An Information-Theoretic Analysis of In-Context Learning

Authors: Hong Jun Jeon, Jason D. Lee, Qi Lei, Benjamin Van Roy

Presented by: Daniel

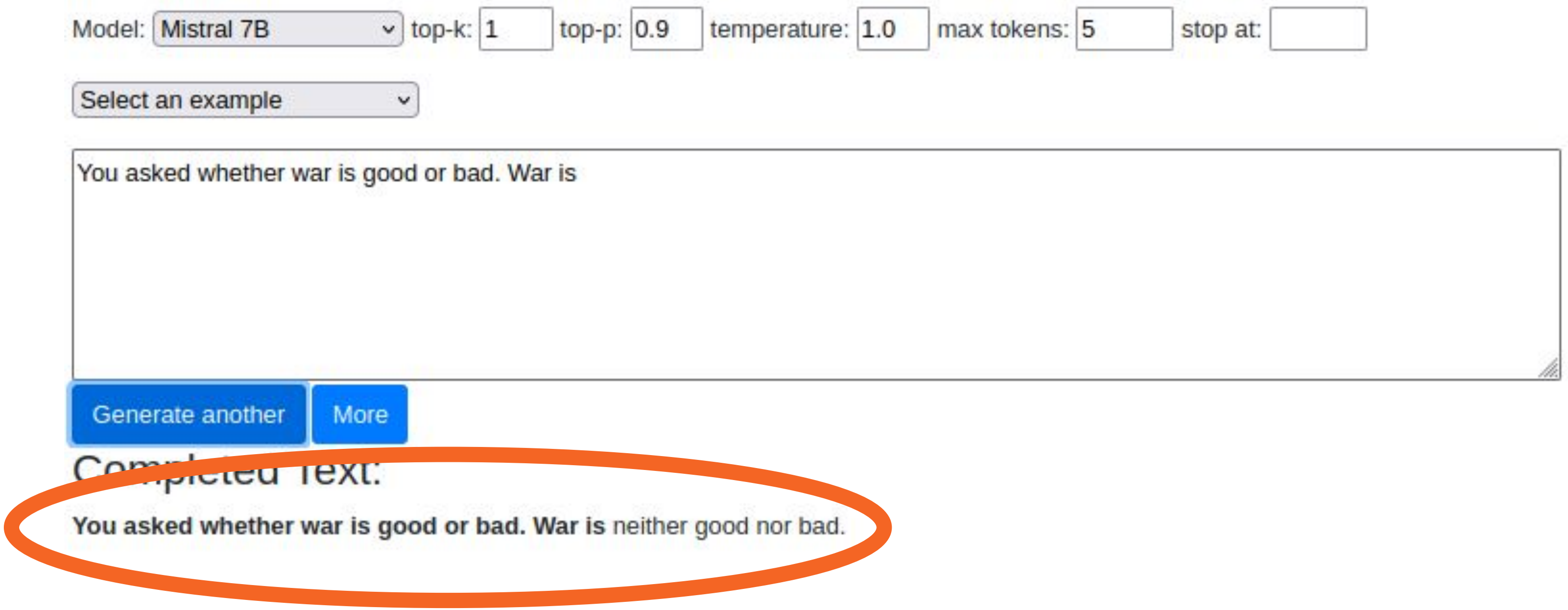
Recap In-Context Learning

Two Example Prompts for LLM Mistral 7B:

You asked whether war is good or bad. War is

freedom: bad,
slavery: good,
justice: bad,
injustice: good,
truth: bad,
falsehood: good,
hate: good,
love: bad,

You asked whether war is good or bad. War is



Source: textsynth.com

Model: Mistral 7B top-k: 1 top-p: 0.9 temperature: 1.0 max tokens: 10 stop at:

Select an example

freedom: bad,
slavery: good,
justice: bad,
injustice: good,
truth: bad,
falsehood: good,
hate: good,
love: bad,

You asked whether war is good or bad. War is

Generate another More

Completed Text:

freedom: bad,
slavery: good,
justice: bad,
injustice: good,
truth: bad,
falsehood: good,
hate: good,
love: bad,

You asked whether war is good or bad. War is good.

