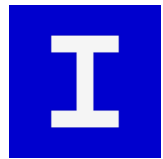
Large Language models as Markov Chains

Oussama Zekri[1]

AI Paper Fest 2024, SSI Club & Arya.ai

[1] ENS Paris-Saclay and Imperial College London, oussama.zekri@ens-paris-saclay.fr

November 14, 2024



Co-authors


Ambroise Odonnat


Abdelhakim Benechehab


Linus Bleistein


Nicolas Boullé


Ievgen Redko

from *Huawei Noah's Ark Lab, Inria, Imperial College London.*

Summary

Introduction

Background on Autoregressive Language Modeling

Goal: Predict the next word based on previous ones.

Background on Autoregressive Language Modeling

Goal: Predict the next word based on previous ones.

Autoregressive Property: Each word depends only on past words.

$$\underbrace{\text{I am the}}_{\text{Previous words (context)}} \quad \underbrace{\text{danger}}_{\text{Word being predicted}}$$

Background on Autoregressive Language Modeling

Goal: Predict the next word based on previous ones.

Autoregressive Property: Each word depends only on past words.

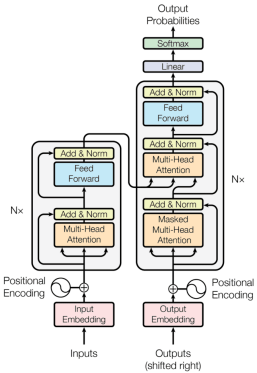
$\underbrace{\text{I am the}}_{\text{Previous words (context)}} \quad \underbrace{\text{danger}}_{\text{Word being predicted}}$

Modelization: Probability of a sequence $(x_1, x_2, \ldots, x_N)$:

$$\begin{aligned}\mathbb{P}(x_1, x_2, \ldots, x_N) &= \mathbb{P}(x_1)\mathbb{P}(x_2 \mid x_1) \cdots \mathbb{P}(x_N \mid x_1, x_2, \ldots, x_{N-1}) \\ &= \prod_{n=1}^{N} \mathbb{P}(x_n \mid x_1, x_2, \ldots, x_{n-1})\end{aligned}$$

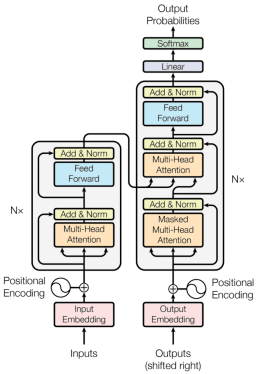
Background on Autoregressive Models

Best models so far : Generative Transformers for Autoregressive Modeling $f_{\Theta}^{T,K}$



Background on Autoregressive Models

Best models so far : Generative Transformers for Autoregressive Modeling $f_{\mathbf{\Theta}}^{T,K}$

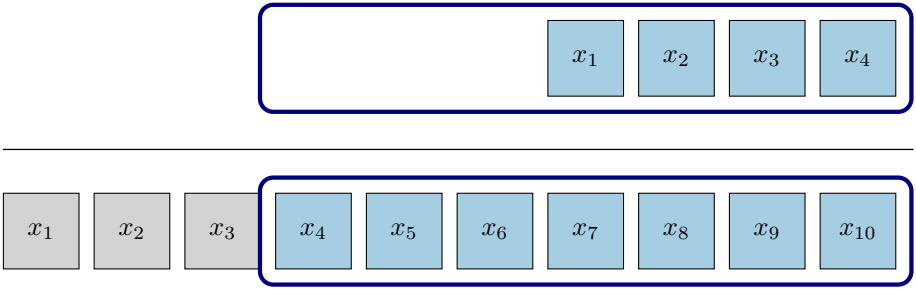


Vocabulary size T.
Context window K.
Parameter set $\mathbf{\Theta}$.

GPT-3 : $T = 50257$, $K = 2048$ and $|\mathbf{\Theta}| \sim 175B$

Context Window K

Context Window $K = 7$ in navy blue.



Top. A sequence of length $N = 4$.
Bottom. A sequence of length $N = 10$.

Background on Markov Chains

Ω discrete finite state-space. $\Omega = \{z_1, ..., z_{|\Omega|}\}$.

