

Large Language Models as Markov Chains

Oussama Zekri^{*1} Ambroise Odonnat^{*2,3}

Abdelhakim BENECHAHAB^{2,4} Linus BLEISTEIN³ Nicolas BOULLÉ⁵ Ievgen REDKO²

Abstract

Large language models (LLMs) are remarkably efficient across a wide range of natural language processing tasks and well beyond them. However, a comprehensive theoretical analysis of the LLMs’ generalization capabilities remains elusive. In our paper, we approach this task by drawing an equivalence between autoregressive transformer-based language models and Markov chains defined on a finite state space. This allows us to study the multi-step inference mechanism of LLMs from first principles. We relate the obtained results to the pathological behavior observed with LLMs such as repetitions and incoherent replies with high temperature. Finally, we leverage the proposed formalization to derive pre-training and in-context learning generalization bounds for LLMs under realistic data and model assumptions. Experiments with the most recent Llama and Gemma herds of models show that our theory correctly captures their behavior in practice.

1. Introduction

Large language models (LLMs) (Brown et al., 2020; Touvron et al., 2023a) have reshaped the landscape of machine learning and artificial intelligence. Trained on vast amounts of data, they have reached unprecedented performance in tasks such as machine translation (Brown et al., 2020), text generation, question answering (Roberts et al., 2020), and sentiment analysis (Zhang et al., 2023a), to name a few. Pre-trained LLMs also exhibit an intriguing capability of performing learning from the context, also called in-context learning (ICL), without explicitly updating their parameters.

Despite all these remarkable achievements, the theoretical justification for LLMs’ impressive performance remains elu-

^{*}Equal contribution ¹ENS Paris-Saclay ²Huawei Noah’s Ark Lab ³Inria ⁴Eurecom ⁵Imperial College London. Correspondence to: OZ <oussama.zekri@ens-paris-saclay.fr>, AO <ambroiseodonnatechnologie@gmail.com>.

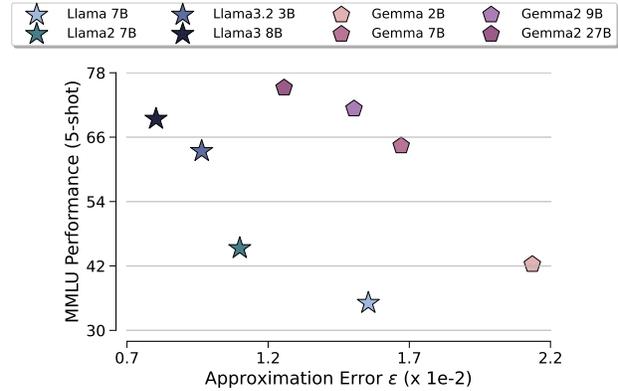


Figure 1: *LLMs’ Sample Complexity*. We plot the Massive Multitask Language Understanding (MMLU) (Hendrycks et al., 2021) performance with respect to the approximation error ϵ predicted by Proposition 4.1. We set N^* equal to the real number of pre-training tokens. Each point represents a model from the Llama or Gemma families (Dubey et al., 2024; Riviere et al., 2024). The approximation error ϵ predicted by our theory correlates with the real performance, with different trends between the models’ families.

sive. The two main reasons for this are 1) the complexity of analyzing modern deep transfer-based architectures used to train LLMs and 2) the non-iid nature of their pre-training data. Prior works usually bypass these challenges by imposing simplifying assumptions. Such assumptions include considering a shallow single-head transformer (Makkuva et al., 2024), only self-attention mechanism (Ildiz et al., 2024) or attention-only transformers (Edelman et al., 2024; Jeon et al., 2024) when studying LLMs. The others are to study the auto-regressive mechanism of the LLMs without taking into account their architecture explicitly (Lotfi et al., 2023; 2024; Xie et al., 2022). When the architecture is studied in full generality (Li et al., 2023; Zhang et al., 2023b), it is also common to relax the non-iid assumption by a certain predefined Markovian process.

This work provides a unified theoretical analysis of LLM’s inference, pre-training, and ICL under realistic assumptions on model architecture and pre-training data. Our core idea

Method	Pre-Train.	ICL	Input	Model-Dep.	Exp. val.
(Xie et al., 2022)	×	✓	HMM	×	✓
(Zhang et al., 2023b)	✓	✓	MC	✓	×
(Li et al., 2023)	✓	✓	MC	✓	✓
(Lotfi et al., 2024)	✓	×	non-iid	×	✓
Ours	✓	✓	non-iid	✓	✓

Table 1: LLMs generalization bounds proposed in the literature. **Pre-train.** stands for pre-training; **Input** refers to the assumptions on the data generating process (Markov Chains (MC); Hidden Markov Model (HMM)); **Model-dep.** means explicit dependence on model’s architecture; **Exp.val.** stands for the experimental validation of the theory.

is to note that despite the seemingly infinite generation capacity of LLMs, their vocabulary and context window are finite, making all possible input and output sequences countable. This leads to an intuitive, yet overlooked, approach that interprets LLMs as Markov chains whose transition matrix operates on a finite state space of such sequences (see Fig. 2). The generalization of LLMs is then their capacity to learn the transition probabilities of language itself.

Main contributions. Our contributions can be summarized as follows.

- 1) We characterize the inference mechanism of any LLM as a *finite-state* Markov chain. This provides a deeper understanding of LLMs’ generation behavior and sheds new light on some of their pathological behaviors.
- 2) We prove sample complexity bounds for deep transformer-based LLMs pre-trained on *non-iid* data. They are more general than existing results (see Table 1 for a full comparison with the literature) and are verified in practice on the Llama and Gemma herd of models from 2B to 27B (see Fig. 1 and Section 4).
- 3) We derive ICL generalization bounds when LLMs are prompted with Markov chains. We validate our results by showing that the most recent open-source LLMs obey the in-context *scaling laws* predicted by our theory.

Organization of the paper. Section 2 provides background material on transformer-based autoregressive LLMs and Markov chains. In Section 3.1, we formalize the multi-step inference mechanism of a pre-trained LLM as a Markov chain. In Section 4, we study how a deep transformer-based LLM learns this inference mechanism through pre-training on non-iid data using the formalism of Marton couplings and also extend the obtained results to the ICL setting. We verify them through a set of experiments in Section 5.

2. Background Knowledge

We recall some facts about Markov chains (Paulin, 2015; Roberts & Rosenthal, 2004) and LLMs. More notations and background materials are available in Appendices A to C.

Markov chains. Let Ω be a discrete finite set of size $|\Omega|$. A discrete-time, time-homogeneous Markov chain $\text{MC}(\Omega, \mathbf{Q})$ defined on a state space $\Omega = \{x_i\}_{i=1}^{|\Omega|}$ with transition matrix $\mathbf{Q} \in \mathbb{R}^{|\Omega| \times |\Omega|}$ with entries $\mathbf{Q}_{ij} = \mathbf{Q}(x_i, x_j) \in [0, 1]$ is a sequence of random variables $(\mathbf{X}_1, \mathbf{X}_2, \dots)$ taking values in Ω such that for any $n \in \mathbb{N}$ and $(x_1, \dots, x_{n+1}) \in \Omega^{n+1}$, we have

$$\begin{aligned} \mathbb{P}(\mathbf{X}_{n+1} = x_{n+1} \mid \mathbf{X}_n = x_n, \dots, \mathbf{X}_1 = x_1) \\ &= \mathbb{P}(\mathbf{X}_{n+1} = x_{n+1} \mid \mathbf{X}_n = x_n) \\ &:= \mathbf{Q}(x_n, x_{n+1}). \end{aligned}$$

A distribution π on Ω is called stationary if $\pi\mathbf{Q} = \pi$. For any $x \in \Omega$, $\text{MC}(\Omega, \mathbf{Q})$ converges to π if $\lim_{n \rightarrow \infty} d_{\text{TV}}(\mathbf{Q}^n(x, \cdot), \pi) = 0$, where $\mathbf{Q}^n(x, \cdot)$ denotes the probability of \mathbf{X}_n conditioned on $\mathbf{X}_1 = x$ and the total variation between two distributions \mathbb{P} and \mathbb{Q} , defined on (Ω, \mathcal{F}) , is $d_{\text{TV}}(\mathbb{P}, \mathbb{Q}) := \sup_{A \in \mathcal{F}} |\mathbb{P}(A) - \mathbb{Q}(A)|$. We recall that the mixing time $t_{\text{mix}}(\varepsilon)$ of a Markov chain is the minimal time needed to be ε -close to its stationary distribution (see Definition C.8). Intuitively, a Markov chain mixes slowly when it remains close to the initial state after a given number of steps and doesn’t explore its state space. A Markov chain with rapid mixing quickly forgets its initial state and transitions more easily to a wider set of states.

Large language models. Let \mathcal{V} denote a dictionary of size T , referred to as the *vocabulary size*, used to encode an arbitrary sequence into a sequence of predefined tokens belonging to \mathcal{V} . We assume that our model admits a maximum of K tokens as input, referred to as the *context window* of the model. The domain of the transformer-based autoregressive LLM is the set of all sequences consisting of at most K elements of \mathcal{V} . We denote this space by \mathcal{V}_K^* , i.e., $\mathcal{V}_K^* := \{v \in \mathcal{V}^*, |v| \leq K\}$ with $|v|$ the length of v . We define an LLM with trainable parameters Θ as a function $f_{\Theta}^{T,K} : \mathcal{V}_K^* \rightarrow \Delta(\mathcal{V})$, where $\Delta(\mathcal{V})$ is the probability simplex over \mathcal{V} . Given a sequence of tokens v , $f_{\Theta}^{T,K}$ outputs a probability distribution over the whole state space indicating the likelihood for each of its elements to appear after v (see Appendix B for more details). Noting that an LLM is defined for a given vocabulary size T and context window K , we will drop the exponents and simply write f_{Θ} to ease the notations. We consider a setting where the learner’s objective is to approximate the probability distribution of sequences over an input vocabulary $\mathbb{P}_{\mathcal{L}} : \mathcal{P}(\mathcal{V}_K^*) \rightarrow [0, 1]$ where $\mathcal{P}(\mathcal{V}_K^*)$ denotes the powerset of \mathcal{V}_K^* . We also assume the existence of a constant $c_0 > 0$ such that for any $n \in [N]$

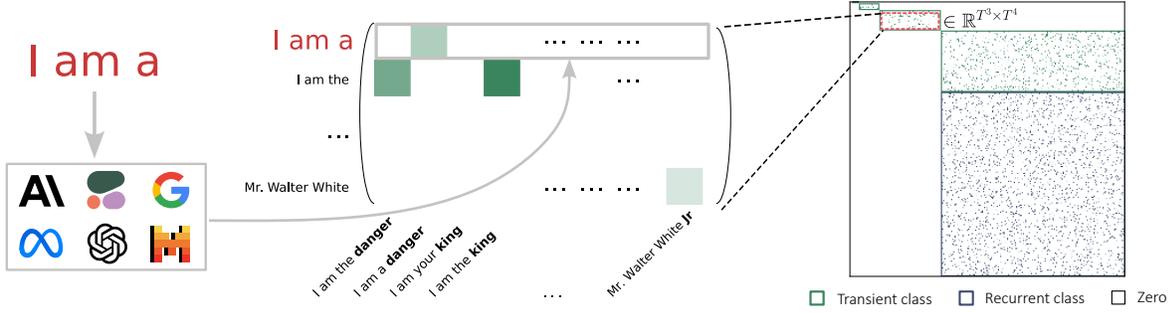


Figure 2: *LLM as a Markov chain.* A large language model with vocabulary size T and context window K is equivalent to a Markov chain with a sparse and block-structured transition matrix of size $\sum_{i \leq K} T^i \sim \mathcal{O}(T^K)$. The latter captures all possible outputs of a given LLM for all possible input sequences allowed by its vocabulary and context window.

and $(x_1, \dots, x_{n+1}) \in \Omega^{n+1}$,

$$\mathbb{P}_{\mathcal{L}}(\mathbf{X}_{n+1} = x_{n+1} \mid \mathbf{X}_n = x_n, \dots, \mathbf{X}_1 = x_1) \geq c_0. \quad (1)$$

This is a common assumption used in (Wies et al., 2024; Xie et al., 2022; Zhang et al., 2023b).

Transformer model. Without loss of generality, f_{Θ} is assumed to be a transformer model with L layers and H heads, consisting of alternating multi-head attention (MHA) and feed-forward blocks (more details in Appendix B). The first layer receives an input $\mathbf{S}^{(0)} = \mathbf{S}$ embedded in an r -dimensional space. To obtain a probability distribution on the vocabulary \mathcal{V} , the output $\mathbf{S}^{(L)} \in \mathbb{R}^{r \times T}$ of the final layer is projected back to the vocabulary size by an “unembedding layer” $\mathbf{W}_U \in \mathbb{R}^{T \times r}$ and averaged over the columns to obtain a vector in \mathbb{R}^T . A softmax layer is finally applied to obtain the probability distribution of the next token $\mathbb{P}_{\Theta}(\cdot \mid \mathbf{S}) := \text{softmax}(\frac{1}{n\tau} \mathbf{W}_U \mathbf{S}^{(L)} \mathbb{1}_n) \in \Delta_T$, where Θ denotes the parameters of the entire network and τ is the softmax temperature (Hinton, 2015). Unless otherwise specified, we assume that the unembedding layer is bounded. The classes of parameters and neural networks it generates respectively write $\mathcal{W} = \{\Theta \text{ s.t. } \|\mathbf{W}_U^T\|_{2,1} \leq B_U\}$ and $\mathcal{F} = \{f_{\Theta} \text{ s.t. } \Theta \in \mathcal{W}\}$.

3. Large Language Models as Markov Chains

To proceed, we formally define a Markov chain that explicitly captures the inference of a given LLM f_{Θ} . We build upon a high-level idea that associates a tokenized input sequence with a state v_i , from which we transition to a new state $v_j = [v_i, v]$ by concatenating the token v predicted by an LLM to it. We then provide a theoretical characterization of this Markov chain. Our analysis holds for any model with finite fixed vocabulary, and context window. This setup includes deep transformer-based LLMs while excluding models such as RNNs and LSTMs.

3.1. Markov Chain Formalization

We first introduce the notion of “incompatible sequences” in the sense of next-token prediction.

Definition 3.1 (Incompatible Sequences). Let $u, v \in \mathcal{V}_K^*$ be two sequences of at most K tokens. We say that v is incompatible with u if v cannot be a plausible completion of u , that is

$$\exists l \in \{1, \dots, |u| - 1\}, \text{ s.t. } (u)_{l+1} \neq (v)_l. \quad (2)$$

We denote by \mathcal{I} the set of incompatible sequences, i.e., $\mathcal{I} := \{(u, v) \text{ s.t. } v \text{ is incompatible with } u\}$.

The order matters since a necessary condition for v to be compatible with u is that v has as many or one more token than u . We now proceed with the definition of the transition matrix associated with an LLM f_{Θ} .

Proposition 3.2. Any large language model f_{Θ} can be equivalently represented by a Markov chain $MC(\mathcal{V}_K^*, \mathbf{Q}_f)$, with a sparse transition matrix $\mathbf{Q}_f \in \mathbb{R}^{|\mathcal{V}_K^*| \times |\mathcal{V}_K^*|}$ that writes for any $v_i, v_j \in \mathcal{V}_K^*$:

$$\mathbf{Q}_f(v_i, v_j) = \begin{cases} 0, & \text{if } (v_i, v_j) \in \mathcal{I} \\ \{f_{\Theta}(v_i)\}_{j_0}, & \text{otherwise,} \end{cases}$$

where j_0 denotes the index in \mathcal{V} of the last token of v_j . We have $|\mathcal{V}_K^*| = T(T^K - 1)/(T - 1)$ and the proportion of non-zero elements in \mathbf{Q}_f is $(T - 1)/(T^K - 1)$.

Proposition 3.2 states that if v_j is incompatible with v_i , then the probability to go from the state v_i to v_j is zero as LLM only transitions between compatible sequences. For such compatible sequences, the probability of going from v_i to v_j is naturally the next-token probability associated with the last token of v_j , i.e., $\{f_{\Theta}(v_i)\}_{j_0}$.

Example: binary sequences. To help unpack Proposition 3.2, we illustrate it with $T = 2$ and $K = 3$ in Fig. 3 (see the general case in Fig. 2). Here, we start with an input prompt consisting of one token. We then transition to sequences of increased size, while attributing a 0 probability to incompatible sequences. Each generation step for different values of $k < K$ defines green rectangular blocks of size $T^k \times T^{k+1}$ in the transition matrix \mathbf{Q}_f .

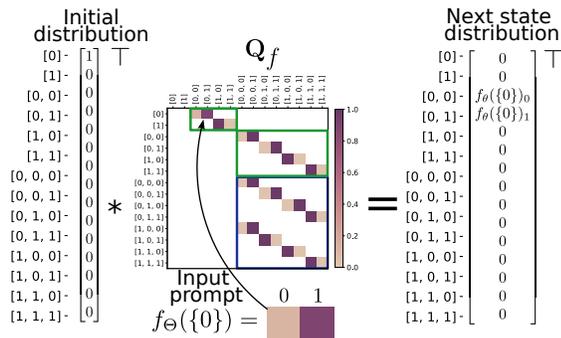


Figure 3: Proposition 3.2 with $T = 2$ and $K = 3$.

Reaching the blue square block in the transition matrix means that the input sequence is of the maximum size K . The model can no longer append tokens to it and has to delete the first token to proceed. This blue block is of size $T^K \times T^K$: it captures transitions between sequences of the maximum admissible length. We define similarly the reference transition matrix \mathbf{Q}^* of the language where the probability of transitions $\{f_{\Theta}(v_i)\}_j$ are replaced by ground-truth probabilities $\mathbb{P}_{\mathcal{L}}(v_j | v_i)$. To use \mathbf{Q}_f as f_{Θ} , it is sufficient to define an input distribution δ_0 of the Markov chain based on input prompt v . It is a one-hot encoding vector of size $|\mathcal{V}_K^*|$ with 1 at the position of the state corresponding to v . Then, the transition to the next state writes $\delta_1 = \delta_0 \mathbf{Q}_f$. The output of $f_{\Theta}(v)$ for individual tokens in \mathcal{V} would then correspond to probabilities in δ_1 for states that are concatenations of v with T tokens from \mathcal{V} . This process is illustrated in Fig. 3. We study the properties of $\text{MC}(\mathcal{V}_K^*, \mathbf{Q}_f)$ below.

Proposition 3.3. *Let $\text{MC}(\mathcal{V}_K^*, \mathbf{Q}_f)$ be a Markov chain defined in Proposition 3.2. Then $\text{MC}(\mathcal{V}_K^*, \mathbf{Q}_f)$ is ergodic and admits a unique stationary distribution.*

The proof builds on showing that such a Markov chain has at most one recurrent class (blue block in Fig. 2) plus some additional transition states (green blocks in Fig. 2). We now characterize how many times one should apply \mathbf{Q}_f to the input to reach the stationary distribution.

Proposition 3.4. *Let $\text{MC}(\mathcal{V}_K^*, \mathbf{Q}_f)$ be as in Proposition 3.3 and $e = (1, 1, \dots, 1)^{\top}$. Then we have that*

$\lim_{n \rightarrow \infty} \mathbf{Q}_f^n = e\pi$, where π is the stationary distribution of the recurrent class \mathcal{R} of states, expanded by 0’s for each transient state of the unichain. Moreover, for all $n \geq K$,

$$|(\mathbf{Q}_f^n)_{i,j} - (e\pi)_{i,j}| \leq (1 - 2\varepsilon)^{\lfloor \frac{n}{K} \rfloor - 1},$$

where $\varepsilon = \min_{i,j \in \mathcal{R}^2} \{(\mathbf{Q}_f^K)_{i,j}\} > 0$.

The stationary distribution is the long-term equilibrium of the multi-step inference of an LLM. While its true value is intractable because of the size of \mathcal{V}_K^* , we discuss below the implications of Proposition 3.4 for LLMs.

- 1. Looping.** The stationary distribution is independent of the initial state (i.e., input prompt), and reaching it should send the LLM into a deterministic loop of repetitions. This behavior is well known and occurs frequently in practice with many existing LLMs (Ivgi et al., 2024), requiring adding a repetition penalty.
- 2. Temperature and coherence.** Increasing temperature of the model leads to a decreasing coherence of its outputs as measured by perplexity (Peeperkorn et al., 2024). Temperature, in turn, impacts ε (the smallest element of the K^{th} power of the transition matrix), as it modifies the probabilities of the predicted next tokens. Consequently, increasing the temperature increases the speed of convergence to the stationary distribution making the output less coherent. We illustrate this on a toy example below.

3.2. Illustration on a Toy Model

We illustrate the results of Section 3 on a toy model trained on a sequence of 0s and 1s. Here, each subsequent token is 0 if the sum of three previous tokens is even and 1 otherwise. Therefore, $T = 2$ and $K = 3$. We generate a sequence of 40 digits, resulting in 37 distinct supervised examples, and train a small ‘‘GPT-like’’ model (Karpathy, 2023) on it. We extract the logits from the model by prompting it with all possible combinations of 0s and 1s of length less than three to obtain the transition matrix $\mathbf{Q}_f \in \mathbb{R}^{14 \times 14}$ depicted in Fig. 4(a). The transition matrix’s structure (e.g., presence of transient and recurrent classes) matches the one presented in Fig. 2. Fig. 4(b) displays the stationary distribution of the trained model obtained by raising \mathbf{Q}_f to power 10^5 . We note that it has a strong bias toward seen training samples in accordance with our intuition behind the stationary distribution presented earlier. Finally, Fig. 4(c) illustrates the convergence rate of the toy model, predicted by Proposition 3.4, and compares it to models with larger dictionary size T and context window K . In Fig. 4(c), we set $\varepsilon = 10^{-6}$ and note that this parameter reflects the ability of the LLM to explore the state space.

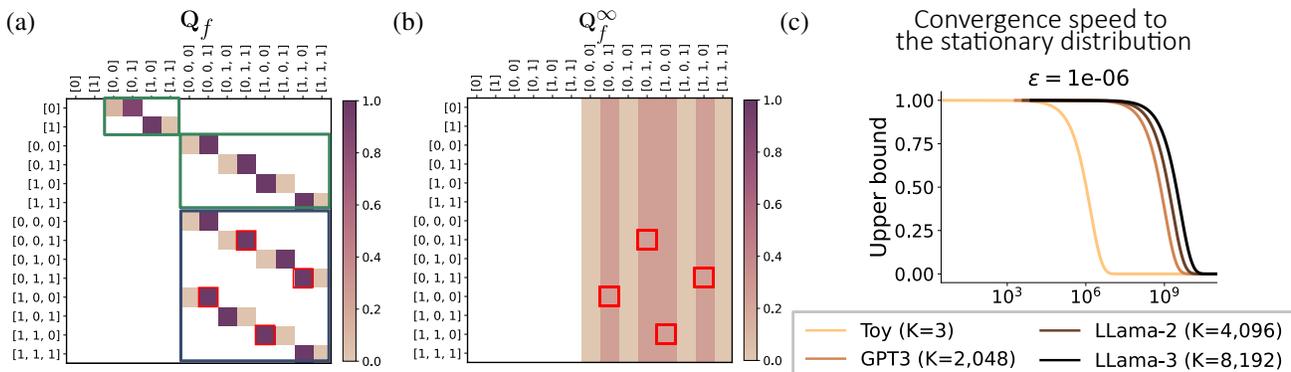


Figure 4: *Markov chain with a small GPT-like model.* (a) Transition matrix \mathbf{Q}_f of the model where \square denotes the examples from the training set. (b) The stationary distribution of the trained model assigns almost uniform probabilities to the states seen during training. (c) Convergence rate to the stationary distribution for the considered toy model along with three LLMs, highlighting the dependence on K . The y-axis is the upper bound in Proposition 3.4.

Role of the temperature. To better illustrate the role of ϵ , we now plot the transition matrix of the studied Markov chain obtained when applying different temperature scaling to the logits returned by the trained model. As the temperature is commonly linked to the ability of LLMs to transition more freely to a large set of states (Chen & Ding, 2023), we expect that lower temperatures should negatively impact the convergence speed to the stationary distribution. In Fig. 5(a), we show that for a low temperature (0.2), the Markov chain mixes slowly and is unable to reach its stationary distribution (same line in the transition matrix as in Fig. 4(c)) even after 10^6 steps. In the case of a more commonly used temperature equal to 1 (Fig. 5(b)), the model requires only 300 steps to converge. Finally, setting the model’s temperature to 2 (Fig. 5(c)) makes the convergence extremely fast, reaching the stationary distribution after only 30 steps. The interplay between ϵ and the model’s temperature is displayed in Fig. 5(d), increasing the temperature leads to a drastic improvement in the convergence speed.

We have demonstrated that any autoregressive transformer-based LLM admits an equivalent Markov chain formulation. This equivalence connects nicely to some of the pathological behaviors of LLMs observed in practice. We will now use this formalization to provide a fine-grained analysis of LLMs’ generalization capabilities that may be of interest to both theorists and practical users.

4. Sample Complexity and Generalization

We now study the generalization of an LLM f_Θ as its capacity to infer correctly all the elements of \mathbf{Q}_f , while approximating the true reference matrix of transition probabilities \mathbf{Q}^* . The hardness of this task is highlighted by the fact that an LLM observes a negligible amount of \mathbf{Q}^* ’s elements during its pre-training. Indeed, for GPT-3 (Brown et al.,

2020), this represents 5×10^{11} training tokens, which pales in comparison with the number of non-zero elements in \mathbf{Q}_f , given by $T^{K+1} \approx 10^{9632}$. In this section, we determine the number of pre-training tokens required to be ϵ -close to perfect generalization. This result stems from a novel pre-training generalization bound on *non-iid* random variables in Section 4.2. We extend this bound to the in-context learning in Section 4.3.

4.1. Main Result: Pre-training Sample Complexity

Non-iid data. We denote by $X = (\mathbf{X}_1, \dots, \mathbf{X}_{N_{\text{train}}})$ the N_{train} tokens in \mathcal{V} that f_Θ observes during pre-training. The training sequences of tokens can be written as $\mathbf{S}_n = (\mathbf{X}_1, \dots, \mathbf{X}_n)$ if $n \leq K$ and $\mathbf{S}_n = (\mathbf{X}_{n-K+1}, \dots, \mathbf{X}_n)$ otherwise due to the deletion process (see Definition B.2). In particular, the \mathbf{S}_n are elements of \mathcal{V}_K^* . We assume that the pre-training data $S = (\mathbf{S}_1, \dots, \mathbf{S}_{N_{\text{train}}})$ is a sequence of *dependent* random variables with a mild coupling structure, namely that a Marton coupling with mixing matrix Γ exists for $S = (\mathbf{S}_1, \dots, \mathbf{S}_{N_{\text{train}}})$ (more details on Marton coupling can be found in Appendix C.3). This ensures that our setting remains very broad as it subsumes the case of independent variables, m -dependent variables, language bigrams (Bietti et al., 2023), and the Markov chain setting considered in state-of-the-art ICL analysis of LLMs (Hu et al., 2024; Zhang et al., 2023b).

We state our main result on the pre-training sample complexity below. The proof is deferred to Appendix F.4.

Proposition 4.1 (Sample complexity). *Let $\delta \in [0, 1]$ and $\epsilon > 0$. Assuming a perfect pre-training of f_Θ and N_{train} pre-training tokens with $N_{\text{train}} \geq N^* :=$*

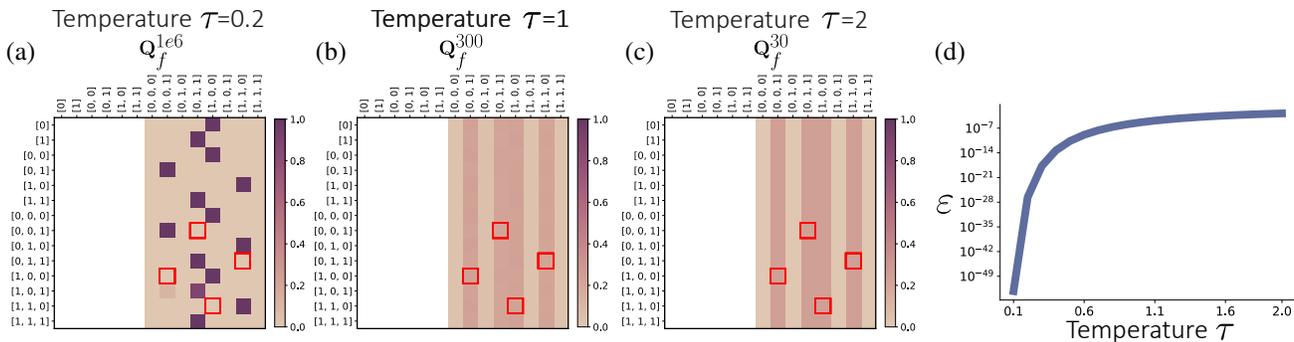


Figure 5: *Dependence of ϵ on the temperature of the model.* (a) For low temperatures, ϵ becomes too small to achieve convergence to the stationary distribution. (b)-(c) Increasing the temperature from 1 to 2 leads to a $\times 10$ faster convergence. (d) ϵ (log-scale) increase for temperature values in $[0.1, 2]$.

$\lceil \frac{4\bar{B}^2}{\epsilon^2} \log(\frac{2}{\delta}) \rceil$, we have with probability at least $1 - \delta$,

$$\mathbb{E}_{\mathbf{S} \sim \mathbb{P}_{\mathcal{L}}} \|\mathbf{Q}^*(\mathbf{S}, \cdot) - \mathbf{Q}_f(\mathbf{S}, \cdot)\|_1 \leq \epsilon,$$

where

$$\bar{B} = 2\|\Gamma\| \max\{\log(T) + 2B_U/\tau, \log(1/c_0)\}^{1/2}$$

is a model- and data-dependent constant.

Proposition 4.1 shows that the number of pre-training tokens needed to achieve a good approximation mostly depends on the problem’s parameters captured by \bar{B} . This constant includes two main ingredients. On the one hand, it accounts for the model’s architecture: the context window size K , the temperature scaling of the output logits τ , and the norm of the unembedding layer B_U . On the other hand, it captures – via the mixing matrix Γ – how interdependent the tokens are, including the special cases of *iid* data ($\|\Gamma\| = 1$) and *strongly dependent* sequences of tokens. Thus, the bound is both model and data-dependent, contrary to the previously proposed non-vacuous bounds of Lotfi et al. (2024) where the model’s architecture was not fully taken into account.