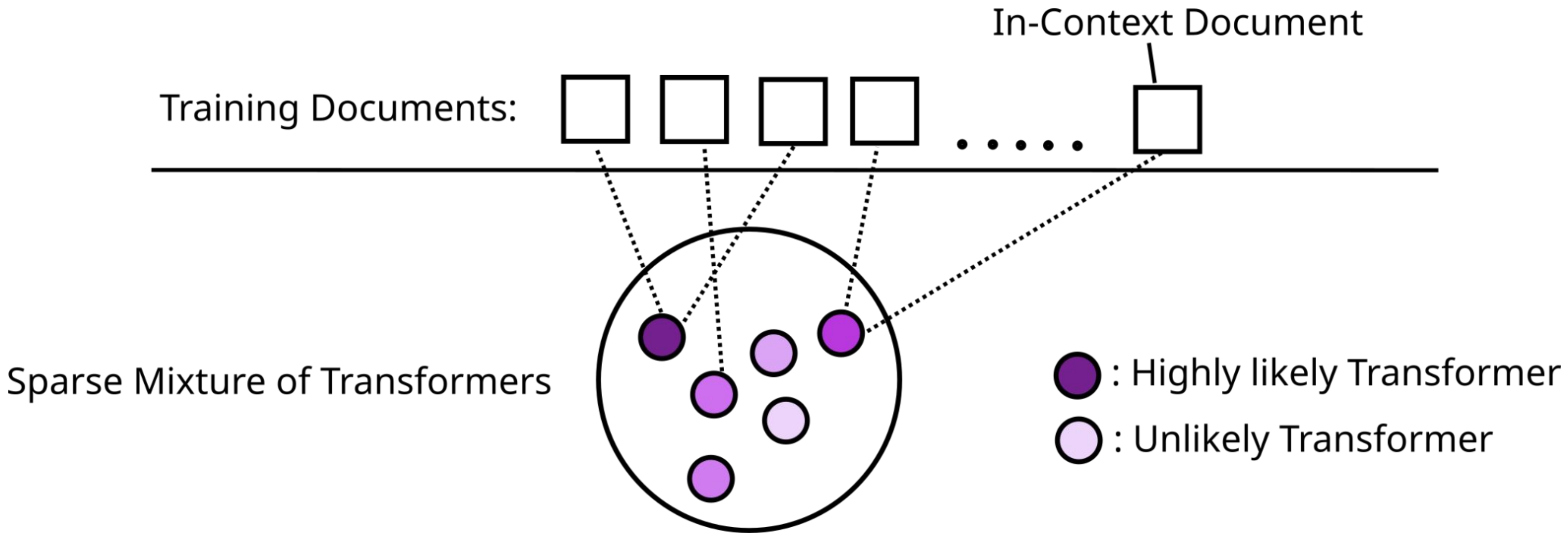
Source: textsynth.com

Why does it work?

Approach from Jeon et al.

They made an assumption about the **probability space** the **training data** for transformers stems from.

Assumption from Jeon et al.



Detailed Approach

Starting explanation by examining simpler case, **without in-context learning**

Question: which π minimizes that loss?

θ : Random vector parameterizing model A_θ

$X_1, X_2, ...X_T$: Random sequence, generated by A_θ

$H_t := X_1, ..., X_t$

$P_{t,\pi} := \pi(H_t),$: estimated probability distribution of X_{t+1} given H_t using algorithm π

$$\mathbb{L}_T(\pi) = \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[-\ln P_{\pi,t}(X_{t+1})]\text{: Loss function of } \pi$$

Reminders:

$$H_t := X_1, ..., X_t$$

$$P_{t,\pi} := \pi(H_t)$$

$$\mathbb{L}_T(\pi) = \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[-\ln P_{\pi,t}(X_{t+1})]$$