Markov Chain: A sequence of states where each state depends only on the previous one.

Background on Markov Chains

Ω discrete finite state-space. $\Omega = \{z_1, ..., z_{|\Omega|}\}$.

Markov Chain: A sequence of states where each state depends only on the previous one.

Mathematical Formulation: A process $(Z_n)_{n\geq 0}$ supported on Ω is a Markov chain if

$$\mathbb{P}(Z_{n+1} \mid Z_n, Z_{n-1}, \ldots, Z_0) = \mathbb{P}(Z_{n+1} \mid Z_n)$$

This is called the *Markov property*.

Are LLMs really Markov chains?

State space: $\Omega = \{z_1, ..., z_{|\Omega|}\}$

Markov Chain: $\mathbb{P}(Z_{n+1} \mid Z_n, Z_{n-1}, \ldots, Z_0) = \mathbb{P}(Z_{n+1} \mid Z_n)$

$\underbrace{\text{I am the}}_{\text{Previous words (context)}}$ $\underbrace{\text{danger}}_{\text{Word being predicted}}$

Are LLMs really Markov chains?

State space: $\Omega = \{z_1, ..., z_{|\Omega|}\}$

Markov Chain: $\mathbb{P}(Z_{n+1} \mid Z_n, Z_{n-1}, \ldots, Z_0) = \mathbb{P}(Z_{n+1} \mid Z_n)$

$\underbrace{\text{I am the}}_{\text{Previous words (context)}}$ $\underbrace{\text{danger}}_{\text{Word being predicted}}$

× LLMs are clearly not Markov chains at the token level ($|\Omega| = T$).

Are LLMs really Markov chains?

State space: $\Omega = \{z_1, ..., z_{|\Omega|}\}$

Markov Chain: $\mathbb{P}(Z_{n+1} \mid Z_n, Z_{n-1}, \ldots, Z_0) = \mathbb{P}(Z_{n+1} \mid Z_n)$

$\underbrace{\text{I am the}}_{\text{Previous words (context)}}$ $\underbrace{\text{danger}}_{\text{Word being predicted}}$

- ✗ LLMs are clearly not Markov chains at the token level ($|\Omega| = T$).
- ✓ Considering the whole sequence, there is something to be done... ($|\Omega| = ?$).

Large Language Models as Markov Chains

Correct State Space

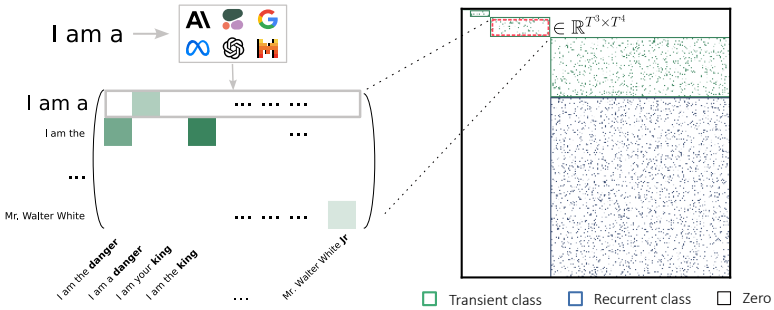
Vocabulary space $\mathcal{V}$ of size T.

- $\Omega = \mathcal{V}_K^*$ is the set of all sequences consisting of elements from $\mathcal{V}$ with up to K elements.

Correct State Space

Vocabulary space $\mathcal{V}$ of size T.

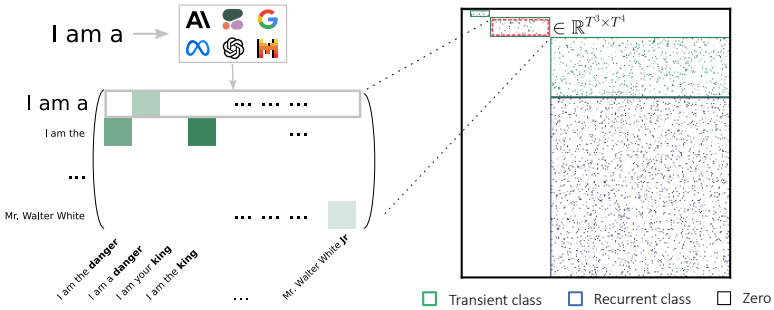
- $\Omega = \mathcal{V}_K^*$ is the set of all sequences consisting of elements from $\mathcal{V}$ with up to K elements.