Practical insights. Proposition 4.1 can be used to predict how well an LLM understands language given a fixed number of pre-training tokens. Letting $N_{\text{train}} = N^*$ in Proposition 4.1 leads to an approximation error $\epsilon = \frac{2\bar{B}}{\sqrt{N_{\text{train}}}} \sqrt{\log(\frac{2}{\delta})}$. Since \bar{B} can be approximated as $\bar{B} \sim 2(\log(T) + 2T\sqrt{r}/\tau)^{1/2}$, we can verify the correlation between ϵ and the reported performance of the Llama (Dubey et al., 2024; Touvron et al., 2023a;b) and Gemma models (Mesnard et al., 2024; Riviere et al., 2024) (see Appendix E.1.1 for the experimental details). We use the values of T , r and N_{train} given in the technical reports of the open-source LLMs and plot the MMLU performance with respect to the approximation error ϵ in Fig. 1. This shows that our theory is model-specific as it results in distinctly different trends for the Llama and Gemma models.

The difference is due to the larger T of the Gemmas that compensates for the generally smaller embedding dimension r and leads to higher values of \bar{B} . The predicted approximation error correlates with the performance for each family of models: the larger the predicted error, the worse the models perform. Finally, we note that in this work we didn’t estimate the norm of the Marton coupling as it requires access to the pre-training data. Studying this quantity and its approximation is of independent interest and we leave it for future work.

4.2. Pre-Training Generalization Bound

Our main result on sample complexity is derived from the generalization bound that we present below.

Risk definition. We recall that for any $n \geq 1$, the true probability of next token \mathbf{X}_{n+1} given a past sequence \mathbf{S}_n is $\mathbb{P}_{\mathcal{L}}(\cdot | \mathbf{S}_n) \in \Delta_T$ and the probability estimated by the model is denoted by $\mathbb{P}_{\Theta}(\cdot | \mathbf{S}_n)$. Following the Markov chain formalization introduced in Section 3.1, we define the theoretical and empirical risks for any $\Theta \in \mathcal{W}$ as

$$\begin{aligned} \mathcal{R}(\Theta) &:= \mathbb{E}_{\mathbf{S} \sim \mathbb{P}_{\mathcal{L}}} [d_{\text{TV}}(\mathbf{Q}^*(\mathbf{S}, \cdot), \mathbf{Q}_f(\mathbf{S}, \cdot))] \\ \widehat{\mathcal{R}}(\Theta) &:= \frac{1}{N} \sum_{n=1}^N d_{\text{TV}}(\mathbb{P}_{\mathcal{L}}(\cdot | \mathbf{S}_n), \mathbb{P}_{\Theta}(\cdot | \mathbf{S}_n)). \end{aligned} \quad (3)$$

Formally, the expected risk can be written as

$$\begin{aligned} \mathcal{R}(\Theta) &= \mathbb{E}_{\mathbf{S} \sim \mathbb{P}_{\mathcal{L}}} [d_{\text{TV}}(\mathbf{Q}^*(\mathbf{S}, \cdot), \mathbf{Q}_f(\mathbf{S}, \cdot))] \\ &= \mathbb{E}_{\mathbf{S} \sim \mathbb{P}_{\mathcal{L}}} [d_{\text{TV}}(\mathbb{P}_{\mathcal{L}}(\cdot | \mathbf{S}), \mathbb{P}_{\Theta}(\cdot | \mathbf{S}))] \\ &= \mathbb{E}[\widehat{\mathcal{R}}(\Theta)]. \end{aligned}$$

The generalization problem consists of bounding the difference $\mathcal{R}(\Theta) - \widehat{\mathcal{R}}(\Theta)$.

Remark 4.1 (Choice of risk). *Our risk definition departs from empirical risk minimization commonly used in statistical learning theory (Bach, 2024; Marion, 2023; Redko et al., 2019; Vapnik, 1999). To assess how well the model*

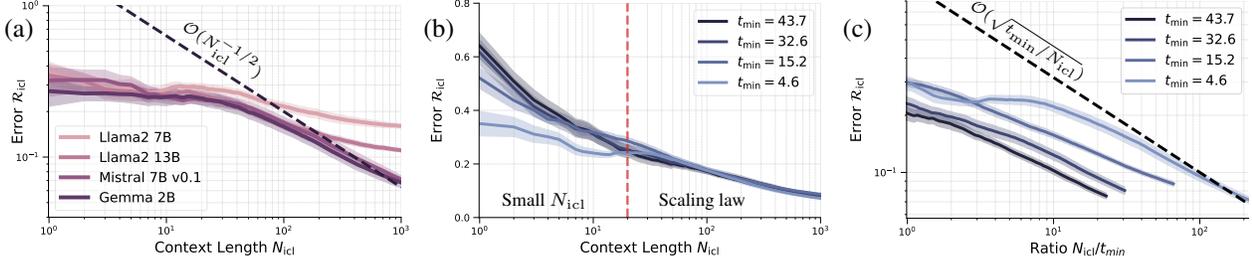


Figure 6: **In-context scaling laws.** The risk \mathcal{R}_{icl} as functions of N_{icl} , with 95% confidence intervals. (a) Risks for different LLMs along with the scaling law of Theorem 4.3. (b)-(c) Risks with Mistral 7B v0.1 for random 3-state transition matrices and different t_{min} as functions of N_{icl} and N_{icl}/t_{min} .

estimates the probability distribution of the next token, we follow (Hu et al., 2024; Zhang et al., 2023b) and study the TV distance used in learning and identity testing of Markov chains literature (Wolfner & Kontorovich, 2019; 2023). Appendix E.3 provides extended results with the KL divergence.

Generalization bound. We denote the risks by $\mathcal{R}_{pre}(\Theta)$ and $\widehat{\mathcal{R}}_{pre}(\Theta)$ to indicate that we take $N = N_{train}$ in Eq. (3). Below, we state a generalization bound on the estimation risk of pre-training (see Appendix F.5 for the proof).

Theorem 4.2 (Pre-training generalization bound). *Consider an LLM $f_{\Theta} \in \mathcal{F}$. We denote by Γ the mixing matrix of the pre-training sequences of tokens $(\mathbf{S}_1, \dots, \mathbf{S}_{N_{train}})$. Let $0 < \delta < 1$, then with probability at least $1 - \delta$,*

$$\mathcal{R}_{pre}(\Theta) \leq \widehat{\mathcal{R}}_{pre}(\Theta) + \frac{\bar{B}}{\sqrt{N_{train}}} \sqrt{\log\left(\frac{2}{\delta}\right)},$$

where \bar{B} is the constant of Proposition 4.1 and writes

$$\bar{B} = 2\|\Gamma\| \max\{\log(T) + 2B_U/\tau, \log(1/c_0)\}^{1/2}.$$

The obtained result has a desirable dependency on the amount of the pre-training data $\mathcal{O}(N_{train}^{-1/2})$. Additionally, it showcases an interplay between the model architecture and the unknown reference constant c_0 . If $B_U \approx \mathcal{O}(T\sqrt{r})$ (due to the normalization of the unembedding layer), then the model’s hidden dimension r and vocabulary size T should be large enough to ensure $\log(T) + 2B_U/\tau \geq \log(1/c_0)$. Below this threshold, the architecture of f_{Θ} is not expressive enough to have any tangible impact on its generalization, although it may affect the training error $\widehat{\mathcal{R}}_{pre}(\Theta)$.

4.3. In-Context Learning of Markov Chains

Although insightful, the analysis presented above is related to the pre-training of LLMs – a process that is hard and extremely costly to reproduce in practice. Similarly, we do not

have access to the ground-truth matrix \mathbf{Q}^* to reason about LLM’s ability to infer it in practice. To provide theoretical results that can be confirmed experimentally, we now turn our attention to ICL on Markov chains: a setup where one feeds a pre-trained LLM with an input sequence formed by a Markov chain of size N_{icl} defined over a state space Ω of size d^1 . The risks $\mathcal{R}_{icl}(\Theta)$ and $\widehat{\mathcal{R}}_{icl}(\Theta)$ are then defined similarly to Eq. (3) where transition kernel \mathbb{P} of the Markov chain replaces the language reference distribution $\mathbb{P}_{\mathcal{L}}$ and we have $N = N_{icl}$ (see Appendix F.7 for more details). To relate the generalization error to the pre-training error, we quantify the discrepancy between an LLM pre-trained mostly on textual data, and a hypothetical LLM with parameters in \mathcal{W}_{mc} that is pre-trained on a dataset of Markov chains with the same data distribution as the Markov chain used as an input during in-context inference. We define the divergence between two models with weights Θ_1, Θ_2 and estimated probability distributions $\mathbb{P}_{\Theta_1}, \mathbb{P}_{\Theta_2}$ as

$$\mathcal{K}(\Theta_1, \Theta_2) := \frac{1}{N} \sum_{n=1}^N \mathbb{E}_{\mathbf{S}_n} [d_{TV}(\mathbb{P}_{\Theta_1}(\cdot | \mathbf{S}_n), \mathbb{P}_{\Theta_2}(\cdot | \mathbf{S}_n))].$$

The operator \mathcal{K} is akin to a distance (the separation property is only verified almost surely, see Appendix C.4 for more details). The next result, whose proof is deferred to Appendix F.7, provides a generalization bound on the in-context learning phase.

Theorem 4.3 (In-Context Learning generalization bound). *Consider an LLM $f_{\Theta} \in \mathcal{F}$. We provide as input of f_{Θ} a d -state Markov chain $X = (\mathbf{X}_1, \dots, \mathbf{X}_{N_{icl}})$. The sequence of subsequences of the first n terms is denoted by $S = (\mathbf{S}_1, \dots, \mathbf{S}_n)$. S is also a Markov chain, and we denote by $t_{mix}(\varepsilon)$ its mixing time. Let $t_{min} := \inf_{0 < \varepsilon < 1} t_{mix}(\frac{\varepsilon}{2}) (\frac{2-\varepsilon}{1-\varepsilon})^2$. Let $\delta > 0$.*

¹This is different from another variation of ICL where supervised (x, y) pairs are provided in-context. Rather, the supervision is provided by observing transitions between states $(x_i, x_{i+1} = f(x_i))$ as discussed in (Li et al., 2023, Fig.1).

Then, with probability at least $1 - \delta$,

$$\mathcal{R}_{\text{icl}}(\Theta) \leq \inf_{\vartheta \in \mathcal{W}_{\text{mc}}} \{ \widehat{\mathcal{R}}_{\text{icl}}(\vartheta) + \mathcal{K}(\vartheta, \Theta) \} + \bar{B} \sqrt{\frac{t_{\text{min}}}{N_{\text{icl}}}} \sqrt{\log\left(\frac{2}{\delta}\right)}, \quad (4)$$

with a constant depending on the problem’s parameters

$$\bar{B} = 2 \max\{\log(d) + 2B_U/\tau, \log(1/p_{\text{min}})\}^{1/2}.$$

We note that instead of $\|\Gamma\|$ seen before, we now have an explicit dependency on t_{min} , which is related to the mixing time of the input Markov chain. This, together with the availability of the ground-truth transition matrix, allows us to use Theorem 4.3 to verify experimentally the in-context learning scaling laws for popular LLMs. Theorem 4.3 also suggests that an LLM pre-trained on diverse data sequences different from Markov chains should exhibit a certain degree of invariance to correctly infer the transition probabilities of the latter. This is reminiscent of the domain adaptation bounds (Redko et al., 2019) that commonly involve a distribution shift (i.e., a distance or a divergence) term that vanishes if the model is invariant to classes of transformations linking the distribution of the input data with that of test data. A recent success of LLMs in time series analysis (Gruver et al., 2023) suggests that this term may be small for certain types of data not used during pre-training.

Finally, we note that Theorem 4.3 implies that the LLM’s ability to learn Markov chains exceeds the frequentist method² (Wolfer & Kontorovich, 2019) that consists of counting the occurrences of different states to fill in the matrix \mathbf{Q}_f . This is experimentally confirmed in Fig. 7 of Section 5 (more details in Appendix E.1.2).

5. Numerical Experiments

We now evaluate the ability of recent LLMs, namely Mistral 7Bv0.1 (Jiang et al., 2023), Llama2 7B & 13B (Touvron et al., 2023b), and Gemma 2B (Mesnard et al., 2024) to infer transition probabilities of Markov chains in-context based on Theorem 4.3. We associate each state in the d -state Markov chain with a token from the set $\{0, \dots, d - 1\}$, concatenated to obtain a prompt of length N_{icl} . We add comas whenever necessary to ensure that each state is tokenized separately. Details on the experimental setup and experiments with more Markov chains and Llama3.2 (Dubey et al., 2024) are in Appendix D.1.

Dependence on N_{icl} . We first analyze the effect of N_{icl} on the risk calculated for a randomly generated 3-state Markov transition matrix. From the results presented in

²To the best of our knowledge, this is the only existing approach of Markov chains learning with theoretical guarantees.

Fig. 6 (left), we note that Llama2 models deviate from our $\mathcal{O}(N_{\text{icl}}^{-1/2})$ theoretical scaling law, while most recent models (Mistral and Gemma) stay much closer to Theorem 4.3, similarly to what was observed by Cabannes et al. (2024). Being randomly generated, the Markov chains provided to the models have not been seen during training, and older (weaker) models naturally struggle to generalize.

Dependence on t_{min} . Theorem 4.3 states that Markov chains with slow mixing (higher t_{min}) are slower to learn. We now plot the true risk for a single model with different values of t_{min} highlighting in Fig. 6 (right) a two-stage regime of ICL. In a first stage, the bound in Eq. (4) is dominated by $\sqrt{t_{\text{min}}/N_{\text{icl}}}$ for small N_{icl} , and depends strongly on t_{min} , while the scaling law $\mathcal{O}(N_{\text{icl}}^{-1/2})$ dominates as N_{icl} increases beyond $N_{\text{icl}} \approx 20$.

Dependence on d . We now verify Theorem 4.3 for Markov chains with a different state space size (previously $d = 3$). We also consider a baseline given by the frequentist method mentioned before. For the latter, its dependence on d behaves like $\mathcal{O}(\sqrt{d}/N_{\text{icl}})$ (Wolfer & Kontorovich, 2019), while Theorem 4.3 gives $\mathcal{O}(\sqrt{\log(d)}/N_{\text{icl}})$. For Markov chains with a small number of states d , there is no clear difference between the frequentist estimator and a LLM. However, as d grows the frequentist estimator struggles to estimate the transition matrix due to the $\mathcal{O}(\sqrt{d})$ scaling factor. This is verified experimentally in Fig. 7, where we vary the parameter d from 3 (left) to 700 (right). We observe that the LLM follows the theoretical neural scaling law $\mathcal{O}(N_{\text{icl}}^{-1/2})$ and outperforms the frequentist method for $d = 700$, while being close to it for $d = 3$. We conclude that our analysis gives theoretical insights on the ICL neural scaling law observed empirically in (Liu et al., 2024). The additional experiments conducted in Appendix D.5 show that our bounds remain valid for large values of d .

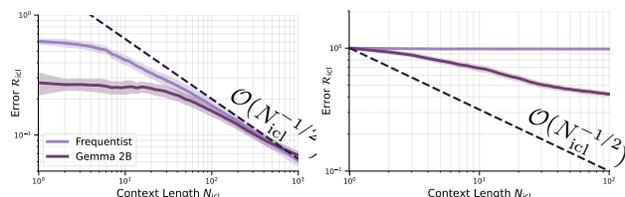


Figure 7: **Impact of the number of states.** We plot the risks \mathcal{R}_{icl} as functions of N_{icl} for Gemma 2B and the frequentist approach (Wolfer & Kontorovich, 2019) with 95% confidence intervals. **Left.** The input sequence is a random 3-state Markov chain. **Right.** The input sequence is a Brownian motion discretized as a 700-state Markov chain, similarly to Liu et al. (2024).

6. Conclusion

This paper proposed an explicit characterization of the inference mechanism in large language models through an equivalent finite-state Markov chain. We provided an insightful theoretical analysis based on the established characterization and the ability of the LLM to infer the transition kernel approximating the true transition probabilities of language. Experiments on commonly used LLMs validate our findings. We adapted our results to in-context learning where experiments confirm our theoretical insights. In the future, we hope that the proposed equivalence will have far-reaching implications on our understanding of LLMs and allow for a more fine-grained understanding of their expressiveness.

Impact Statement

This paper presents work that aims to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which must be specifically highlighted here.

Acknowledgements

The authors would like to thank Alexander Hägele for insightful feedback on the experiments with the large language models as well as Alain Durmus, Vasilii Feofanov, Giuseppe Paolo, Albert Thomas, Malik Tiomoko and Aladin Virmaux for fruitful discussions on early versions of this work. Nicolas Boullé was supported by the Office of Naval Research (ONR), under grant N00014-23-1-2729.

References

- Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- Agarwal, A., Kakade, S., Krishnamurthy, A., and Sun, W. Flambe: Structural complexity and representation learning of low rank mdps. In *Advances in Neural Information Processing Systems*, volume 33, pp. 20095–20107, 2020.
- Ali, M., Fromm, M., Thellmann, K., et al. Tokenizer Choice For LLM Training: Negligible or Crucial? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 3907–3924. Association for Computational Linguistics, 2024.
- Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., and Hauth, A. e. a. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- Bach, F. *Learning Theory from First Principles*. MIT Press, 2024.
- Bietti, A., Cabannes, V., Bouchacourt, D., Jegou, H., and Bottou, L. Birth of a Transformer: A Memory Viewpoint. In *Advances on Neural Information Processing Systems*, 2023.
- Blondel, M., Martins, A., and Niculae, V. Learning classifiers with fenchel-young losses: Generalized entropies, margins, and algorithms. In Chaudhuri, K. and Sugiyama, M. (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 606–615. PMLR, 16–18 Apr 2019. URL <https://proceedings.mlr.press/v89/blondell19a.html>.
- Brown, T., Mann, B., Ryder, N., et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020.
- Cabannes, V., Dohmatob, E., and Bietti, A. Scaling Laws for Associative Memories. In *International Conference on Learning Representations*, 2024.
- Chen, H. and Ding, N. Probing the “creativity” of large language models: Can models produce divergent semantic association? In *EMNLP*, pp. 12881–12888. ACL, 2023.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T. (eds.), *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, 2019.
- Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- Edelman, B. L., Goel, S., Kakade, S., and Zhang, C. Inductive biases and variable creation in self-attention mechanisms. In *International Conference on Machine Learning*, pp. 5793–5831. PMLR, 2022.
- Edelman, B. L., Edelman, E., Goel, S., Malach, E., and Tsilivis, N. The evolution of statistical induction heads: In-context learning markov chains. *arXiv preprint arXiv:2402.11004*, 2024.
- Folland, G. B. *Real analysis: modern techniques and their applications*, volume 40. John Wiley & Sons, 1999.
- Furuya, T., de Hoop, M. V., and Peyré, G. Transformers are universal in-context learners. *arXiv preprint arXiv:2408.01367*, 2024.

- Gallager, R. G. Finite state Markov chains. In *Discrete Stochastic Processes*, pp. 103–147. Springer, 1996.
- Golkar, S., Pettee, M., Eickenberg, M., Bietti, A., Cranmer, M., Krawezik, G., Lanasse, F., McCabe, M., Ohana, R., Parker, L., et al. xval: A continuous number encoding for large language models. *arXiv preprint arXiv:2310.02989*, 2023.
- Gruver, N., Finzi, M., Qiu, S., and Wilson, A. G. Large language models are zero-shot time series forecasters. *Advances in Neural Information Processing Systems*, 36, 2023.
- Hao, Y., Orlitsky, A., and Pichapati, V. On learning markov chains. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/d34ab169b70c9dcd35e62896010cd9ff-Paper.pdf.
- Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL <https://openreview.net/forum?id=d7KBjmI3GmQ>.
- Hinton, G. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531*, 2015.
- Hu, X., Zhang, F., Chen, S., and Yang, Z. Unveiling the Statistical Foundations of Chain-of-Thought Prompting Methods. *arXiv preprint arXiv:2408.14511*, 2024.
- Ildiz, M. E., Huang, Y., Li, Y., Rawat, A. S., and Oymak, S. From Self-Attention to Markov Models: Unveiling the Dynamics of Generative Transformers. *arXiv preprint arXiv:2402.13512*, 2024.
- Ivgi, M., Yoran, O., Berant, J., and Geva, M. From loops to oops: Fallback behaviors of language models under uncertainty, 2024. URL <https://arxiv.org/abs/2407.06071>.
- Jeon, H. J., Lee, J. D., Lei, Q., and Van Roy, B. An Information-Theoretic Analysis of In-Context Learning. In *International Conference on Machine Learning*, volume 235, pp. 21522–21554, 2024.
- Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.
- Jurafsky, D. and Martin, J. H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. Pearson, 3rd edition, 2024.
- Karpathy, A. minGPT: A minimal PyTorch reimplementation of the GPT (Generative Pretrained Transformer). <https://github.com/karpathy/minGPT>, 2023. GitHub repository.
- Kim, J., Nakamaki, T., and Suzuki, T. Transformers are Minimax Optimal Nonparametric In-Context Learners. *arXiv preprint arXiv:2408.12186*, 2024.
- Li, Y., Ildiz, M. E., Papailiopoulos, D., and Oymak, S. Transformers as algorithms: Generalization and stability in in-context learning. In *International Conference on Machine Learning*, pp. 19565–19594. PMLR, 2023.
- Liu, T. J., Boullé, N., Sarfati, R., and Earls, C. J. LLMs learn governing principles of dynamical systems, revealing an in-context neural scaling law. *arXiv preprint arXiv:2402.00795*, 2024.
- Lotfi, S., Finzi, M., Kuang, Y., Rudner, T. G., Goldblum, M., and Wilson, A. G. Non-vacuous generalization bounds for large language models. *arXiv preprint arXiv:2312.17173*, 2023.
- Lotfi, S., Kuang, Y., Amos, B., Goldblum, M., Finzi, M., and Wilson, A. G. Unlocking tokens as data points for generalization bounds on larger language models. *arXiv preprint arXiv:2407.18158*, 2024.
- Makkuva, A. V., Bondaschi, M., Girish, A., Nagle, A., Jaggi, M., Kim, H., and Gastpar, M. Attention with Markov: A framework for principled analysis of transformers via markov chains. *arXiv preprint arXiv:2402.04161*, 2024.
- Marion, P. Generalization bounds for neural ordinary differential equations and deep residual networks. In *Advances on Neural Information Processing Systems*, volume 36, 2023.
- Marton, K. Measure concentration for Euclidean distance in the case of dependent random variables. *Ann. Probab.*, 32(3):2526–2544, 2004.
- Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M. S., and Love, J. e. a. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: an imperative style,

- high-performance deep learning library. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2019. Curran Associates Inc.
- Paulin, D. Concentration inequalities for Markov chains by Marton couplings and spectral methods. *Electron. J. Probab.*, 20:79, 2015.
- Peeperkorn, M., Kouwenhoven, T., Brown, D., and Jordanous, A. Is temperature the creativity parameter of large language models?, 2024. URL <https://arxiv.org/abs/2405.00492>.
- Redko, I., Habrard, A., Morvant, E., Sebban, M., and Benani, Y. State of the Art of Statistical Learning Theory. In *Advances in Domain Adaptation Theory*, pp. 1–19. Elsevier, 2019.
- Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., and et al., J. F. Gemma 2: Improving open language models at a practical size, 2024. URL <https://arxiv.org/abs/2408.00118>.
- Roberts, A., Raffel, C., and Shazeer, N. How much knowledge can you pack into the parameters of a language model? In *Empirical Methods in Natural Language Processing*, pp. 5418–5426. Association for Computational Linguistics, 2020.
- Roberts, G. O. and Rosenthal, J. S. General state space Markov chains and MCMC algorithms. *Probab. Surveys*, 1:20 – 71, 2004.
- Sennrich, R., Haddow, B., and Birch, A. Neural Machine Translation of Rare Words with Subword Units. In Erk, K. and Smith, N. A. (eds.), *Association for Computational Linguistics*, pp. 1715–1725, 2016.
- Singh, A. K. and Strouse, D. Tokenization counts: the impact of tokenization on arithmetic in frontier llms. *arXiv preprint arXiv:2402.14903*, 2024.
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.
- Tsybakov, A. B. *Introduction to Nonparametric Estimation*. Springer, 1st edition, 2008.
- Vapnik, V. N. An overview of statistical learning theory. *IEEE Trans. Neural Netw.*, 10(5):988–999, 1999.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.
- Veličković, P., Perivolaropoulos, C., Barbero, F., and Pascanu, R. softmax is not enough (for sharp out-of-distribution), 2024. URL <https://arxiv.org/abs/2410.01104>.
- Wies, N., Levine, Y., and Shashua, A. The learnability of in-context learning. In *Advances in Neural Information Processing Systems*, volume 36, 2024.
- Wolfer, G. and Kontorovich, A. Minimax learning of ergodic Markov chains. In *Algorithmic Learning Theory*, pp. 904–930. PMLR, 2019.
- Wolfer, G. and Kontorovich, A. Learning and identity testing of Markov chains. In *Handbook of Statistics*, volume 49, pp. 85–102. Elsevier, 2023.
- Wu, Z., Qi, Q., Zhuang, Z., Sun, H., and Wang, J. Pre-tokenization of numbers for large language models. In *The Second Tiny Papers Track at ICLR 2024*, 2024.
- Xie, R., Odonnat, A., Feofanov, V., Deng, W., Zhang, J., and An, B. Mano: Exploiting matrix norm for unsupervised accuracy estimation under distribution shifts, 2024. URL <https://arxiv.org/abs/2405.18979>.
- Xie, S. M., Raghunathan, A., Liang, P., and Ma, T. An Explanation of In-context Learning as Implicit Bayesian Inference. In *International Conference on Learning Representations*, 2022.
- Zhang, B. and Sennrich, R. Root Mean Square Layer Normalization. In *Advances in Neural Information Processing Systems*, volume 32, 2019a.
- Zhang, B. and Sennrich, R. Root mean square layer normalization. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019b. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/1e8a19426224ca89e83cef47f1e7f53b-Paper.pdf.
- Zhang, W., Deng, Y., Liu, B., Pan, S. J., and Bing, L. Sentiment analysis in the era of large language models: A reality check. *arXiv preprint arXiv:2305.15005*, 2023a.

Zhang, Y., Zhang, F., Yang, Z., and Wang, Z. What and how does in-context learning learn? Bayesian model averaging, parameterization, and generalization. *arXiv preprint arXiv:2305.19420*, 2023b.

Appendix

Roadmap. In Appendix A, we first recall our notations. We provide additional details on large language models and transformers in Appendix B. Important notions and definitions related to Markov chains and Marton couplings are given in Appendix C. The detailed proofs of our theoretical results are given in Appendix F. Finally, we provide additional experiments in Appendix D.

Table of Contents

A	Notations	14
B	Background on Large Language Models	14
B.1	Token Generation and Deletion Process	14
B.2	Autoregressive Transformer-Based LLMs	15
C	Background on Markov Chains	16
C.1	Basic Notions	16
C.2	Ergodic Unichains	18
C.3	Marton Couplings	18
C.4	An (Almost) Distance between Markov Chains	19
D	Additional Experiments	20
D.1	Experimental Setup and Tokenization	20
D.2	Impact of the Number of States d	21
D.3	More Structured Markov Chains	21
D.4	Recent Models: Impact of the Tokenization	25
D.5	Dynamical Systems	26
E	Additional Theoretical Results	26
E.1	Sample Complexity	26
E.2	Depth-Dependent Generalization Bounds	29
E.3	Generalization Bounds with the KL Divergence	29
F	Proofs	31
F.1	Proof of Proposition 3.2	31
F.2	Proof of Proposition 3.3	32
F.3	Proof of Proposition 3.4	33
F.4	Proof of Proposition 4.1	34
F.5	Proof of Theorem 4.2	35
F.6	Proof of Corollary E.3	43
F.7	Proof of Theorem 4.3	46

A. Notations

We denote $\{1, \dots, N\}$ as $[N]$. We represent scalar values with regular letters (e.g., parameter λ), vectors with bold lowercase letters (e.g., vector \mathbf{x}), and matrices with bold capital letters (e.g., matrix \mathbf{A}). The i -th row of the matrix \mathbf{A} is denoted by \mathbf{A}_i , its j -th column is denoted by $\mathbf{A}_{.j}$ and its transpose is denoted by \mathbf{A}^\top . The identity matrix of size n is denoted by $\mathbf{I}_n \in \mathbb{R}^{n \times n}$. The vector of size n with each entry equal to 1 is denoted by $\mathbb{1}_n$. We denote by $\|\mathbf{A}\|_{p,q}$ the $L_{p,q}$ matrix norm where the p -norm is over columns and the q -norm is over rows. We denote by $\|\mathbf{A}\|$ the operator norm of \mathbf{A} induced by the ℓ_2 norm and by $\|\mathbf{A}\|_\infty = \max_{i,j} |\mathbf{A}_{ij}|$ the operator norm induced by the ℓ_∞ -norm. Similarly, \mathbf{x}^\top is the transpose of the vector \mathbf{x} and $\|\mathbf{x}\|_p$ is its ℓ_p -norm. The total variation between two probability distributions \mathbb{P}, \mathbb{Q} is denoted by $d_{\text{TV}}(\mathbb{P}, \mathbb{Q})$. The term ‘‘almost surely’’ is denoted by the notation ‘‘a.s.’’ while the term *random variable* is denoted by the notation ‘‘r.v.’’. $\Delta_n := \{\mathbf{p} \in [0, 1]^n \mid \sum_{i=1}^n p_i = 1\}$ is the probability simplex of \mathbb{R}^n .

B. Background on Large Language Models

We first recall important notions regarding large language models before focusing on the most widely used ones, namely the transformer-based LLMs. We describe the components of the vanilla transformer architecture before describing the whole network at the heart of such a model and formally defining the class of parameters and neural networks considered in our work.

B.1. Token Generation and Deletion Process

In this section, we recall how the sequences of tokens are processed by the large language model notably regarding the next token generation and the deletion process.

Definition B.1 (Generation process). *Given an input $s \in \mathcal{V}_K^*$ of size p , an large language model outputs a probability mass function $f_{\Theta}^{T,K}(s)$ over the discrete vocabulary space. A next token x is then sampled from $f_{\Theta}^{T,K}(s)$, to construct a new sequence (s, x) of size $p + 1$.*

Generation can be repeated by considering (s, x) as new input sequence and iterating this process. Since these models are designed to handle only sequences of size at most K , a deletion process is required.

