound $\mathbb{P}(\beta > \epsilon)$.

For any nonnegative numbers $b_1, \ldots, b_m$, we have

$$\mathbf{1}_{\{\sum_{\ell=1}^m b_\ell > \epsilon\}} \leq \sum_{\ell=1}^m \min\left\{1, \frac{b_\ell}{\epsilon}\right\}. \tag{49}$$

Indeed, if some $b_\ell \geq \epsilon$, the right-hand side is at least one; otherwise, it equals $\epsilon^{-1} \sum_{\ell=1}^m b_\ell$. Conditionally on $(M, k^\star)$, the remaining $(K-1)$ energies have the common conditional law $\frac{\mathbf{1}_{\{t > M\}}\, dF_n(t)}{1 - F_n(M)}$ almost surely. Applying (49) to the terms of β and taking conditional expectations gives, almost surely,

$$\begin{aligned}
\mathbb{P}(\beta > \epsilon \mid M, k^\star) &\leq \mathbb{E}\left[\sum_{k \neq k^\star} \min\left\{1, \frac{\exp(-(n-1)(Y_k - M))}{\epsilon}\right\} \,\middle|\, M, k^\star\right] \\
&= \frac{K-1}{1 - F_n(M)} \int_M^\infty \min\left\{1, \frac{\exp(-(n-1)(t-M))}{\epsilon}\right\} dF_n(t) \\
&\leq \frac{c_2(K-1)}{1 - F_n(M)} \int_M^\infty \min\left\{1, \frac{\exp(-(n-1)(t-M))}{\epsilon}\right\} dt \\
&= \frac{c_2(K-1)}{1 - F_n(M)} \int_0^\infty \min\left\{1, \frac{\exp(-(n-1)u)}{\epsilon}\right\} du,
\end{aligned}$$

where c_2 is the constant in Lemma C.4, and the second inequality follows from (38) in Lemma C.4. Set $u_\epsilon := \log(1/\epsilon)/(n-1)$. The last integral satisfies

$$\begin{aligned}
\int_0^\infty \min\left\{1, \frac{\exp(-(n-1)u)}{\epsilon}\right\} du &= \int_0^{u_\epsilon} 1\, du + \frac{1}{\epsilon} \int_{u_\epsilon}^\infty \exp(-(n-1)u)\, du \\
&= u_\epsilon + \frac{\exp(-(n-1)u_\epsilon)}{(n-1)\epsilon} = \frac{1}{n-1}\left(1 + \log \frac{1}{\epsilon}\right).
\end{aligned}$$

Consequently, it holds almost surely that

$$\mathbb{P}(\beta > \epsilon \mid M, k^\star) \le \frac{c_2(K-1)}{(n-1)(1-F_n(M))}\left(1+\log\frac{1}{\epsilon}\right), \tag{50}$$

To remove the conditioning, we compute the expected reciprocal of the conditional normalizing factor. By independence, we have $\mathbb{P}(M > y) = (1-F_n(y))^K$. Thus, writing F_M for the distribution function of M, we have $dF_M(y) = K(1-F_n(y))^{K-1}\,dF_n(y)$. Since $K \ge 2$, we obtain

$$\begin{aligned}\mathbb{E}\left[\frac{1}{1-F_n(M)}\right] &= K\int_0^\infty (1-F_n(y))^{K-2}\,dF_n(y)\\ &= K\int_0^1 (1-u)^{K-2}\,du = \frac{K}{K-1},\end{aligned} \tag{51}$$

where the first equality follows from the expression for $dF_M(y)$, and the second follows by the substitution $u = F_n(y)$, using the absolute continuity of F_n, $F_n(0) = 0$, and $\lim_{y\to\infty} F_n(y) = 1$. Integrating (50) over the joint distribution of $(M, k^\star)$ and applying (51), we conclude that

$$\begin{aligned}\mathbb{P}(1-\alpha_{k^\star} > \epsilon) &\le \mathbb{P}(\beta > \epsilon) = \mathbb{E}[\mathbb{P}(\beta > \epsilon \mid M, k^\star)]\\ &\le \frac{c_2(K-1)}{n-1}\left(1+\log\frac{1}{\epsilon}\right)\mathbb{E}\left[\frac{1}{1-F_n(M)}\right] = \frac{c_2 K}{n-1}\left(1+\log\frac{1}{\epsilon}\right).\end{aligned}$$

where the first inequality uses (48). By Lemma C.3, there exists $c_Z > 0$, depending only on (η, γ), such that $Z_n(s_{1:n}) \ge c_Z/(n-1)$. Therefore, we have $K/(n-1) \le KZ_n(s_{1:n})/c_Z$. Taking $c_3 := c_2/c_Z$ and using $\log(1/\epsilon) \le \log(2/\epsilon)$ gives (46). □

The following lemma bounds the difference between the Bayes predictor based on K independently sampled transition matrices and that under the continuous uniform prior, for any fixed prefix.