Correct State Space

Vocabulary space $\mathcal{V}$ of size T.

- $\Omega = \mathcal{V}_K^*$ is the set of all sequences consisting of elements from $\mathcal{V}$ with up to K elements.



- $|\Omega| = T\left(\frac{T^K - 1}{T - 1}\right)$. This is $\sim 50257^{2048} = 1.123456 \times 10^{9628}$ for GPT-3.

Formalization

Any autoregressive model $f_{\mathbf{\Theta}}^{T,K}$ can be equivalently represented by a **finite** Markov chain, with a sparse transition matrix $\mathbf{Q}_f \in \mathbb{R}^{|\mathcal{V}_K^*| \times |\mathcal{V}_K^*|}$ defined as:

$$\forall v_i, v_j \in \mathcal{V}_K^*,$$

$$\mathbf{Q}_f(v_i, v_j) = \begin{cases} 0, & \text{if } \exists l \in \{1, \dots, |v_i| - 1\}, \text{s.t. } (v_i)_{l+1} \neq (v_j)_l, \\ \{f_{\mathbf{\Theta}}^{T,K}(v_i)\}_j, & \text{otherwise.} \end{cases}$$

Formalization

Any autoregressive model $f_{\mathbf{\Theta}}^{T,K}$ can be equivalently represented by a **finite** Markov chain, with a sparse transition matrix $\mathbf{Q}_f \in \mathbb{R}^{|\mathcal{V}_K^*| \times |\mathcal{V}_K^*|}$ defined as:

$$\forall v_i, v_j \in \mathcal{V}_K^*,$$

$$\mathbf{Q}_f(v_i, v_j) = \begin{cases} 0, & \text{if } \exists l \in \{1, \dots, |v_i| - 1\}, \text{s.t. } (v_i)_{l+1} \neq (v_j)_l, \\ \{f_{\mathbf{\Theta}}^{T,K}(v_i)\}_j, & \text{otherwise.} \end{cases}$$

The proportion of non-zero elements in $\mathbf{Q}_f$ is $\dfrac{T-1}{T^K - 1}$.

Is this Markov Chain point of view is useful?

- ✗ $|\Omega| = T\left(\frac{T^K - 1}{T - 1}\right)$, grows exponentially with K. $|\Omega| \sim 10^{9628}$ for GPT-3. $\mathbf{Q}_f$ cannot be stored.
- ∼ Model weights, a few GPUs and a single forward pass are all you need to access the row you want in the matrix!
- ✓ Connection to the rich theory of **finite** Markov Chain.
- ✓ Gives an insight into the dynamics of LLMs.

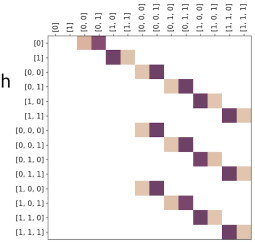
Toy example

Toy example "Baby" LLM.

- $\mathcal{V} = \{0, 1\}$ (i.e. $T = 2$),
- $K = 3$,
- $|\mathbf{\Theta}| = 12688$ and
- $|\Omega| = 14$.

The matrix product yields a probability distribution from which we sample the next token.

Output : 0111010



Stationary distribution

- A stationary distribution π represents the long-term behavior of a Markov chain.

Stationary distribution

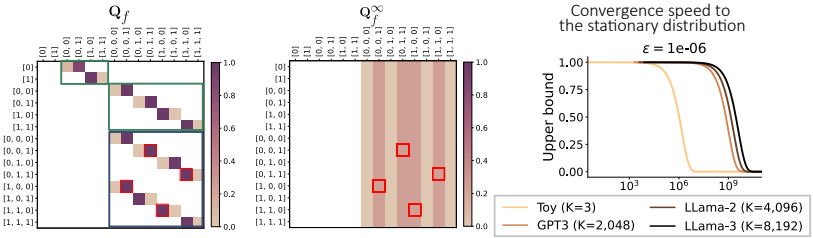
- A stationary distribution π represents the long-term behavior of a Markov chain.
- It verifies $\pi \mathbf{Q}_f = \pi$ and each row of $\mathbf{Q}_f^n$ tends to π when $n \to \infty$.