Definition B.2 (Deletion process). *Given an input s of size $p > K$, an large language model outputs a probability mass function $f_{\Theta}^{T,K}(s_K)$ where s_K is a truncation of K tokens of the sequence s . Large language models implement **front truncation**, which is done by setting s_K as the last K tokens of s .*

As shown in Fig. 8, only the last K tokens of a long input sequence are used. This is why we speak of *deletion*, since we ignore the first tokens.

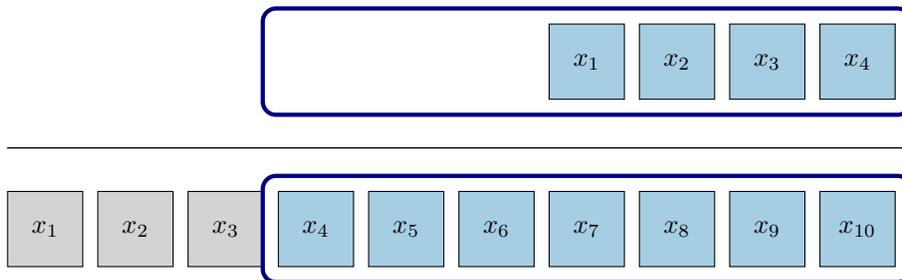


Figure 8: **Deletion process, front truncation.** A large language model with context window $K = 7$ in navy blue, processing sequences of different lengths. **Top.** A sequence of length 4. **Bottom.** Front truncation of a sequence of length 10.

Note that it is possible to implement other kinds of truncation, but large language models usually do not (Brown et al., 2020; Touvron et al., 2023a), however, in models like BERT (Devlin et al., 2019), which are not autoregressive, back truncation as described in Fig. 9 is also an option.

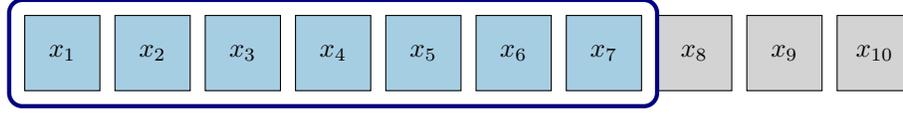


Figure 9: **Back truncation.** A large language model with context window $K = 7$ in navy blue, processing back truncation of a sequence of a sequence of length 10.

B.2. Autoregressive Transformer-Based LLMs

The most popular autoregressive LLMs rely on the transformer architecture (Vaswani et al., 2017) which we describe below following (Brown et al., 2020; Edelman et al., 2022; Zhang et al., 2023b). An autoregressive transformer-based LLM takes as input a sequence of length n , with $n \leq K$ and K is the context window, tokens with values in a vocabulary \mathcal{V} of size T . The tokens are embedded into a r -dimensional space and the input can be written as $\mathbf{S} \in \mathbb{R}^{r \times n}$. We consider a transformer model with L layers and h heads. The output of the ℓ -th layer writes $\mathbf{S}^{(\ell)}$ and is fed as input of the $(\ell + 1)$ -th layer. The input of the whole model is $\mathbf{S}^{(0)} = \mathbf{S}$. Below, we describe the operations performed by the model, including the embeddings of the tokens.

- **Token embeddings.** The tokens are embedded in a r -dimensional space via an embedding layer \mathbf{W}_E which results in an input of the form $\mathbf{S}^{r \times n}$;
- **Positional embeddings.** (Learnable) positional embeddings are added to each token depending on its position in the input sequence. This breaks the permutation-invariance of the transformer architecture and leads, by abuse of notation, to an output $\mathbf{S} \in \mathbb{R}^{r \times n}$;
- **Multi-head attention (MHA).** Given an input sequence $\mathbf{S} \in \mathbb{R}^{r \times n}$, query, key, and value matrices $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{r \times r}$ (here the value and output matrices are merged for ease of notations), the self-attention module computes

$$\mathcal{A}(\mathbf{S}; \mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V) := \text{softmax}\left(\frac{\mathbf{W}_Q \mathbf{S} (\mathbf{W}_K \mathbf{S})^\top}{\sqrt{r}}\right) (\mathbf{W}_V \mathbf{S}) \in \mathbb{R}^{r \times n},$$

with softmax: $\mathbf{x} \in \mathbb{R}^n \rightarrow \exp(\mathbf{x}) / \sum_i \exp(\mathbf{x})_i \in \Delta_n$. The operation described below corresponds to single-head self-attention. In practice, multi-head attention (MHA) is used with H heads and the query and key matrices are in $\mathbb{R}^{\frac{r}{H} \times \frac{r}{H}}$ and the value matrix is in $\mathbb{R}^{\frac{r}{H} \times r}$ (r, H are taken such that $\frac{r}{H}$ is an integer). The MHA module concatenates on the row dimension the outputs of \mathcal{A} for each head and then projects it back to the embedding dimension r with an output matrix $\mathbf{W}_O \in \mathbb{R}^{r \times r}$. By abuse of notation, we also denote by \mathcal{A} this operation which results in an output of dimension $r \times n$, and we include the output matrix in the argument of the operator. The ℓ -th layer of the transformer applies attention with layer-specific weight matrices and a residual connection that leads to an output

$$\mathbf{Z}^{(\ell)} = \mathbf{S}^{(\ell-1)} + \mathcal{A}\left(\mathbf{S}^{(\ell-1)}; \mathbf{W}_Q^{(\ell)}, \mathbf{W}_K^{(\ell)}, \mathbf{W}_V^{(\ell)}, \mathbf{W}_O^{(\ell)}\right).$$

This is followed by a layer normalization (Zhang & Sennrich, 2019a) that projects each token into the ℓ_2 -unit ball, i.e., each column $\mathbf{S}_{:,n}^{(\ell)}$ has an ℓ_2 -norm lower than 1;

- **Feed-forward block (FF).** Finally, a feed-forward block is applied, consisting of a two-layer MLP with hidden dimension m , layer-specific weight matrices $\mathbf{W}_1^{(\ell)} \in \mathbb{R}^{m \times r}$, $\mathbf{W}_2^{(\ell)} \in \mathbb{R}^{r \times m}$ and ReLU activation denoted by $\text{ReLU}(x) = \max\{0, x\}$ and applied entry-wise. The output of this layer reads

$$\mathbf{Y}^{(\ell)} = \mathbf{W}_2^{(\ell)} \text{ReLU}\left(\mathbf{W}_1^{(\ell)} \mathbf{Z}^{(\ell)}\right).$$

It is followed by a residual connection to produce the output

$$\mathbf{S}^{(\ell)} = \mathbf{Z}^{(\ell)} + \mathbf{W}_2^{(\ell)} \text{ReLU}\left(\mathbf{W}_1^{(\ell)} \mathbf{Z}^{(\ell)}\right) \in \mathbb{R}^{r \times n},$$

on which layer normalization (Zhang & Sennrich, 2019a) is applied ensuring that each column $\mathbf{S}_{:,n}^{(\ell)}$ has an ℓ_2 -norm lower than 1.

- **softmax output layer.** In the autoregressive setting, the model outputs a probability distribution on the vocabulary \mathcal{V} . To that end, the output $\mathbf{S}^{(L)} \in \mathbb{R}^{r \times n}$ of the final layer is projected back to the vocabulary size by an “unembedding layer” $\mathbf{W}_U \in \mathbb{R}^{T \times r}$ and averaged over the columns to obtain a vector in \mathbb{R}^T . A softmax layer is finally applied on top of it to obtain the probability distribution of the next token $\mathbb{P}_{\Theta}(\cdot | \mathbf{S})$. Formally, we have

$$\mathbb{P}_{\Theta}(\cdot | \mathbf{S}) = \text{softmax}\left(\frac{1}{n\tau} \mathbf{W}_U \mathbf{S}^{(L)} \mathbb{1}_n\right) \in \Delta_T,$$

n is the length (i.e., number of columns) of the input sequence \mathbf{S} (and thus of the last layer output $\mathbf{S}^{(L)}$), Θ denotes the parameters of the whole network that subsume the parameters of each layer and each block and τ is the softmax temperature (Hinton, 2015).

Theory faithful to the practice. The architecture described above is used in most of the transformer-based autoregressive LLM (Anil et al., 2023; Brown et al., 2020; Dubey et al., 2024; Jiang et al., 2023). In the theoretical analysis of Section 4, and unless specified otherwise, we remain faithful to their practical implementation and only make the following mild assumption: we assume that the unembedding layer is bounded. The class of parameters and the class of neural networks it generates respectively writes

$$\mathcal{W} = \{\Theta \mid \|\mathbf{W}_U^T\|_{2,1} \leq B_U\} \quad \text{and} \quad \mathcal{F} = \{f_{\Theta} \mid \Theta \in \mathcal{W}\}.$$

It should be noted that this assumption is significantly weaker than what is usually done in the literature (Edelman et al., 2022; Zhang et al., 2023b).

C. Background on Markov Chains

We recall below some important notions related to Markov chains based on (Paulin, 2015; Roberts & Rosenthal, 2004) and that will be used in our proofs.

C.1. Basic Notions

Consider two distribution probabilities \mathbb{P} and \mathbb{Q} defined on a measurable space (Ω, \mathcal{F}) .

Definition C.1. *The total variation between \mathbb{P} and \mathbb{Q} is defined as*

$$d_{\text{TV}}(\mathbb{P}, \mathbb{Q}) := \sup_{A \in \mathcal{F}} |\mathbb{P}(A) - \mathbb{Q}(A)|.$$

In the setting considered in the main paper, we consider Markov chains with finite discrete state space Ω . In this section, we refer to Ω as a general Polish space, whose elements are referred to as *states*.

Informally, a discrete-time, time-homogeneous Markov chain with state space Ω is a sequence of random variables $(\mathbf{X}_1, \mathbf{X}_2, \dots)$ taking values in Ω such that the next observation is independent of the past given the present. This property is referred to as the Markov property and is defined below.

Definition C.2. *A sequence of random variables $(\mathbf{X}_1, \mathbf{X}_2, \dots)$ is said to satisfy the Markov property if for all $n \geq 1$ and any $(x_1, \dots, x_{n+1}) \in \Omega^{n+1}$*

$$\mathbb{P}(\mathbf{X}_{n+1} = x_{n+1} \mid \mathbf{X}_n = x_n, \dots, \mathbf{X}_1 = x_1) = \mathbb{P}(\mathbf{X}_{n+1} = x_{n+1} \mid \mathbf{X}_n = x_n).$$

To a given Markov chain, we associate its *transition kernel* $\mathbf{Q} : \Omega^2 \rightarrow [0, 1]$ which collects the transition probabilities from one state to another

$$\forall n \in \mathbb{N}, (x, y) \in \Omega^2, \quad \mathbf{Q}(x, y) = \mathbb{P}(\mathbf{S}_{n+1} = y \mid \mathbf{S}_n = x).$$

In the main text, we refer to \mathbf{Q} as a transition *matrix* as the Markov chains we consider are of finite state space.

Definition C.3. A distribution π on Ω is said to be a stationary distribution if the action of the transition kernel leaves π unchanged, that is

$$(\pi\mathbf{Q})(A) := \int_{y \in A} \mathbf{Q}(x, y) d\pi(x) = \pi(A)$$

for all $A \in \mathcal{F}$.

A natural question is whether such a distribution exists for a generic Markov chain. Before stating an existence theorem, we introduce a classification of states below.

Class of states. All definitions below are borrowed from (Gallager, 1996)

Definition C.4 (Accessibility and communication). A state x is accessible from y (abbreviated as $x \rightarrow y$) if there exists $n > 0$ such that $\mathbf{Q}^n(x, y) > 0$. Two distinct states x and y communicate (abbreviated $x \leftrightarrow y$) if x is accessible from y and y is accessible from x .

Accessibility and communication concepts define how states can reach each other within a Markov chain. This leads to an important classification of states into transient and recurrent categories.

Definition C.5 (Recurrent and transient states). For finite-state Markov chains, a recurrent state is a state i that is accessible from all states that are accessible from i (i is recurrent if $i \rightarrow j$ implies that $j \rightarrow i$). A transient state is a state that is not recurrent.

With the distinction between recurrent and transient states established, we can now group states into classes based on their communication properties.

Definition C.6 (Class of states). A class \mathcal{C} of states is a non-empty set of states such that each $i \in \mathcal{C}$ communicates with every other state $j \in \mathcal{C}$ and communicates with no $j \notin \mathcal{C}$

Aperiodicity and Ergodicity. Aperiodicity ensures that the system does not exhibit cyclic behavior, which is a key condition for understanding the asymptotic behavior of states.

Definition C.7 (Aperiodicity). The period of a state i , denoted $d(i)$, is the greatest common divisor (gcd) of those values of n for which $\mathbf{Q}^n(i, i) > 0$. If the period is 1, the state is aperiodic.

Under some conditions on the Markov chain (aperiodicity and irreducibility (Roberts & Rosenthal, 2004)), it can be proven that the chain converges to its stationary distribution i.e. for any $x \in \Omega$, $\lim_{n \rightarrow \infty} d_{\text{TV}}(Q^n(x, \cdot), \pi) = 0$, where $Q^n(x, \cdot)$ denotes the probability of \mathbf{S}_n conditioned on $\mathbf{S}_1 = x$.

We recall below the notion of mixing time that assesses the time taken by the Markov chain to be ε -close to its stationary distribution (see Definition C.8).

Definition C.8 (Mixing time for time-homogeneous Markov chains (Paulin, 2015)). Let $X := (\mathbf{X}_1, \mathbf{X}_2, \dots)$ be a time-homogeneous Markov chain with a state space Ω , a transition kernel Q , and a stationary distribution π . Its mixing time is defined for any $\varepsilon \in [0, 1]$ as

$$t_{\text{mix}}(\varepsilon) := \min \{t \mid d(t) \leq \varepsilon\} \text{ where } d(t) := \sup_{x \in \Omega} d_{\text{TV}}(Q^t(x, \cdot), \pi).$$

We also introduce the quantity

$$t_{\text{min}} := \inf_{0 \leq \varepsilon < 1} t_{\text{mix}}\left(\frac{\varepsilon}{2}\right) \cdot \left(\frac{2 - \varepsilon}{1 - \varepsilon}\right)^2$$

which will be useful later on.

Remark C.1 (Well-posedness of t_{\min}). *As we only consider finite state-space Markov chains in our work, we know that a stationary distribution always exists. However, its uniqueness and the convergence to it require additional assumptions (see Appendix C.2). In particular, not all Markov chains admit a finite $t_{\text{mix}}(\varepsilon)$, t_{\min} for some $\varepsilon < \frac{1}{2}$. In such case, t_{\min} can be infinite. In our practical experimentation, this is never the case despite considering various Markov chains.*

C.2. Ergodic Unichains

We are now ready to state the following theorem, which formalizes the classification of states into recurrent, transient, and aperiodic classes.

Theorem C.9 (Recurrent and transient classes). *For finite state Markov chains, either all states in a class are transient or all are recurrent. We refer to these classes as transient and recurrent, respectively.*

For any Markov chain, all states in a class have the same period. If the period is 1, the class is said to be aperiodic

Having categorized states into recurrent, transient, and aperiodic classes, we can now define ergodicity.

Definition C.10 (Ergodicity). *For a finite-state Markov chain, an ergodic class of states is a class that is both recurrent and aperiodic. A Markov chain consisting entirely of one ergodic class is called an ergodic chain.*

Unichains. We now introduce the concept of unichains.

Definition C.11 (Unichains and ergodic unichains). *A unichain is a finite-state Markov chain containing a single recurrent class and transient states. An ergodic unichain is a unichain for which the recurrent class is ergodic.*

C.3. Marton Couplings

While we consider Markov chain inputs in Section 4.3, we consider less structured inputs during the pre-training phase Section 4.2.

More specifically, we model the sequences of tokens used during pre-training as generic dependent random variables. To derive meaningful results, we rely on the notion of Marton couplings introduced by [Marton \(2004\)](#). A Marton coupling can be seen as a weak dependency structure between random variables. The associated notion of the mixing matrix, analogous to the mixing time of a Markov chain, is used to assess the strength of the dependence between those variables.

This minimal modeling choice is made to remain as faithful as possible to the pre-training considered in practical applications of LLMs, for which the pre-training data is not public and may contain arbitrary data points such as video, code snippets, text and images ([Achiam et al., 2023](#); [Anil et al., 2023](#); [Brown et al., 2020](#); [Dubey et al., 2024](#); [Jiang et al., 2023](#); [Touvron et al., 2023a](#)).

As shown in [Paulin \(2015, Remark 2.2.\)](#), considering sequences of random variables linked through a Marton coupling is a weaker assumption than what is usually done in the literature on generalization bounds, which typically relies on independent random variables and Markov chains ([Hu et al., 2024](#); [Marion, 2023](#); [Wolfer & Kontorovich, 2019](#); [Zhang et al., 2023b](#)).

In particular, the results stated in Section 4.2 encompass the case where the pre-training input sequences of tokens are independent random variables ([Kim et al., 2024](#)) or Markov chains ([Zhang et al., 2023b](#)). We also note that Markov chains can model bigrams used in natural language ([Bietti et al., 2023](#); [Jurafsky & Martin, 2024](#)). We do not provide an exhaustive review of Marton couplings. We will simply recall its definition and introduce the associated mixing matrix. We refer the interested reader to [Marton \(2004\)](#) and [Paulin \(2015\)](#). Consider a sequence of dependent random variables $S = (\mathbf{S}_1, \dots, \mathbf{S}_N)$ taking values in a polish space $\Omega = \Omega_1 \times \dots \times \Omega_N$. We will denote by $\mathbb{P}(\mathbf{S}_1, \dots, \mathbf{S}_N)$ the distribution of S .

Definition C.12 (Marton coupling). We define a Marton coupling for S as a set of couplings

$$\left(S^{(s_1, \dots, s_i, s'_i)}, S'^{(s_1, \dots, s_i, s'_i)} \right) \in \Omega \times \Omega,$$

for every $i \in [N]$, every $s_1 \in \Omega_1, \dots, s_i \in \Omega_i, s'_i \in \Omega_i$, satisfying the following conditions.

$$(i) \quad \begin{aligned} \mathbf{S}_1^{(s_1, \dots, s_i, s'_i)} &= s_1, & \dots, & & \mathbf{S}_i^{(s_1, \dots, s_i, s'_i)} &= s_i, \\ \mathbf{S}'_1^{(s_1, \dots, s_i, s'_i)} &= s_1, & \dots, & & \mathbf{S}'_{i-1}^{(s_1, \dots, s_i, s'_i)} &= s_{i-1}, & \mathbf{S}'_i^{(s_1, \dots, s_i, s'_i)} &= s'_i. \end{aligned}$$

$$(ii) \quad \begin{aligned} &\left(\mathbf{S}_{i+1}^{(s_1, \dots, s_i, s'_i)}, \dots, \mathbf{S}_N^{(s_1, \dots, s_i, s'_i)} \right) \\ &\sim \mathbb{P}(\mathbf{S}_{i+1}, \dots, \mathbf{S}_N \mid \mathbf{S}_1 = s_1, \dots, \mathbf{S}_i = s_i), \\ &\left(\mathbf{S}'_{i+1}^{(x_1, \dots, x_i, x'_i)}, \dots, \mathbf{S}'_N^{(x_1, \dots, x_i, x'_i)} \right) \\ &\sim \mathbb{P}(\mathbf{S}_{i+1}, \dots, \mathbf{S}_N \mid \mathbf{S}_1 = x_1, \dots, \mathbf{S}_{i-1} = x_{i-1}, \mathbf{S}_i = x'_i). \end{aligned}$$

(iii) If $x_i = x'_i$, then $S^{(x_1, \dots, x_i, x'_i)} = S'^{(x_1, \dots, x_i, x'_i)}$.

Definition C.13 (Mixing matrix (Paulin, 2015)). For a Marton coupling, we define the mixing matrix $\Gamma \in \mathbb{R}^{N \times N}$ as an upper diagonal matrix with

$$\forall 1 \leq i < j \leq N, \quad \begin{cases} \Gamma_{i,i} := 1, \\ \Gamma_{j,i} := 0 \\ \Gamma_{i,j} := \sup_{s_1, \dots, s_i, s'_i} \mathbb{P} \left[\mathbf{S}_j^{(s_1, \dots, s_i, s'_i)} \neq \mathbf{S}'_j^{(s_1, \dots, s_i, s'_i)} \right]. \end{cases}$$

For independent random variables, one can define a Marton coupling with a mixing matrix equal to the identity (see Paulin, 2015, Remark 2.2). In particular, it means that for independent variables, we have the operator norm of the mixing matrix equal to 1, i.e., $\|\Gamma\| = 1$.

C.4. An (Almost) Distance between Markov Chains

In Theorem F.23, We state elementary properties of \mathcal{K} in the proposition below.

Proposition C.14 (Properties of \mathcal{K}). \mathcal{K} is an almost-distance between transition matrices in the sense that it satisfies the properties below:

1. **Non-negativity.** For any Θ_1, Θ_2 , $\mathcal{K}(\Theta_1, \Theta_2) \geq 0$.
2. **Almost sure positivity.** $\mathcal{K}(\Theta_1, \Theta_2) = 0 \iff \forall n \in [N], \mathbb{P}_{\Theta_1}(\cdot \mid \mathbf{S}_n) = \mathbb{P}_{\Theta_2}(\cdot \mid \mathbf{S}_n)$ a.s..
3. **Symmetry.** For any Θ_1, Θ_2 , $\mathcal{K}(\Theta_1, \Theta_2) = \mathcal{K}(\Theta_2, \Theta_1)$.
4. **Triangular inequality.** For any $\Theta_1, \Theta_2, \Theta_3$, $\mathcal{K}(\Theta_1, \Theta_3) \leq \mathcal{K}(\Theta_1, \Theta_2) + \mathcal{K}(\Theta_2, \Theta_3)$.

Proof of Proposition C.14. We first recall the following technical lemma.

Lemma C.15 (Proposition 2.16 in (Folland, 1999)). Let Y be a non-negative random variable defined on a probability space Ω with probability function \mathbb{P} . If $\mathbb{E}[Y] = 0$, then $Y = 0$ almost surely, i.e.,

$$\mathbb{P}(\{\omega \in \Omega \mid Y(\omega) = 0\}) = 1$$

The non-negativity and symmetry of \mathcal{K} directly come from the symmetry and non-negativity of the total variation distance. The triangular inequality follows from the fact that the total variation is a distance and that the expectation respects inequalities. For the almost positivity, consider Θ_1, Θ_2 such that $\mathcal{K}(\Theta_1, \Theta_2) = 0$. By non-negativity of all the terms in the sum, it means that for all $n \in [N]$, we have

$$\mathbb{E}_{\mathbf{S}_n} [d_{\text{TV}}(\mathbb{P}_{\Theta_1}(\cdot | \mathbf{S}_n), \mathbb{P}_{\Theta_2}(\cdot | \mathbf{S}_n))] = 0.$$

As the total variation is a distance, we know that the random variable under the expectation is non-negative. Applying Lemma C.15 leads to

$$d_{\text{TV}}(\mathbb{P}_{\Theta_1}(\cdot | \mathbf{S}_n), \mathbb{P}_{\Theta_2}(\cdot | \mathbf{S}_n)) = 0 \quad \text{almost surely.}$$

On the probability space, deprived of the set where the distance is non-zero (which is of null measure), the total variation is equal to zero and as a distance between probability distributions, it means that on this subset of the probability space, the probabilities are equal. Again, as the set on which they are not equal is of null measure, we have

$$\mathbb{P}_{\Theta_1}(\cdot | \mathbf{S}_n) = \mathbb{P}_{\Theta_2}(\cdot | \mathbf{S}_n) \quad \text{almost surely.}$$

Putting everything together, we have

$$\forall n \in [N], \mathbb{P}_{\Theta_1}(\cdot | \mathbf{S}_n) = \mathbb{P}_{\Theta_2}(\cdot | \mathbf{S}_n) \quad \text{a.s.,} \tag{5}$$

which concludes the direct sense. The converse sense is proved by assuming that Eq. (5) holds and using the distance properties of the total variation. This concludes the proof. \square

D. Additional Experiments

In this section, we present additional numerical experiments that confirm that our theory correctly predicts the in-context learning behavior of LLMs.

D.1. Experimental Setup and Tokenization

Experimental setup. To ensure a fair validation of our theoretical results, we conduct our experiments on some of the most recent and widely used LLMs: Gemma 2B (Mesnard et al., 2024), Llama2 7B & 13B (Touvron et al., 2023b), Llama3 8B, Llama3.2 1B & 3B (Dubey et al., 2024), and Mistral 7Bv0.1 (Jiang et al., 2023).

Tokenization. As the models we consider have different tokenizations, we need to do this step with extra care as it is a crucial part of the experimental procedure. Indeed, LLMs’ ability to handle numerical values has been proved to be dependent on the tokenization algorithm (Ali et al., 2024; Gruver et al., 2023; Singh & Strouse, 2024). The most widely used tokenization algorithm to-date, *BPE* (Sennrich et al., 2016), tends to assign tokens to arbitrary 3-digits numbers based on their occurrences in large-scale corpora, and the tokenizer’s vocabulary size. As highlighted by (Gruver et al., 2023), this artifact severely hinders LLMs’ ability to predict numerical values in-context. This is the case for popular LLMs such as GPT-3 (Brown et al., 2020). Newer models (LLama3, GPT-3.5, GPT-4) however, tend to have hard-coded rules on top of *BPE*, making them able to encode all 3-digits numbers with their own token. Although this feature would accelerate the ICL procedure by eliminating the need for the *Hierarchy-PDF* algorithm in (Liu et al., 2024), the under-representability of larger numbers in the training data could be an issue. Other tokenization techniques that are numerical values-focused has been presented in the literature (Golkar et al., 2023; Wu et al., 2024), paving the way for another research direction that may benefit our method.

Rodmap. In the rest of this section, we extend our experiments to study the following setups:

- In Appendix D.2: impact of the number of states d ;
- In Appendix D.3: extension to Markov chains with $p_{\min} = 0$;
- In Appendix D.4: impact of the tokenization;
- In Appendix D.5: extension to dynamical systems.

D.2. Impact of the Number of States d

We further analyze the effect of the number of states d on the risk and consider randomly generated d -state transition matrices in Fig. 10. After a first stage of stagnation, the risk tends to take the correct scaling law coefficient. As in (Liu et al., 2024), we notice that considering randomly generated transition matrices seems to be difficult for an LLM to learn when there are more than 9 states. We interpret this behavior as the distribution shift term in Theorem 4.3. Indeed, the lack of structure in these transition matrices can hinder the correct decay of this term. Note also that the increase in d tends to implicitly increase t_{\min} , which could have an impact on the upper bound on \mathcal{R}_{icl} (both in the generalization term and in the distribution shift term). We will now consider more structured Markov chains, and look at their impact on decay.

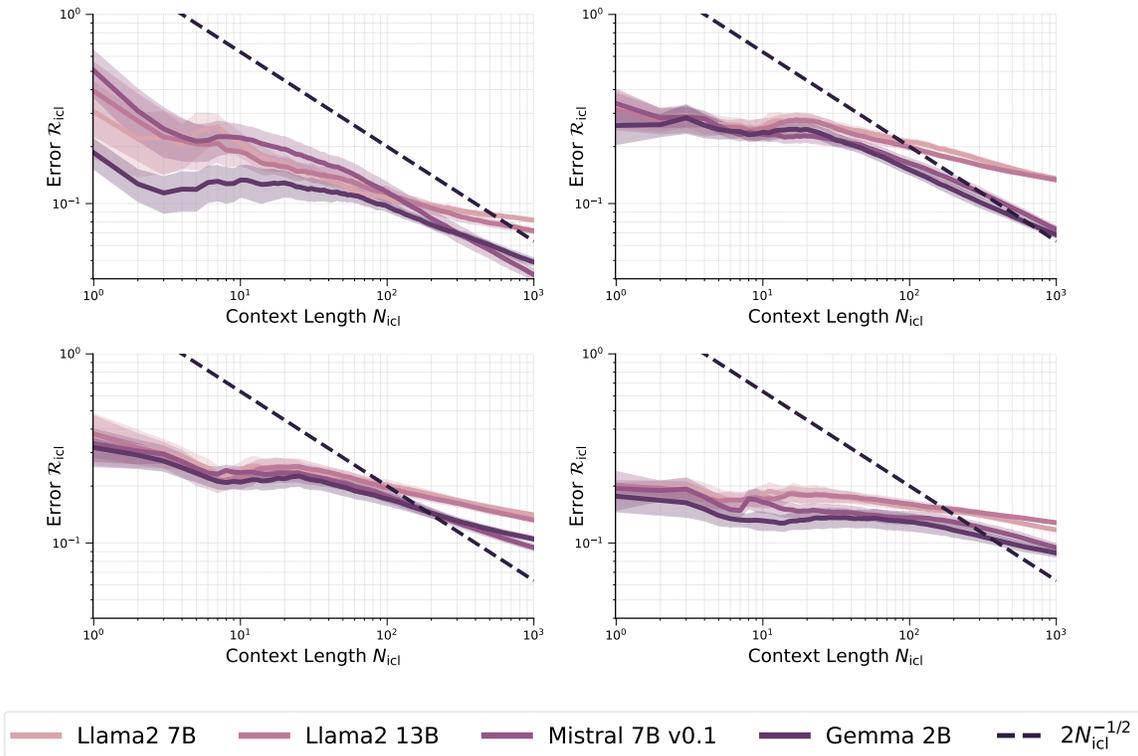


Figure 10: **Impact of the number of states d .** We plot the risk \mathcal{R}_{icl} as functions of N_{icl} , with 95% confidence intervals. **Upper Left.** 2–states Markov transition matrices. **Upper Right.** 4–states Markov transition matrices. **Lower Left.** 6–states Markov transition matrices. **Lower Right.** 8–states Markov transition matrices.

D.3. More Structured Markov Chains

In this section, we empirically verify our theoretical results on more general Markov chains that do not verify $p_{\min} > 0$.

D.3.1. RANDOM WALKS

Random walks are a simple example of more structured Markov chains. Although we still have the possibility of discretizing the kernel of Markov chains with infinite state spaces as it is done in (Liu et al., 2024), we consider two types of random walks on finite state spaces.