Lemma C.6. *Fix a prefix $s_{1:n}$ and draw $\{(p^{(k)}, q^{(k)})\}_{k=1}^K$ independently from $\mathrm{Unif}(0,1)^2$. For every $\delta \in (0,1)$, if*

$$KZ_n(s_{1:n}) \ge 8\log(3/\delta), \tag{52}$$

then, with probability at least $1-\delta$, the following inequalities hold simultaneously:

$$|f_K^*(s_{1:n}) - f_{\mathrm{unif}}^*(s_{1:n})| \le 2\sqrt{\frac{\log(3/\delta)}{KZ_n(s_{1:n})(2N_{s_n}+3)}}, \tag{53}$$

$$\rho_K(s_{1:n}) \ge \frac{1}{2}\rho_{\mathrm{unif}}(s_{1:n}). \tag{54}$$

Proof. Unless otherwise specified, probabilities and expectations are taken over the independent draws of the transition parameters. For ease of exposition, we write $Z := Z_n(s_{1:n})$ and $w_k := w_n(p^{(k)}, q^{(k)})$ for all $k \in [K]$. We define

$$H_k := w_k\left(\boldsymbol{P}^{(k)}_{s_n,1} - f_{\mathrm{unif}}^*(s_{1:n})\right).$$

By (25) in Lemma C.1, we have

$$\begin{aligned}f_{\mathrm{unif}}^*(s_{1:n}) &= \frac{\displaystyle\int_{(0,1)^2} \boldsymbol{P}(p,q)_{s_n,1}\mathbb{P}(s_{1:n} \mid p,q)\,dp\,dq}{\displaystyle\int_{(0,1)^2}\mathbb{P}(s_{1:n} \mid p,q)\,dp\,dq}\\ &= \frac{\displaystyle\int_{(0,1)^2} w_n(p,q)\boldsymbol{P}(p,q)_{s_n,1}\,dp\,dq}{\displaystyle\int_{(0,1)^2} w_n(p,q)\,dp\,dq} = \frac{\mathbb{E}[w_k\boldsymbol{P}^{(k)}_{s_n,1}]}{Z}.\end{aligned}$$

Using this and $\mathbb{E}[w_k] = Z$, we have $\mathbb{E}[H_k] = 0$. Moreover, Lemma 3.2 gives $f_K^*(s_{1:n}) = \sum_{k=1}^K w_k \boldsymbol{P}_{s_n,1}^{(k)} / \left(\sum_{k=1}^K w_k\right)$. It follows that

$$f_K^*(s_{1:n}) - f_{\text{unif}}^*(s_{1:n}) = \frac{\sum_{k=1}^K H_k}{\sum_{k=1}^K w_k}. \tag{55}$$

We next establish second-moment and range bounds for H_k. By (8) in Lemma 3.2, we have

$$f_{\text{unif}}^*(s_{1:n}) = \frac{N_{s_n,1}+1}{N_{s_n}+2}.$$

Define the probability density

$$\widetilde{\pi}(p,q) := \frac{w_n(p,q)^2}{\mathbb{E}[w_k^2]} = \frac{p^{2N_{01}}(1-p)^{2N_{00}}}{\text{Beta}(2N_{01}+1, 2N_{00}+1)} \frac{q^{2N_{10}}(1-q)^{2N_{11}}}{\text{Beta}(2N_{10}+1, 2N_{11}+1)}. \tag{56}$$

Indeed, the constants introduced by the maximum normalization cancel after normalization. This identity also holds when $N_i = 0$, in which case the corresponding factor is the uniform density on $(0,1)$. Thus, under $\widetilde{\pi}$,

$$p \sim \text{BetaDist}(2N_{01}+1, 2N_{00}+1), \qquad q \sim \text{BetaDist}(2N_{10}+1, 2N_{11}+1),$$

independently.

Since $\boldsymbol{P}(p,q)_{s_n,1} = p$ when $s_n = 0$, while $\boldsymbol{P}(p,q)_{s_n,1} = 1-q$ when $s_n = 1$, in both cases

$$\boldsymbol{P}(p,q)_{s_n,1} \sim \text{BetaDist}\left(2N_{s_n,1}+1,\ 2N_{s_n} - 2N_{s_n,1} + 1\right)$$

under $\widetilde{\pi}$. Therefore,

$$\begin{aligned}
\mathbb{E}[H_k^2] &= \mathbb{E}[w_k^2]\, \mathbb{E}_{\widetilde{\pi}}\left[\left(\boldsymbol{P}(p,q)_{s_n,1} - f_{\text{unif}}^*(s_{1:n})\right)^2\right] \\
&= \mathbb{E}[w_k^2]\left[\frac{(2N_{s_n,1}+1)(2N_{s_n,1}+2)}{(2N_{s_n}+2)(2N_{s_n}+3)} - \frac{2(2N_{s_n,1}+1) f_{\text{unif}}^*(s_{1:n})}{2N_{s_n}+2} + \left(f_{\text{unif}}^*(s_{1:n})\right)^2\right] \\
&= \mathbb{E}[w_k^2]\, \frac{f_{\text{unif}}^*(s_{1:n})\left(1 - f_{\text{unif}}^*(s_{1:n})\right)}{2N_{s_n}+3} \le \frac{Z}{4(2N_{s_n}+3)}.
\end{aligned} \tag{57}$$

Here, the equality uses $f_{\text{unif}}^*(s_{1:n}) = (N_{s_n,1}+1)/(N_{s_n}+2)$, and the inequality follows from $0 \le w_k \le 1$, $\mathbb{E}[w_k] = Z$, and $f_{\text{unif}}^*(s_{1:n})\left(1 - f_{\text{unif}}^*(s_{1:n})\right) \le 1/4$.

