ICLR

中国人民大学高瓴人工智能学院
Gaoling School of Artificial Intelligence,Renmin University of China

Towards Auto-Regressive Next-Token Prediction: In-Context Learning Emerges from Generalization

Zixuan Gong[*], Xiaolin Hu[*], Huayi Tang, Yong Liu[†]

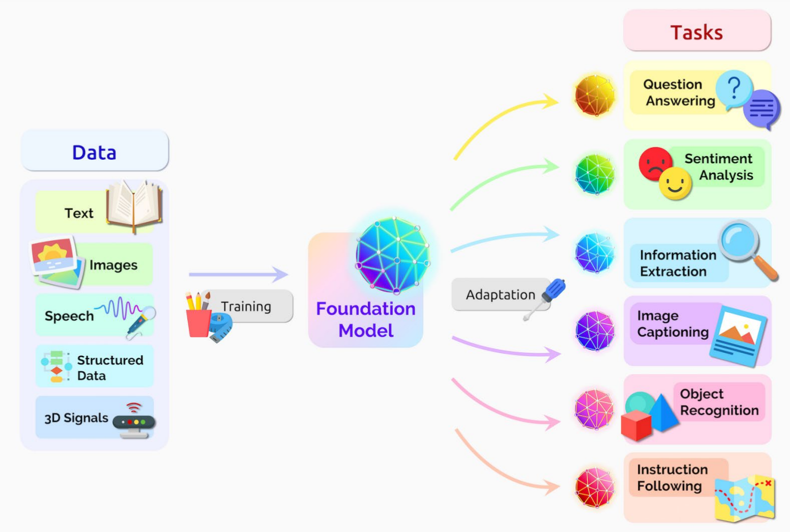
Gaoling School of Artificial Intelligence, Renmin University of China

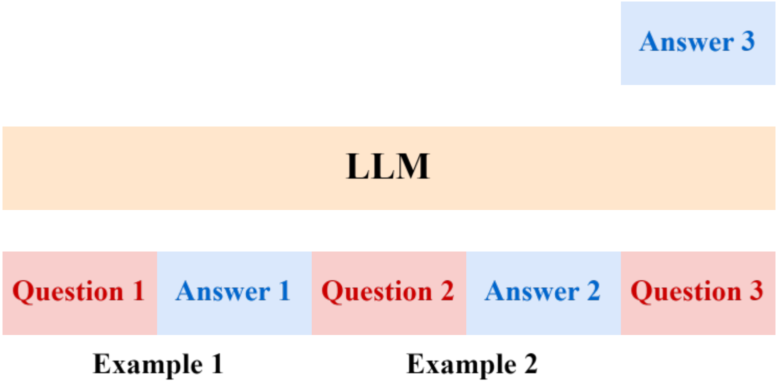
Large Language Models

- **Large language models (LLMs)** have demonstrated remarkable in-context learning (ICL) abilities.
  - ICL means that the model solves new tasks based on prompts without further parameter fine-tuning.
- **Two Limitations in ICL Theory**
  - ☐ **Limited *i.i.d.* Setting**

    Most studies focus on supervised function learning tasks where prompts are constructed with *i.i.d.* input-label pairs.
  - ☐ **Lack of Emergence Explanation**

    Most literature answers **what** ICL does but falls short in explaining **how** pre-trained LLMs can be good enough to emerge ICL ability.
- **The following fundamental questions remain relatively underexplored:**





(a) How can we model language tasks with token-dependency, going beyond the i.i.d. limitation?
(b) How can ICL emerge from pre-trained LLMs?