Detailed Approach

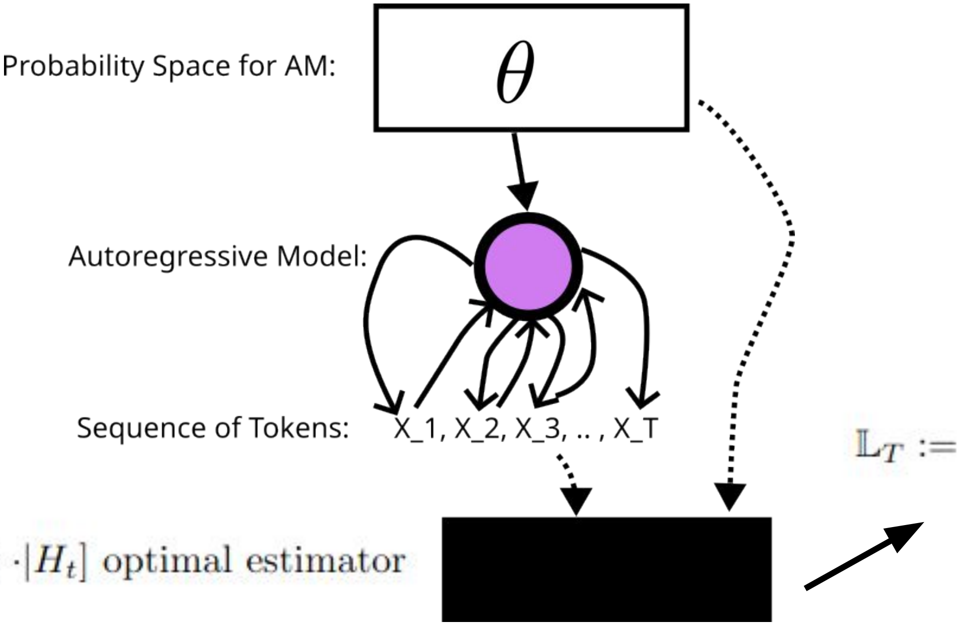
Starting explanation by examining simpler case, **without in-context learning**

$$\hat{P} := \mathbb{P}[X_{t+1} \in \cdot | H_t] \text{ optimal estimator}$$

$$\mathbb{L}_T := \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[-\ln \hat{P}_t(X_{t+1})] \text{ optimal achievable Bayesian error}$$

Detailed Approach

Starting explanation by examining simpler case, **without in-context learning**



$$\hat{P} := \mathbb{P}[X_{t+1} \in \cdot | H_t] \text{ optimal estimator}$$

$$\mathbb{L}_T := \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[-\ln \hat{P}_t(X_{t+1})]$$

Interlude

Recap of Entropy and Information

Entropy of random variable X, $\mathbb{H}(X)$: least expected amount of "bits" required to encode X, given decoder and encoder know distribution of X.

The quantity, which is described by the unit of "minimal required amount of bits", is called **information**.

Mutual Information of two random variables X, Y, $\mathbb{I}(X;Y)$: Expected reduction of entropy of X, if Y is known.

Interlude

Recap of Entropy and Information

QUIZ:
Which of these expressions equals $\mathbb{I}(X;Y)$ (if X, Y are discrete RVs):

1. $\mathbb{H}(X|Y)$
2. $\mathbb{H}(X) - \mathbb{H}(X|Y)$
3. $\mathbb{H}(X) - \mathbb{H}(Y|X)$

Entropy of random variable X, $\mathbb{H}(X)$: least expected amount of “bits” required to encode X, given decoder and encoder know distribution of X.

How "random" is X

The quantity, which is described by the unit of "minimal required amount of bits", is called **information**.

Mutual Information of two random variables X, Y, $\mathbb{I}(X;Y)$: Expected reduction of entropy of X, if Y is known.