For the range bound, consider first $N_{s_n} > 0$. The normalized likelihood factor associated with the other transition coordinate is at most one. Hence,

$$|H_k| \le \max_{x \in [0,1]} \frac{x^{N_{s_n,1}}(1-x)^{N_{s_n} - N_{s_n,1}}}{\left(\frac{N_{s_n,1}}{N_{s_n}}\right)^{N_{s_n,1}} \left(1 - \frac{N_{s_n,1}}{N_{s_n}}\right)^{N_{s_n} - N_{s_n,1}}} \left| x - f_{\text{unif}}^*(s_{1:n}) \right|.$$

Here, x represents the relevant transition probability $\boldsymbol{P}(p,q)_{s_n,1}$. Under the convention $0^0 = 1$, the expression extends continuously to $[0,1]$ and therefore attains its maximum. Consider an interior maximizer. If $x = f_{\text{unif}}^*(s_{1:n})$, the desired bound is immediate. Otherwise, the maximand is positive, and its first-order condition is

$$0 = \frac{N_{s_n,1}}{x} - \frac{N_{s_n} - N_{s_n,1}}{1-x} + \frac{1}{x - f_{\text{unif}}^*(s_{1:n})}.$$

Multiplying by $x(1-x)(x - f_{\text{unif}}^*(s_{1:n}))$ and using $N_{s_n,1} = (N_{s_n}+2) f_{\text{unif}}^*(s_{1:n}) - 1$ by (8) gives

$$0 = f_{\text{unif}}^*(s_{1:n})\left(1 - f_{\text{unif}}^*(s_{1:n})\right) - (N_{s_n}+1)\left(x - f_{\text{unif}}^*(s_{1:n})\right)^2.$$

Consequently, we obtain

$$\left(x - f_{\text{unif}}^*(s_{1:n})\right)^2 = \frac{f_{\text{unif}}^*(s_{1:n})\left(1 - f_{\text{unif}}^*(s_{1:n})\right)}{N_{s_n}+1}.$$

Since the normalized likelihood factor is at most one, every interior maximum is bounded by $\sqrt{\frac{f^*_{\text{unif}}(s_{1:n})\left(1-f^*_{\text{unif}}(s_{1:n})\right)}{N_{s_n}+1}}$. It remains to check the endpoints. If $0 < N_{s_n,1} < N_{s_n}$, both endpoint values vanish. If $N_{s_n,1} \in \{0, N_{s_n}\}$, direct substitution shows that the only nonzero endpoint value is $1/(N_{s_n}+2)$, which equals $\sqrt{\frac{f^*_{\text{unif}}(s_{1:n})\left(1-f^*_{\text{unif}}(s_{1:n})\right)}{N_{s_n}+1}}$.

If $N_{s_n} = 0$, the transition coordinate determining the next-step prediction does not appear in the likelihood. In this case, we have $f^*_{\text{unif}}(s_{1:n}) = 1/2$, and since $0 \le w_k \le 1$,

$$|H_k| = w_k \left| \boldsymbol{P}^{(k)}_{s_n,1} - \frac{1}{2} \right| \le \frac{1}{2}.$$

Thus, in all cases,

$$|H_k| \le \sqrt{\frac{f^*_{\text{unif}}(s_{1:n})\big(1 - f^*_{\text{unif}}(s_{1:n})\big)}{N_{s_n}+1}} \le \frac{1}{2\sqrt{N_{s_n}+1}}. \tag{58}$$

Now set $u := \log(3/\delta)$. Since the weights are independent, take values in $[0,1]$, and have mean Z, the multiplicative Chernoff bound gives

$$\mathbb{P}\left(\frac{1}{K}\sum_{k=1}^{K} w_k < \frac{Z}{2} \right) \le \exp(-KZ/8) \le \exp(-u),$$

where the last inequality uses $KZ \ge 8u$. Moreover, Bernstein's inequality (see, e.g., (Vershynin, 2018, Theorem 2.8.4)), together with (57) and (58), implies that, with probability at least $1 - 2\exp(-u)$,

$$\left| \frac{1}{K}\sum_{k=1}^{K} H_k \right| \le \sqrt{\frac{Zu}{2K(2N_{s_n}+3)}} + \frac{u}{3K\sqrt{N_{s_n}+1}}.$$

By a union bound, both estimates hold simultaneously with probability at least $1 - 3\exp(-u) = 1 - \delta$. On this event, $\sum_{k=1}^{K} w_k \ge KZ/2$. Hence, (55) gives

$$\begin{aligned} |f^*_K(s_{1:n}) - f^*_{\text{unif}}(s_{1:n})| &\le \sqrt{\frac{2u}{KZ(2N_{s_n}+3)}} + \frac{2u}{3KZ\sqrt{N_{s_n}+1}} \\ &= \sqrt{\frac{u}{KZ(2N_{s_n}+3)}} \left(\sqrt{2} + \frac{2}{3}\sqrt{\frac{u(2N_{s_n}+3)}{KZ(N_{s_n}+1)}} \right) \le 2\sqrt{\frac{u}{KZ(2N_{s_n}+3)}}, \end{aligned}$$

where the last inequality follows from $u/(KZ) \le 1/8$, $(2N_{s_n}+3)/(N_{s_n}+1) \le 3$, and $\sqrt{2} + \sqrt{6}/6 < 2$. Substituting $Z = Z_n(s_{1:n})$ and $u = \log(3/\delta)$ proves (53). Moreover, by the definitions of w_k and Z, on the same event,

$$\frac{\rho_K(s_{1:n})}{\rho_{\text{unif}}(s_{1:n})} = \frac{1}{KZ}\sum_{k=1}^{K} w_k \ge \frac{1}{2},$$

which proves (54). Thus, both inequalities hold simultaneously with probability at least $1 - \delta$. □

Armed with the above Lemma C.3-C.6, we are ready to prove Theorem 3.3.