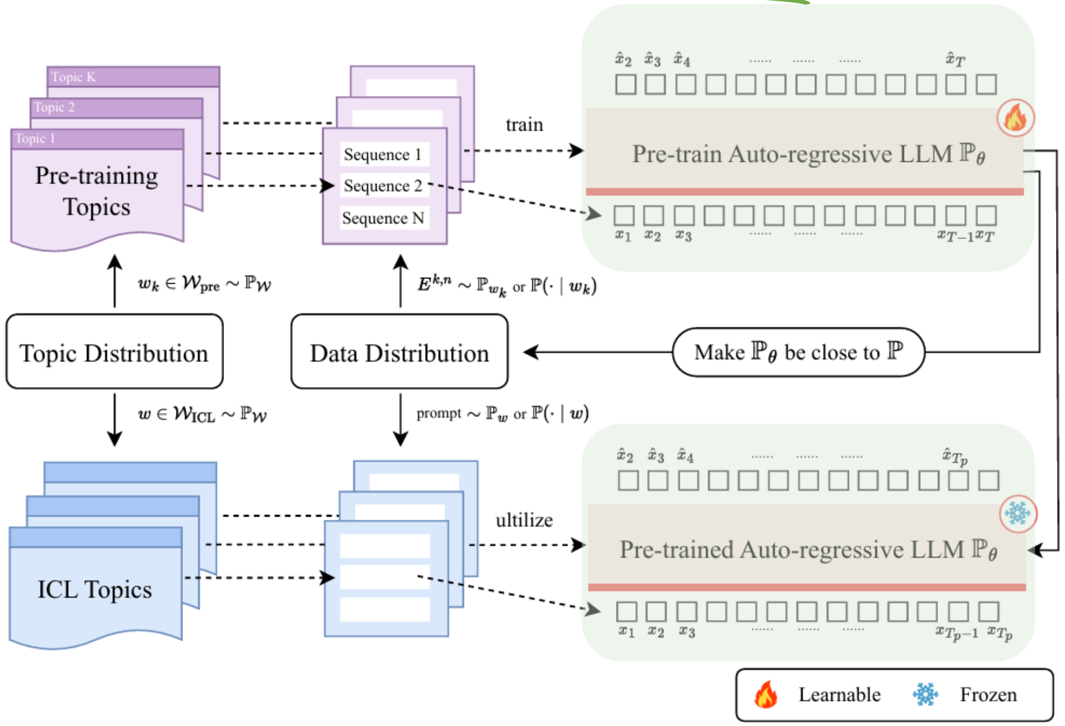
Formalization of Pre-training and ICL Framework

(a) How can we model language tasks with token-dependency, going beyond the i.i.d. limitation?

- **Auto-Regressive Next-Token Prediction (AR-NTP)**

  - **Dependent Tokens**

Each subsequent token in sequences is generated based on the preceding tokens.



Formalization of Pre-training and ICL Framework

(a) How can we model language tasks with token-dependency, going beyond the i.i.d. limitation?

- **Auto-Regressive Next-Token Prediction (AR-NTP)**

  ☐ **Dependent Tokens**

  Each subsequent token in sequences is generated based on the preceding tokens.

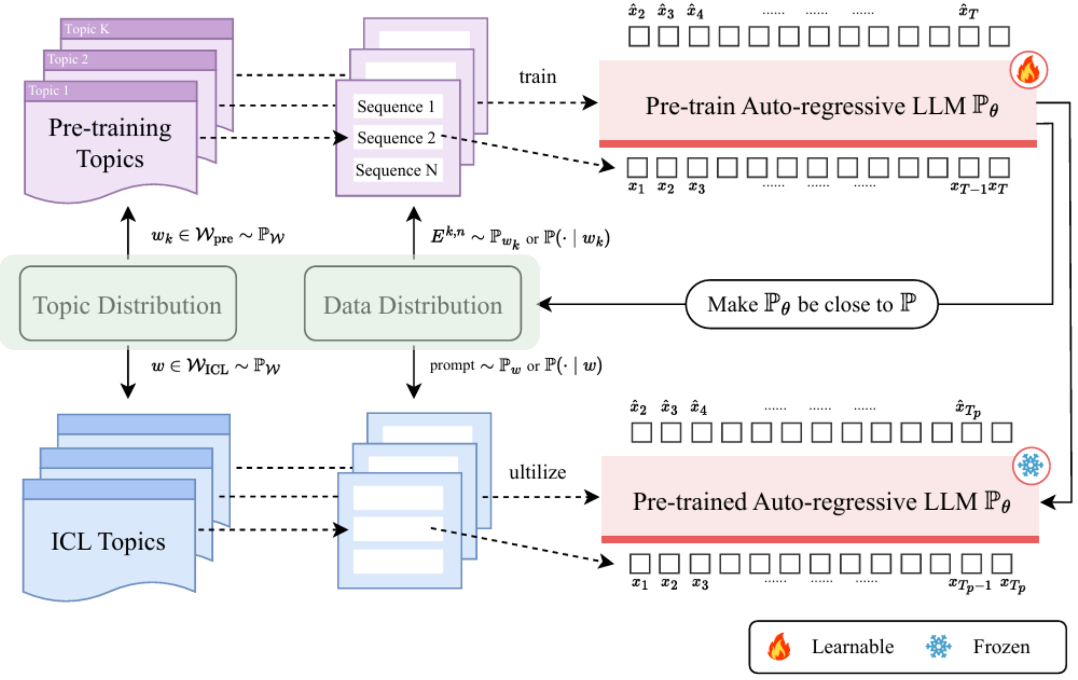
- **Two-Level Distribution**

  ☐ **Topic Distribution**

  The pre-trained topics and downstream topics satisfy the same topic distribution.