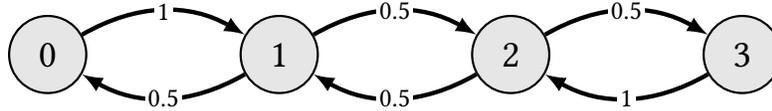


Figure 11: Constrained random walk with $d = 3$.

Constrained random walk. We define the transition matrix P of a constrained random walk of d states as in Eq. (6). We draw the probabilistic graph in Fig. 11 for the case $d = 3$.

$$P_{ij} = \begin{cases} 1, & \text{if } i = 0 \text{ and } j = 1, \\ 1, & \text{if } i = d - 1 \text{ and } j = d - 2, \\ 0.5, & \text{if } 1 \leq i \leq d - 2 \text{ and } j = i - 1, \\ 0.5, & \text{if } 1 \leq i \leq d - 2 \text{ and } j = i + 1, \\ 0, & \text{otherwise.} \end{cases} \quad (6)$$

Fig. 12 highlights the scaling laws of Theorem 4.3, as well as the $\log(d)$ dependency. As before, the best-performing models generalize almost perfectly.

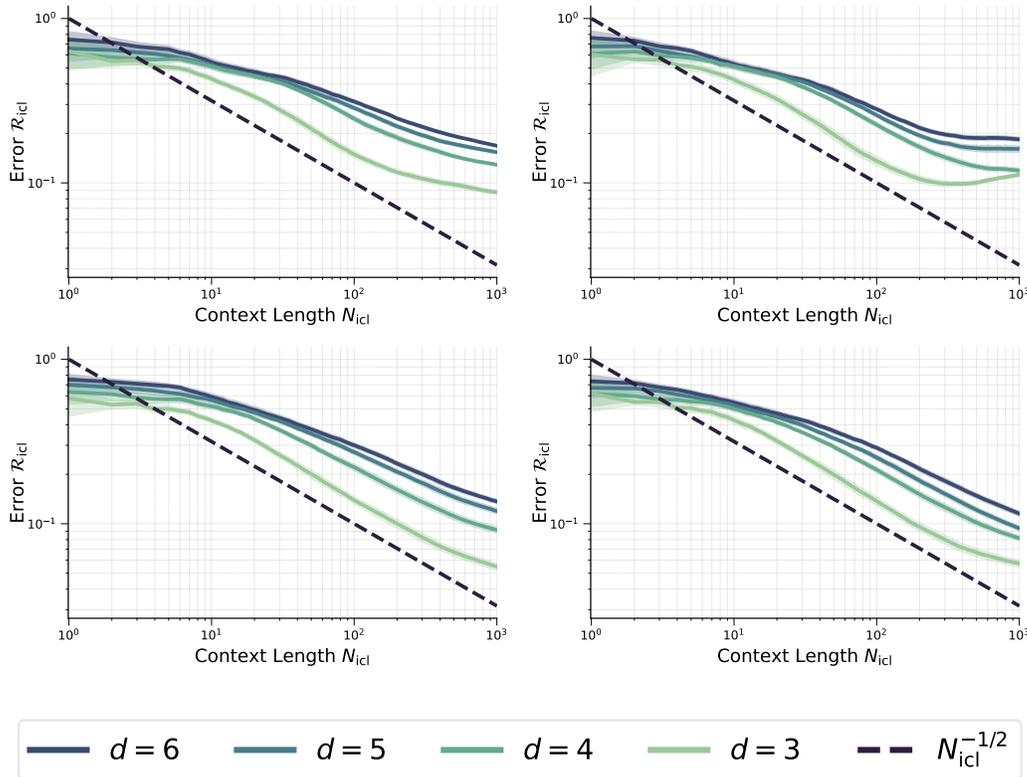


Figure 12: **Constrained random walks.** We plot the risk \mathcal{R}_{icl} as functions of N_{icl} , with 99% confidence intervals. We consider different size d . **Upper Left.** Llama2 7B **Upper Right.** Llama2 13B **Lower Left.** Mistral 7Bv0.1 **Lower Right.** Gemma 2B

Polygonal random walk. We define the transition matrix \mathbb{P} of a polygonal random walk of d states as in Eq. (7). We draw the probabilistic graph in Fig. 13 for the case $d = 4$.

$$P_{ij} = \begin{cases} 0.5, & \text{if } j = (i + 1) \pmod{d} \text{ (clockwise transition),} \\ 0.5, & \text{if } j = (i - 1) \pmod{d} \text{ (counterclockwise transition),} \\ 0, & \text{otherwise.} \end{cases} \quad (7)$$

We draw the same conclusions as above for this second type of random walk, in Fig. 14.

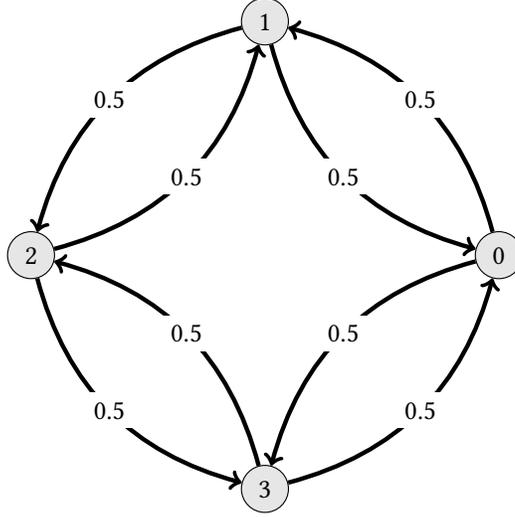


Figure 13: Polygonal random walk with $d = 4$.

D.3.2. INNER CLIQUES AND OUTER RIMS

Inner Cliques and Outer Rims. We also want to test our method on the class of Markov chain put forward in (Wolfer & Kontorovich, 2019) to derive their lower bound. Let $\eta > 0$ and $d = 3k$ for some $k \in \mathbb{N}$, and define the collection of Markov matrices $\mathcal{H}_\eta = \{M_{\eta, \tau} : \tau \in \{0, 1\}^{d/3}\}$. Every element of this set consists of an *inner clique* and an *outer rim*. $M_{\eta, \tau}$ is the block matrix defined as follows,

$$M_{\eta, \tau} = \begin{pmatrix} C_\eta & R_\tau \\ R_\tau^\top & L_\tau \end{pmatrix},$$

where $C_\eta \in \mathbb{R}^{d/3 \times d/3}$, $L_\tau \in \mathbb{R}^{2d/3 \times 2d/3}$, and $R_\tau \in \mathbb{R}^{d/3 \times 2d/3}$ are given by

$$L_\tau = \frac{1}{8} \text{diag} (7 - 4\tau_1\varepsilon, 7 + 4\tau_1\varepsilon, \dots, 7 - 4\tau_{d/3}\varepsilon, 7 + 4\tau_{d/3}\varepsilon),$$

$$C_\eta = \begin{pmatrix} \frac{3}{4} - \eta & \frac{\eta}{d/3-1} & \cdots & \frac{\eta}{d/3-1} \\ \frac{\eta}{d/3-1} & \frac{3}{4} - \eta & \ddots & \vdots \\ \vdots & \ddots & \ddots & \frac{\eta}{d/3-1} \\ \frac{\eta}{d/3-1} & \cdots & \frac{\eta}{d/3-1} & \frac{3}{4} - \eta \end{pmatrix},$$

$$R_\tau = \frac{1}{8} \begin{pmatrix} 1 + 4\tau_1\varepsilon & 1 - 4\tau_1\varepsilon & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & 0 & 1 + 4\tau_2\varepsilon & 1 - 4\tau_2\varepsilon & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & \cdots & \cdots & 0 & 1 + 4\tau_{d/3}\varepsilon & 1 - 4\tau_{d/3}\varepsilon \end{pmatrix}.$$

We provide in Fig. 15 a probabilistic graph of the case $M_{\eta, 0}$ and $d = 9$.

Fig. 16 compares different LLMs with the frequentist method, on the case depicted in Fig. 16 with $\eta = 0.02$. Although the frequentist method achieves a lower loss, the power laws seem to be the same with LLMs.

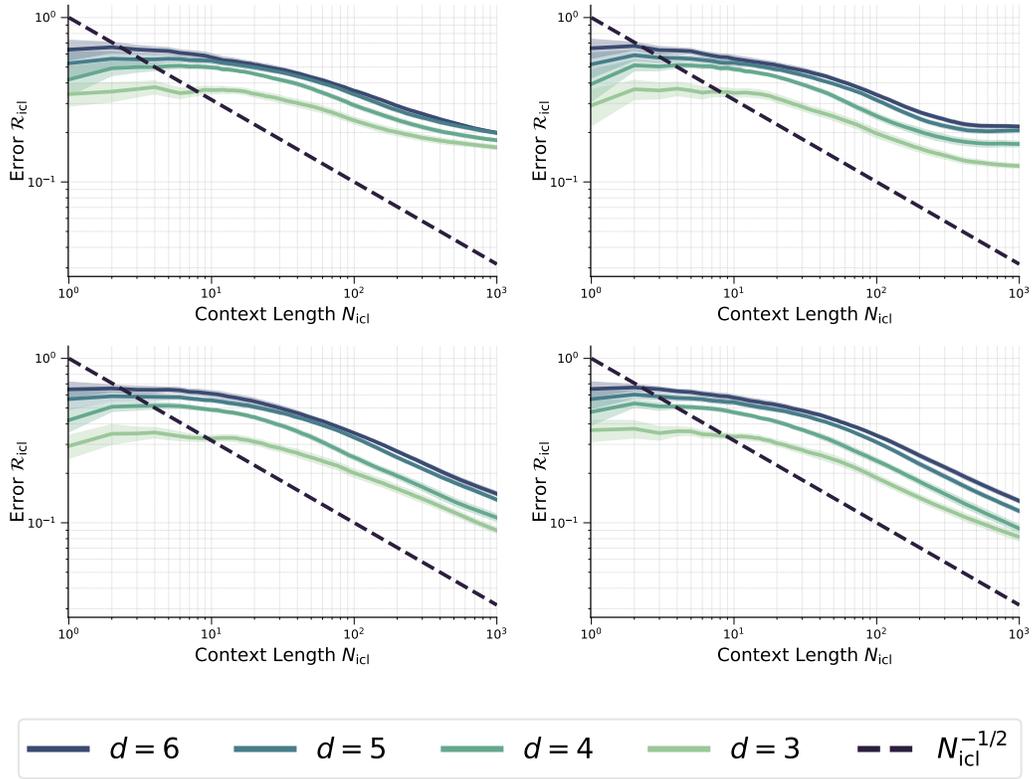


Figure 14: **Polygonal random walks.** We plot the risk \mathcal{R}_{icl} as functions of N_{icl} , with 99% confidence intervals. We consider different size d . **Upper Left.** Llama2 7B **Upper Right.** Llama2 13B **Lower Left.** Mistral 7Bv0.1 **Lower Right.** Gemma 2B

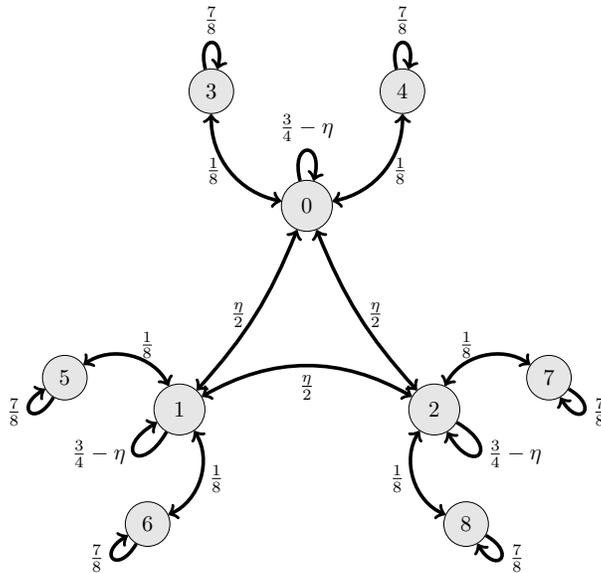


Figure 15: Probabilistic graph of $M_{\eta,0}$ when $d = 9$.

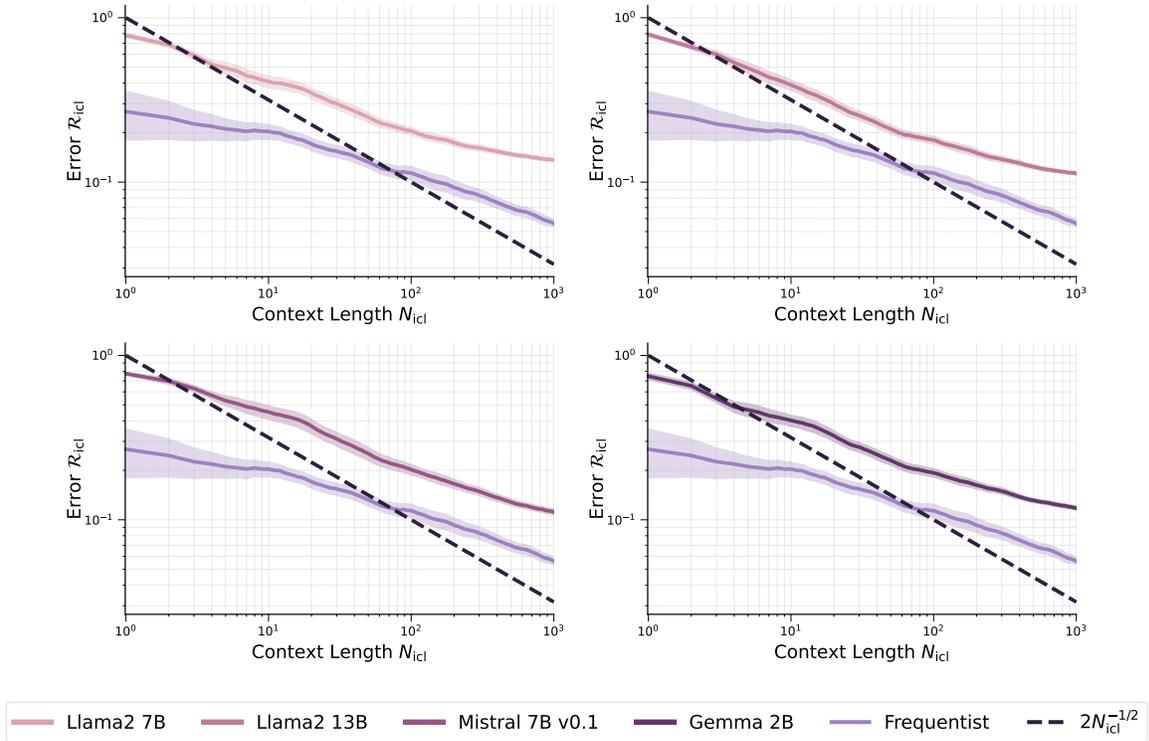


Figure 16: We plot the risk \mathcal{R}_{icl} as functions of N_{icl} , with 95% confidence intervals. **Upper Left.** Llama2 7B **Upper Right.** Llama2 13B **Lower Left.** Mistral 7Bv0.1 **Lower Right.** Gemma 2B

D.4. Recent Models: Impact of the Tokenization

As explained in Appendix D.1, models like Llama 3 tokenize 3-digit numbers with a single token. This saves a lot of inference compute time, but not necessarily in terms of performance when considering Markov chains with a few number of states d , since we have to separate the states by a comma to force tokenization into a single digit (e.g. the transitions $1 \rightarrow 0 \rightarrow 1$ will be prompted as $1, 0, 1$ (5 tokens) instead of 101 (1 token)). In Fig. 17, we reproduce the same experiment as in Fig. 6(left), but with Llama 3 models. The scaling laws are quite good, but much less so than those obtained with Gemma 2B and Mistral 7Bv0.1 on the same inputs. On the other hand, with these models, it can be extremely interesting to consider Markov chains with many states, as we did in Fig. 7(right). In the next section, we will use Llama3 to learn other dynamic systems presented in (Liu et al., 2024).

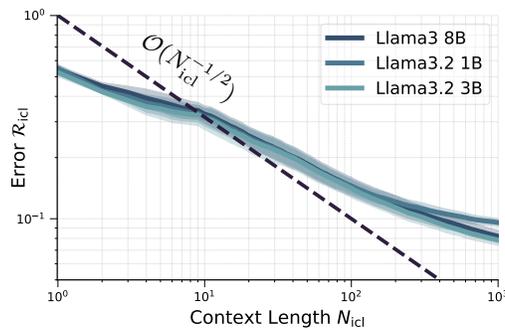


Figure 17: **In-context scaling laws for LLama3 herd of models.** We plot the risk \mathcal{R}_{icl} as functions of N_{icl} , with 95% confidence intervals.

D.5. Dynamical Systems

We consider four of the dynamic systems highlighted in (Liu et al., 2024) : a geometric Brownian motion, a correlated Gaussian, an uncorrelated Gaussian, and an uncorrelated uniform processe. We display in Fig. 18 the risks of LLama3 8B and the frequentist method, which once again highlights the emerging capacity of in-context learning.

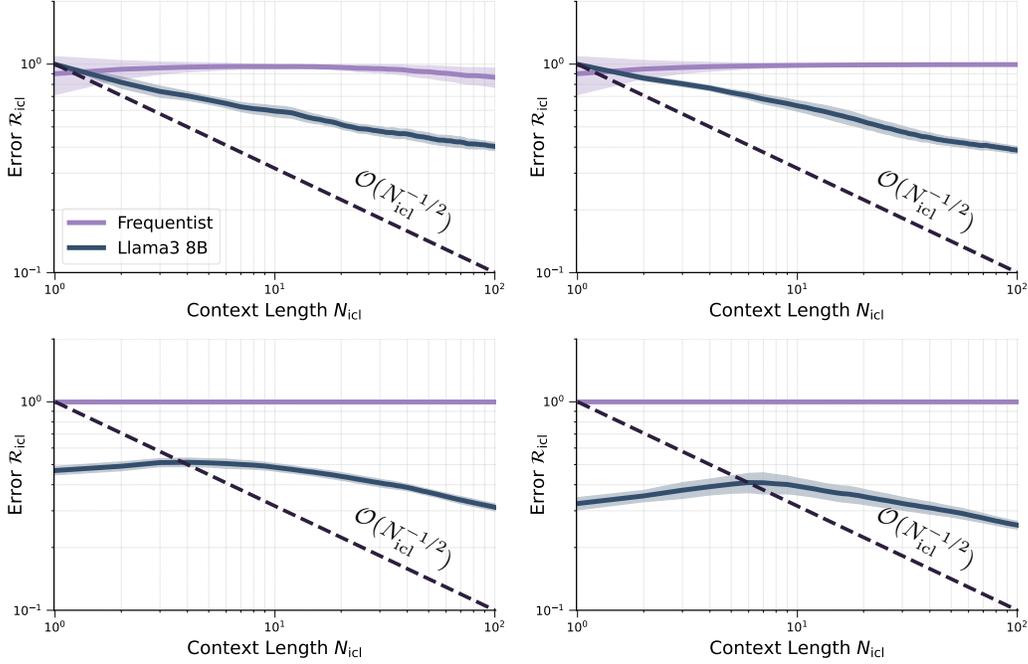


Figure 18: **LLama3 8B on dynamical systems.** We plot the risks \mathcal{R}_{icl} as functions of N_{icl} for LLama3 8B and the frequentist approach (Wolfer & Kontorovich, 2019) with 95% confidence intervals. **Upper Left.** Geometric Brownian motion. **Upper Right.** Correlated Gaussian. **Lower Left.** Uncorrelated Gaussian. **Lower Right.** Uncorrelated Uniform.

E. Additional Theoretical Results

In this section, we present additional theoretical results on the sample complexity and generalization capabilities of LLMs.

E.1. Sample Complexity

In this section, we provide more details to Section 4.1, including the experimental setting, the connections to the Markov chain literature, and the extension of Proposition 4.1 to the in-context learning setting.

E.1.1. PRE-TRAINING

We recall that Proposition 4.1 states

Proposition E.1 (Restatement of Proposition 4.1). *Let $\delta \in [0, 1]$ and $\epsilon > 0$. Assuming a perfect pre-training of f_{Θ} and N_{train} pre-training tokens with $N_{\text{train}} \geq N^* := \lceil \frac{4\bar{B}^2}{\epsilon^2} \log(\frac{2}{\delta}) \rceil$, we have with probability at least $1 - \delta$,*

$$\mathbb{E}_{\mathbf{S} \sim \mathbb{P}_{\mathcal{L}}} \|\mathbf{Q}^*(\mathbf{S}, \cdot) - \mathbf{Q}_f(\mathbf{S}, \cdot)\|_1 \leq \epsilon.$$

with a constant depending on the problem's parameters

$$\bar{B} = 2\|\mathbf{\Gamma}\| \max\{\log(T) + 2B_U/\tau, \log(1/c_0)\}^{1/2}.$$

To determine the approximation error of open-source LLM, it suffices to plug $N^* = N_{\text{train}}$ in Proposition 4.1. Then, the approximation error writes

$$\epsilon = \frac{2\bar{B}}{\sqrt{N_{\text{train}}}} \sqrt{\log\left(\frac{2}{\delta}\right)}.$$

We recall that $\bar{B} = 2\|\Gamma\| \sqrt{\max\{\log(T) + 2B_U/\tau, \log(1/c_0)\}}$. To numerically compute ϵ , we proceed to the following simplifications. Since we do not have access to the training data of the Gemmas and Llamas, we cannot compute Γ that captures the data dependency. Even if we had access to the data, the amount of it prevents us from numerically computing Γ for all the pre-training sequences. The same goes for c_0 . We circumvent these issues using the fact that T is very high in practice and compute $\bar{B} = \mathcal{O}(2\sqrt{\log(T) + 2B_U/\tau})$. Finally, we notice that the softmax temperature is of order 1 in those models (we note that it evolves in $[0.2 - 1]$ depending on the downstream tasks (Dubey et al., 2024)) and that $B_U \sim T\sqrt{r}$. This comes from the fact that B_U controls the norm of the unembedding matrix $\mathbf{W}_U \in \mathbb{R}^{T \times r}$ (see Appendix B), i.e.

$$\|\mathbf{W}_U^\top\|_{2,1} \leq B_U.$$

When looking at this norm for the unembedding layer of Llama3 8B, we observe that it behaves like $T\sqrt{r}$. This is not surprising since the layer normalizations project the outputs' columns on the unit-ball (notably for the RMSNorm (Zhang & Sennrich, 2019b) used in the Llamas models (Touvron et al., 2023b)) and \mathbf{W}_U is initialized following $\mathcal{N}(0, 1)$ (in PyTorch (Paszke et al., 2019)). This should lead the entries of \mathbf{W}_U to be in $[0, 1]$ at the end of training which would imply, using the fact that $\mathbf{W}_U^\top \in \mathbb{R}^{r \times T}$,

$$\|\mathbf{W}_U^\top\|_{2,1} = \sum_{j=1}^T \sqrt{\sum_{i=1}^r (\mathbf{W}_U^\top)_{ij}^2} = \sum_{j=1}^T \sqrt{\sum_{i=1}^r \underbrace{(\mathbf{W}_U)_{ji}^2}_{\leq 1}} \leq T\sqrt{r}.$$

In summary, we can indeed consider $B_U = T\sqrt{r}$ and we obtain

$$\epsilon = \frac{2\bar{B}}{\sqrt{N_{\text{train}}}} \sqrt{\log\left(\frac{2}{\delta}\right)}.$$

where

$$\bar{B} = 2\sqrt{\log(T) + 2T\sqrt{r}/\tau}.$$

In the technical reports of the Gemmas (Mesnard et al., 2024; Riviere et al., 2024) and Llamas (Dubey et al., 2024; Touvron et al., 2023a;b) models, the vocabulary sizes T , the embedding dimension r , the number of pre-training tokens are given and the MMLU results are given. We summarize it in Table 2. This enables us to plot the evolution of the MMLU with respect to the predicted approximation error ϵ in Fig. 1.

Model	Nb. Pre-Training Tokens N_{train}	Vocabulary Size T	Embedding Dimension r
Llama 7B (Touvron et al., 2023a)	10^{12}	32000	4096
Llama2 7B (Touvron et al., 2023b)	2×10^{12}	32000	4096
Llama3 8B (Dubey et al., 2024)	1.5×10^{13}	128000	4096
Llama3.2 3B (Dubey et al., 2024)	1.5×10^{13}	128000	3072
Gemma 2B (Mesnard et al., 2024)	3×10^{12}	256128	2048
Gemma 7B (Mesnard et al., 2024)	6×10^{12}	256128	3072
Gemma2 9B (Riviere et al., 2024)	8×10^{12}	256128	3584
Gemma2 27B (Riviere et al., 2024)	1.3×10^{13}	256128	4608

Table 2: LLMs' parameters reported in the technical reports of the Llamas and Gemmas models.

E.1.2. CONNECTION OF PROPOSITION 4.1 TO MARKOV CHAIN LITERATURE

Proposition 4.1 allows us to contextualize LLMs' ability to learn Markov chains with respect to the existing literature. To the best of our knowledge, the only existing approach with theoretical guarantees for learning Markov chains is the

frequentist method (Wolfer & Kontorovich, 2019): counting the number of occurrences of different states to fill in the matrix \mathbf{Q}_f . Wolfer & Kontorovich (2019) show that the sample complexity of approximating \mathbf{Q}^* up to ϵ with such approach is $N^* = \mathcal{O}(\max\{|\mathcal{V}_K^*|/\epsilon^2\gamma_s, 1/\gamma_s\pi^*\} \log(\frac{1}{\delta}))$ samples, where $|\mathcal{V}_K^*| = \mathcal{O}(T^K)$ in our setting, γ_s is a (pseudo) spectral gap of the Markov chain and π^* is the smallest element of its stationary distribution. The authors state that the frequentist approach is minimax optimal (up to logarithmic factors). Our sample complexity in Proposition E.1 has a dependence that behaves as $\bar{B}^2 = \mathcal{O}(\max\{\log(T) + \frac{2T\sqrt{r}}{\tau}, \log(1/c_0)\})$, using $B_U \sim T\sqrt{r}$. Given that in practice $T > r$, it then simplifies to $N^* = \mathcal{O}(\max\{T/\epsilon^2\tau, 1/\epsilon^2\} \log(\frac{1}{\delta}))$. In particular, our LLMs' sample complexity is linear in the vocabulary size T , which is remarkable compared to the sample complexity of the frequentist approach, which scales as $\mathcal{O}(T^K)$ with K the context window.

Extension to In-Context Learning. Using Theorem 4.3, we can extend Proposition 4.1 to the in-context learning framework. The following proposition gives the sample complexity of in-context learning.

Proposition E.2 (Sample complexity of in-context learning). *Let $\delta \in [0, 1]$ and $\epsilon > 0$. The model f_{Θ} receives as inputs a d -state Markov chain $X = (\mathbf{X}_1, \dots, \mathbf{X}_{N_{\text{icl}}})$ with transition matrix \mathbf{Q} . Assuming that there is no distribution shift and $N_{\text{icl}} \geq N^* := \lceil \frac{4\bar{B}^2}{\epsilon^2} \log(\frac{2}{\delta}) \rceil$, we have with probability at least $1 - \delta$,*

$$\mathbb{E}_{\mathbf{S} \sim \mathbb{P}_{\mathcal{L}}} \|\mathbf{Q}(\mathbf{S}, \cdot) - \mathbf{Q}_f(\mathbf{S}, \cdot)\|_1 \leq \epsilon.$$

with a constant depending on the problem's parameters

$$\bar{B} = 2 \max\{\log(d) + 2B_U/\tau, \log(1/p_{\min})\}^{1/2}.$$

Proof. We recall that the probability distribution associated with the input Markov chain is \mathbb{P} and its transition matrix writes \mathbb{Q} (see Appendix C for the connection between \mathbb{P} and \mathbb{Q}). We first note that by definition of the total variation distance (Wolfer & Kontorovich, 2019), we have

$$\begin{aligned} \mathbb{E}_{\mathbf{S} \sim \mathbb{P}} \|\mathbf{Q}(\mathbf{S}, \cdot) - \mathbf{Q}_f(\mathbf{S}, \cdot)\|_1 &= \mathbb{E}_{\mathbf{S} \sim \mathbb{P}} [2 \cdot d_{\text{TV}}(\mathbb{Q}(\mathbf{S}, \cdot), \mathbf{Q}_f(\mathbf{S}, \cdot))] \\ &= 2 \cdot \mathbb{E}_{\mathbf{S} \sim \mathbb{P}} [d_{\text{TV}}(\mathbb{Q}(\mathbf{S}, \cdot), \mathbf{Q}_f(\mathbf{S}, \cdot))] \\ &= 2 \cdot \mathcal{R}_{\text{icl}}(\Theta). \end{aligned} \quad (\text{by definition of the risk Eq. (3)})$$

Applying Theorem 4.3, we know that

$$\mathcal{R}_{\text{icl}}(\Theta) \leq \inf_{\vartheta \in \mathcal{W}_{\text{mc}}} \{\widehat{\mathcal{R}}_{\text{icl}}(\vartheta) + \mathcal{K}(\vartheta, \Theta)\} + \frac{\bar{B}}{\sqrt{N_{\text{icl}}}} \sqrt{\log\left(\frac{2}{\delta}\right)},$$

where \bar{B} is formally defined in Theorem 4.3. Assuming no distribution shift amounts to considering the infimum attained and equal to 0. We denote by N^* the integer such that the error is equal to $\frac{\epsilon}{2}$, i.e.,

$$\frac{\bar{B}}{\sqrt{N^*}} \sqrt{\log\left(\frac{2}{\delta}\right)} = \frac{\epsilon}{2} \iff \frac{\bar{B}^2}{N^*} \log\left(\frac{2}{\delta}\right) = \frac{\epsilon^2}{4} \iff N^* = \left(\frac{2\bar{B}}{\epsilon}\right)^2 \log\left(\frac{2}{\delta}\right).$$

Taking the ceiling function ensures that N^* is an integer. Hence, taking $N_{\text{train}} \geq N^* = \lceil \left(\frac{2\bar{B}}{\epsilon}\right)^2 \log\left(\frac{2}{\delta}\right) \rceil$ ensures that

$$\frac{\bar{B}}{\sqrt{N_{\text{icl}}}} \sqrt{\log\left(\frac{2}{\delta}\right)} \leq \frac{\bar{B}}{\sqrt{N^*}} \sqrt{\log\left(\frac{2}{\delta}\right)} = \frac{\epsilon}{2}.$$

Putting everything together, taking $N_{\text{icl}} \geq N^*$ leads to

$$\mathbb{E}_{\mathbf{S} \sim \mathbb{P}} \|\mathbf{Q}(\mathbf{S}, \cdot) - \mathbf{Q}_f(\mathbf{S}, \cdot)\|_1 \leq 2 \cdot \mathcal{R}_{\text{icl}}(\Theta) \leq 2 \cdot \frac{\epsilon}{2} = \epsilon,$$

which concludes the proof. \square

Similarly to the analysis above, we observe that our sample complexity of in-context learning scales logarithmically with the number of states d while the frequentist’s one (Wolfer & Kontorovich, 2019) scales in $\mathcal{O}(d)$. In Fig. 7 of Section 5, we show that this is confirmed experimentally: LLM’s ability to learn Markov chains exceeds the frequentist approach for Markov chains with a large state space.