Proof of Theorem 3.3. Fix an (η, γ)-regular prefix $s_{1:n}$ and an integer $K \ge 1$. By Lemma C.3, there exist constants $c_Z, C_Z > 0$, depending only on (η, γ), such that

$$\frac{c_Z}{n} \le Z_n(s_{1:n}) \le \frac{C_Z}{n}. \tag{59}$$

(i) According to Lemma 3.2, we have $f_K^*(s_{1:n}) = \sum_{k=1}^K \alpha_k \boldsymbol{P}_{s_n,1}^{(k)}$, where α_k is defined in (7). Since $\sum_{k=1}^K \alpha_k = 1$, we have

$$\begin{aligned}\left|f_K^*(s_{1:n}) - \boldsymbol{P}_{s_n,1}^{(k^\star)}\right| &= \left|\sum_{k\neq k^\star} \alpha_k \left(\boldsymbol{P}_{s_n,1}^{(k)} - \boldsymbol{P}_{s_n,1}^{(k^\star)}\right)\right| \\ &\leq \sum_{k\neq k^\star} \alpha_k \left|\boldsymbol{P}_{s_n,1}^{(k)} - \boldsymbol{P}_{s_n,1}^{(k^\star)}\right| \leq \sum_{k\neq k^\star} \alpha_k = 1 - \alpha_{k^\star},\end{aligned}$$

where the last inequality uses $\boldsymbol{P}_{s_n,1}^{(k)} \in [0,1]$ for every k.

Set

$$\epsilon := 2e \exp\left(-\frac{\delta n}{\kappa_1 K}\right).$$

If $\epsilon \geq 1$, (9) holds deterministically by the preceding bound. We therefore assume that $\epsilon \in (0,1)$. Applying Lemma C.5 and (59), we obtain

$$\begin{aligned}\mathbb{P}\left(1 - \alpha_{k^\star} > \epsilon\right) &\leq c_3 K Z_n(s_{1:n}) \left(1 + \log\frac{2}{\epsilon}\right) \\ &\leq \frac{c_3 C_Z K}{n}\left(1 + \log\frac{2}{\epsilon}\right) = \frac{\kappa_1 K}{n}\frac{\delta n}{\kappa_1 K} = \delta,\end{aligned}$$

where $\kappa_1 := c_3 C_Z$ and the last equality uses $1 + \log\frac{2}{\epsilon} = \frac{\delta n}{\kappa_1 K}$. Thus, with probability at least $1-\delta$,

$$1 - \alpha_{k^\star} \leq \epsilon = 2e \exp\left(-\frac{\delta n}{\kappa_1 K}\right). \tag{60}$$

Combining this with the deterministic prediction bound above proves (9).

(ii) By $\frac{n}{K} \leq \frac{c}{\log(3/\delta)}$, (59), and the choice $c := c_Z/8$, we have

$$K Z_n(s_{1:n}) \geq \frac{c_Z K}{n} \geq \frac{c_Z}{c}\log(3/\delta) = 8\log(3/\delta).$$

This, together with Lemma C.6, yields that with probability at least $1-\delta$,

$$|f_K^*(s_{1:n}) - f_{\text{unif}}^*(s_{1:n})| \leq 2\sqrt{\frac{\log(3/\delta)}{K Z_n(s_{1:n})(2N_{s_n}+3)}} \leq 2\sqrt{\frac{\log(3/\delta)n}{c_Z(2N_{s_n}+3)K}}, \tag{61}$$

where the second inequality uses (59). According to Definition 3.1, we have $N_0 = \hat{c}_0(n-1)$, $N_1 = (1-\hat{c}_0)(n-1)$, and $\hat{c}_0 \in [\gamma, 1-\gamma]$. Then, we have $N_{s_n} \geq \gamma(n-1)$. This implies $2N_{s_n} + 3 \geq 2\gamma(n-1) + 3 = 2\gamma n + (3 - 2\gamma) \geq 2\gamma n$. Substituting this bound into (61) yields

$$|f_K^*(s_{1:n}) - f_{\text{unif}}^*(s_{1:n})| \leq \sqrt{\frac{2}{c_Z\gamma}}\sqrt{\frac{\log(3/\delta)}{K}}.$$

Finally, setting $\kappa_2 := \sqrt{\frac{2}{c_Z\gamma}}$ completes the proof.

□

D Proofs in Section 3.3

In this section, we provide the proof for Theorem 3.4. Towards this goal, we first establish some supporting results in Section D.1 and Section D.2, and then prove Theorem 3.4 in Section D.3.