Stationary distribution

- A stationary distribution π represents the long-term behavior of a Markov chain.
- It verifies $\pi \mathbf{Q}_f = \pi$ and each row of $\mathbf{Q}_f^n$ tends to π when $n \to \infty$.
- A finite state unichain has a unique stationary distribution.

Stationary distribution

- A stationary distribution π represents the long-term behavior of a Markov chain.
- It verifies $\pi \mathbf{Q}_f = \pi$ and each row of $\mathbf{Q}_f^n$ tends to π when $n \to \infty$.
- A finite state unichain has a unique stationary distribution.



Speed of convergence

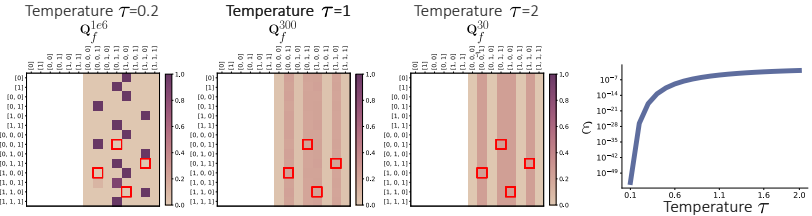
- Convergence speed to the stationary distribution.
- Impact of the temperature.

Proposition

For all $n \geq K$,

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

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



Generalization bounds on pre-training

Setup

Question : How far our matrix $\mathbf{Q}_f$ is from the reference matrix $\mathbf{Q}^*$?

Setup

Question : How far our matrix $\mathbf{Q}_f$ is from the reference matrix $\mathbf{Q}^*$?

- For GPT-3, only 5×10^{11} training tokens, but $T^{K+1} \approx 10^{9632}$ non zero elements in $\mathbf{Q}_f$.

Setup

Question : How far our matrix $\mathbf{Q}_f$ is from the reference matrix $\mathbf{Q}^*$?

- For GPT-3, only 5×10^{11} training tokens, but $T^{K+1} \approx 10^{9632}$ non zero elements in $\mathbf{Q}_f$.
- Generalization capacity of the model?

Setup

Question : How far our matrix $\mathbf{Q}_f$ is from the reference matrix $\mathbf{Q}^*$?

- For GPT-3, only 5×10^{11} training tokens, but $T^{K+1} \approx 10^{9632}$ non zero elements in $\mathbf{Q}_f$.
- Generalization capacity of the model?

Generalization problem

$$\mathcal{R}(\mathbf{\Theta}) := \mathbb{E}[\widehat{\mathcal{R}}(\mathbf{\Theta})], \quad \widehat{\mathcal{R}}(\mathbf{\Theta}) := \frac{1}{N}\sum_{n=1}^{N} d_{\mathrm{TV}}(\mathbf{Q}^*(\mathbf{S}_n, \cdot), \mathbf{Q}_f(\mathbf{S}_n, \cdot)), \tag{1}$$

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

Pre-training generalization bound

Assumptions

✓ Pre-training data $S = (\mathbf{S}_1, \ldots, \mathbf{S}_{N_{\text{train}}})$ a sequence of **dependent** random variables with a Marton coupling matrix $\mathbf{\Gamma}$.

Pre-training generalization bound

Assumptions

- ✓ Pre-training data $S = (\mathbf{S}_1, \ldots, \mathbf{S}_{N_{\text{train}}})$ a sequence of **dependent** random variables with a Marton coupling matrix $\mathbf{\Gamma}$.
- ✓ Assumption only on the last transformer layer: Bounded unembedding matrix, i.e. $\|\mathbf{W}_U^\top\|_{2,1} \leq B_U$.

Pre-training generalization bound

Assumptions

- ✓ Pre-training data $S = (\mathbf{S}_1, \ldots, \mathbf{S}_{N_{\text{train}}})$ a sequence of **dependent** random variables with a Marton coupling matrix $\mathbf{\Gamma}$.
- ✓ Assumption only on the last transformer layer: Bounded unembedding matrix, i.e. $\|\mathbf{W}_U^\top\|_{2,1} \leq B_U$.