How much information about X does Y contain

Reminders:

Optimal loss:

$\mathbb{L}_T := \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[-\ln \hat{P}_t(X_{t+1})]$

Detailed Approach

Starting explanation by examining simpler case, **without in-context learning**

$\mathbb{P}(X_{t+1} \in \cdot|H_t)$: Function for P to minimize the Loss

$$\mathbb{L}_T = \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[-\ln \hat{P}_t(X_{t+1})\right].$$

: Optimal Bayesian Error

Theorem 3.2. (Bayesian error) *For all* $T \in \mathbb{Z}_+$,

$$\mathbb{L}_T = \underbrace{\frac{\mathbb{H}(H_T|\theta)}{T}}_{\substack{\textit{irreducible}\\ \textit{error}}} + \underbrace{\frac{\mathbb{I}(H_T;\theta)}{T}}_{\substack{\textit{estimation}\\ \textit{error}}}.$$

$$\mathcal{L}_T = \frac{\mathbb{I}(H_T;\theta)}{T},$$

Reminders:

Optimal loss:

$$\mathbb{L}_T := \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[-\ln \hat{P}_t(X_{t+1})]$$

$$\mathbb{L}_T = \underbrace{\frac{\mathbb{H}(H_T|\theta)}{T}}_{\substack{irreducible\\ error}} + \underbrace{\frac{\mathbb{I}(H_T;\theta)}{T}}_{\substack{estimation\\ error}}$$

$$\mathcal{L}_T = \frac{\mathbb{I}(H_T;\theta)}{T},$$

Goal: find upper bound of L_T

Scenario, where we have a sequence, generated by a Transformer. Elements of θ are independent and Gaussian distributed.

- θ_i: parameters in Layer i
- K: context length
- L: transformer depth
- r: attention dimension
- d: size of vocabulary

Theorem 3.5. (estimation error bound) *For all d, r, L, K, T, if for all t, X_t is generated by the transformer environment, then*

$$\mathcal{L}_T \leq \frac{pL\ln\left(136eK^2\right) + p\ln\left(\frac{2KT^2}{L}\right)}{T},$$

where $p = 2r^2(L-1) + (dr + r^2)$ denotes the parameter count of the transformer.

What about in context learning?

What about in context learning?

Different probabilistic model of the pre-training data needed-

Different probabilistic model of training data needed

What we had until now

- One sequence of tokens/One document $H_t := (X_1, ..., X_t)$
- One Transformer parameterized by θ