D.1 Excess Risk and Prediction Error

We relate pretraining excess risk to prediction error relative to the finite-prior Bayes predictor. Fix the realized pretraining task family and trained predictor. Under the finite training prior, the conditional next-token distribution is $\text{Bern}(f_K^*(s_{1:n}))$. Consequently, the excess risk defined in (12) satisfies

$$\epsilon_{K,n} = \mathbb{E}_{s_{1:n}\sim\rho_K}\left[\text{KL}\left(\text{Bern}\big(f_K^*(s_{1:n})\big) \,\middle\|\, \text{Bern}\big(f_{\widehat{\boldsymbol{\theta}}}(s_{1:n})\big)\right)\right]. \tag{62}$$

Defining $\epsilon_{K,m}$ analogously at each training context length, it follows from (2) that the population excess risk is

$$\mathcal{L}_K(f_{\widehat{\boldsymbol{\theta}}}) - \mathcal{L}_K(f_K^*) = \frac{1}{T-\hat{T}} \sum_{m=\hat{T}}^{T-1} \epsilon_{K,m}. \tag{63}$$

Lemma D.1. *Fix the realized pretraining task family $\{(p^{(k)}, q^{(k)})\}_{k=1}^K$. For every $f \in \mathcal{F}$, every $n \in \{\hat{T}, \dots, T-1\}$, and every prefix $s_{1:n} \in \{0,1\}^n$ with $\rho_K(s_{1:n}) > 0$, we have*

$$|f(s_{1:n}) - f_K^*(s_{1:n})| \le \sqrt{\frac{(T-\hat{T})\big(\mathcal{L}_K(f) - \mathcal{L}_K(f_K^*)\big)}{2\rho_K(s_{1:n})}}. \tag{64}$$

Proof. Fix $f \in \mathcal{F}$ and condition on the realized pretraining task family $\{(p^{(k)}, q^{(k)})\}_{k=1}^K$. By Lemma 3.2, the conditional next-token distribution under the finite training prior is $s_{m+1} \mid s_{1:m} = u \sim \mathrm{Bern}\big(f_K^*(u)\big)$. Hence, at each prefix u, the conditional excess cross-entropy equals $\mathrm{KL}\left(\mathrm{Bern}\big(f_K^*(u)\big) \,\middle\|\, \mathrm{Bern}\big(f(u)\big)\right)$. Averaging over prefixes and training context lengths gives

$$\mathcal{L}_K(f) - \mathcal{L}_K(f_K^*) = \frac{1}{T-\hat{T}} \sum_{m=\hat{T}}^{T-1} \sum_{u \in \{0,1\}^m} \rho_K(u)\ \mathrm{KL}\left(\mathrm{Bern}\big(f_K^*(u)\big) \,\middle\|\, \mathrm{Bern}\big(f(u)\big)\right). \tag{65}$$

Every summand is nonnegative. For any $n \in \{\hat{T}, \dots, T-1\}$ and prefix $s_{1:n}$ with $\rho_K(s_{1:n}) > 0$, retaining its contribution to the sum yields

$$\frac{\rho_K(s_{1:n})}{T-\hat{T}}\, \mathrm{KL}\left(\mathrm{Bern}\big(f_K^*(s_{1:n})\big) \,\middle\|\, \mathrm{Bern}\big(f(s_{1:n})\big)\right) \le \mathcal{L}_K(f) - \mathcal{L}_K(f_K^*). \tag{66}$$

Pinsker's inequality for Bernoulli distributions gives

$$\mathrm{KL}\left(\mathrm{Bern}\big(f_K^*(s_{1:n})\big) \,\middle\|\, \mathrm{Bern}\big(f(s_{1:n})\big)\right) \ge 2\,|f(s_{1:n}) - f_K^*(s_{1:n})|^2.$$

Combining the last two inequalities gives

$$|f(s_{1:n}) - f_K^*(s_{1:n})|^2 \le \frac{(T-\hat{T})\big(\mathcal{L}_K(f) - \mathcal{L}_K(f_K^*)\big)}{2\rho_K(s_{1:n})}.$$

Taking square roots proves (64). □

D.2 Controlling Transformer–Bayes Error on Unseen Tasks

The pretraining excess risk $\epsilon_{K,n}$ is measured under the training prefix distribution ρ_K. The transformer–Bayes prediction error, however, is evaluated under the test distribution ρ_{unif}, conditional on the regularity event $\mathcal{E}_n$. To relate these distributions, define the coverage ratio

$$R_K(s_{1:n}) := \frac{\rho_{\mathrm{unif}}(s_{1:n})}{\rho_K(s_{1:n})}. \tag{67}$$

A large value of R_K means that a prompt receives little probability under the training distribution relative to its probability under the test distribution. For fixed $\eta, \gamma \in (0, 1/2)$ and each integer $n \ge 2$, let $\mathcal{E}_n$ denote the event that the test prefix is (η,γ)-regular, and write

$$r_n := \mathbb{P}_{s_{1:n} \sim \rho_{\mathrm{unif}}}(\mathcal{E}_n). \tag{68}$$

The following proposition controls the prediction error using the achieved excess risk and the probability, among regular test prompts, that R_K exceeds a chosen threshold.