Theorem

Let $0 < \delta < 1$, then with probability at least $1 - \delta$,

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

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

Fine-grained bound

More fine-grained bound with additional assumption.

$$\widetilde{\mathcal{W}} = \left\{ \boldsymbol{\Theta} \mid \forall \ell \in [L], \begin{array}{lll} \|\mathbf{W}_V^{(\ell)}\|_\infty \leq B_V, & \|\mathbf{W}_O^{(\ell)}\|_\infty \leq B_O, & \\ \|\mathbf{W}_1^{(\ell)}\|_\infty \leq B_1, & \|\mathbf{W}_2^{(\ell)}\|_\infty \leq B_2, & \|\mathbf{W}_U^\top\|_{2,1} \leq B_U \end{array} \right\}.$$

Fine-grained bound

More fine-grained bound with additional assumption.

$$\widetilde{\mathcal{W}} = \left\{ \boldsymbol{\Theta} \mid \forall \ell \in [L], \begin{array}{ll} \|\mathbf{W}_V^{(\ell)}\|_\infty \leq B_V, & \|\mathbf{W}_O^{(\ell)}\|_\infty \leq B_O, \\ \|\mathbf{W}_1^{(\ell)}\|_\infty \leq B_1, & \|\mathbf{W}_2^{(\ell)}\|_\infty \leq B_2, \quad \|\mathbf{W}_U^\top\|_{2,1} \leq B_U \end{array} \right\}.$$

Corollary

Let $0 < \delta < 1$, then with probability at least $1 - \delta$,

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

where $\bar{B} = 2\|\boldsymbol{\Gamma}\| \max\{\log(T) + 2(B_{\boldsymbol{\Theta}})^L/\tau, \log(1/c_0)\}^{1/2}$, and $B_{\boldsymbol{\Theta}} = [(1 + rmB_1B_2)(1 + \frac{r^3}{H} B_O B_V)](B_{\text{tok}} B_U)^{1/L}$.

Sample complexity

Question : How much training data do I need for $\mathbf{Q}_f$ to be close to $\mathbf{Q}^*$?

Sample complexity

Question : How much training data do I need for $\mathbf{Q}_f$ to be close to $\mathbf{Q}^*$?

✓ Number of sequences that an LLM requires such that $\mathbf{Q}_f$ is ε-close to $\mathbf{Q}^*$.

Sample complexity

Question : How much training data do I need for $\mathbf{Q}_f$ to be close to $\mathbf{Q}^*$?

- ✓ Number of sequences that an LLM requires such that $\mathbf{Q}_f$ is ε-close to $\mathbf{Q}^*$.
- ✓ Dependency on model parameters.

Sample complexity

Question : How much training data do I need for $\mathbf{Q}_f$ to be close to $\mathbf{Q}^*$?

- ✓ Number of sequences that an LLM requires such that $\mathbf{Q}_f$ is ε-close to $\mathbf{Q}^*$.
- ✓ Dependency on model parameters.

Corollary

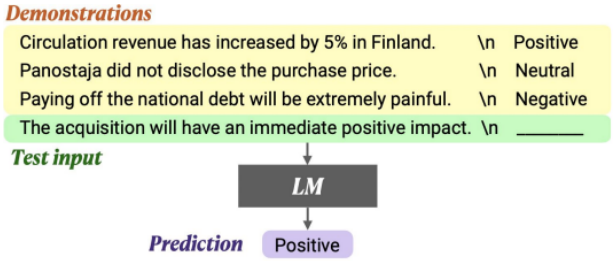
Let $\delta \in [0,1]$ and let $\epsilon > 0$. If $N_{\text{train}} \geq N^* := \lceil \frac{4\bar{B}^2}{\epsilon^2} \log\left(\frac{2}{\delta}\right) \rceil$ and if we assume a perfect pre-training error for $f_{\boldsymbol{\Theta}}$, then we have with probability at least $1-\delta$,

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

In-Context Learning and experiments

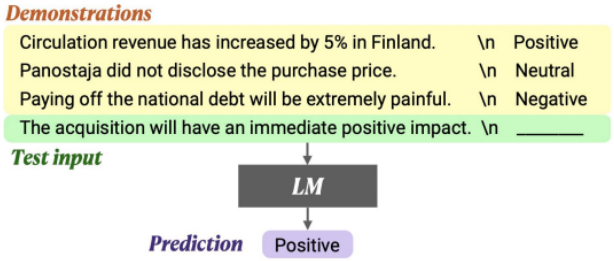
In Context Learning

In-Context Learning : Model's ability to learn and adapt to patterns, without updating its internal parameters.