E.2. Depth-Dependent Generalization Bounds

We extend Theorem 4.2 to make its dependency on f_{Θ} more fine-grained. Rather than assuming that only the norm of the embedding layer’s matrix is bounded, we follow the setting of prior work (Edelman et al., 2022; Furuya et al., 2024; Marion, 2023; Zhang et al., 2023b) and consider the parameter space defined as follows:

$$\begin{aligned} \tilde{\mathcal{W}} = \{ & \Theta \in \mathcal{W} \mid \forall \ell \in [L], \|\mathbf{W}_V^{(\ell)}\|_{\infty} \leq B_V, \\ & \|\mathbf{W}_O^{(\ell)}\|_{\infty} \leq B_O, \|\mathbf{W}_1^{(\ell)}\|_{\infty} \leq B_1, \|\mathbf{W}_2^{(\ell)}\|_{\infty} \leq B_2 \}. \end{aligned}$$

The definition of $\tilde{\mathcal{W}}$ concerns the query, key, and value matrices of all layers and heads. Similarly to Zhang et al. (2023b, Assumption 5.1), we assume that each token has an ℓ_1 -norm bounded by B_{tok} . We have the following generalization bound, whose proof is deferred to Appendix F.6.

Corollary E.3 (Depth-dependent bound). *Consider an LLM $f_{\Theta} \in \tilde{\mathcal{F}} := \{f_{\Theta} \mid \Theta \in \tilde{\mathcal{W}}\}$. With the same assumptions as in Theorem 4.2, we have*

$$\mathcal{R}_{\text{pre}}(\Theta) \leq \hat{\mathcal{R}}_{\text{pre}}(\Theta) + \frac{\bar{B}}{\sqrt{N_{\text{train}}}} \sqrt{\log\left(\frac{2}{\delta}\right)},$$

with constants depending on the problem’s parameters

$$\begin{aligned} \bar{B} &= 2\|\Gamma\| \max\{\log(T) + 2(B_{\Theta})^L/\tau, \log(1/c_0)\}^{1/2}, \\ B_{\Theta} &= [(1 + rmB_1B_2)(1 + \frac{r^3}{H}B_0B_V)](B_{\text{tok}}B_U)^{1/L}. \end{aligned}$$

We note that \bar{B} exhibits an exponential dependence on the depth of the transformer, which also amplifies the hidden dimensionality (width) of the embedding layer r . This contrasts with the dependency in m , the hidden dimensionality of the MLP block, which is linear. All these factors are commonly associated with higher expressive power of transformers suggesting that they should contribute to a better minimization of $\hat{\mathcal{R}}_{\text{pre}}(\Theta)$ at the expense of requiring more training data. The number of heads H can be used as a counterbalance to increasing the width in the cubic term r^3 , suggesting that a good balance between these parameters may lead to more data-efficient models.

E.3. Generalization Bounds with the KL Divergence

As explained in Remark 4.1, the total variation is the natural choice to define the risks in Eq. (3). Another possibility in the Markov chain literature is to use the KL divergence to compare probability distributions (Hao et al., 2018). This is an interesting candidate as the KL divergence is naturally connected to the cross-entropy loss commonly used to train neural networks (the cross-entropy corresponds to the KL divergence between the true distribution and the predicted softmax distribution (Blondel et al., 2019)). In this section, we discuss the extension of the theoretical results of Section 4 by replacing the TV distance with the KL divergence in the risks’ definition, i.e.,

$$\mathcal{R}(\Theta) := \mathbb{E}_{\mathbf{S} \sim \mathbb{P}_{\mathcal{L}}} [d_{\text{KL}}(\mathbf{Q}^*(\mathbf{S}, \cdot) \parallel \mathbf{Q}_f(\mathbf{S}, \cdot))], \hat{\mathcal{R}}(\Theta) := \frac{1}{N} \sum_{n=1}^N d_{\text{KL}}(\mathbb{P}_{\mathcal{L}}(\cdot \mid \mathbf{S}_n) \parallel \mathbb{P}_{\Theta}(\cdot \mid \mathbf{S}_n)). \quad (8)$$

Theorem 4.2 and Corollary E.3 related to the pre-training phase in Section 4.2 can be obtained similarly if the risks are defined with the KL divergence following Eq. (8). Indeed, the key step to derive the proofs is to obtain a similar result to Lemma F.6 but with the KL divergence. The next lemma provides this result.

Lemma E.4. Consider two probability distributions \mathbb{P}, \mathbb{Q} defined on a measure space (Ω, \mathcal{F}) and a σ -finite measure ν on (Ω, \mathcal{F}) . Let p, q be the corresponding probabilities densities, i.e., we have $\mathbb{P}(d\omega) = q(\omega)\nu(d\omega)$ and $\mathbb{Q}(d\omega) = p(\omega)\nu(d\omega)$. If there exists a non-negative constant B such that for any $z \in \Omega$, $\left| \log \sqrt{\frac{\mathbb{P}(z)}{\mathbb{Q}(z)}} \right| \leq B$, then we have

$$d_{\text{KL}}(\mathbb{P}||\mathbb{Q}) \leq B.$$

Proof. We have

$$\begin{aligned} 0 \leq d_{\text{KL}}(\mathbb{P}||\mathbb{Q}) &= |d_{\text{KL}}(\mathbb{P}||\mathbb{Q})| \\ &= \left| \int \mathbb{P}(z) \log\left(\frac{\mathbb{P}(z)}{\mathbb{Q}(z)}\right) dz \right| \\ &\leq \int |\mathbb{P}(z)| \left| \log\left(\frac{\mathbb{P}(z)}{\mathbb{Q}(z)}\right) \right| dz \\ &\leq B \int |\mathbb{P}(z)| dz \\ &= B \int \mathbb{P}(z) dz \\ &= B, \end{aligned}$$

which concludes the proof. \square

We can now state the results similar to Theorem 4.2, Corollary E.3 and Proposition 4.1 from the pre-training phase when the risk is defined according to Eq. (8).

Theorem E.5 (Pre-training generalization bound). Consider an LLM $f_{\Theta} \in \mathcal{F}$. We denote by $\mathbf{\Gamma}$ the mixing matrix of the pre-training sequences of tokens $(\mathbf{S}_1, \dots, \mathbf{S}_{N_{\text{train}}})$. Let $0 < \delta < 1$, then with probability at least $1 - \delta$,

$$\mathcal{R}_{\text{pre}}(\Theta) \leq \widehat{\mathcal{R}}_{\text{pre}}(\Theta) + \frac{\bar{B}}{\sqrt{N_{\text{train}}}} \sqrt{\log\left(\frac{2}{\delta}\right)},$$

where $\bar{B} = \sqrt{2}\|\mathbf{\Gamma}\| \max\{\log(T) + 2B_U/\tau, \log(1/c_0)\}$ is a constant depending on the parameters of the problem.

Proof. The proof simply follows from the proof of Theorem 4.2 by replacing the upper bound $\sqrt{2\bar{B}}$ by B (with the appropriate upper-bound B) when Lemma F.6 is used in the proof. \square

Corollary E.6 (Depth-dependent bound). Consider an LLM $f_{\Theta} \in \tilde{\mathcal{F}} := \{f_{\Theta} \mid \Theta \in \tilde{\mathcal{W}}\}$. With the same assumptions as in Theorem 4.2, we have

$$\mathcal{R}_{\text{pre}}(\Theta) \leq \widehat{\mathcal{R}}_{\text{pre}}(\Theta) + \frac{\bar{B}}{\sqrt{N_{\text{train}}}} \sqrt{\log\left(\frac{2}{\delta}\right)},$$

where $\bar{B} = \sqrt{2}\|\mathbf{\Gamma}\| \max\{\log(T) + 2(B_{\Theta})^L/\tau, \log(1/c_0)\}$ is a constant depending on the parameters of the problem, and $B_{\Theta} = [(1 + rmB_1B_2)(1 + \frac{r^3}{H}B_0B_V)](B_{\text{tok}}B_U)^{1/L}$.

Proof. The proof simply follows from the proof of Theorem 4.2 by replacing the upper bound $\sqrt{2\bar{B}}$ by B (with the appropriate upper-bound B) when Lemma F.6 is used in the proof. \square

Limitations. We recall from Remark 4.1 that the TV distance is a natural choice to compare transition matrices in the Markov chain literature. In addition, while the KL divergence can be used to compare probability distributions, it does not define a metric space. Hence, we cannot straightforwardly extend Theorem 4.3 with the KL divergence because the proof relies on the use of the triangular inequality. As Theorem 4.3 is one of our main results and enables us to show that the theory and the practice align (Section 5), this also contributed to our preference for the TV distance instead of the KL divergence. We also note that the proof of Proposition 4.1 relies on properties of the total variation and hence we cannot extend it straightforwardly with the KL divergence.

F. Proofs

F.1. Proof of Proposition 3.2

We detail below the proof of Proposition 3.2.

Proof of Proposition 3.2. Step 1: Large language models as Markov chains. Given an input $v_i \in \mathcal{V}_K^*$ of p tokens, a large language model outputs a probability mass function $f_{\Theta}^{T,K}(v_i)$ over the discrete vocabulary space. As the temperature is positive, i.e., $\tau > 0$, and as the exponential is positive, we know that all the tokens in the vocabulary will be given a positive mass.

A next sequence $v_j \in \mathcal{V}_K^*$ is then sampled according to $f_{\Theta}^{T,K}(v_i)$. But the v_j sequences that fit necessarily contain the v_i sequence (except possibly the first element of v_i , thanks to Definition B.2), i.e. $\forall l, (v_j)_l = (v_i)_{l+1}$. Note also the size of v_j is $p + 1$ when $p < k$ and k when $p = k$. All other sequences v_j that do not satisfy this condition are not suitable.

In that sense, $f_{\Theta}^{T,K}$ can be represented by a Markov chain $\text{MC}(\mathcal{V}_K^*, \mathbf{Q}_f)$ with transition kernel $\mathbf{Q}_f \in \mathbb{R}^{|\mathcal{V}_K^*| \times |\mathcal{V}_K^*|}$, as defined in Proposition 3.2.

Step 2: Proportion of non-zero elements. We denote by \mathcal{R} the set of states of length K . The set of states of length strictly less than equal K is denoted by \mathcal{T} . We can construct a transition matrix $P_{\mathcal{R}} \in \mathbb{R}^{T^K \times T^K}$ with the states of this class, containing the probabilities of moving from one state of \mathcal{R} to another. $P_{\mathcal{R}}$ corresponds to the blue block in Fig. 2 while green rectangle blocks correspond to part of $P_{\mathcal{T}}$ and $P_{\mathcal{T}\mathcal{R}}$ in the following description of large language models as Markov chains,

$$\mathbf{Q}_f = \left(\begin{array}{c|c} P_{\mathcal{T}} & P_{\mathcal{T}\mathcal{R}} \\ \hline 0 & P_{\mathcal{R}} \end{array} \right). \quad (9)$$

Now, let us count the number of non-zero elements in each of these 4 large blocks.

$P_{\mathcal{T}}$ block : The size of this block is $\left[\frac{T}{T-1} (T^{K-1} - 1) \right] \times \left[\frac{T}{T-1} (T^{K-1} - 1) \right]$. There are $K - 2$ green blocks contained in $P_{\mathcal{T}}$. The block number $i \in [K - 2]$ is of size $T^i \times T^{i+1}$. Since each sentence of size i can be completed with non-zero probability, by any other token, there are a total of $\sum_{p=1}^{T^i} T = T^{i+1}$ non-zero elements. There are therefore $\sum_{i=1}^{K-2} T^{i+1}$ non-zero elements in the entire $P_{\mathcal{T}}$ block.

$P_{\mathcal{T}\mathcal{R}}$ block : The size of this block is $\left[\frac{T}{T-1} (T^{K-1} - 1) \right] \times T^K$. The green block contained in $P_{\mathcal{T}\mathcal{R}}$ that contains non-zero elements is of size $T^{K-1} \times T^K$. Since each sentence of size $K - 1$ can be completed with non-zero probability, by any other token, there are a total of $\sum_{p=1}^{T^{K-1}} T = T^K$ non-zero elements.

$P_{\mathcal{R}}$ block : The size of this block is $T^K \times T^K$. Each sentence $v = (v_1, \dots, v_K)$ of size K is mapped to another sentence $v' = (v'_1, \dots, v'_K)$ of size K with non-zero probability, if and only if $v'_1 = v_2, v'_2 = v_3, \dots, v'_{k-1} = v_K$. The final token v'_K can be any other token in the vocabulary. It means that there are a total of $\sum_{p=1}^{T^K} T = T^{K+1}$ non-zero elements.

$0's$ block : There are no non-zero elements in this block.

Finally, there are

$$\sum_{i=1}^{K-2} T^{i+1} + T^K + T^{K+1} = \sum_{i=1}^K T^{i+1} = T^2 \left(\frac{T^K - 1}{T - 1} \right)$$

non-zero elements. This means that the proportion of non-zero elements in the matrix is exactly

$$\frac{T^2 \left(\frac{T^K - 1}{T - 1} \right)}{\left(T \left(\frac{T^K - 1}{T - 1} \right) \right)^2} = \frac{T - 1}{T^K - 1}.$$

Note that for large T and K we have that

$$\frac{T - 1}{T^K - 1} \sim \frac{1}{T^{K-1}}.$$

□

F.2. Proof of Proposition 3.3

We begin with a preliminary lemma.

Lemma F.1 (Powers of \mathbf{Q}_f greater than K). *For any initial state i , the following hold:*

- $\forall k \geq K, \forall j \in \mathcal{T}, (\mathbf{Q}_f^k)_{i,j} = 0,$
- $\forall k \geq K, \forall j \in \mathcal{R}, (\mathbf{Q}_f^k)_{i,j} > 0.$

Proof. By considering \mathbf{Q}_f as defined in (9), we can compute its powers. For any $k \geq 1,$

$$\mathbf{Q}_f^k = \begin{pmatrix} P_{\mathcal{T}}^k & B_k \\ 0 & P_{\mathcal{R}}^k \end{pmatrix},$$

where $B_k = \sum_{m=0}^{k-1} P_{\mathcal{T}}^m P_{\mathcal{T}\mathcal{R}} P_{\mathcal{R}}^{k-1-m}.$

To prove the first item, we focus on the blocks on the left of \mathbf{Q}_f . Since the lower left block is zero, we have that $\forall k \geq 1, \forall i \in \mathcal{R}, \forall j \in \mathcal{T}, (\mathbf{Q}_f^k)_{i,j} = 0.$ In the upper left block, the element $(P_{\mathcal{T}}^k)_{i,j}$ designates the probability of moving from one transient state $i \in \mathcal{T}$ to another transient state $j \in \mathcal{T}$ after k iterations. According to Definition B.1, if state i is a sequence of $p \geq 1$ tokens, state j is necessarily a sequence of $\min\{K, p + K\} = K$ elements. Thus, $P_{\mathcal{T}}$ is a nilpotent matrix and $\forall k \geq K, \forall i, j, (P_{\mathcal{T}}^k)_{i,j} = 0.$ This proves that $\forall k \geq K, \forall j \in \mathcal{T}, (\mathbf{Q}_f^k)_{i,j} = 0.$

We now move on to the second item. From the above, $\forall k \geq K, B_k = \sum_{m=0}^{K-1} P_{\mathcal{T}}^m P_{\mathcal{T}\mathcal{R}} P_{\mathcal{R}}^{k-1-m}.$ Note that this sum is finite, but there is still a dependence on $k,$ in the powers of the matrix $P_{\mathcal{R}}.$ In the lower right block, the element $(P_{\mathcal{R}}^k)_{i,j}$ designates the probability of moving from one recurrent state $i \in \mathcal{R}$ to another recurrent state $j \in \mathcal{R}$ after k iterations. According to the definition of \mathbf{Q}_f in Proposition 3.2 and Definition B.2, $\forall k \geq K, \forall i, j \in \mathcal{R}^2, (P_{\mathcal{R}}^k)_{i,j}$ is nonzero. Exploiting this also in $B_k,$ we obtain the result, i.e. $\forall k \geq K, \forall j \in \mathcal{R}, (\mathbf{Q}_f^k)_{i,j} > 0.$ □

We are now ready to prove Proposition 3.3, which is inspired by Gallager (1996).

Proof of Proposition 3.3. The states of length strictly less than equal K (elements of \mathcal{T}) are transient, because of Definition B.1. To discuss the nature of states of length K (elements of \mathcal{R}), let us introduce a result regarding the powers of the \mathbf{Q}_f matrix as defined in (9). Thanks to Lemma F.1, the set of states \mathcal{R} (i.e. the states of length K) form a class. Lemma F.1 gives us also that \mathcal{R} is ergodic. In fact, every state in \mathcal{R} only communicates with all the other states in $\mathcal{R},$ which proves the recurrence. Since $\forall i, j \in \mathcal{R}^2, (\mathbf{Q}_f^K)_{i,j} > 0,$ we can move between any two states in exactly K steps, regardless of the initial position. This ensures that \mathcal{R} is aperiodic because the transition probabilities do not depend on a specific cycle, and states can be revisited at various time steps, not just multiples of a particular number. More simply, by considering a token $x,$ the state defined as $i = \underbrace{xx \dots x}_{K \text{ times}}$ has period 1, i.e. $(\mathbf{Q}_f)_{i,i} > 0.$ This is a consequence of Definition B.2 and Proposition 3.2.

Thanks to Theorem C.9, it means that the whole class \mathcal{R} is aperiodic. Finally, this means that $\text{MC}(\mathcal{V}_K^*, \mathbf{Q}_f)$ are ergodic unichains, in the sense of Definition C.11. □

E.3. Proof of Proposition 3.4

We start by introducing three technical lemmas that will be useful in the proof of Proposition 3.4. We start with the Chapman–Kolmogorov equation.

Lemma F.2 (Chapman-Kolmogorov equation). *Let P be a matrix of size d . Then, P satisfies*

$$\forall i, j \in [d]^2, \forall n, n' \in \mathbb{N}^2, (P^{n+n'})_{i,j} = \sum_{k=1}^d (P^n)_{i,k} (P^{n'})_{k,j}.$$

Proof. The result follows from the fact that $\forall n, n' \in \mathbb{N}^2, P^{n+n'} = P^n P^{n'}$. \square

Then, we introduce a simple but useful result of monotonicity.

Lemma F.3 (Lemma 3.3.1. in Gallager (1996)). *Let the transition matrix P of a finite state Markov chain. Then, for all states j and $n \geq 1$, we have*

$$\max_i (P^{n+1})_{i,j} \leq \max_i (P^n)_{i,j}, \quad \text{and} \quad \min_i (P^{n+1})_{i,j} \geq \min_i (P^n)_{i,j}.$$

We now refer to a result on Markov chains with positive transition matrices.

Lemma F.4 (Lemma 3.3.2. in Gallager (1996)). *Let the transition matrix P of a finite state Markov chain satisfy $\forall i, j, P_{i,j} > 0$, and let $\alpha = \min_{i,j} P_{i,j} > 0$. Then, for all states j and $n \geq 1$,*

$$\begin{aligned} \max_i (P^n)_{i,j} - \min_i (P^n)_{i,j} &\leq (1 - 2\alpha) \left(\max_i (P^n)_{i,j} - \min_i (P^n)_{i,j} \right), \\ \max_i (P^n)_{i,j} - \min_i (P^n)_{i,j} &\leq (1 - 2\alpha)^n, \\ \lim_{n \rightarrow \infty} \max_i (P^n)_{i,j} &= \lim_{n \rightarrow \infty} \min_i (P^n)_{i,j} > 0. \end{aligned}$$

We are now ready to prove Proposition 3.4 using a similar argument as in (Gallager, 1996).

Proof of Proposition 3.4. Let \mathcal{T} and \mathcal{R} denote respectively the sets of transient and recurrent states. For any state j , we define $\pi_j := \lim_{n \rightarrow \infty} \max_i (\mathbf{Q}_f^n)_{i,j} = \lim_{n \rightarrow \infty} \min_i (\mathbf{Q}_f^n)_{i,j}$. Then $\boldsymbol{\pi} = (\pi_j)_{j \in \Omega}$ is the stationary distribution for \mathbf{Q}_f .

Step 1: Stationary distribution for transient states. Lemma F.1 gives us that $\forall i, \forall k \geq K, \forall j \in \mathcal{T}, (\mathbf{Q}_f^k)_{i,j} = 0$. This means that $\forall j \in \mathcal{T}, \pi_j = 0$ and hence the limit is reached at most after K iteration.

Step 2: Stationary distribution for recurrent states. Lemma F.1 gives us $\forall i, j \in \mathcal{R}^2, (\mathbf{Q}_f^K)_{i,j} > 0$. By defining $\varepsilon := \min_{i,j \in \mathcal{R}^2} (\mathbf{Q}_f^K)_{i,j}$, Lemma F.4 shows that for any integer $\ell \geq 1$,

$$\max_{i \in \mathcal{R}} (\mathbf{Q}_f^{\ell K})_{i,j} - \min_{i \in \mathcal{R}} (\mathbf{Q}_f^{\ell K})_{i,j} \leq (1 - 2\varepsilon) \left(\max_{i \in \mathcal{R}} (\mathbf{Q}_f^{\ell K})_{i,j} - \min_{i \in \mathcal{R}} (\mathbf{Q}_f^{\ell K})_{i,j} \right), \quad (10)$$

$$\max_{i \in \mathcal{R}} (\mathbf{Q}_f^{\ell K})_{i,j} - \min_{i \in \mathcal{R}} (\mathbf{Q}_f^{\ell K})_{i,j} \leq (1 - 2\varepsilon)^\ell, \quad (11)$$

$$\lim_{\ell \rightarrow \infty} \max_{i \in \mathcal{R}} (\mathbf{Q}_f^{\ell K})_{i,j} = \lim_{\ell \rightarrow \infty} \min_{i \in \mathcal{R}} (\mathbf{Q}_f^{\ell K})_{i,j} > 0. \quad (12)$$

Thanks to Lemma F.3, $\max_i (\mathbf{Q}_f^{n+1})_{i,j}$ is non-decreasing in n , so the limit on the left in Eq. (12) can be replaced with a limit in n . The same argument for the limit on the right gives that, $\forall j \in \mathcal{R}$,

$$\max_{i \in \mathcal{R}} (\mathbf{Q}_f^n)_{i,j} - \min_{i \in \mathcal{R}} (\mathbf{Q}_f^n)_{i,j} \leq (1 - 2\varepsilon)^{\lfloor n/K \rfloor},$$

$$\lim_{n \rightarrow \infty} \max_{i \in \mathcal{R}} (\mathbf{Q}_f^n)_{i,j} = \lim_{n \rightarrow \infty} \min_{i \in \mathcal{R}} (\mathbf{Q}_f^n)_{i,j} > 0,$$

where we have taken the floor function to also convert the result of (11). Since π_j lies between the minimum and maximum $(\mathbf{Q}_f^n)_{i,j}$ for each n , we have that $\forall i, j \in \mathcal{R}^2$,

$$|(\mathbf{Q}_f^n)_{i,j} - \pi_j| \leq (1 - 2\varepsilon)^{\lfloor \frac{n}{K} \rfloor}.$$

It means that $\forall i, j \in \mathcal{R}^2, \pi_j = \lim_{n \rightarrow \infty} (\mathbf{Q}_f^n)_{i,j}$. This also gives us the convergence rate when the initial state i is recurrent. In the next step, we consider the general convergence rate, regardless of the nature of the initial state i .

Step 3: Convergence bound. We proceed to the remaining case, i.e. the case where the initial state $i \in \mathcal{T}$ and the final state $j \in \mathcal{R}$. Lemma F.2 says that $\forall n \geq K$,

$$(\mathbf{Q}_f^n)_{i,j} = \sum_{k \in \mathcal{T}} (\mathbf{Q}_f^K)_{i,k} (\mathbf{Q}_f^{n-K})_{k,j} + \sum_{k \in \mathcal{R}} (\mathbf{Q}_f^K)_{i,k} (\mathbf{Q}_f^{n-K})_{k,j}.$$

We then have that $\forall i \in \mathcal{T}, \forall n \in \mathbb{N}$,

$$\begin{aligned} |(\mathbf{Q}_f^n)_{i,j} - \pi_j| &\leq \left| \sum_{k \in \mathcal{T}} (\mathbf{Q}_f^K)_{i,k} [(\mathbf{Q}_f^{n-K})_{k,j} - \pi_j] + \sum_{k \in \mathcal{R}} (\mathbf{Q}_f^K)_{i,k} [(\mathbf{Q}_f^{n-K})_{k,j} - \pi_j] \right| \\ &\leq \sum_{k \in \mathcal{T}} (\mathbf{Q}_f^K)_{i,k} |(\mathbf{Q}_f^{n-K})_{k,j} - \pi_j| + \sum_{k \in \mathcal{R}} (\mathbf{Q}_f^K)_{i,k} |(\mathbf{Q}_f^{n-K})_{k,j} - \pi_j| \\ &\leq \sum_{k \in \mathcal{T}} (\mathbf{Q}_f^K)_{i,k} + \sum_{k \in \mathcal{R}} (\mathbf{Q}_f^K)_{i,k} |(\mathbf{Q}_f^{n-K})_{k,j} - \pi_j| \\ &\leq (1 - 2\varepsilon)^{\lfloor \frac{n-K}{K} \rfloor}, \end{aligned}$$

where the first sum vanishes and $\sum_{k \in \mathcal{R}} (\mathbf{Q}_f^K)_{i,k} \leq 1$. Finally, $\forall i \in \mathcal{T}, \forall n \geq K$,

$$|(\mathbf{Q}_f^n)_{i,j} - \pi_j| \leq (1 - 2\varepsilon)^{\lfloor \frac{n}{K} \rfloor - 1}.$$

Combining this with the result of Step 2 concludes the proof. \square

E.4. Proof of Proposition 4.1

We detail the proof of Proposition 4.1 below.

Proof. We first note that by definition of the total variation distance (Wolfer & Kontorovich, 2019), we have

$$\begin{aligned} \mathbb{E}_{\mathbf{S} \sim \mathbb{P}_{\mathcal{L}}} \|\mathbf{Q}^*(\mathbf{S}, \cdot) - \mathbf{Q}_f(\mathbf{S}, \cdot)\|_1 &= \mathbb{E}_{\mathbf{S} \sim \mathbb{P}_{\mathcal{L}}} [2 \cdot d_{\text{TV}}(\mathbf{Q}^*(\mathbf{S}, \cdot), \mathbf{Q}_f(\mathbf{S}, \cdot))] \\ &= 2 \cdot \mathbb{E}_{\mathbf{S} \sim \mathbb{P}_{\mathcal{L}}} [d_{\text{TV}}(\mathbf{Q}^*(\mathbf{S}, \cdot), \mathbf{Q}_f(\mathbf{S}, \cdot))] \\ &= 2 \cdot \mathcal{R}_{\text{pre}}(\Theta). \end{aligned} \quad (\text{by definition of the risk Eq. (3)})$$

Applying Theorem 4.2 (a similar result can be derived using Corollary E.3), we know that

$$\mathcal{R}_{\text{pre}}(\Theta) \leq \widehat{\mathcal{R}}_{\text{pre}}(\Theta) + \frac{\bar{B}}{\sqrt{N_{\text{train}}}} \sqrt{\log \left(\frac{2}{\delta} \right)},$$

where \bar{B} is formally defined in Theorem 4.2 (respectively Corollary E.3). Assuming a perfect pre-training error amounts to consider $\widehat{\mathcal{R}}_{\text{pre}}(\Theta) = 0$. We denote by N^* the integer such that the error is equal to $\frac{\epsilon}{2}$, i.e.,

$$\frac{\bar{B}}{\sqrt{N^*}} \sqrt{\log \left(\frac{2}{\delta} \right)} = \frac{\epsilon}{2} \iff \frac{\bar{B}^2}{N^*} \log \left(\frac{2}{\delta} \right) = \frac{\epsilon^2}{4} \iff N^* = \left(\frac{2\bar{B}}{\epsilon} \right)^2 \log \left(\frac{2}{\delta} \right).$$

Taking the ceiling function ensures that N^* is an integer. Hence, taking $N_{\text{train}} \geq N^* = \lceil \left(\frac{2\bar{B}}{\epsilon}\right)^2 \log\left(\frac{2}{\delta}\right) \rceil$ ensures that

$$\frac{\bar{B}}{\sqrt{N_{\text{train}}}} \sqrt{\log\left(\frac{2}{\delta}\right)} \leq \frac{\bar{B}}{\sqrt{N^*}} \sqrt{\log\left(\frac{2}{\delta}\right)} = \frac{\epsilon}{2}.$$

Putting everything together, taking $N_{\text{train}} \geq N^*$ leads to

$$\mathbb{E}_{\mathbf{S} \sim \mathbb{P}_{\mathcal{L}}} \|\mathbf{Q}^*(\mathbf{S}, \cdot) - \mathbf{Q}_f(\mathbf{S}, \cdot)\|_1 \leq 2 \cdot \mathcal{R}_{\text{pre}}(\Theta) \leq 2 \cdot \frac{\epsilon}{2} = \epsilon,$$

which concludes the proof. \square

F.5. Proof of Theorem 4.2

In this section, we detail the proof of Theorem 4.2. We provide below an overview of the proof before detailing it.

Overview of the proof. We are going to use McDiarmid’s inequality for dependent random variables of Paulin (2015, Theorem 2.9). To adapt the arguments of Paulin (2015, Theorem 2.9) to our setting, we bound the total variation between the true probability of the next token and the one estimated by the LLM. The rest of this section is organized as follows. First in Appendix F.5.1, we adapt the concentration inequality of Paulin (2015, Theorem 2.9). Then in Appendix F.5.2, we show how to bound the total variation between the true and the estimated probability of the next token. , in Appendix F.5.3, we restate Theorem 4.2 and conclude the proof.