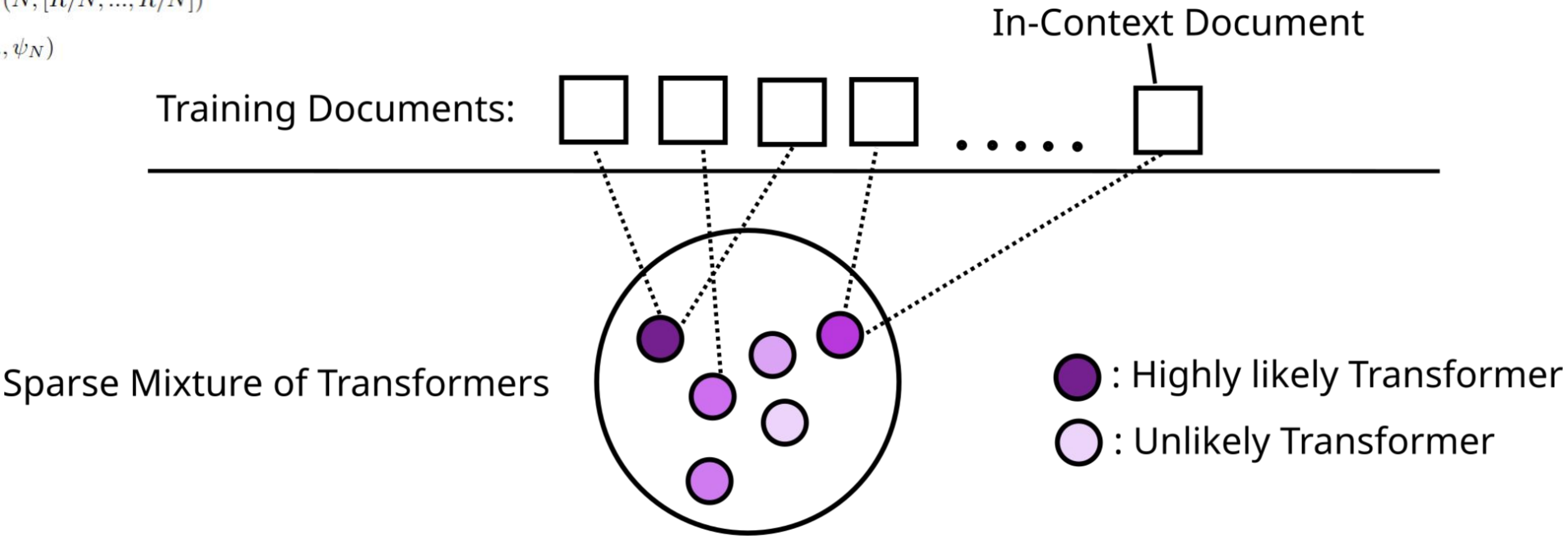
How we model in-context learning

- M training documents $\{D_1, ...D_M\}$, $H_{m,t} := (D_1, ..., D_{m-1}, X_1^{(m)}, ..., X_t^{(m-1)})$
- D_{M+1}: in-context document
- generated by Transformers, parameterized by $\theta_1, ..., \theta_M$
- $\theta_1, ..., \theta_M$ share common variable ψ, e.g. $\theta_1, ..., \theta_M | \psi$ is iid
- $\theta - \psi$ relationship realized by sparse mixture of N random versions of θ: $\psi_1, ...\psi_N$.
- α is the parameter controlling the selection probabilities of $\psi_1, ...\psi_N$. $\alpha \sim$ Dirichlet $(N, [R/N, ..., R/N])$
- $\psi := (\alpha, \psi_1, ..., \psi_N)$

"A further rigorous investigation into the mechanisms by which transformers may be implementing a mixture of models would provide stronger credence to the hypothesis and results provided in this work."

Different probabilistic model of training data needed

- M training documents $\{D_1, ...D_M\}$, $H_{m,t} := (D_1, ..., D_{m-1}, X_1^{(m)}, ..., X_t^{(m-1)})$
- D_{M+1}: in-context document
- generated by Transformers, parameterized by $\theta_1, ..., \theta_M$
- $\theta_1, ..., \theta_M$ share common variable ψ, e.g. $\theta_1, ..., \theta_M | \psi$ is iid
- $\theta - \psi$ relationship realized by sparse mixture of N random versions of θ: $\psi_1, ...\psi_N$.
- α is the parameter controlling the selection probabilities of $\psi_1, ...\psi_N$. $\alpha \sim$ Dirichlet $(N, [R/N, ..., R/N])$
- $\psi := (\alpha, \psi_1, ..., \psi_N)$



NEXT SLIDES MAIN RESULT

Reminders:

$\theta_1, ..., \theta_M | \psi$ is iid

M training documents $\{D_1, ...D_M\}$, $H_{m,t} := (D_1, ..., D_{m-1}, X_1^m, ..., X_t^{m-1})$

Results for Sparse Mixture

$$\mathbb{L}_{M,T} = \frac{1}{MT} \sum_{m=1}^{M} \sum_{t=0}^{T-1} \mathbb{E}\left[-\ln \hat{P}_{m,t}\left(X_{t+1}^{(m)} \right) \right].$$

Theorem 4.2. (Main Result) *For all* $M, T \in \mathbb{Z}_+$ *and* $m \in \{1, 2, \ldots, M\}$,

$$\mathbb{L}_{M,T} = \underbrace{\frac{\mathbb{H}(H_{M,T}|\theta_{1:M})}{MT}}_{\substack{\text{irreducible} \\ \text{error}}} + \underbrace{\frac{\mathbb{I}(H_{M,T};\psi)}{MT}}_{\substack{\text{meta} \\ \text{estimation} \\ \text{error}}} + \underbrace{\frac{\mathbb{I}(D_m;\theta_m|\psi)}{T}}_{\substack{\text{intra-document} \\ \text{estimation} \\ \text{error}}}.$$

$$\mathcal{L}_{M,T} = \frac{\mathbb{I}(H_{M,T};\psi)}{MT} + \frac{\mathbb{I}(D_m;\theta_m|\psi)}{T}$$