In Context Learning

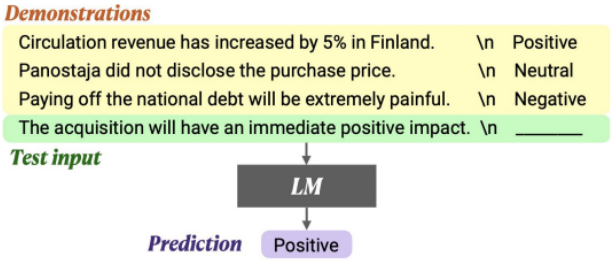
In-Context Learning : Model's ability to learn and adapt to patterns, without updating its internal parameters.



- Not as costly as model pre-training.

In Context Learning

In-Context Learning : Model's ability to learn and adapt to patterns, without updating its internal parameters.



- Not as costly as model pre-training.
- Access to a ground truth matrix $\mathbf{Q}^*$.

In Context Learning of Markov chains

Setup

- A $d-$state Markov chain $X = (\mathbf{X}_1, \ldots, \mathbf{X}_{N_{\text{icl}}})$, and the sequence of its first n terms is denoted by $S = (\mathbf{S}_1, \ldots, \mathbf{S}_n)$.

In Context Learning of Markov chains

Setup

- A $d-$state Markov chain $X = (\mathbf{X}_1, \dots, \mathbf{X}_{N_{\text{icl}}})$, and the sequence of its first n terms is denoted by $S = (\mathbf{S}_1, \dots, \mathbf{S}_n)$.
- Mixing time of S, denoted as $t_{\text{mix}}(\varepsilon)$.

In Context Learning of Markov chains

Setup

- A $d-$state Markov chain $X = (\mathbf{X}_1, \ldots, \mathbf{X}_{N_{\text{icl}}})$, and the sequence of its first n terms is denoted by $S = (\mathbf{S}_1, \ldots, \mathbf{S}_n)$.
- Mixing time of S, denoted as $t_{\text{mix}}(\varepsilon)$.
- Almost distance $K(\mathbf{\Theta}_1, \mathbf{\Theta}_2) \coloneqq \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}_{\mathbf{S}_n}[d_{\text{TV}}(\mathbb{P}_{\mathbf{\Theta}_1}(\cdot \mid \mathbf{S}_n), \mathbb{P}_{\mathbf{\Theta}_2}(\cdot \mid \mathbf{S}_n))]$.

In Context Learning of Markov chains

Setup

- A $d-$state Markov chain $X = (\mathbf{X}_1, \ldots, \mathbf{X}_{N_{\mathrm{icl}}})$, and the sequence of its first n terms is denoted by $S = (\mathbf{S}_1, \ldots, \mathbf{S}_n)$.
- Mixing time of S, denoted as $t_{\mathrm{mix}}(\varepsilon)$.
- Almost distance $K(\mathbf{\Theta}_1, \mathbf{\Theta}_2) := \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}_{\mathbf{S}_n}[d_{\mathrm{TV}}(\mathbb{P}_{\mathbf{\Theta}_1}(\cdot \mid \mathbf{S}_n), \mathbb{P}_{\mathbf{\Theta}_2}(\cdot \mid \mathbf{S}_n))]$.