  ☐ **Data Distribution**

  All sequences under a given topic satisfy the same data distribution.



ICL Emerges from **Generalization** of Pre-trained LLMs

(b) How can ICL emerge from pre-trained LLMs?

★ ***ICL Emerges!***

Population / Expected Risk Minimization (Two-Level Expectation)

Start From Empirical Risk Minimization

$\underbrace{K \text{ pre-training topics},}_{\text{Outer Expectation}}\ \underbrace{N \text{ pre-training sequences per topic}, T \text{ sequence length}}_{\text{Inner Expectation}}$

Take expectation over topics $\mathbb{E}_{w_k}$ Take expectation over sequences $\mathbb{E}_{E^{k,n}}$

* Division on inner expectation

(prompt token-dependency)

- ✓ Expectation over each token when given prefix sequences $\mathbb{E}_{x_{t+1}^{k,n} \sim \mathbb{P}(\cdot|E_t^{k,n}, w_k)}$
- ✓ Expectation over prefix sequences $\mathbb{E}_{E_t^{k,n}}$

Empirical Loss

$$L_E(\theta, W_{pre}) = \frac{1}{KNT} \sum_{k=1}^{K} \sum_{n=1}^{N} \sum_{t=1}^{T} \log \frac{\mathbb{P}(x_{t+1}^{k,n} \mid E_t^{k,n}, w_k)}{\mathbb{P}_\theta(x_{t+1}^{k,n} \mid E_t^{k,n}, w_k)}$$

Population Loss

$$L(\theta) = \frac{1}{T_p} \sum_{t=1}^{T_p} \mathbb{E}_w \mathbb{E}_{\text{prompt}_t}\left[D_{KL}\left(\mathbb{P}(\cdot \mid \text{prompt}_t, w) \,||\, \mathbb{P}_\theta(\cdot \mid \text{prompt}_t, w)\right)\right]$$

ICL Emerges from **Generalization** of Pre-trained LLMs

(b) How can ICL emerge from pre-trained LLMs?

★ ICL Emerges!

Theorem (Informal): Under some mild assumptions, for any $0 < \delta < 1$, with probability at least $1 - \delta$, the population loss (Two-level expected loss) $L(\theta)$ obeys,

$$\mathbb{E}_{\mu}[L(\theta)] = \mathcal{O}\left\{\sqrt{\frac{1}{(K-K')T_p}\left[D_{KL}(\mu\|\nu_J) + \log\frac{1}{\delta}\right]} + U(\mathcal{W}_{\text{pre}}, K, N, N', T)\right\}$$

μ: posterior distribution

topic-dependent prior distribution

from Generalization of sequences

$$\mathcal{O}\left\{\sqrt{\frac{1}{K(N-N')T}\left[D_{KL}(\mu\|\nu_J) + \log\frac{1}{\delta}\right] - \epsilon_{\text{opt}}} + \sqrt{\frac{\log 1/\delta}{K(N-N')T}}\right\}$$

data-dependent prior distribution

- ✓ Refine the representation of KL divergence, and provide optimization-dependent generalization bounds.
- ✓ Through continuous analysis techniques on SGD.

ICL Emerges from **Generalization** of Pre-trained LLMs

(b) How can ICL emerge from pre-trained LLMs?

★ ***ICL Emerges!***

Theorem (Informal): Under some mild assumptions, for any $0 < \delta < 1$, with probability at least $1 - \delta$, the population loss (Two-level expected loss) $L(\theta)$ obeys,

$$\mathbb{E}_{\mu}[L(\theta)] = \mathcal{O}\left\{\sqrt{\frac{1}{(K-K')T_p}}\left[D_{KL}(\mu||\nu_J) + \log\frac{1}{\delta}\right] + U(\mathcal{W}_{\text{pre}}, K, N, N', T)\right\}$$

Theoretical Insights

- The impact of pre-training topics, sequences and sequence length.
- The impact of parameter size.
- The data-dependent and topic-dependent prior uniquely enhances optimization (origin from the KL part).
- May provide practical guidance on model training, data selection and deduplication (origin from the KL part).