Theorem 4.2. **(Main Result)** *For all $M, T \in \mathbb{Z}_+$ and $m \in \{1, 2, \ldots, M\}$,*

$$\mathbb{L}_{M,T} = \underbrace{\frac{\mathbb{H}(H_{M,T}|\theta_{1:M})}{MT}}_{\substack{\text{irreducible}\\ \text{error}}} + \underbrace{\frac{\mathbb{I}(H_{M,T};\psi)}{MT}}_{\substack{\text{meta}\\ \text{estimation}\\ \text{error}}} + \underbrace{\frac{\mathbb{I}(D_m;\theta_m|\psi)}{T}}_{\substack{\text{intra-document}\\ \text{estimation}\\ \text{error}}}.$$

$$\theta_1, \ldots, \theta_M | \psi \text{ is iid}$$

$$\mathbb{L}_{M,T} = \frac{1}{MT} \sum_{m=1}^{M} \sum_{t=0}^{T-1} \mathbb{E}\left[-\ln \hat{P}_{m,t}\left(X_{t+1}^{(m)}\right)\right].$$

Theorem 3.2. **(Bayesian error)** *For all $T \in \mathbb{Z}_+$,*

$$\mathbb{L}_T = \underbrace{\frac{\mathbb{H}(H_T|\theta)}{T}}_{\substack{\text{irreducible}\\ \text{error}}} + \underbrace{\frac{\mathbb{I}(H_T;\theta)}{T}}_{\substack{\text{estimation}\\ \text{error}}}.$$

Results for Sparse Mixture

Theorem 4.5. **(estimation error bound)** *For all $d, r, K, L, M, N, R, T \in \mathbb{Z}_{++}$, if for all $(m,t) \in [M] \times [T]$, $X_t^{(m)}$ is generated according to the sparse mixture of transformers environment, then*

$$\mathcal{L}_{M,T} \leq \frac{pRL \ln\left(1 + \frac{M}{R}\right) \ln(136eK^2)}{MT} + \frac{pR \ln\left(1 + \frac{M}{R}\right) \ln\left(\frac{2KMT^2}{L}\right)}{MT} + \frac{\ln(N)}{T},$$

where $p = 2r^2(L-1) + (dr + r^2)$ denotes the parameter count of each transformer in the mixture.

$$\mathcal{L}_{M,T} = \frac{\mathbb{I}(H_{M,T};\psi)}{MT} + \frac{\mathbb{I}(D_m;\theta_m|\psi)}{T}$$

Reminders:

M training documents $\{D_1, ...D_M\}$, $H_{m,t} := (D_1, ..., D_{m-1}, X_1^m, ..., X_t^{m-1})$

Theorem 4.5. (estimation error bound) *For all* $d, r, K, L, M, N, R, T \in \mathbb{Z}_{++}$, *if for all* $(m, t) \in [M] \times [T]$, $X_t^{(m)}$ *is generated according to the sparse mixture of transformers environment, then*

$$\mathcal{L}_{M,T} \leq \frac{pRL \ln\left(1 + \frac{M}{R}\right) \ln(136eK^2)}{MT} + \frac{pR \ln\left(1 + \frac{M}{R}\right) \ln\left(\frac{2KMT^2}{L}\right)}{MT} + \frac{\ln(N)}{T},$$

where $p = 2r^2(L-1) + (dr + r^2)$ *denotes the parameter count of each transformer in the mixture.*

Results of In-Context Learning Analysis

$$\mathbb{L}_{M,T,\tau} = \frac{1}{\tau} \sum_{t=0}^{\tau-1} \mathbb{E}\left[-\log \hat{P}_t(X_{t+1}^{(M+1)})\right].$$

Theorem 4.7. (in context learning error bound) *For all* $M, T, \tau \in \mathbb{Z}_{++}$, *if* $\tau \leq T$, *then*