Theorem

Let $\delta > 0$. Then, with probability at least $1 - \delta$,

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

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

Experiments : In-Context Scaling Laws

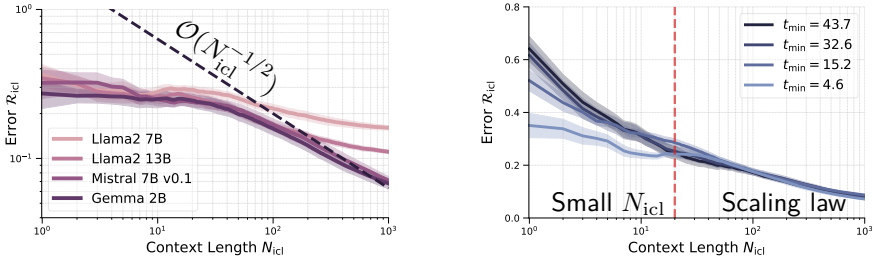


Figure: **In-context scaling laws.** We plot the risk $\mathcal{R}_{\text{icl}}$ as functions of N_{icl}, with 95% confidence intervals.

Experiments : In-Context Scaling Laws

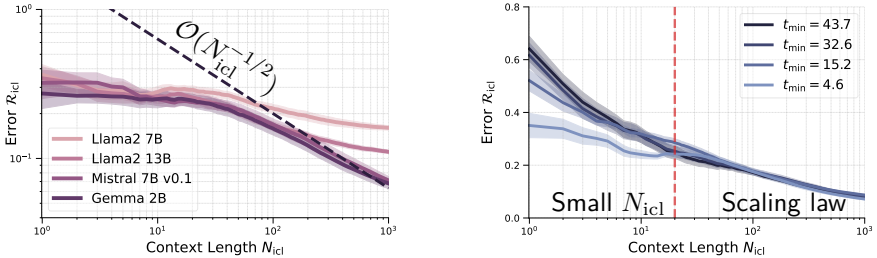


Figure: **In-context scaling laws.** We plot the risk $\mathcal{R}_{\text{icl}}$ as functions of N_{icl}, with 95% confidence intervals.

✓ Randomly generated datas : not seen during training.

Experiments : In-Context Scaling Laws

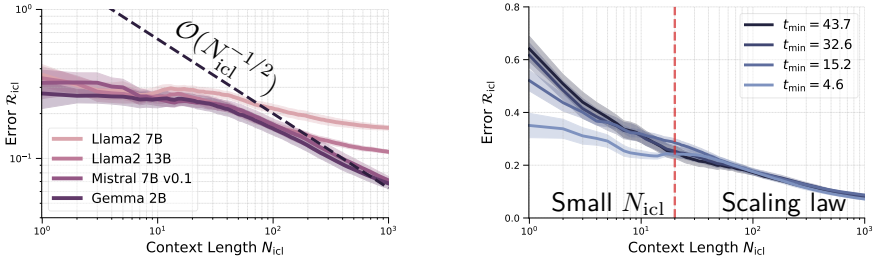
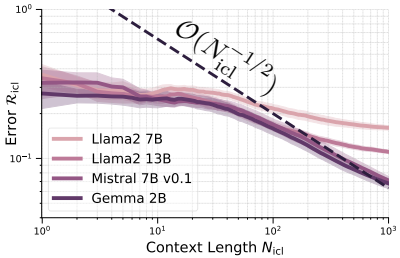


Figure: **In-context scaling laws.** We plot the risk $\mathcal{R}_{\text{icl}}$ as functions of N_{icl}, with 95% confidence intervals.

- ✓ Randomly generated datas : not seen during training.
- ✓ N_{icl} and t_{min} dependence in line with theoretical result.

Experiments : In-Context Scaling Laws



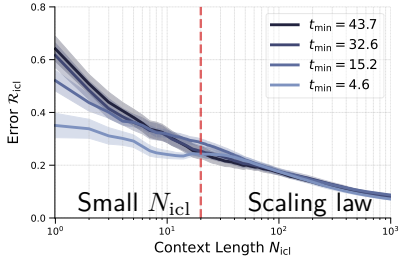


Figure: **In-context scaling laws.** We plot the risk $\mathcal{R}_{\text{icl}}$ as functions of N_{icl}, with 95% confidence intervals.

- ✓ Randomly generated datas : not seen during training.
- ✓ N_{icl} and t_{min} dependence in line with theoretical result.
- ✓ Most recent models stay much closer to the theoretical result.