F.5.1. CONCENTRATION INEQUALITIES FOR DEPENDENT RANDOM VARIABLES

We first state a concentration inequality for dependent random variables that will be used to obtain our final bound.

Proposition F.5 (McDiarmid’s inequality for dependent random variables). *Let $S := (\mathbf{S}_1, \dots, \mathbf{S}_N)$ be a sequence of random variables that take values in $\Omega = \Omega_1 \times \dots \times \Omega_N$. Assume there exists a Marton coupling for S with mixing matrix Γ . Let $\|\Gamma\|$ be the operator norm of Γ . If $f: \Omega \rightarrow \mathbb{R}$ is such that there exists $\mathbf{c} \in \mathbb{R}^N$ satisfying*

$$\forall \mathbf{x}, \mathbf{y} \in \Omega, \quad f(\mathbf{x}) - f(\mathbf{y}) \leq \sum_{i=1}^N \mathbf{c}_i \mathbb{1}_{\{\mathbf{x}_i \neq \mathbf{y}_i\}},$$

then we have for any $u \geq 0$,

$$\mathbb{P}(|f(S) - \mathbb{E}_S[f(S)]| \geq u) \leq 2 \exp\left(\frac{-2u^2}{\|\Gamma\|^2 \|\mathbf{c}\|_2^2}\right).$$

Proof. Consider a function f verifying the properties of Proposition F.5. Paulin (2015, Theorem 2.9) ensures that for a partition \hat{S} of S (see Paulin, 2015, Definition 2.3) the following inequality holds

$$\forall u \geq 0, \quad \mathbb{P}\left(|f(\hat{S}) - \mathbb{E}[f(\hat{S})]| \geq u\right) \leq 2 \exp\left(\frac{-2u^2}{\|\Gamma \cdot C(\mathbf{c})\|_2^2}\right), \quad (13)$$

where $C(\mathbf{c})$ is a vector of \mathbb{R}^N whose i -th element is the sum of the \mathbf{c}_j such that j is an index of the elements of $\hat{\mathbf{S}}_i$. Taking the trivial partition $\hat{S} = S$ implies that the index of the elements in $\hat{\mathbf{S}}_i$ are reduced to $\{i\}$. Hence the i -th entry of $C(\mathbf{c})$ is equal to \mathbf{c}_i and $C(\mathbf{c}) = \mathbf{c}$. By definition of the operator norm (naturally induced by the ℓ_2 -norm), we have

$$\|\Gamma \cdot \mathbf{c}\|_2 = \frac{\|\Gamma \mathbf{c}\|_2}{\|\mathbf{c}\|_2} \cdot \|\mathbf{c}\|_2 \leq \underbrace{\sup_{\mathbf{x} \neq 0} \frac{\|\Gamma \mathbf{x}\|_2}{\|\mathbf{x}\|_2}}_{=\|\Gamma\|} \cdot \|\mathbf{c}\|_2 \leq \|\Gamma\| \cdot \|\mathbf{c}\|_2,$$

where the first inequality comes from the fact that \mathbf{c} is non-zero (otherwise the only possible f is the zero function which is not of great interest). Using the fact that the function $x \rightarrow \exp\left(-\frac{2u^2}{x}\right)$ is increasing, we obtain

$$\exp\left(\frac{-2u^2}{\|\Gamma \cdot \mathbf{c}\|_2^2}\right) \leq \exp\left(\frac{-2u^2}{\|\Gamma\|^2 \cdot \|\mathbf{c}\|_2^2}\right),$$

which concludes the proof. \square

By looking at the definition of the risk $\widehat{\mathcal{R}}_{\text{pre}}(\Theta)$, we can see that applying Proposition F.5 to the function

$$f: (\mathbf{S}_1, \dots, \mathbf{S}_{N_{\text{train}}}) = \frac{1}{N_{\text{train}}} \sum_{n=1}^{N_{\text{train}}} d_{\text{TV}}(\mathbb{P}_{\mathcal{L}}(\cdot | \mathbf{S}_n), \mathbb{P}_{\Theta}(\cdot | \mathbf{S}_n)),$$

would lead to the desired bound as we already know S admits a Marton coupling with mixing matrix Γ . We investigate in the next section how to find the bounding vector \mathbf{c} to apply Proposition F.5.

F.5.2. FINDING THE BOUNDING VECTOR

Technical lemmas. We first recall the following important notions from (Tsybakov, 2008). Let (Ω, \mathcal{F}) be a measure space and consider two probability distributions \mathbb{P}, \mathbb{Q} defined on (Ω, \mathcal{F}) . For any σ -finite measure ν on (Ω, \mathcal{F}) such that \mathbb{P}, \mathbb{Q} are absolutely continuous with respect to ν , we can define $p = \frac{d\mathbb{P}}{d\nu}, q = \frac{d\mathbb{Q}}{d\nu}$ which can also be written as $\mathbb{P}(d\omega) = p(\omega)\nu(d\omega)$ and $\mathbb{Q}(d\omega) = q(\omega)\nu(d\omega)$. We will adopt both notations interchangeably. It should be noted that there always exists at least one such measure ν as one can take $\nu = \mathbb{P} + \mathbb{Q}$. With these notations, the squared Hellinger distance between \mathbb{P} and \mathbb{Q} is defined as

$$H(\mathbb{P}, \mathbb{Q})^2 := \int_{\omega \in \Omega} \left(\sqrt{p(\omega)} - \sqrt{q(\omega)} \right)^2 \nu(d\omega) = \int_{\omega \in \Omega} \left(\sqrt{\mathbb{P}(d\omega)} - \sqrt{\mathbb{Q}(d\omega)} \right)^2.$$

The lemma below shows that the total variation between two probability distributions is controlled from above by the absolute value of the logarithm of their ratio.

Lemma F.6. Consider two probability distributions \mathbb{P}, \mathbb{Q} defined on a measure space (Ω, \mathcal{F}) and a σ -finite measure ν on (Ω, \mathcal{F}) . Let p, q be the corresponding probabilities densities, i.e., we have $\mathbb{P}(d\omega) = p(\omega)\nu(d\omega)$ and $\mathbb{Q}(d\omega) = q(\omega)\nu(d\omega)$, the total variation between \mathbb{P} and \mathbb{Q} satisfies

$$d_{\text{TV}}(\mathbb{P}, \mathbb{Q}) \leq \left(2 \int_{\omega \in \Omega} \left| \log \sqrt{\frac{\mathbb{P}(d\omega)}{\mathbb{Q}(d\omega)}} \right| q(\omega) d\nu(d\omega) \right)^{1/2}.$$

If there exists a non-negative constant B such that for any $z \in \Omega$, $\left| \log \sqrt{\frac{\mathbb{P}(z)}{\mathbb{Q}(z)}} \right| \leq B$, then we have

$$d_{\text{TV}}(\mathbb{P}, \mathbb{Q}) \leq \sqrt{2B}.$$

Proof. We have the following relation between the total variation and the Hellinger distance (cf. Tsybakov, 2008, Lem. 2.3, Chapt. 2, p. 86):

$$d_{\text{TV}}(\mathbb{P}, \mathbb{Q})^2 \leq H(\mathbb{P}, \mathbb{Q})^2 \cdot \left(1 - \underbrace{H(\mathbb{P}, \mathbb{Q})^2/4}_{\geq 0} \right) \leq H(\mathbb{P}, \mathbb{Q})^2, \quad (14)$$

where the last inequality uses the positivity of the Hellinger distance. Inspired by the decomposition of the Hellinger distance in (Agarwal et al., 2020, Lem. 25), we have

$$\begin{aligned} H(\mathbb{P}, \mathbb{Q})^2 &= \int_{\omega \in \Omega} \left(\sqrt{\mathbb{P}(d\omega)} - \sqrt{\mathbb{Q}(d\omega)} \right)^2 = \int_{\omega \in \Omega} \left(\mathbb{P}(d\omega) + \mathbb{Q}(d\omega) - 2\sqrt{\mathbb{P}(d\omega)}\sqrt{\mathbb{Q}(d\omega)} \right) \\ &= 2 \cdot \left(1 - \int_{\omega \in \Omega} \sqrt{\mathbb{P}(d\omega)}\sqrt{\mathbb{Q}(d\omega)} \right) = 2 \cdot \left(1 - \int_{\omega \in \Omega} \sqrt{\frac{\mathbb{P}(d\omega)}{\mathbb{Q}(d\omega)}} \mathbb{Q}(d\omega) \right) \\ &= 2 \cdot \left(1 - \int_{\omega \in \Omega} \sqrt{\frac{\mathbb{P}(d\omega)}{\mathbb{Q}(d\omega)}} q(\omega) d\nu(d\omega) \right) \quad (\text{by definition of } \mathbb{Q}(d\omega)) \end{aligned}$$

$$\leq -2 \log \left(\int_{\omega \in \Omega} \sqrt{\frac{\mathbb{P}(d\omega)}{\mathbb{Q}(d\omega)}} q(\omega) d\nu(d\omega) \right) \quad (\text{using } 1 - x \leq -\log(x))$$

It follows using Eq. (14)

$$\begin{aligned} d_{\text{TV}}(\mathbb{P}, \mathbb{Q})^2 &\leq H(\mathbb{P}, \mathbb{Q})^2 \\ &\leq 2 \int_{\omega \in \Omega} -\log \left(\sqrt{\frac{\mathbb{P}(d\omega)}{\mathbb{Q}(d\omega)}} \right) q(\omega) d\nu(d\omega) && (\text{by Jensen as } -\log \text{ is convex}) \\ &\leq 2 \left| \int_{\omega \in \Omega} -\log \left(\sqrt{\frac{\mathbb{P}(d\omega)}{\mathbb{Q}(d\omega)}} \right) q(\omega) d\nu(d\omega) \right| \\ &\leq 2 \int_{\omega \in \Omega} \left| -\log \left(\sqrt{\frac{\mathbb{P}(d\omega)}{\mathbb{Q}(d\omega)}} \right) \right| q(\omega) d\nu(d\omega) && (\text{by Jensen as } |\cdot| \text{ is convex}) \\ &\leq 2 \int_{\omega \in \Omega} \underbrace{\left| \log \left(\sqrt{\frac{\mathbb{P}(d\omega)}{\mathbb{Q}(d\omega)}} \right) \right|}_{\leq B} q(\omega) d\nu(d\omega) && (\text{first part of Lemma F.6}) \\ &\leq 2B \underbrace{\int_{\omega \in \Omega} q(\omega) d\nu(d\omega)}_{=1} \leq 2B. && (\text{second part of Lemma F.6}) \end{aligned}$$

This concludes both parts of the proof. \square

The next lemma provides a lower bound on the softmax output if its input is upper-bounded (in ℓ_1 -norm).

Lemma F.7. *Let $\mathbf{x} \in \mathbb{R}^m$ be such that $\|\mathbf{x}\|_1 \leq c_1$ for some $c_1 > 0$. Then, we have*

$$\text{softmax}(\mathbf{u}) \geq \frac{1}{m \exp(2c_1)},$$

where the inequality holds for each component of $\text{softmax}(\mathbf{u})$.

Proof. Using the fact that

$$\|\mathbf{x}\|_1 = \sum_{i=1}^m |\mathbf{x}_i| \leq c_1,$$

we know that for any $i \in [m]$, we have

$$-c_1 \leq \mathbf{x}_i \leq c_1.$$

Hence, using the fact that the exponential is increasing, we have for any $i \in [m]$

$$\exp(-c_1) \leq \exp(\mathbf{x}_i) \leq \exp(c_1). \quad (15)$$

Summing and taking the inverse leads to

$$\begin{aligned} \sum_{i=1}^m \exp(-c_1) &\leq \sum_{i=1}^m \exp(\mathbf{x}_i) \leq \sum_{i=1}^m \exp(c_1) \\ \iff \frac{1}{\sum_{j=1}^m \exp(c_1)} &\leq \frac{1}{\sum_{j=1}^m \exp(\mathbf{x}_j)} \leq \frac{1}{\sum_{j=1}^m \exp(-c_1)}. \end{aligned} \quad (16)$$

Combining Eq. (15) and Eq. (16) yields

$$\frac{\exp(-c_1)}{\sum_{j=1}^m \exp(c_1)} \leq \frac{\exp(\mathbf{x}_i)}{\sum_{j=1}^m \exp(\mathbf{x}_j)} \leq \frac{\exp(c_1)}{\sum_{j=1}^m \exp(-c_1)}.$$

As we desire a lower bound, we only focus on the left-hand side of the previous inequality. Multiplying the numerator and denominator by $\exp(c_1)$ leads to

$$\forall i \in [m], \quad \text{softmax}(\mathbf{x})_i = \frac{\exp(\mathbf{x}_i)}{\sum_{j=1}^m \exp(\mathbf{x}_j)} \geq \frac{1}{m \exp(2c_1)},$$

which concludes the proof. While we only need the lower bound of Eq. (16) to obtain Lemma F.7, both bounds can be used, for instance in Xie et al. (2024, Lemma E.7) and Veličković et al. (2024, Lemma 2.1) to show that, under a mild assumption on $\mathbf{x} \in \mathbb{R}^m$, $\text{softmax}(\mathbf{x})$ behaves as $\mathcal{O}(\frac{1}{m})$ when m grows to infinity. \square

Upper-bounding the total variation. We now proceed with finding an upper bound on the total variation between the true probability of the next token and the one estimated by the LLM f_Θ . It will enable us to find the bounding vector \mathbf{c} . The next lemma shows that the input of the softmax layer of the model is bounded.

Lemma F.8. Consider an LLM $f_\Theta \in \mathcal{F}$. For any input sequence $\mathbf{S} \in \mathbb{R}^{r \times n}$, the following inequality holds

$$\left\| \frac{1}{n\tau} \mathbf{W}_U \mathbf{S}^{(L)} \mathbb{1}_n \right\|_1 \leq \frac{1}{\tau} \|\mathbf{W}_U^\top\|_{2,1},$$

where τ is the temperature, \mathbf{W}_U is the unembedding matrix (which is bounded as stated in the definition of the parameters space \mathcal{W}), and $\mathbf{S}^{(L)}$ is the output of the last transformer layer.

Proof. We recall that the layer normalization ensures that at each layer, the tokens are in the unit ℓ_2 -ball. This is, in particular, the case for the output of the last layer $\mathbf{S}^{(L)}$. It means that the columns of $\mathbf{S}^{(L)}$ verifies

$$\forall k \in [n], \quad \|\mathbf{S}_{:,k}^{(L)}\|_2 \leq 1, \quad (17)$$

which implies

$$\max_{1 \leq k \leq n} \|\mathbf{S}_{:,i}^{(L)}\|_2 \leq 1. \quad (18)$$

Recalling that the $L_{p,q}$ -norm of a matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$ can be rewritten as

$$\|\mathbf{A}\|_{p,q} := \left(\sum_{j=1}^m \left(\sum_{i=1}^n |\mathbf{A}_{ij}|^p \right)^{\frac{q}{p}} \right)^{\frac{1}{q}} = \|(\|\mathbf{A}_{:,j}\|_p)_{j=1}^m\|_q,$$

which corresponds to

$$\|\mathbf{A}\|_{2,1} = \sum_{j=1}^m \|\mathbf{A}_{:,j}\|_2 \quad (19)$$

with $(p, q) = (2, 1)$. Hence, the ℓ_1 -norm of the last layer before the softmax layer satisfies

$$\begin{aligned} \left\| \frac{1}{n\tau} \mathbf{W}_U \mathbf{S}^{(L)} \mathbb{1}_n \right\|_1 &= \frac{1}{n\tau} \sum_{i=1}^T \left| \sum_{j=1}^r \mathbf{W}_{ij} \sum_{k=1}^n \mathbf{S}_{jk} \right| = \frac{1}{n\tau} \sum_{i=1}^T \left| \sum_{j=1}^r \sum_{k=1}^n \mathbf{W}_{ij} \mathbf{S}_{jk} \right| \\ &\leq \frac{1}{n\tau} \sum_{i=1}^T \sum_{j=1}^r \sum_{k=1}^n |\mathbf{W}_{ij} \mathbf{S}_{jk}| \quad (\text{triangular inequality}) \\ &\leq \frac{1}{n\tau} \sum_{i=1}^T \sum_{k=1}^n |\mathbf{W}_i^\top \mathbf{S}_{:,k}| \\ &\leq \frac{1}{n\tau} \sum_{i=1}^T \sum_{k=1}^n \|\mathbf{W}_i\|_2 \|\mathbf{S}_{:,k}\|_2 \quad (\text{Cauchy-Schwartz inequality}) \end{aligned}$$

$$\begin{aligned}
 &\leq \frac{1}{n\tau} \sum_{i=1}^T \sum_{k=1}^n \|\mathbf{W}_i\|_2 \max_{1 \leq k \leq n} \|\mathbf{S}_{\cdot,k}\|_2 \\
 &\leq \frac{1}{n\tau} n \max_{1 \leq k \leq n} \|\mathbf{S}_{\cdot,k}\|_2 \sum_{i=1}^T \|\mathbf{W}_i\|_2 \\
 &\leq \frac{1}{\tau} \sum_{i=1}^T \|\mathbf{W}_i\|_2 \\
 &\leq \frac{1}{\tau} \sum_{i=1}^T \|(\mathbf{W}_U^\top)_{\cdot,i}\|_2 \\
 &\leq \frac{1}{\tau} \|\mathbf{W}_U^\top\|_{2,1} \quad (\text{by Eq. (18) and the def. of } L_{2,1} \text{ in Eq. (19))
 \end{aligned}$$

where we dropped the subscript and superscript on \mathbf{W}_U and $\mathbf{S}^{(L)}$ to ease the notations. This concludes the proof. \square

The previous lemma can be used to show that the logarithm of the ratio between the true probability of the next token and the one estimated by the LLM f_Θ is upper bounded as a function of the vocabulary size T , the temperature, the upper-bound on \mathbf{W}_U and some constant related to the ambiguity of language (see Eq. (1)).

Proposition F.9 (Upper-bound on the logarithm). *Consider an LLM $f_\Theta \in \mathcal{F}$ with vocabulary size T . We recall that B_U is the upper bound on the norm of \mathbf{W}_U in the definition of parameter space \mathcal{W} , τ is the softmax temperature and c_0 is the constant related to the ambiguity of language (see Eq. (1)). We have*

$$\forall n \in [N], \quad \left| \log \left(\frac{\mathbb{P}_{\mathcal{L}}(\mathbf{X}_{n+1} \mid \mathbf{S}_n)}{\mathbb{P}_{\Theta}(\mathbf{X}_{n+1} \mid \mathbf{S}_n)} \right) \right| \leq \bar{B} = \max \left\{ \log(T) + \frac{2B_U}{\tau}, \log \left(\frac{1}{c_0} \right) \right\}.$$

Proof. The main idea of the proof is to bound the probability ratio and use the fact that \log is non-decreasing. Let $n \in [N]$. The model f_Θ receives as input sequences of tokens \mathbf{S}_n of size $n \leq K$. We first lower-bound each term of the probability ratio. From Eq. (1), we have

$$\mathbb{P}_{\mathcal{L}}(\mathbf{X}_{n+1} \mid \mathbf{S}_n) \geq c_0. \quad (20)$$

We want to obtain a similar inequality for $\mathbb{P}_{\Theta}(\mathbf{X}_{n+1} \mid \mathbf{S}_n)$. As the parameters Θ of the LLM are in \mathcal{W} , we know that $\|\mathbf{W}_U^\top\|_{2,1} \leq B_U$. Lemma F.8 ensures that

$$\left\| \frac{1}{n\tau} \mathbf{W}_U \mathbf{S}^{(L)} \mathbb{1}_n \right\|_1 \leq \frac{1}{\tau} \|\mathbf{W}_U^\top\|_{2,1} \leq \frac{B_U}{\tau}.$$

We can then apply Lemma F.7 with $c_1 = \frac{B_U}{\tau}$ and given that $\frac{1}{n\tau} \mathbf{W}_U \mathbf{S}^{(L)} \mathbb{1}_n \in \mathbb{R}^T$, it leads to

$$\mathbb{P}_{\Theta}(\cdot \mid \mathbf{S}_n) = \text{softmax} \left(\frac{1}{n\tau} \mathbf{W}_U \mathbf{S}^{(L)} \mathbb{1}_n \right) \geq \frac{1}{T \exp(2B_U/\tau)},$$

where the inequality holds for each component of $\mathbb{P}_{\Theta}(\cdot \mid \mathbf{S}_n)$. This is in particular the case for $\mathbb{P}_{\Theta}(\mathbf{X}_{n+1} \mid \mathbf{S}_n)$ which is the entry we are interested in, i.e., we have

$$\mathbb{P}_{\Theta}(\mathbf{X}_{n+1} \mid \mathbf{S}_n) \geq \frac{1}{T \exp(2B_U/\tau)}. \quad (21)$$

Going back to the ratio of probability, consider the situation where we have

$$\frac{\mathbb{P}_{\mathcal{L}}(\mathbf{X}_{n+1} \mid \mathbf{S}_n)}{\mathbb{P}_{\Theta}(\mathbf{X}_{n+1} \mid \mathbf{S}_n)} \geq 1.$$

Then, using Eq. (21), we have

$$1 \leq \frac{\mathbb{P}_{\mathcal{L}}(\mathbf{X}_{n+1} \mid \mathbf{S}_n)}{\mathbb{P}_{\Theta}(\mathbf{X}_{n+1} \mid \mathbf{S}_n)} \leq \frac{1}{\mathbb{P}_{\Theta}(\mathbf{X}_{n+1} \mid \mathbf{S}_n)} \leq T \exp(2B_U/\tau),$$

which implies, as the log is non-decreasing monotonically,

$$0 \leq \log \left(\frac{\mathbb{P}_{\mathcal{L}}(\mathbf{X}_{n+1} | \mathbf{S}_n)}{\mathbb{P}_{\Theta}(\mathbf{X}_{n+1} | \mathbf{S}_n)} \right) \leq \log(T \exp(2B_U/\tau)) = \log(T) + \frac{2B_U}{\tau}. \quad (22)$$

Similarly, consider the case where we have

$$\frac{\mathbb{P}_{\mathcal{L}}(\mathbf{X}_{n+1} | \mathbf{S}_n)}{\mathbb{P}_{\Theta}(\mathbf{X}_{n+1} | \mathbf{S}_n)} \leq 1.$$

Then, we have

$$\frac{\mathbb{P}_{\Theta}(\mathbf{X}_{n+1} | \mathbf{S}_n)}{\mathbb{P}_{\mathcal{L}}(\mathbf{X}_{n+1} | \mathbf{S}_n)} \geq 1,$$

and similarly to above, we can use Eq. (20) to obtain

$$1 \leq \frac{\mathbb{P}_{\Theta}(\mathbf{X}_{n+1} | \mathbf{S}_n)}{\mathbb{P}_{\mathcal{L}}(\mathbf{X}_{n+1} | \mathbf{S}_n)} \leq \frac{1}{\mathbb{P}_{\mathcal{L}}(\mathbf{X}_{n+1} | \mathbf{S}_n)} \leq \frac{1}{c_0}.$$

This implies

$$0 \leq \log \left(\frac{\mathbb{P}_{\Theta}(\mathbf{X}_{n+1} | \mathbf{S}_n)}{\mathbb{P}_{\mathcal{L}}(\mathbf{X}_{n+1} | \mathbf{S}_n)} \right) \leq \log \left(\frac{1}{c_0} \right),$$

which also rewrites

$$0 \leq -\log \left(\frac{\mathbb{P}_{\mathcal{L}}(\mathbf{X}_{n+1} | \mathbf{S}_n)}{\mathbb{P}_{\Theta}(\mathbf{X}_{n+1} | \mathbf{S}_n)} \right) \leq \log \left(\frac{1}{c_0} \right). \quad (23)$$

By definition of the absolute value, combining Eq. (22) and Eq. (23) leads to

$$\left| \log \left(\frac{\mathbb{P}_{\mathcal{L}}(\mathbf{X}_{n+1} | \mathbf{S}_n)}{\mathbb{P}_{\Theta}(\mathbf{X}_{n+1} | \mathbf{S}_n)} \right) \right| \leq \max \left\{ \log(T) + \frac{2B_U}{\tau}, \log \left(\frac{1}{c_0} \right) \right\}.$$

This concludes the proof. \square

We are now ready to upper-bound the total variation.

Corollary F.10 (Upper-bound on the total variation). *Consider an LLM $f_{\Theta} \in \mathcal{F}$ with vocabulary size T . We recall that B_U is the upper bound on the norm of \mathbf{W}_U in the definition of parameter space \mathcal{W} , τ is the softmax temperature and c_0 is the constant related to the ambiguity of language (see Eq. (1)). For $n \in [N]$, we have*

$$d_{\text{TV}}(\mathbb{P}_{\mathcal{L}}(\cdot | \mathbf{S}_n), \mathbb{P}_{\Theta}(\cdot | \mathbf{S}_n)) \leq \sqrt{2 \max \left\{ \log(T) + \frac{2B_U}{\tau}, \log \left(\frac{1}{c_0} \right) \right\}} := c_2. \quad (24)$$

Proof. Using Proposition F.9, we can directly apply Lemma F.6 with $B = \max \left\{ \log(T) + \frac{2B_U}{\tau}, \log \left(\frac{1}{c_0} \right) \right\}$ for any $n \in [N]$. This leads to

$$\forall n \in [N], \quad d_{\text{TV}}(\mathbb{P}_{\mathcal{L}}(\cdot | \mathbf{S}_n), \mathbb{P}_{\Theta}(\cdot | \mathbf{S}_n)) \leq \sqrt{2 \max \left\{ \log(T) + \frac{2B_U}{\tau}, \log \left(\frac{1}{c_0} \right) \right\}}.$$

This concludes the proof. \square

F.5.3. CONCLUDING THE PROOF

We are now ready to state our main result.