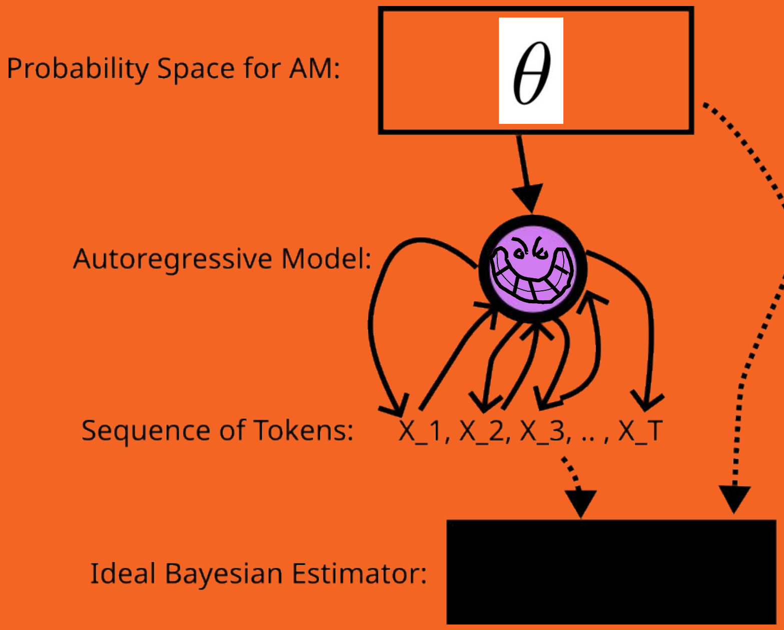
$$\mathbb{L}_{M,T,\tau} \leq \underbrace{\frac{\mathbb{H}\left(D_{M+1} | \theta_{M+1}\right)}{\tau}}_{\substack{\text{irreducible} \\ \text{error}}} + \underbrace{\frac{\mathbb{I}(H_{M,T}; \psi)}{M\tau}}_{\substack{\text{meta} \\ \text{estimation} \\ \text{error}}} + \underbrace{\frac{\mathbb{I}(D_{M+1}; \theta_{M+1} | \psi)}{\tau}}_{\substack{\text{in-context} \\ \text{estimation} \\ \text{error}}}.$$

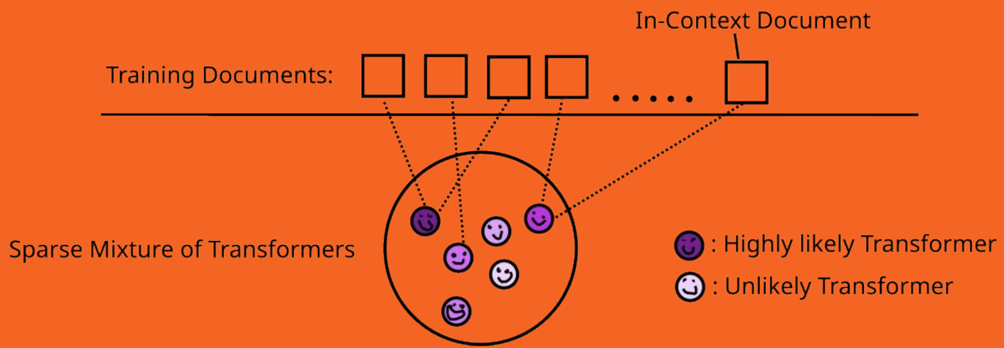
$$\frac{\mathbb{I}(D_{M+1}; \theta_{M+1} | \psi)}{\tau} \leq \frac{\log(N)}{\tau}$$

Take Home Messages

- In Juan et al. **probabilistic assumption**s about the training data and the in-context window were made.
    - Namely: training data and in-context window stem from same distribution(!), which can be expressed by a **sparse mixture (SM) of transformers**.
- Given SM has good hyperparameters, optimal bayesian estimator achieves low error on in-context document
- If transformers imitate bayesian estimator well, paper provides possible explanation/view.
- Good hyperparameters of SM are not guaranteed as far is I can see it, further investigation would increase plausibility.

The End


Probability Space for AM:
θ
Autoregressive Model:
Sequence of Tokens:
X_1, X_2, X_3, .. , X_T
Ideal Bayesian Estimator:
[redacted]


In-Context Document
Training Documents:
Sparse Mixture of Transformers
: Highly likely Transformer
: Unlikely Transformer