Proposition D.2. *Fix the realized pretraining task family $\{(p^{(k)}, q^{(k)})\}_{k=1}^K$ and a trained predictor $f_{\widehat{\boldsymbol{\theta}}} \in \mathcal{F}$. Fix $\eta, \gamma \in (0, 1/2)$ and a context length $n \in \{\hat{T}, \dots, T-1\}$ with $n \ge 2$ and $r_n > 0$. With $\epsilon_{K,n}$ and R_K defined in (12) and (67), respectively, for every $a > 0$, we have*

$$\mathrm{RMSE}_n^2(f_{\widehat{\boldsymbol{\theta}}}, f_K^*) \le \frac{a\epsilon_{K,n}}{2r_n} + \mathbb{P}_{s_{1:n} \sim \rho_{\mathrm{unif}}}\left(R_K(s_{1:n}) > a \,\middle|\, \mathcal{E}_n\right). \tag{69}$$

Proof of Proposition D.2. Condition on the realized pretraining task family and trained predictor. For brevity, write

$$s = s_{1:n}, \qquad e(s) := f_{\widehat{\boldsymbol{\theta}}}(s) - f_K^*(s).$$

By (62) and the Bernoulli KL lower bound in Lemma F.1, we have

$$\mathbb{E}_{s\sim\rho_K}[e(s)^2] \le \frac{\epsilon_{K,n}}{2}.$$

Using the definitions of R_K and r_n, we obtain

$$\begin{aligned}\mathbb{E}_{s\sim\rho_{\mathrm{unif}}}\left[\frac{e(s)^2}{R_K(s)}\,\middle|\,\mathcal{E}_n\right] &= \frac{1}{r_n}\sum_{s\in\mathcal{E}_n}\rho_K(s)e(s)^2 \\ &\le \frac{1}{r_n}\mathbb{E}_{s\sim\rho_K}[e(s)^2] \le \frac{\epsilon_{K,n}}{2r_n}.\end{aligned}$$

Fix $a > 0$. On $\{R_K(s) \le a\}$, we have $e(s)^2 \le ae(s)^2/R_K(s)$. On $\{R_K(s) > a\}$, we use $e(s)^2 \le 1$, since both predictions lie in $(0,1)$. Splitting the conditional expectation over these two events therefore gives

$$\begin{aligned}\mathrm{RMSE}_n^2(f_{\widehat{\boldsymbol{\theta}}}, f_K^*) &= \mathbb{E}_{s\sim\rho_{\mathrm{unif}}}\left[e(s)^2 \mid \mathcal{E}_n\right] \\ &\le a\,\mathbb{E}_{s\sim\rho_{\mathrm{unif}}}\left[\frac{e(s)^2}{R_K(s)}\,\middle|\,\mathcal{E}_n\right] + \mathbb{P}_{s\sim\rho_{\mathrm{unif}}}\left(R_K(s) > a \mid \mathcal{E}_n\right) \\ &\le \frac{a\epsilon_{K,n}}{2r_n} + \mathbb{P}_{s\sim\rho_{\mathrm{unif}}}\left(R_K(s) > a \mid \mathcal{E}_n\right).\end{aligned}$$

□

D.3 Proof of Theorem 3.4

We first establish a uniform positive lower bound on the probability of regular test prompts over all context lengths for which this probability is positive.

Lemma D.3. *Fix $\eta, \gamma \in (0, 1/2)$, and let r_n be defined in equation 68. Under the uniform task prior and a fixed initial-state distribution $\boldsymbol{\mu}$ independent of the task,*

$$\lim_{n\to\infty} r_n = r_{\eta,\gamma} := (1-2\eta)^2 - \frac{(\gamma-\eta)_+^2}{\gamma(1-\gamma)} > 0, \tag{70}$$

where $x_+ := \max\{x, 0\}$. Consequently, there exists a constant $c_{\mathrm{reg}} > 0$, depending only on (η, γ), such that

$$r_n \ge c_{\mathrm{reg}} \qquad \textit{for every } n \ge 2 \textit{ with } r_n > 0. \tag{71}$$

Proof. For each $(p, q) \in (0,1)^2$, the corresponding Markov chain is irreducible and aperiodic, with stationary probability $q/(p+q)$ for state 0. The ergodic theorem therefore gives, almost surely,

$$\frac{N_0}{n-1} \longrightarrow \frac{q}{p+q}, \qquad \widehat{p} \longrightarrow p, \qquad \widehat{q} \longrightarrow q.$$

Define

$$A_{\eta,\gamma} := \left\{(p,q) \in [\eta, 1-\eta]^2 : \gamma \le \frac{q}{p+q} \le 1-\gamma\right\}.$$

Except on the boundary of $A_{\eta,\gamma}$, which has Lebesgue measure zero, these limits imply

$$\mathbb{P}(\mathcal{E}_n \mid p, q) \longrightarrow \mathbf{1}_{A_{\eta,\gamma}}(p,q).$$

By dominated convergence,

$$r_n = \int_{(0,1)^2} \mathbb{P}(\mathcal{E}_n \mid p, q)\, dp\, dq \longrightarrow \mathrm{Leb}(A_{\eta,\gamma}).$$

If $\gamma \le \eta$, the occupancy constraint holds throughout $[\eta, 1-\eta]^2$, so $\mathrm{Leb}(A_{\eta,\gamma}) = (1-2\eta)^2$. If $\gamma > \eta$, the two excluded regions are congruent triangles, each with area $\frac{(\gamma-\eta)^2}{2\gamma(1-\gamma)}$. Subtracting their

areas proves (70). The limiting area is strictly positive because $A_{\eta,\gamma}$ contains a neighborhood of $(1/2, 1/2)$.