Theorem F.11 (Restatement of Theorem 4.2). *Consider an LLM $f_{\Theta} \in \mathcal{F}$ with vocabulary size T . We denote by $\mathbf{\Gamma}$ the mixing matrix of the pretraining sequences of tokens $(\mathbf{S}_1, \dots, \mathbf{S}_{N_{\text{train}}})$. Let $\delta > 0$. Then, with probability at least $1 - \delta$,*

$$\mathcal{R}_{\text{pre}}(\Theta) \leq \widehat{\mathcal{R}}_{\text{pre}}(\Theta) + \frac{\bar{B}}{\sqrt{N_{\text{train}}}} \sqrt{\log \left(\frac{2}{\delta} \right)},$$

where \bar{B} is a constant depending on the parameters of the problem. More precisely,

$$\bar{B} = 2\|\Gamma\| \sqrt{\max\{\log(T) + \frac{2B_U}{\tau}, \log\left(\frac{1}{c_0}\right)\}}.$$

Proof of Theorem 4.2. By definition of the risk, we have

$$\begin{aligned} \widehat{\mathcal{R}}_{\text{pre}}(\Theta) &= \frac{1}{N_{\text{train}}} \sum_{n=1}^{N_{\text{train}}} \underbrace{d_{\text{TV}}(\mathbb{P}_{\mathcal{L}}(\cdot | \mathbf{S}_n), \mathbb{P}_{\Theta}(\cdot | \mathbf{S}_n))}_{=g_n(\mathbf{S}_n)} = \frac{1}{N_{\text{train}}} \sum_{n=1}^{N_{\text{train}}} g_n(\mathbf{S}_n) \\ &= f(\mathbf{S}_1, \dots, \mathbf{S}_{N_{\text{train}}}) = f(S). \end{aligned}$$

Using Corollary F.10, we know that

$$|g_n(\mathbf{S}_n)| \leq \sqrt{2 \max\{\log(T) + \frac{2B_U}{\tau}, \log\left(\frac{1}{c_0}\right)\}} := c_2.$$

By definition, each sequence of tokens \mathbf{S}_n takes its values in \mathcal{V}^n (again by abuse of notation, $n = \min\{n, K\}$) and S takes its values in $\mathcal{V}^1 \times \dots \times \mathcal{V}^{N_{\text{train}}}$. For any two sequences ζ, Σ with values in $\mathcal{V}^1 \times \dots \times \mathcal{V}^{N_{\text{train}}}$, we have

$$\begin{aligned} f(\zeta) - f(\Sigma) &= \frac{1}{N_{\text{train}}} \sum_{n=1}^{N_{\text{train}}} \left(\underbrace{d_{\text{TV}}(\mathbb{P}_{\mathcal{L}}(\cdot | \zeta_n), \mathbb{P}_{\Theta}(\cdot | \zeta_n))}_{=g_n(\zeta_n)} - \underbrace{d_{\text{TV}}(\mathbb{P}_{\mathcal{L}}(\cdot | \Sigma_n), \mathbb{P}_{\Theta}(\cdot | \Sigma_n))}_{=g_n(\Sigma_n)} \right) \\ &= \frac{1}{N_{\text{train}}} \sum_{n=1}^{N_{\text{train}}} (g_n(\zeta_n) - g_n(\Sigma_n)) \\ &= \frac{1}{N_{\text{train}}} \sum_{n=1}^{N_{\text{train}}} (g_n(\zeta_n) - g_n(\Sigma_n)) \mathbb{1}_{\{\zeta_n \neq \Sigma_n\}} \quad (\text{removing the zero terms}) \\ &\leq \left| \frac{1}{N_{\text{train}}} \sum_{n=1}^{N_{\text{train}}} (g_n(\zeta_n) - g_n(\Sigma_n)) \mathbb{1}_{\{\zeta_n \neq \Sigma_n\}} \right| \\ &\leq \frac{1}{N_{\text{train}}} \sum_{n=1}^{N_{\text{train}}} |(g_n(\zeta_n) - g_n(\Sigma_n)) \mathbb{1}_{\{\zeta_n \neq \Sigma_n\}}| \\ &\leq \frac{1}{N_{\text{train}}} \sum_{n=1}^{N_{\text{train}}} |g_n(\zeta_n) - g_n(\Sigma_n)| \mathbb{1}_{\{\zeta_n \neq \Sigma_n\}} \\ &\leq \frac{1}{N_{\text{train}}} \sum_{n=1}^{N_{\text{train}}} \left(\underbrace{|g_n(\zeta_n)|}_{\leq c_2} + \underbrace{|g_n(\Sigma_n)|}_{\leq c_2} \right) \mathbb{1}_{\{\zeta_n \neq \Sigma_n\}} \quad (\text{Corollary F.10}) \\ &\leq \frac{1}{N_{\text{train}}} \sum_{n=1}^{N_{\text{train}}} 2c_2 \mathbb{1}_{\{\zeta_n \neq \Sigma_n\}} = \sum_{n=1}^{N_{\text{train}}} \left(\frac{2c_2}{N_{\text{train}}} \right) \mathbb{1}_{\{\zeta_n \neq \Sigma_n\}}, \end{aligned}$$

where $c_2 = \sqrt{2 \max\{\log(T) + \frac{2B_U}{\tau}, \log\left(\frac{1}{c_0}\right)\}}$. As ζ and Σ were taken arbitrary, choosing $\mathbf{c} \in \mathbb{R}^{N_{\text{train}}}$ with all entries equal to $\frac{2c_2}{N_{\text{train}}}$ ensures that f verifies the condition in Proposition F.5, i.e.,

$$\forall \zeta, \Sigma, \quad f(\zeta) - f(\Sigma) \leq \sum_{n=1}^{N_{\text{train}}} \mathbf{c}_n \mathbb{1}_{\{\zeta_n \neq \Sigma_n\}}.$$

We assumed in Section 4.2 that the sequences \mathbf{S}_n were related via a Marton coupling with mixing matrix $\mathbf{\Gamma}$. Putting everything together, we can apply Proposition F.5 which leads to

$$\forall u \geq 0, \quad \mathbb{P}(|f(S) - \mathbb{E}_S[f(S)]| \geq u) \leq 2 \exp\left(\frac{-2u^2}{\|\mathbf{\Gamma}\|^2 \|\mathbf{c}\|_2^2}\right). \quad (25)$$

Let $u \geq 0$. We have the following events ordering

$$\begin{aligned} (\mathbb{E}_S[f(S)] - f(S) \geq u) &\subseteq (\mathbb{E}_S[f(S)] - f(S) \geq u) \cup (f(S) - \mathbb{E}_S[f(S)] \geq u) \\ &= (|f(S) - \mathbb{E}_S[f(S)]| \geq u). \end{aligned}$$

Hence, as u was taken arbitrary and using Eq. (25), we have

$$\forall u \geq 0, \quad \mathbb{P}(\mathbb{E}_S[f(S)] - f(S) \geq u) \leq 2 \exp\left(\frac{-2u^2}{\|\mathbf{\Gamma}\|^2 \|\mathbf{c}\|_2^2}\right).$$

We recall that by definition

$$f(S) = \widehat{\mathcal{R}}_{\text{pre}}(\Theta) \text{ and } \mathcal{R}_{\text{pre}}(\Theta) = \mathbb{E}_S[\widehat{\mathcal{R}}_{\text{pre}}(\Theta)].$$

Since the previous inequality holds for any $u \geq 0$, we can hence choose u such that

$$\begin{aligned} \delta = 2 \exp\left(\frac{-2u^2}{\|\mathbf{\Gamma}\|^2 \|\mathbf{c}\|_2^2}\right) &\iff \frac{-2u^2}{\|\mathbf{\Gamma}\|^2 \|\mathbf{c}\|_2^2} = \log\left(\frac{\delta}{2}\right) \iff u^2 = \frac{1}{2} \|\mathbf{\Gamma}\|^2 \|\mathbf{c}\|_2^2 \log\left(\frac{2}{\delta}\right) \\ &\iff u = \frac{1}{\sqrt{2}} \|\mathbf{\Gamma}\| \|\mathbf{c}\|_2 \sqrt{\log\left(\frac{2}{\delta}\right)}. \end{aligned}$$

Using the fact that

$$\|\mathbf{c}\|_2 = \sqrt{\sum_{n=1}^{N_{\text{train}}} \mathbf{c}_n^2} = \sqrt{\sum_{n=1}^{N_{\text{train}}} \left(\frac{2c_2}{N_{\text{train}}}\right)^2} = \sqrt{\sum_{n=1}^{N_{\text{train}}} \frac{4c_2^2}{N_{\text{train}}^2}} = \sqrt{\frac{4c_2^2}{N_{\text{train}}}} = \frac{2c_2}{\sqrt{N_{\text{train}}}}$$

and using the fact that $c_2 = \sqrt{2 \max\{\log(T) + \frac{2B_U}{\tau}, \log\left(\frac{1}{c_0}\right)\}}$ from Corollary F.10, we obtain

$$\begin{aligned} u &= \frac{1}{\sqrt{2}} \frac{2c_2}{\sqrt{N_{\text{train}}}} \|\mathbf{\Gamma}\| \sqrt{\log\left(\frac{2}{\delta}\right)} = \frac{\sqrt{2}c_2}{\sqrt{N_{\text{train}}}} \|\mathbf{\Gamma}\| \sqrt{\log\left(\frac{2}{\delta}\right)} \\ &= \frac{2\|\mathbf{\Gamma}\| \sqrt{\max\{\log(T) + \frac{2B_U}{\tau}, \log\left(\frac{1}{c_0}\right)\}}}{\sqrt{N_{\text{train}}}} \sqrt{\log\left(\frac{2}{\delta}\right)} = \frac{\bar{B}}{\sqrt{N_{\text{train}}}} \sqrt{\log\left(\frac{2}{\delta}\right)}, \end{aligned}$$

where we define

$$\bar{B} = 2\|\mathbf{\Gamma}\| \sqrt{\max\{\log(T) + \frac{2B_U}{\tau}, \log\left(\frac{1}{c_0}\right)\}}.$$

Putting everything together, we have

$$\mathbb{P}\left(\mathcal{R}_{\text{pre}}(\Theta) - \widehat{\mathcal{R}}_{\text{pre}}(\Theta) \geq \frac{\bar{B}}{\sqrt{N_{\text{train}}}} \sqrt{\log\left(\frac{2}{\delta}\right)}\right) \leq \delta.$$

Taking the opposite event leads to the following inequality with probability at least $1 - \delta$

$$\mathcal{R}_{\text{pre}}(\Theta) \leq \widehat{\mathcal{R}}_{\text{pre}}(\Theta) + \frac{\bar{B}}{\sqrt{N_{\text{train}}}} \sqrt{\log\left(\frac{2}{\delta}\right)},$$

which concludes the proof. \square

F.6. Proof of Corollary E.3

As the layer norm is not applied anymore, each token is no longer in the ℓ_2 -unit ball, and Lemma F.8 does not hold anymore. We want to provide an analogous lemma for our setting. We first prove the following technical lemmas.

Lemma F.12. *The ReLU is a norm-decreasing operator, i.e., we have*

$$\forall \mathbf{A} \in \mathbb{R}^{n \times m}, \quad \|\text{ReLU}(\mathbf{A})\|_{1,1} \leq \|\mathbf{A}\|_{1,1},$$

where the ReLU is applied entry-wise.

Proof. Recalling that $\text{ReLU}(x) = \max\{0, x\}$ is applied entry-wise, using the fact that $|\max\{0, x\}| \leq |x|$ and considering \mathbf{A} and $\tilde{\mathbf{A}} = \text{ReLU}(\mathbf{A})$, we have

$$\|\tilde{\mathbf{A}}\|_{1,1} = \sum_{i,j} |\tilde{\mathbf{A}}_{i,j}| = \sum_{i,j} |\max\{0, \tilde{\mathbf{A}}_{i,j}\}| \leq \sum_{i,j} |\mathbf{A}_{i,j}| \leq \|\mathbf{A}\|_{1,1},$$

which concludes the proof. \square

Lemma F.13. *The $L_{1,1}$ -norm verifies the following property:*

$$\forall \mathbf{A} \in \mathbb{R}^{n \times m}, \mathbf{B} \in \mathbb{R}^{m \times p}, \quad \|\mathbf{AB}\|_{1,1} \leq n \|\mathbf{A}\|_{\infty} \|\mathbf{B}\|_{1,1}.$$

Proof. We have

$$\begin{aligned} \|\mathbf{AB}\|_{1,1} &= \sum_{j=1}^p \sum_{i=1}^n |(\mathbf{AB})_{ij}| = \sum_{j=1}^p \sum_{i=1}^n \left| \sum_{k=1}^m \mathbf{A}_{ik} \mathbf{B}_{kj} \right| \leq \sum_{j=1}^p \sum_{i=1}^n \sum_{k=1}^m |\mathbf{A}_{ik} \mathbf{B}_{kj}| \\ &\leq \sum_{j=1}^p \sum_{i=1}^n \sum_{k=1}^m |\mathbf{A}_{ik}| |\mathbf{B}_{kj}| \leq \max_{ik} |\mathbf{A}_{ik}| \sum_{j=1}^p \sum_{i=1}^n \sum_{k=1}^m |\mathbf{B}_{kj}| \\ &\leq n \|\mathbf{A}\|_{\infty} \sum_{j=1}^p \sum_{k=1}^m |\mathbf{B}_{kj}| \leq n \|\mathbf{A}\|_{\infty} \|\mathbf{B}\|_{1,1}, \end{aligned}$$

which concludes the proof. \square

Lemma F.14. *The $L_{2,1}$ and $L_{\infty,1}$ -norms verify the following relation*

$$\forall \mathbf{A} \in \mathbb{R}^{n \times m}, \quad \|\mathbf{A}\|_{\infty,1} \leq \|\mathbf{A}\|_{2,1}.$$

Proof. By definition of the $L_{p,q}$ -norm, we have

$$\begin{aligned} \|\mathbf{A}\|_{\infty,1} &= \sum_{j=1}^M \max_{1 \leq i \leq n} |\mathbf{A}_{ij}| = \sum_{j=1}^M \sqrt{\max_{1 \leq i \leq n} |\mathbf{A}_{ij}^2|} && \text{(as } x \rightarrow x^2 \text{ is increasing)} \\ &\leq \sum_{j=1}^M \sqrt{\sum_{i=1}^n |\mathbf{A}_{ij}^2|} \leq \sum_{j=1}^M \|\mathbf{A}_{\cdot,j}\|_2 \leq \|\mathbf{A}\|_{2,1}, \end{aligned}$$

where the first inequality comes from adding non-negative terms. \square

We are now ready to state the lemma analogous to Lemma F.8.

Lemma F.15. Consider an LLM $f_{\Theta} \in \tilde{\mathcal{F}}$ with L layers. For any input sequence $\mathbf{S} \in \mathbb{R}^{r \times n}$, the following inequality holds

$$\left\| \frac{1}{n\tau} \mathbf{W}_U \mathbf{S}^{(L)} \mathbb{1}_n \right\|_1 \leq \frac{c_3}{\tau} \|\mathbf{W}_U^\top\|_{2,1},$$

where τ is the temperature and c_3 is a constant depending on the parameters upper-bound. More precisely,

$$c_3 = \left[(1 + rmB_1B_2) \cdot \left(1 + \frac{r^3}{H} B_O B_V \right) \right]^L \cdot B_{\text{tok}}.$$

\mathbf{W}_U is the unembedding matrix (which is bounded as stated in the definition of the parameters space \mathcal{W}), and $\mathbf{S}^{(L)}$ is the output of the last transformer layer.

Proof of Lemma F.15. Our model $f_{\Theta} \in \tilde{\mathcal{F}}$ is given as input a sequence $\mathbf{S} \in \mathbb{R}^{r \times n}$. With similar computations than in Lemma F.8, we have

$$\begin{aligned} \frac{1}{n\tau} \|\mathbf{W}_U \mathbf{S}^{(L)} \mathbb{1}_n\|_1 &= \frac{1}{n\tau} \sum_{i=1}^T \left| \sum_{j=1}^r \mathbf{W}_{ij} \sum_{k=1}^n \mathbf{S}_{jk} \right| = \frac{1}{n\tau} \sum_{i=1}^T \left| \sum_{j=1}^r \sum_{k=1}^n \mathbf{W}_{ij} \mathbf{S}_{jk} \right| \\ &\leq \frac{1}{n\tau} \sum_{i=1}^T \sum_{j=1}^r \sum_{k=1}^n |\mathbf{W}_{ij} \mathbf{S}_{jk}| && \text{(triangular inequality)} \\ &\leq \frac{1}{n\tau} \sum_{i=1}^T \sum_{k=1}^n |\mathbf{W}_i^\top \mathbf{S}_{\cdot,k}| \leq \frac{1}{n\tau} \sum_{i=1}^T \sum_{k=1}^n \|\mathbf{W}_i\|_\infty \|\mathbf{S}_{\cdot,k}\|_1 && \text{(Hölder inequality)} \\ &\leq \frac{1}{n\tau} \left(\sum_{i=1}^T \|\mathbf{W}_i\|_\infty \right) \cdot \left(\sum_{k=1}^n \|\mathbf{S}_{\cdot,k}\|_1 \right) \leq \frac{1}{n\tau} \|\mathbf{W}_U^\top\|_{\infty,1} \|\mathbf{S}^{(L)}\|_{1,1} \\ &\leq \frac{1}{n\tau} \|\mathbf{W}_U^\top\|_{2,1} \|\mathbf{S}^{(L)}\|_{1,1}, && \text{(Lemma F.14)} \end{aligned}$$

where, again, we dropped the subscript and superscript on \mathbf{W}_U and $\mathbf{S}^{(L)}$ to ease the notations. We obtain

$$\left\| \frac{1}{n\tau} \mathbf{W}_U \mathbf{S}^{(L)} \mathbb{1}_n \right\|_1 \leq \frac{1}{n\tau} \|\mathbf{W}_U^\top\|_{2,1} \|\mathbf{S}^{(L)}\|_{1,1}. \quad (26)$$

As we do not use layer normalization, we want to find another way to bound $\mathbf{S}^{(L)}$. To that end, we will first express $\mathbf{S}^{(\ell)}$, the output of the (ℓ) -th layer of the transformer, as a function of $\mathbf{S}^{(\ell-1)}$, the output of the $(\ell-1)$ -th layer. Using the definition of the transformer model (see Appendix B), we have

$$\begin{cases} \mathbf{Z}^{(\ell)} = \mathbf{S}^{(\ell-1)} + \mathcal{A}\left(\mathbf{S}^{(\ell-1)}; \mathbf{W}_Q^{(\ell)}, \mathbf{W}_K^{(\ell)}, \mathbf{W}_V^{(\ell)}, \mathbf{W}_O^{(\ell)}\right), \\ \mathbf{Y}^{(\ell)} = \mathbf{W}_2^{(\ell)} \text{ReLU}\left(\mathbf{W}_1^{(\ell)} \mathbf{Z}^{(\ell)}\right), \\ \mathbf{S}^{(\ell)} = \mathbf{Z}^{(\ell)} + \mathbf{Y}^{(\ell)}. \end{cases}$$

We will compute each layer's $L_{1,1}$ -norm.

Step 1: MHA. By definition, denoting the number of heads by H , we know that $\mathcal{A}\left(\mathbf{S}^{(\ell-1)}; \mathbf{W}_Q^{(\ell)}, \mathbf{W}_K^{(\ell)}, \mathbf{W}_V^{(\ell)}, \mathbf{W}_O^{(\ell)}\right) \in \mathbb{R}^{r \times n}$ multiplies $\mathbf{W}^{(\ell)} \in \mathbb{R}^{r \times r}$ with the concatenation on the rows of the H softmax layers that each writes

$$\text{softmax}\left(\mathbf{W}_Q^{(\ell)} \mathbf{S}^{(\ell)} \left(\mathbf{W}_K^{(\ell)} \mathbf{S}^{(\ell-1)}\right)^\top / \sqrt{r}\right) \left(\mathbf{W}_V^{(\ell)} \mathbf{S}^{(\ell-1)}\right) \in \mathbb{R}^{\frac{r}{H} \times n},$$

We keep the notations ℓ without explicating the index of the head to ease notations. Denoting the concatenation on the rows by $\mathbf{C}^{(\ell)} \in \mathbb{R}^{r \times n}$, we have

$$\|\mathcal{A}\left(\mathbf{S}^{(\ell-1)}; \mathbf{W}_Q^{(\ell)}, \mathbf{W}_K^{(\ell)}, \mathbf{W}_V^{(\ell)}, \mathbf{W}_O^{(\ell)}\right)\|_{1,1} = \|\mathbf{W}_O^{(\ell)} \mathbf{C}^{(\ell)}\|_{1,1} \leq r \cdot \|\mathbf{W}_O^{(\ell)}\|_\infty \|\mathbf{C}^{(\ell)}\|_{1,1}$$

$$\leq rB_O\|\mathbf{C}^{(\ell)}\|_{1,1}. \quad (\text{definition of } \tilde{\mathcal{W}})$$

Moreover, by definition of $\mathbf{C}^{(\ell)}$, we have

$$\|\mathbf{C}^{(\ell)}\|_{1,1} = \sum_{j=1}^r \sum_{i=1}^r |\mathbf{C}_{ij}^{(\ell)}| = \sum_{j=1}^r \sum_{i=1}^{r/H} \sum_{h=1}^H |\mathbf{C}_{ij}^{(\ell,h)}| = \sum_{h=1}^H \|\mathbf{C}^{(\ell,h)}\|_{1,1}, \quad (27)$$

where $\mathbf{C}^{(\ell,h)} \in \mathbb{R}^{\frac{r}{H} \times n}$ is the softmax matrix of the h -th layer. We recall that the softmax matrix is a row-stochastic matrix of $\mathbb{R}^{\frac{r}{H} \times r}$ so it has all values lower than 1. In the next computations, we drop the h index on the query, key, and value matrices to ease the notations. Using Lemma F.13 on the softmax matrix and on the value matrix $\mathbf{W}_V^{(\ell)} \in \mathbb{R}^{\frac{r}{H} \times r}$, we have

$$\begin{aligned} \|\mathbf{C}^{(\ell,h)}\|_{1,1} &= \|\text{softmax}\left(\mathbf{W}_Q^{(\ell)}\mathbf{S}^{(\ell)}\left(\mathbf{W}_K^{(\ell)}\mathbf{S}^{(\ell-1)}\right)^\top / \sqrt{r}\right)\left(\mathbf{W}_V^{(\ell)}\mathbf{S}^{(\ell-1)}\right)\|_{1,1} \\ &\leq \frac{r}{H} \cdot \|\text{softmax}\left(\mathbf{W}_Q^{(\ell)}\mathbf{S}^{(\ell)}\left(\mathbf{W}_K^{(\ell)}\mathbf{S}^{(\ell-1)}\right)^\top / \sqrt{r}\right)\|_\infty \cdot \|\mathbf{W}_V^{(\ell)}\mathbf{S}^{(\ell-1)}\|_{1,1} \\ &\leq \frac{r}{H} \cdot \|\mathbf{W}_V^{(\ell)}\mathbf{S}^{(\ell-1)}\|_{1,1} \quad (\text{the softmax matrix is row-stochastic}) \\ &\leq \frac{r}{H} \cdot \frac{r}{H} \|\mathbf{W}_V^{(\ell)}\|_\infty \|\mathbf{S}^{(\ell-1)}\|_{1,1} \leq \left(\frac{r}{H}\right)^2 B_V \|\mathbf{S}^{(\ell-1)}\|_{1,1}. \quad (\text{definition of } \tilde{\mathcal{W}}) \end{aligned}$$

Combining the previous inequality with Eq. (27) leads to

$$\|\mathbf{C}^{(\ell)}\|_{1,1} \leq \frac{r^2}{H} B_V \|\mathbf{S}^{(\ell-1)}\|_{1,1}.$$

In the end, the multi-head attention norm verifies

$$\|\mathcal{A}(\mathbf{S}^{(\ell-1)}; \mathbf{W}_Q^{(\ell)}, \mathbf{W}_K^{(\ell)}, \mathbf{W}_V^{(\ell)}, \mathbf{W}_O^{(\ell)})\|_{1,1} \leq \frac{r^3}{H} B_O B_V \|\mathbf{S}^{(\ell-1)}\|_{1,1}.$$

Using the triangular inequality, we obtain

$$\|\mathbf{Z}^{(\ell)}\|_{1,1} \leq \left(1 + \frac{r^3}{H} B_O B_V\right) \cdot \|\mathbf{S}^{(\ell-1)}\|_{1,1}. \quad (28)$$

Step 2: FF. We recall that $\mathbf{W}_1 \in \mathbb{R}^{m \times r}$ and $\mathbf{W}_2 \in \mathbb{R}^{r \times m}$. Using similar arguments to the above, we have

$$\begin{aligned} \|\mathbf{Y}^{(\ell)}\|_{1,1} &= \|\mathbf{W}_2^{(\ell)} \text{ReLU}\left(\mathbf{W}_1^{(\ell)}\mathbf{Z}^{(\ell)}\right)\|_{1,1} \\ &\leq r \cdot \|\mathbf{W}_2^{(\ell)}\|_\infty \|\text{ReLU}\left(\mathbf{W}_1^{(\ell)}\mathbf{Z}^{(\ell)}\right)\|_{1,1} \quad (\text{Lemma F.13}) \\ &\leq r \cdot \|\mathbf{W}_2^{(\ell)}\|_\infty \|\mathbf{W}_1^{(\ell)}\mathbf{Z}^{(\ell)}\|_{1,1} \quad (\text{Lemma F.12}) \\ &\leq r \cdot m \cdot \|\mathbf{W}_2^{(\ell)}\|_\infty \|\mathbf{W}_1^{(\ell)}\|_\infty \|\mathbf{Z}^{(\ell)}\|_{1,1} \quad (\text{Lemma F.13}) \\ &\leq rmB_1B_2\|\mathbf{Z}^{(\ell)}\|_{1,1}. \quad (\text{definition of } \tilde{\mathcal{W}}) \end{aligned}$$

Step 3: output layer. Again, applying the triangular inequality and using the previous inequality and Eq. (28), we have

$$\begin{aligned} \|\mathbf{S}^{(\ell)}\|_{1,1} &\leq \|\mathbf{Z}^{(\ell)}\|_{1,1} + \|\mathbf{Y}^{(\ell)}\|_{1,1} \leq (1 + rmB_1B_2)\|\mathbf{Z}^{(\ell)}\|_{1,1} \\ &\leq (1 + rmB_1B_2) \left(1 + \frac{r^3}{H} B_O B_V\right) \|\mathbf{S}^{(\ell-1)}\|_{1,1}. \end{aligned}$$

Iterating through the layers, recalling that $\mathbf{S}^{(0)} = \mathbf{S}$, we finally obtain

$$\|\mathbf{S}^{(L)}\|_{1,1} \leq \left[(1 + rmB_1B_2) \left(1 + \frac{r^3}{H} B_O B_V\right) \right]^L \|\mathbf{S}\|_{1,1},$$

where \mathbf{S} is the input sequence. Combining this inequality with Eq. (26) leads to

$$\left\| \frac{1}{n\tau} \mathbf{W}_U \mathbf{S}^{(L)} \mathbb{1}_n \right\|_1 \leq \left[(1 + rmB_1B_2) \left(1 + \frac{r^3}{H} B_O B_V \right) \right]^L \frac{\|\mathbf{S}\|_{1,1}}{n} \left(\frac{1}{\tau} \|\mathbf{W}_U^\top\|_{2,1} \right).$$

Using the fact that each token has a ℓ_1 -norm bounded by B_{tok} . Hence, each column of \mathbf{S} is too and we have

$$\frac{1}{n} \|\mathbf{S}\|_{1,1} = \frac{1}{n} \sum_{j=1}^n \sum_{i=1}^r |\mathbf{S}_{ij}| = \frac{1}{n} \sum_{j=1}^n \underbrace{\|\mathbf{S}_{:,j}\|_1}_{\leq B_{\text{tok}}} \leq B_{\text{tok}}.$$

Combining the last two inequalities concludes the proof. \square

We can now restate Corollary E.3.

Corollary F.16 (Restatement of Corollary E.3). *Consider an LLM $f_\Theta \in \tilde{\mathcal{F}}$ with vocabulary size T composed of L transformer blocks and H attention heads. We denote by $\mathbf{\Gamma}$ the mixing matrix of the pretraining sequences of tokens $(\mathbf{S}_1, \dots, \mathbf{S}_{N_{\text{train}}})$. Let $\delta > 0$. Then, with probability at least $1 - \delta$,*

$$\mathcal{R}_{\text{pre}}(\Theta) \leq \hat{\mathcal{R}}_{\text{pre}}(\Theta) + \frac{\bar{B}}{\sqrt{N_{\text{train}}}} \sqrt{\log \left(\frac{2}{\delta} \right)},$$

where \bar{B} is a constant depending on the parameters of the problem. More precisely,

$$\bar{B} = 2 \|\mathbf{\Gamma}\| \sqrt{\max \left\{ \log(T) + \frac{2(B_\Theta)^L}{\tau}, \log \left(\frac{1}{c_0} \right) \right\}},$$

with $B_\Theta = \left[(1 + rmB_1B_2) \left(1 + \frac{r^3}{H} B_O B_V \right) \right] (B_{\text{tok}} B_U)^{1/L}$.

Proof of Corollary E.3. We first note that the only change from Lemma F.8 to Lemma F.15 is the multiplicative constant $c_3 = \left[(1 + rmB_1B_2) \left(1 + \frac{r^3}{H} B_O B_V \right) \right]^L B_{\text{tok}}$ in front of $\frac{1}{\tau} \|\mathbf{W}_U^\top\|_{2,1}$. In particular, as we know that $\tilde{\mathcal{W}} \subset \mathcal{W}$, we also have $\|\mathbf{W}_U^\top\|_{2,1} \leq B_U$. Hence, we can apply the proof of Theorem 4.2 in a straightforward manner by changing $\frac{B_U}{\tau}$ by $c_3 \cdot \frac{B_U}{\tau}$. This concludes the proof. \square

F.7. Proof of Theorem 4.3

In this section, we detail the proof of Theorem 4.3. We first recall the problem setup.

Markov chains inputs. In this section, we give as input of the model a single Markov chain $X = (\mathbf{X}_1, \dots, \mathbf{X}_{N_{\text{icl}}})$ with finite, discrete state space Ω of size d with transition probability \mathbb{P} . We assume the \mathbf{X}_n are already tokenized and thus we have $\Omega \subset \mathcal{V}$. We denote the sequence of tokens the LLM receives by $\mathbf{S}_n = (\mathbf{X}_1, \dots, \mathbf{X}_n)$ if $n \leq K$ and $\mathbf{S}_n = (\mathbf{X}_{n-K+1}, \dots, \mathbf{X}_n)$ otherwise due to the deletion process (see Definition B.2). In particular, the \mathbf{S}_n are elements of \mathcal{V}_K^* . We note that $S = (\mathbf{S}_1, \dots, \mathbf{S}_{N_{\text{icl}}})$ is also a Markov chain (see Appendix F.7.1). By definition of \mathbb{P} , we know that for any $n \in [N_{\text{icl}}]$, the next token \mathbf{X}_{n+1} follows the distribution $\mathbb{P}(\cdot | \mathbf{S}_n)$. We assume that there exists a positive constant p_{\min} that lower bounds all the transition probability between states, i.e., $\forall n \in [N_{\text{icl}}], \forall x, y \in \Omega, \mathbb{P}(\mathbf{X}_{n+1} = y | \mathbf{X}_n = x) \geq p_{\min} > 0$. This is akin to the ambiguity of language constant c_0 considered in the previous section and in Hu et al. (2024); Wies et al. (2024); Xie et al. (2022); Zhang et al. (2023b).

Next token probability distribution. An important difference with the setting considered in Theorem 4.2 is that here, we predict a probability distribution on the state space Ω of the Markov chain and not on the vocabulary of the LLM \mathcal{V} . To that end, we restrict the predicted probability given the past tokens \mathbf{S}_n to the state space Ω . Formally, denoting the output of the last layer of f_Θ by $\mathbf{S}^{(L)}$, the last layer before the softmax outputs a vector $\mathbf{u} = \frac{1}{n\tau} \mathbf{W}_U \mathbf{S}^{(L)} \mathbb{1}_n \in \mathbb{R}^T$. We first extract

the entries of \mathbf{u} whose index i are such that the i -th element of the vocabulary space \mathcal{V} is in Ω . This can be formalized as follows. We denote by $\mathbb{I}_d = (i_1 \leq i_2 \leq \dots \leq i_d) \in [T]^d$ the subset of d distinct elements of $[T]$ and consider the matrix $\mathbf{M}_j = \mathbf{e}_{i_j}^\top$, where $\mathbf{e}_{i_j} \in \mathbb{R}^T$ has value 1 at entry $i_j \in \mathbb{I}$ and 0 elsewhere. Extracting only the d entries of \mathbf{u} that corresponds to the state space yields a vector in \mathbb{R}^d that writes $\mathbf{v} = \frac{1}{n\tau} \mathbf{M} \mathbf{W}_U \mathbf{S}^{(L)} \mathbb{1}_n \in \mathbb{R}^T$. Similarly to Appendix B, the probability distribution of next token \mathbf{X}_{n+1} provided by the LLM f_{Θ} now writes

$$\mathbb{P}_{\Theta}(\cdot | \mathbf{S}_n) = \text{softmax}\left(\frac{1}{n\tau} \mathbf{M} \mathbf{W}_U \mathbf{S}^{(L)} \mathbb{1}_n\right) \in \Delta_d.$$

We aim to obtain a similar generalization bound than in Theorem 4.2 where the reference probability distribution is the Markov chain transition probability \mathbb{P} instead of the probability distribution of language $\mathbb{P}_{\mathcal{L}}$. In particular, \mathbb{P} will replace $\mathbb{P}_{\mathcal{L}}$ in the definition of the risks in Eq. (3). We provide below an overview of the proof before detailing it.

Overview of the proof. We are going to use McDiarmid’s inequality for Markov chains of Paulin (2015, Corollary 2.11). To adapt their arguments to our setting, we bound the total variation between the true probability of the next token and the one estimated by the LLM. The rest of this section is organized as follows. First, in Appendix F.7.1, we show that $S = (\mathbf{S}_1, \dots, \mathbf{S}_{N_{\text{icl}}})$ is a Markov chain. Then in Appendix F.7.2, we adapt the concentration inequality of Paulin (2015, Corollary 2.11). Afterwards in Appendix F.7.3, we show how to bound the total variation between the true and the estimated probability of the next token. Finally Appendix F.7.4 concludes the proof.

F.7.1. CONNECTION BETWEEN TOKENS AND SEQUENCES OF TOKENS MARKOV CHAINS

We first show that $S = (\mathbf{S}_1, \dots, \mathbf{S}_{N_{\text{icl}}})$ is also a Markov chain.