Experiments : Number of states

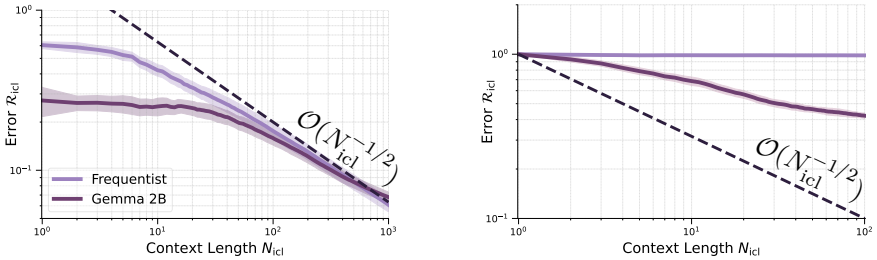


Figure: **Impact of the number of states. Left.** Random 3-state Markov chain. **Right.** Brownian motion discretized as a 700-state Markov chain.

Experiments : Number of states

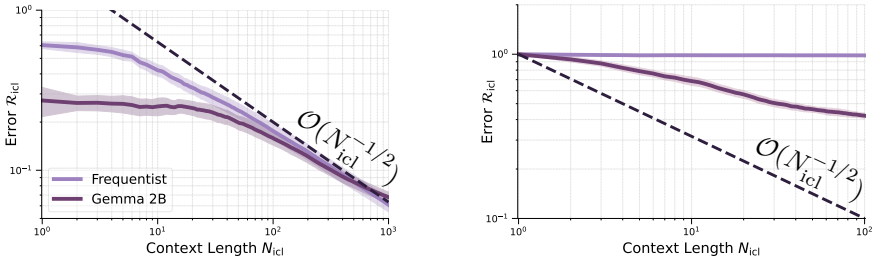


Figure: **Impact of the number of states. Left.** Random 3-state Markov chain. **Right.** Brownian motion discretized as a 700-state Markov chain.

✓ Frequentist (baseline method) $\mathcal{O}(\sqrt{d/N_{\text{icl}}})$ vs. LLM's bound $\mathcal{O}(\sqrt{\log(d)/N_{\text{icl}}})$.

Experiments : Number of states

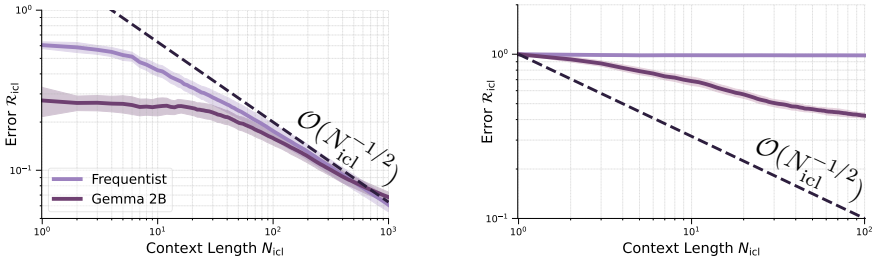


Figure: **Impact of the number of states. Left.** Random 3-state Markov chain. **Right.** Brownian motion discretized as a 700-state Markov chain.

- ✓ Frequentist (baseline method) $\mathcal{O}(\sqrt{d/N_{\text{icl}}})$ vs. LLM's bound $\mathcal{O}(\sqrt{\log(d)/N_{\text{icl}}})$.
- ✓ As d grows, frequentist method struggles.

Take home message

Take home message

- ✓ Explicit characterization of the inference mechanism in LLMs through an equivalent finite-state Markov chain.

Take home message

- ✓ Explicit characterization of the inference mechanism in LLMs through an equivalent finite-state Markov chain.
- ✓ Existence and uniqueness of a stationary distribution. Generalization bounds on pre-training and in-context learning (ICL) phases.

Take home message

- ✓ Explicit characterization of the inference mechanism in LLMs through an equivalent finite-state Markov chain.
- ✓ Existence and uniqueness of a stationary distribution. Generalization bounds on pre-training and in-context learning (ICL) phases.
- ✓ Experiments validate our theory with Llama2 7B & 13B, Gemma 2B, Mistral 7B, and even Llama 3.2.

Thank you for your attention !

You can follow me on social networks!

Thank you for your attention !

My website : www.oussamazekri.fr