Experiments on Real-World Language Dataset

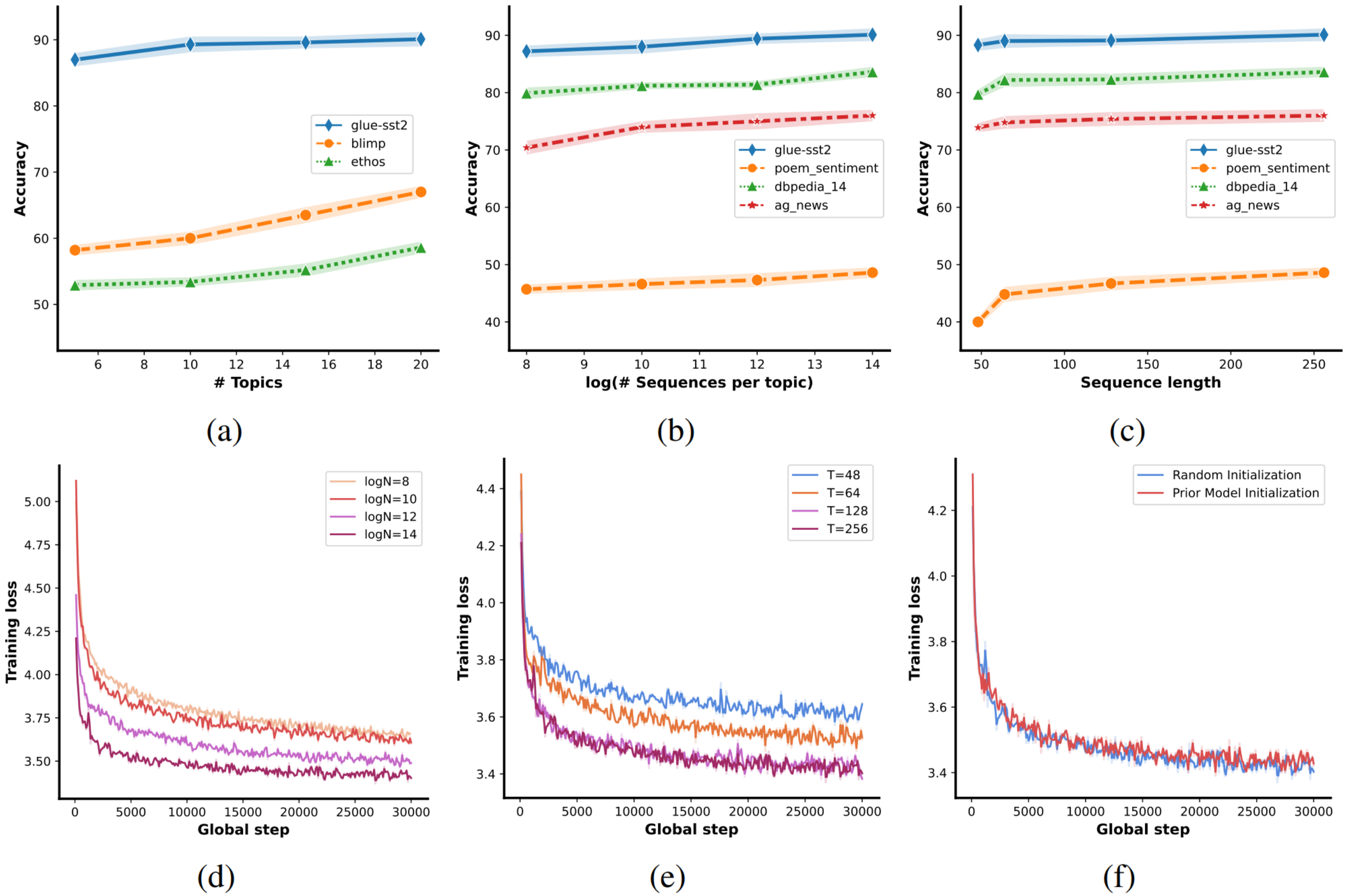
- Experiments on Linear Dynamic System.
- Experiments on Synthetic Language Dataset.
- **Experiments on Real-World Language Dataset.**
  - **Observation (1): Separate Effects of K, N, T.**
  - **Observation (2): Optimization Process.**

  Faster training leads to better generalization.
  - **Observation (3): Prior Model Initialization.**

  1. Random Initialization Regime. ***Nearly 7 hours.***

  2. Prior model initialization Regime. ***Nearly 4.5 hours.***

  Prior model initialization not only accelerates training but also stabilizes the training process (especially in the early stages), leading to comparable model performance.



Towards Auto-Regressive Next-Token Prediction: In-Context Learning Emerges from Generalization

Zixuan Gong*, Xiaolin Hu*, Huayi Tang, Yong Liu†

Gaoling School of Artificial Intelligence Renmin University of China

Thanks!