Lemma F.17. *Consider a sequence (not necessarily a Markov chain) $X = (\mathbf{X}_1, \dots, \mathbf{X}_N)$ with values in Ω and let $\mathbf{S}_n = (\mathbf{X}_1, \dots, \mathbf{X}_n)$ if $n < K$ and $\mathbf{S}_n = (\mathbf{X}_{n-K+1}, \dots, \mathbf{X}_n)$ otherwise. Then, the sequence $S = (\mathbf{S}_1, \dots, \mathbf{S}_N)$ is a Markov chain with state space Ω_K^* that contains the sequence of elements in Ω of length smaller than K .*

Proof. By definition of the \mathbf{S}_n , we know that they take values in Ω_K^* . Let $x_1, \dots, x_{n+1} \in \Omega$. We first assume that $n > K$ and denote $s_i = (x_{n-K+1}, \dots, x_i)$. We have

$$\begin{aligned} & \mathbb{P}(\mathbf{S}_{n+1} = s_{n+1} | \mathbf{S}_n = s_n, \dots, \mathbf{S}_{n-K+1} = s_{n-K+1}) \\ &= \mathbb{P}(\mathbf{S}_{n+1} = s_{n+1} | \mathbf{X}_n = x_n, \dots, \mathbf{X}_{n-K+1} = x_{n-K+1}) \\ &= \mathbb{P}(\mathbf{S}_{n+1} = s_{n+1} | \mathbf{S}_n = s_n). \end{aligned} \quad (\text{by definition of } \mathbf{S}_n)$$

Similarly, we assume $n < K$ and denote $s_i = (x_1, \dots, x_i)$. We have

$$\begin{aligned} & \mathbb{P}(\mathbf{S}_{n+1} = s_{n+1} | \mathbf{S}_n = s_n, \dots, \mathbf{S}_1 = s_1) \\ &= \mathbb{P}(\mathbf{S}_{n+1} = s_{n+1} | \mathbf{X}_n = x_n, \dots, \mathbf{X}_1 = x_1) \\ &= \mathbb{P}(\mathbf{S}_{n+1} = s_{n+1} | \mathbf{S}_n = s_n). \end{aligned} \quad (\text{by definition of } \mathbf{S}_n)$$

Finally, for $n = K$, we denote $s_i = (x_1, \dots, x_i)$ for $i \leq K$ and $s_{K+1} = (x_2, \dots, x_{K+1})$. We have

$$\begin{aligned} & \mathbb{P}(\mathbf{S}_{K+1} = s_{K+1} | \mathbf{S}_n = s_n, \dots, \mathbf{S}_2 = s_2) \\ &= \mathbb{P}(\mathbf{S}_{K+1} = s_{K+1} | \mathbf{X}_K = x_K, \dots, \mathbf{X}_1 = x_1) \\ &= \mathbb{P}(\mathbf{S}_{K+1} = s_{K+1} | \mathbf{S}_K = s_K). \end{aligned} \quad (\text{by definition of } \mathbf{S}_K)$$

This establishes the Markov property for \mathbf{S} . □

F.7.2. CONCENTRATION INEQUALITIES FOR MARKOV CHAINS

We first state a concentration inequality for time-homogeneous Markov chains that will be used to obtain our final bound.

Proposition F.18 (McDiarmid’s inequality for time-homogeneous Markov chains). *Let $S := (\mathbf{S}_1, \dots, \mathbf{S}_N)$ be a Markov chain with value in a discrete, finite state space Ω and mixing time $t_{\text{mix}}(\varepsilon)$. Let $t_{\text{min}} := \inf_{0 \leq \varepsilon < 1} t_{\text{mix}}(\frac{\varepsilon}{2}) \cdot \left(\frac{2-\varepsilon}{1-\varepsilon}\right)^2$. If $f: \Omega \rightarrow \mathbb{R}$ is such that there exists $\mathbf{c} \in \mathbb{R}^N$ satisfying*

$$\forall \mathbf{x}, \mathbf{y} \in \Omega, \quad f(\mathbf{x}) - f(\mathbf{y}) \leq \sum_{i=1}^N \mathbf{c}_i \mathbb{1}_{\{\mathbf{x}_i \neq \mathbf{y}_i\}},$$

then we have for any $u \geq 0$,

$$\mathbb{P}(|f(S) - \mathbb{E}_S[f(S)]| \geq u) \leq 2 \exp\left(\frac{-2u^2}{\|\mathbf{c}\|_2^2 \cdot t_{\text{min}}}\right).$$

Proof. We recall that Corollary 2.11 of [Paulin \(2015\)](#) ensures that for such a function f , we have

$$\mathbb{P}(|f(S) - \mathbb{E}[f(S)]| \geq u) \leq 2 \exp\left(\frac{-2u^2}{\|\mathbf{c}\|_2^2 \cdot \tau_{\text{min}}}\right), \quad (29)$$

where τ_{min} is defined as

$$\tau_{\text{min}} := \inf_{0 \leq \varepsilon < 1} \tau(\varepsilon) \left(\frac{2-\varepsilon}{1-\varepsilon}\right)^2,$$

with $\tau(\varepsilon)$ being the mixing time of a Markov chain *without assuming time homogeneity* (see [Paulin \(2015, Definition 1.4\)](#)). As in our case, we assume the time homogeneity, this inequality in Eq. (29) has to be adapted. Following Remark 1.5 of [Paulin \(2015\)](#), we notice that

$$\forall \varepsilon \in [0, 1], \quad \tau(2\varepsilon) \leq t_{\text{mix}}(\varepsilon) \leq \tau(\varepsilon).$$

Let $0 \leq \varepsilon < 1$. Using the fact that $\left(\frac{2-\varepsilon}{1-\varepsilon}\right)^2 > 0$, the previous inequality ensures

$$\tau(\varepsilon) \leq t_{\text{mix}}\left(\frac{\varepsilon}{2}\right) \iff \tau(\varepsilon) \left(\frac{2-\varepsilon}{1-\varepsilon}\right)^2 \leq t_{\text{mix}}\left(\frac{\varepsilon}{2}\right) \left(\frac{2-\varepsilon}{1-\varepsilon}\right)^2.$$

Taking the infimum on the left-hand side leads to

$$\tau_{\text{min}} = \inf_{0 \leq \varepsilon < 1} \tau(\varepsilon) \left(\frac{2-\varepsilon}{1-\varepsilon}\right)^2 \leq t_{\text{mix}}\left(\frac{\varepsilon}{2}\right) \left(\frac{2-\varepsilon}{1-\varepsilon}\right)^2.$$

As we took ε arbitrary in $[0, 1)$, we can take the infimum on the right-hand side, which leads to

$$\tau_{\text{min}} \leq t_{\text{min}}.$$

As the function $x \rightarrow \exp\left(\frac{-2u^2}{\|\mathbf{c}\|_2^2 x}\right)$ is decreasing, we finally obtain

$$\exp\left(\frac{-2u^2}{\|\mathbf{c}\|_2^2 \tau_{\text{min}}}\right) \leq \exp\left(\frac{-2u^2}{\|\mathbf{c}\|_2^2 t_{\text{min}}}\right). \quad (30)$$

Combining Eqs. (29) and (30) concludes the proof. \square

Similarly to Theorem 4.2, we want to apply Proposition F.18 to a function f that consists of sums of total variation. We investigate in the next section how to find the bounding vector \mathbf{c} to apply Proposition F.18.

F.7.3. FINDING THE BOUNDING VECTOR

We want to apply the same arguments as in the proof of Theorem 4.2 to find the bounding vector \mathbf{c} . The only difference in terms of setting is the definition of the probability of the next token. Indeed, in our case, we apply an extraction matrix $\mathbf{M} \in \mathbb{R}^{d \times T}$ to recover the d states of the input Markov chain. We first prove the following technical lemma.

Lemma F.19. Let $d \leq T$ and consider a subset of d distinct elements of $[T]$ that writes $\mathbb{I}_d = (i_1 \leq i_2 \leq \dots \leq i_d) \in [T]^d$. We denote by $\mathbf{M} \in \mathbb{R}^{d \times T}$ the matrix with rows $\mathbf{M}_j = \mathbf{e}_{i_j}^\top$, where $\mathbf{e}_{i_j} \in \mathbb{R}^T$ has value 1 at entry $i_j \in \mathbb{I}$ and 0 elsewhere. For any vector $\mathbf{u} \in \mathbb{R}^T$, we have

$$\|\mathbf{M}\mathbf{u}\|_1 \leq \|\mathbf{u}\|_1.$$

Proof. By definition of the ℓ_1 -norm, we have

$$\|\mathbf{M}\mathbf{u}\|_1 = \sum_{k=1}^d \left| \sum_{l=1}^T \mathbf{M}_{kl} \mathbf{u}_l \right| \leq \sum_{k=1}^d \sum_{l=1}^T |\mathbf{M}_{kl} \mathbf{u}_l| \leq \sum_{l=1}^T |\mathbf{u}_l| \sum_{k=1}^d |\mathbf{M}_{kl}|.$$

Moreover, each column of \mathbf{M} contains at most one non-zero entry (with value 1). Otherwise, it means that two \mathbf{e}_{i_j} are identical (as they only have one non-zero entry with value 1, having it at the same position ensures their equality) which contradicts the fact that the i_j were taken distinct. Hence, for all l , we have $\sum_{k=1}^d |\mathbf{M}_{kl}| \leq 1$, which concludes the proof. \square

We now prove a lemma analogous to Lemma F.8.

Lemma F.20. Let $\mathbf{S} \in \mathbb{R}^{r \times n}$ denote the entry of the LLM f_Θ and $\mathbf{S}^{(L)}$ denote the output of the last layer before the softmax. Let $d \leq T$ and consider a subset of d distinct elements of $[T]$ that writes $\mathbb{I}_d = (i_1 \leq i_2 \leq \dots \leq i_d) \in [T]^d$. We denote by $\mathbf{M} \in \mathbb{R}^{d \times T}$ the matrix with rows $\mathbf{M}_j = \mathbf{e}_{i_j}^\top$, where $\mathbf{e}_{i_j} \in \mathbb{R}^T$ has value 1 at entry $i_j \in \mathbb{I}$ and 0 elsewhere. Then, the following inequality holds

$$\frac{1}{n\tau} \|\mathbf{M}\mathbf{W}_U \mathbf{S}^{(L)} \mathbb{1}_n\|_1 \leq \frac{1}{\tau} \|\mathbf{W}_U^\top\|_{2,1}.$$

Proof. Applying Lemma F.19 with the matrix $\mathbf{M} \in \mathbb{R}^d$ and the vector $\frac{1}{n\tau} \mathbf{W}_U \mathbf{X}^{(L)} \mathbb{1}_n \in \mathbb{R}^T$ leads to

$$\frac{1}{n\tau} \|\mathbf{M}\mathbf{W}_U \mathbf{S}^{(L)} \mathbb{1}_n\|_1 \leq \frac{1}{n\tau} \|\mathbf{W}_U \mathbf{X}^{(L)} \mathbb{1}_n\|_1.$$

Applying Lemma F.8 concludes the proof. \square

The previous lemma can be used to show that the logarithm of the ratio between the true probability of the next token and the one estimated by the LLM f_Θ is upper bounded as a function of the number of states of the Markov chain d , the temperature τ , the upper-bound on \mathbf{W}_U and some constant related to the ambiguity of language (see Eq. (1)).

Proposition F.21 (Upper-bound on the logarithm). Consider an LLM $f_\Theta \in \mathcal{F}$ and an input Markov chain $X = (\mathbf{X}_1, \dots, \mathbf{X}_{N_{\text{icl}}})$ with d states. We recall that B_U is the upper bound on the norm of \mathbf{W}_U in the definition of parameter space \mathcal{W} , τ is the softmax temperature, and p_{\min} is the constant related to the minimal transition probability between states. We have

$$\forall n \in [N], \quad \left| \log \left(\frac{\mathbb{P}(\mathbf{X}_{n+1} | \mathbf{S}_n)}{\mathbb{P}_\Theta(\mathbf{X}_{n+1} | \mathbf{S}_n)} \right) \right| \leq \bar{B} = \max \left\{ \log(d) + \frac{2B_U}{\tau}, \log \left(\frac{1}{p_{\min}} \right) \right\}.$$

Proof. The main idea of the proof is to bound the probability ratio and use the non-decreasing monotonicity of the log. Let $n \in [N]$. The model f_Θ receives as input sequences of tokens \mathbf{S}_n of size $n \leq K$. We first lower-bound each term of the probability ratio. By definition of p_{\min} , we have

$$\mathbb{P}(\mathbf{X}_{n+1} | \mathbf{S}_n) = \mathbb{P}(\mathbf{X}_{n+1} | \mathbf{X}_n) \geq p_{\min} > 0, \quad (31)$$

where we used the Markov property for the first equality. We want to obtain a similar inequality for $\mathbb{P}_\Theta(\mathbf{X}_{n+1} | \mathbf{S}_n)$. As the parameters Θ of the LLM are in \mathcal{W} , we know that $\|\mathbf{W}_U^\top\|_{2,1} \leq B_U$. Lemma F.20 ensures that

$$\left\| \frac{1}{n\tau} \mathbf{M}\mathbf{W}_U \mathbf{S}^{(L)} \mathbb{1}_T \right\|_1 \leq \frac{1}{\tau} \|\mathbf{W}_U^\top\|_{2,1} \leq \frac{B_U}{\tau}.$$

We can then apply Lemma F.7 with $c_1 = \frac{B_U}{\tau}$ and given that $\frac{1}{T\tau} \mathbf{M} \mathbf{W}_U \mathbf{S}^{(L)} \mathbb{1}_T \in \mathbb{R}^d$, it leads to

$$\mathbb{P}_{\Theta}(\cdot | \mathbf{S}_n) = \text{softmax}\left(\frac{1}{n\tau} \mathbf{M} \mathbf{W}_U \mathbf{S}^{(L)} \mathbb{1}_n\right) \geq \frac{1}{d \exp(2B_U/\tau)},$$

where the inequality holds for each component of $\mathbb{P}_{\Theta}(\cdot | \mathbf{S}_n)$. This is in particular the case $\mathbb{P}_{\Theta}(\mathbf{X}_{n+1} | \mathbf{S}_n)$ which is the entry we are interested in, i.e., we have

$$\mathbb{P}_{\Theta}(\mathbf{X}_{n+1} | \mathbf{S}_n) \geq \frac{1}{d \exp(2B_U/\tau)}. \quad (32)$$

Going back to the ratio of probability, consider the situation where we have

$$\frac{\mathbb{P}(\mathbf{X}_{n+1} | \mathbf{S}_n)}{\mathbb{P}_{\Theta}(\mathbf{X}_{n+1} | \mathbf{S}_n)} \geq 1.$$

Then, using Eq. (32), we have

$$1 \leq \frac{\mathbb{P}(\mathbf{X}_{n+1} | \mathbf{S}_n)}{\mathbb{P}_{\Theta}(\mathbf{X}_{n+1} | \mathbf{S}_n)} \leq \frac{1}{\mathbb{P}_{\Theta}(\mathbf{X}_{n+1} | \mathbf{S}_n)} \leq d \exp(2B_U/\tau),$$

which implies, as the log is non-decreasing monotonically,

$$0 \leq \log\left(\frac{\mathbb{P}(\mathbf{X}_{n+1} | \mathbf{S}_n)}{\mathbb{P}_{\Theta}(\mathbf{X}_{n+1} | \mathbf{S}_n)}\right) \leq \log(d \exp(2B_U/\tau)) = \log(d) + \frac{2B_U}{\tau}. \quad (33)$$

Similarly, consider the case where we have

$$\frac{\mathbb{P}(\mathbf{X}_{n+1} | \mathbf{S}_n)}{\mathbb{P}_{\Theta}(\mathbf{X}_{n+1} | \mathbf{S}_n)} \leq 1.$$

Then, we have

$$\frac{\mathbb{P}_{\Theta}(\mathbf{X}_{n+1} | \mathbf{S}_n)}{\mathbb{P}(\mathbf{X}_{n+1} | \mathbf{S}_n)} \geq 1,$$

and similarly to above, we can use Eq. (31) to obtain

$$1 \leq \frac{\mathbb{P}_{\Theta}(\mathbf{X}_{n+1} | \mathbf{S}_n)}{\mathbb{P}(\mathbf{X}_{n+1} | \mathbf{S}_n)} \leq \frac{1}{\mathbb{P}(\mathbf{X}_{n+1} | \mathbf{S}_n)} \leq \frac{1}{p_{\min}}.$$

This implies

$$0 \leq \log\left(\frac{\mathbb{P}_{\Theta}(\mathbf{X}_{n+1} | \mathbf{S}_n)}{\mathbb{P}(\mathbf{X}_{n+1} | \mathbf{S}_n)}\right) \leq \log\left(\frac{1}{p_{\min}}\right),$$

which also rewrites

$$0 \leq -\log\left(\frac{\mathbb{P}(\mathbf{X}_{n+1} | \mathbf{S}_n)}{\mathbb{P}_{\Theta}(\mathbf{X}_{n+1} | \mathbf{S}_n)}\right) \leq \log\left(\frac{1}{p_{\min}}\right). \quad (34)$$

By definition of the absolute value, combining Eq. (33) and Eq. (34) leads to

$$\left| \log\left(\frac{\mathbb{P}(\mathbf{X}_{n+1} | \mathbf{S}_n)}{\mathbb{P}_{\Theta}(\mathbf{X}_{n+1} | \mathbf{S}_n)}\right) \right| \leq \max\left\{\log(d) + \frac{2B_U}{\tau}, \log\left(\frac{1}{p_{\min}}\right)\right\}.$$

This concludes the proof. \square

We are now ready to upper-bound the total variation.

Corollary F.22 (Upper-bound on the total variation). *Consider an LLM $f_{\Theta} \in \mathcal{F}$ and an input Markov chain $X = (\mathbf{X}_1, \dots, \mathbf{X}_{N_{\text{icl}}})$ with d states. We recall that B_U is the upper bound on the norm of \mathbf{W}_U in the definition of parameter space \mathcal{W} , τ is the softmax temperature, and p_{\min} is the constant related to the minimal transition probability*

between states. We have

$$\forall n \in [N], \quad d_{\text{TV}}(\mathbb{P}(\cdot | \mathbf{S}_n), \mathbb{P}_{\Theta}(\cdot | \mathbf{S}_n)) \leq \sqrt{2 \max\{\log(d) + \frac{2B_U}{\tau}, \log\left(\frac{1}{p_{\min}}\right)\}} := c_4. \quad (35)$$

Proof. Using Proposition F.21, we can directly apply Lemma F.6 with $B = \max\{\log(d) + \frac{2B_U}{\tau}, \log\left(\frac{1}{p_{\min}}\right)\}$ for any $n \in [N]$. It leads to

$$\forall n \in [N], \quad d_{\text{TV}}(\mathbb{P}(\cdot | \mathbf{S}_n), \mathbb{P}_{\Theta}(\cdot | \mathbf{S}_n)) \leq \sqrt{2 \max\{\log(d) + \frac{2B_U}{\tau}, \log\left(\frac{1}{p_{\min}}\right)\}}.$$

This concludes the proof. \square

F.7.4. CONCLUDING THE PROOF

We are now ready to state our main result.

Theorem F.23 (Restatement of Theorem 4.3). *Consider an LLM $f_{\Theta} \in \mathcal{F}$. We provide as input of f_{Θ} a d -state Markov chain $X = (\mathbf{X}_1, \dots, \mathbf{X}_{N_{\text{icl}}})$. The sequence of subsequences of the first n terms is denoted by $S = (\mathbf{S}_1, \dots, \mathbf{S}_{N_{\text{icl}}})$. S is also a Markov chain, and we denote by $t_{\text{mix}}(\varepsilon)$ its mixing time. Let $t_{\min} := \inf_{0 \leq \varepsilon < 1} t_{\text{mix}}(\frac{\varepsilon}{2}) \cdot \left(\frac{2-\varepsilon}{1-\varepsilon}\right)^2$. Let $\delta > 0$. Then, with probability at least $1 - \delta$,*

$$\mathcal{R}_{\text{icl}}(\Theta) \leq \inf_{\vartheta \in \mathcal{W}_{\text{mc}}} \{\widehat{\mathcal{R}}_{\text{icl}}(\vartheta) + K(\vartheta, \Theta)\} + \bar{B} \sqrt{\frac{t_{\min}}{N_{\text{icl}}}} \sqrt{\log\left(\frac{2}{\delta}\right)},$$

where \bar{B} is a constant depending on the parameters of the problem. More precisely,

$$\bar{B} = 2 \sqrt{\max\{\log(d) + \frac{2B_U}{\tau}, \log\left(\frac{1}{p_{\min}}\right)\}}.$$

Proof. Let $\vartheta \in \mathcal{W}_{\text{mc}}$. We first benefit from the metric properties of the total variation to decompose the risk.

$$\begin{aligned} \mathcal{R}_{\text{icl}}(\Theta) &= \frac{1}{N_{\text{icl}}} \sum_{n=1}^{N_{\text{icl}}} \mathbb{E}_{\mathbf{S}_n} [d_{\text{TV}}(\mathbb{P}(\cdot | \mathbf{S}_n), \mathbb{P}_{\Theta}(\cdot | \mathbf{S}_n))] \\ &\leq \frac{1}{N_{\text{icl}}} \sum_{n=1}^{N_{\text{icl}}} \mathbb{E}_{\mathbf{S}_n} [d_{\text{TV}}(\mathbb{P}(\cdot | \mathbf{S}_n), \mathbb{P}_{\vartheta}(\cdot | \mathbf{S}_n)) + d_{\text{TV}}(\mathbb{P}_{\vartheta}(\cdot | \mathbf{S}_n), \mathbb{P}_{\Theta}(\cdot | \mathbf{S}_n))] \\ &\leq \frac{1}{N_{\text{icl}}} \sum_{n=1}^{N_{\text{icl}}} \mathbb{E}_{\mathbf{S}_n} [d_{\text{TV}}(\mathbb{P}(\cdot | \mathbf{S}_n), \mathbb{P}_{\vartheta}(\cdot | \mathbf{S}_n))] \\ &\quad + \frac{1}{N_{\text{icl}}} \sum_{n=1}^{N_{\text{icl}}} \mathbb{E}_{\mathbf{S}_n} [d_{\text{TV}}(\mathbb{P}_{\vartheta}(\cdot | \mathbf{S}_n), \mathbb{P}_{\Theta}(\cdot | \mathbf{S}_n))] \\ &\leq \mathcal{R}_{\text{icl}}(\vartheta) + K(\vartheta, \Theta). \end{aligned} \quad (36)$$

By definition of the risk, we have

$$\widehat{\mathcal{R}}_{\text{icl}}(\vartheta) = \frac{1}{N_{\text{icl}}} \sum_{n=1}^{N_{\text{icl}}} \underbrace{d_{\text{TV}}(\mathbb{P}(\cdot | \mathbf{S}_n), \mathbb{P}_{\vartheta}(\cdot | \mathbf{S}_n))}_{=g_n(\mathbf{S}_n)} = \frac{1}{N_{\text{icl}}} \sum_{n=1}^{N_{\text{train}}} g_n(\mathbf{S}_n) = f(\mathbf{S}_1, \dots, \mathbf{S}_{N_{\text{icl}}}) = f(S).$$

Using Corollary F.22, we know that

$$|g_n(\mathbf{S}_n)| \leq \sqrt{2 \max\{\log(d) + \frac{2B_U}{\tau}, \log\left(\frac{1}{p_{\min}}\right)\}} := c_4.$$

Similarly to Theorem 4.2, and using the fact that $S = (\mathbf{S}_1, \dots, \mathbf{S}_{N_{\text{icl}}})$ is a Markov chain, we can show that choosing $\mathbf{c} \in \mathbb{R}^{N_{\text{icl}}}$ with all entries equal to $\frac{2c_4}{N_{\text{icl}}}$ ensures that f verifies the condition in Proposition F.5, i.e.,

$$\forall S, \Sigma, \quad f(S) - f(\Sigma) \leq \sum_{n=1}^{N_{\text{icl}}} \mathbf{c}_n \mathbb{1}_{\{\mathbf{S}_n \neq \Sigma_n\}}.$$

Putting everything together, we can apply Proposition F.18 which leads to

$$\forall u \geq 0, \quad \mathbb{P}(|f(S) - \mathbb{E}_S[f(S)]| \geq u) \leq 2 \exp\left(\frac{-2u^2}{t_{\min} \|\mathbf{c}\|_2^2}\right). \quad (37)$$

Let $u \geq 0$. We have the following events ordering

$$\begin{aligned} (\mathbb{E}_S[f(S)] - f(S) \geq u) &\subseteq (\mathbb{E}_S[f(S)] - f(S) \geq u) \cup (f(S) - \mathbb{E}_S[f(S)] \geq u) \\ &= (|f(S) - \mathbb{E}_S[f(S)]| \geq u). \end{aligned}$$

Hence, as u was taken arbitrary and using Eq. (37), we have

$$\forall u \geq 0, \quad \mathbb{P}(\mathbb{E}_S[f(S)] - f(S) \geq u) \leq 2 \exp\left(\frac{-2u^2}{t_{\min} \|\mathbf{c}\|_2^2}\right).$$

We recall that by definition

$$f(S) = \widehat{\mathcal{R}}_{\text{icl}}(\boldsymbol{\vartheta}) \text{ and } \mathcal{R}_{\text{icl}}(\boldsymbol{\vartheta}) = \mathbb{E}_S[\widehat{\mathcal{R}}_{\text{icl}}(\boldsymbol{\vartheta})].$$

Moreover, the inequality on the probability holds for any $u \geq 0$, we can choose u such that

$$\begin{aligned} \delta = 2 \exp\left(\frac{-2u^2}{t_{\min} \|\mathbf{c}\|_2^2}\right) &\iff \frac{-2u^2}{t_{\min} \|\mathbf{c}\|_2^2} = \log\left(\frac{\delta}{2}\right) \iff u^2 = \frac{1}{2} t_{\min} \|\mathbf{c}\|_2^2 \log\left(\frac{2}{\delta}\right) \\ &\iff u = \frac{1}{\sqrt{2}} \sqrt{t_{\min} \|\mathbf{c}\|_2^2} \sqrt{\log\left(\frac{2}{\delta}\right)}. \end{aligned}$$

Using the fact that

$$\|\mathbf{c}\|_2 = \sqrt{\sum_{n=1}^{N_{\text{icl}}} \mathbf{c}_n^2} = \sqrt{\sum_{n=1}^{N_{\text{icl}}} \left(\frac{2c_4}{N_{\text{icl}}}\right)^2} = \sqrt{\sum_{n=1}^{N_{\text{icl}}} \frac{4c_4^2}{N_{\text{icl}}^2}} = \sqrt{\frac{4c_4^2}{N_{\text{icl}}}} = \frac{2c_4}{\sqrt{N_{\text{icl}}}}.$$

Using the fact that $c_4 = \sqrt{2 \max\{\log(d) + \frac{2B_U}{\tau}, \log\left(\frac{1}{p_{\min}}\right)\}}$ (Corollary F.22), we obtain

$$\begin{aligned} u &= \frac{1}{\sqrt{2}} \frac{2c_4}{\sqrt{N_{\text{icl}}}} \sqrt{t_{\min}} \sqrt{\log\left(\frac{2}{\delta}\right)} = \frac{\sqrt{2}c_4}{\sqrt{N_{\text{icl}}}} \sqrt{t_{\min}} \sqrt{\log\left(\frac{2}{\delta}\right)} \\ &= \frac{2\sqrt{t_{\min}} \sqrt{\max\{\log(d) + \frac{2B_U}{\tau}, \log\left(\frac{1}{p_{\min}}\right)\}}}{\sqrt{N_{\text{train}}}} \sqrt{\log\left(\frac{2}{\delta}\right)} \\ &= \bar{B} \sqrt{\frac{t_{\min}}{N_{\text{icl}}}} \sqrt{\log\left(\frac{2}{\delta}\right)}, \end{aligned}$$

where we define

$$\bar{B} = 2 \sqrt{\max\{\log(d) + \frac{2B_U}{\tau}, \log\left(\frac{1}{p_{\min}}\right)\}}.$$

Putting everything together, we have

$$\mathbb{P}\left(\mathcal{R}_{\text{icl}}(\boldsymbol{\vartheta}) - \widehat{\mathcal{R}}_{\text{icl}}(\boldsymbol{\vartheta}) \geq \bar{B} \sqrt{\frac{t_{\min}}{N_{\text{icl}}}} \sqrt{\log\left(\frac{2}{\delta}\right)}\right) \leq \delta.$$

Taking the opposite event leads to the following inequality with probability at least $1 - \delta$

$$\mathcal{R}_{\text{icl}}(\boldsymbol{\vartheta}) \leq \widehat{\mathcal{R}}_{\text{icl}}(\boldsymbol{\vartheta}) + \bar{B} \sqrt{\frac{t_{\min}}{N_{\text{icl}}}} \sqrt{\log\left(\frac{2}{\delta}\right)}.$$

Going back to the decomposition of the risk in Eq. (36) and rearranging the terms, we obtain

$$\mathcal{R}_{\text{icl}}(\boldsymbol{\Theta}) \leq \widehat{\mathcal{R}}_{\text{icl}}(\boldsymbol{\vartheta}) + K(\boldsymbol{\Theta}, \boldsymbol{\vartheta}) + \bar{B} \sqrt{\frac{t_{\min}}{N_{\text{icl}}}} \sqrt{\log\left(\frac{2}{\delta}\right)}.$$

As the left-hand side and the bound function of \bar{B} do not depend on $\boldsymbol{\vartheta}$, we can put them both on the left side of the inequality and then take the infimum on $\boldsymbol{\vartheta}$. Rearranging the terms to keep only $\widehat{\mathcal{R}}_{\text{icl}}(\boldsymbol{\Theta})$ on the left side of the inequality leads to

$$\mathcal{R}_{\text{icl}}(\boldsymbol{\Theta}) \leq \inf_{\boldsymbol{\vartheta} \in \mathcal{W}_{\text{mc}}} \{\widehat{\mathcal{R}}_{\text{icl}}(\boldsymbol{\vartheta}) + K(\boldsymbol{\vartheta}, \boldsymbol{\Theta})\} + \bar{B} \sqrt{\frac{t_{\min}}{N_{\text{icl}}}} \sqrt{\log\left(\frac{2}{\delta}\right)},$$

which concludes the proof. □