Finally, exchanging the two states maps (p, q) to (q, p) and preserves both the uniform prior and regularity. Hence r_n is the same for either initial state and is independent of the task-independent initial-state distribution. Since $r_n \to r_{\eta,\gamma} > 0$, there exists an integer $N \geq 2$, depending only on (η, γ), such that $r_n \geq r_{\eta,\gamma}/2$ for all $n \geq N$. Set

$$c_{\mathrm{reg}} := \min\left(\left\{\frac{r_{\eta,\gamma}}{2}\right\} \cup \{r_m : 2 \leq m < N,\ r_m > 0\}\right).$$

This is the minimum of a finite nonempty set of positive numbers and depends only on (η, γ). Thus $c_{\mathrm{reg}} > 0$, and (71) follows. □

Proof of Theorem 3.4. Fix the context length $n \geq 2$ with $r_n > 0$ and let $u := \log(12n^2/\delta)$. By Lemma D.3, there exists a constant $c_{\mathrm{reg}} > 0$, depending only on (η, γ), such that $r_n \geq c_{\mathrm{reg}}$. By Lemma C.3, there exists a constant $c_Z > 0$, depending only on (η, γ), such that $Z_n(s_{1:n}) \geq c_Z/n$ for every (η, γ)-regular prefix of length n. Set

$$c_1 := 8/c_Z, \qquad c_2 := \max\left\{\sqrt{\frac{2}{\gamma c_Z}}, \frac{1}{\sqrt{c_{\mathrm{reg}}}}\right\}.$$

The assumed condition $K \geq c_1 n u$ then ensures that

$$K Z_n(s_{1:n}) \geq \frac{K c_Z}{n} \geq 8u$$

for every such prefix.

We first obtain bounds that hold simultaneously for all regular prefixes of length n. Let $v := N_{01} + N_{10}$ denote the number of state changes. The tuple (s_n, v, N_{00}) determines the initial state and all four transition counts. Indeed, s_n and the parity of v determine s_1, while

$$N_{01} - N_{10} = s_n - s_1, \quad N_{01} + N_{10} = v, \quad N_{11} = n - 1 - v - N_{00}.$$

For each $s_n \in \{0, 1\}$ and $v \in \{0, \ldots, n-1\}$, there are at most $n - v$ possible values of N_{00}. Consequently, the number M of distinct regular types satisfies

$$M \leq 2\sum_{v=0}^{n-1}(n - v) = n(n+1) \leq 2n^2.$$

Prefixes of the same type have the same likelihood function, regularity status, and terminal state, and therefore the same values of ρ_K, ρ_{unif}, Z_n, f_K^*, and f_{unif}^*.

Choose one deterministic representative from each nonempty regular type and apply Lemma C.6 with $\delta_0 := \delta/4n^2$. Since $\log(3/\delta_0) = u$ and $KZ_n \geq 8u$, each of the two inequalities below fails with probability at most δ_0 for each representative. A union bound over both inequalities and all regular types gives an event $\mathcal{G}_{K,n}$, depending only on the sampled pretraining task family, with

$$\mathbb{P}(\mathcal{G}_{K,n}) \geq 1 - 2M\delta_0 \geq 1 - \delta,$$

on which

$$\rho_K(s_{1:n}) \geq \frac{1}{2}\rho_{\mathrm{unif}}(s_{1:n}), \tag{72}$$

$$|f_K^*(s_{1:n}) - f_{\mathrm{unif}}^*(s_{1:n})|^2 \leq \frac{4u}{KZ_n(s_{1:n})(2N_{s_n} + 3)} \tag{73}$$

hold simultaneously for every regular prefix of length n.

For the remainder of the proof, fix any task family in $\mathcal{G}_{K,n}$ and any realized trained predictor $f_{\widehat{\boldsymbol{\theta}}}$. By (72), $R_K(s_{1:n}) \leq 2$ on $\mathcal{E}_n$, so

$$\mathbb{P}_{s_{1:n}\sim\rho_{\mathrm{unif}}}\left(R_K(s_{1:n}) > 2 \,|\, \mathcal{E}_n\right) = 0.$$

Applying Theorem D.2 with $a = 2$ yields

$$\mathrm{RMSE}_n^2\left(f_{\widehat{\boldsymbol{\theta}}}, f_K^*\right) \leq \frac{\epsilon_{K,n}}{r_n}. \tag{74}$$

Taking square roots and using $r_n \geq c_{\mathrm{reg}}$ gives

$$\mathrm{RMSE}_n\big(f_{\widehat{\boldsymbol{\theta}}}, f_K^*\big) \leq \frac{1}{\sqrt{c_{\mathrm{reg}}}}\sqrt{\epsilon_{K,n}} \leq c_2\sqrt{\epsilon_{K,n}},$$

which proves the first inequality in (13).

To prove the second inequality, we first control the difference between the two Bayes predictors. For every regular prefix, $N_{s_n} \geq \gamma(n-1)$, and hence

$$Z_n(s_{1:n})(2N_{s_n}+3) \geq \frac{c_Z}{n}\big(2\gamma(n-1)+3\big) \geq 2\gamma c_Z.$$

Substituting this bound into (73) gives

$$|f_K^*(s_{1:n}) - f_{\mathrm{unif}}^*(s_{1:n})|^2 \leq \frac{2u}{\gamma c_Z K}$$

uniformly over regular prefixes. Taking the conditional expectation under ρ_{unif} given $\mathcal{E}_n$, followed by the square root, yields

$$\mathrm{RMSE}_n(f_K^*, f_{\mathrm{unif}}^*) \leq c_2\sqrt{\frac{u}{K}}. \tag{75}$$

Finally, the triangle inequality in $L^2\big(\rho_{\mathrm{unif}}(\cdot \mid \mathcal{E}_n)\big)$, together with (74), (75), and $r_n \geq c_{\mathrm{reg}}$, gives

$$\mathrm{RMSE}_n\big(f_{\widehat{\boldsymbol{\theta}}}, f_{\mathrm{unif}}^*\big) \leq \mathrm{RMSE}_n\big(f_{\widehat{\boldsymbol{\theta}}}, f_K^*\big) + \mathrm{RMSE}_n(f_K^*, f_{\mathrm{unif}}^*) \leq c_2\left(\sqrt{\epsilon_{K,n}} + \sqrt{\frac{u}{K}}\right).$$

Substituting $u = \log(12n^2/\delta)$ proves the second inequality in (13). Both constants c_1 and c_2 depend only on (η, γ), and both bounds hold on the same event $\mathcal{G}_{K,n}$, whose probability is at least $1-\delta$. This completes the proof. □

E Additional Experimental Setups

E.1 Transition from Task Retrieval and Task Learning

Data generation. We generate pretraining data following Definition 2.1. Specifically, the pretraining prior is a discrete uniform mixture over K Markov chains, whose transition parameters are sampled independently from $\mathrm{Unif}(0.01, 0.99)^2$ for numerical stability unless otherwise specified. The initial distribution is fixed as $\boldsymbol{\mu} = (0.5, 0.5)$, and the sequence length is set to $T = 128$.

Training setup. We train a 4-layer causal transformer with relative positional encoding, using $d_{\mathrm{model}} = 128$ and $d_{\mathrm{ff}} = 128$. The model is trained from scratch with AdamW, batch size 512, peak learning rate 10^{-3}, and a linear warmup over 10 epochs. To ensure sufficiently long contexts, the next-token prediction loss is computed only at positions $n \geq 20$. To study the transition from task retrieval to task learning, we vary the number of pretraining Markov chains over $K \in \{2, 4, 8, 16, 32, 64, 128, 256, 512, 1024\}$. For each K, we generate $N = 200K$ binary sequences, keeping the number of training sequences per task fixed.

Evaluation setup. After training, we evaluate on 50 unseen Markov tasks, for a total of 1,000 test sequences. We compute the ICL next-token prediction loss for prefixes with $n \geq 50$, and compare the PT with two theoretical baselines: the K-prior Bayes predictor in (7), which corresponds to task retrieval, and the uniform-prior Bayes predictor in (8), which corresponds to task learning. As shown in Figure 1(a), the PT closely follows the K-prior Bayes predictor in the low-diversity regime, indicating task-retrieval behavior. As K increases, its performance shifts toward the uniform-prior Bayes predictor, suggesting that the model increasingly infers the transition rule from the in-context prompt and exhibits task-learning behavior. Figure 1(b) shows an analogous transition on four-state Markov data.

F Auxiliary Results

Lemma F.1 (Pinsker's inequality for Bernoulli distributions)**.** *For any $x, y \in (0,1)$, we have*

$$2(x-y)^2 \leq \mathrm{KL}(\mathrm{Bern}(x) \,\|\, \mathrm{Bern}(y)) \leq \frac{(x-y)^2}{y(1-y)}. \tag{76}$$

Proof. By definition, we have

$$\mathrm{KL}(\mathrm{Bern}(x) \,\|\, \mathrm{Bern}(y)) = x \log \frac{x}{y} + (1-x) \log \frac{1-x}{1-y}. \tag{77}$$

We first prove the lower bound. Fix $x \in (0,1)$ and define

$$f(y) := x \log \frac{x}{y} + (1-x) \log \frac{1-x}{1-y} - 2(x-y)^2, \quad \forall y \in (0,1).$$

Differentiating with respect to y, we obtain

$$f'(y) = -\frac{x}{y} + \frac{1-x}{1-y} + 4(x-y) = (y-x)\left(\frac{1}{y(1-y)} - 4\right).$$

Since $y(1-y) \le 1/4$, it follows that $f'(y) \le 0$ for $y < x$ and $f'(y) \ge 0$ for $y > x$. Hence, f attains its global minimum at $y = x$. Since $f(x) = 0$, we obtain

$$\mathrm{KL}(\mathrm{Bern}(x) \,\|\, \mathrm{Bern}(y)) \ge 2(x-y)^2.$$

For the upper bound, applying $\log z \le z - 1$ to (77) gives

$$\begin{aligned}\mathrm{KL}(\mathrm{Bern}(x) \,\|\, \mathrm{Bern}(y)) &\le x\left(\frac{x}{y} - 1\right) + (1-x)\left(\frac{1-x}{1-y} - 1\right) \\ &= \frac{x(x-y)}{y} + \frac{(1-x)(y-x)}{1-y} = (x-y)\left(\frac{x}{y} - \frac{1-x}{1-y}\right) = \frac{(x-y)^2}{y(1-y)}.\end{aligned}$$

This proves both inequalities in (76). □