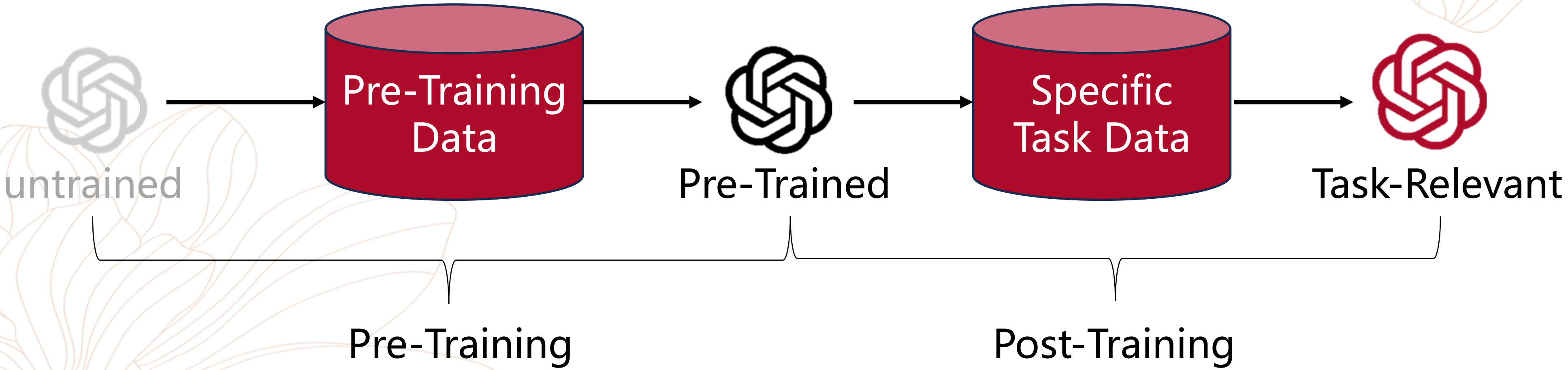
中国人民大學
RENMIN UNIVERSITY OF CHINA

Towards A Theoretical Understanding of Synthetic Data in LLM Post-Training: A Reverse-Bottleneck Perspective

Zeyu Gan

- ## Post-Training of LLMs

The training of LLMs can be divided as Pre-Training and Post-Training

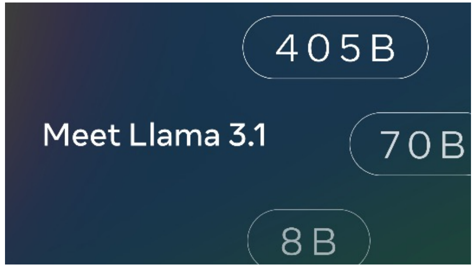


Synthetic Data

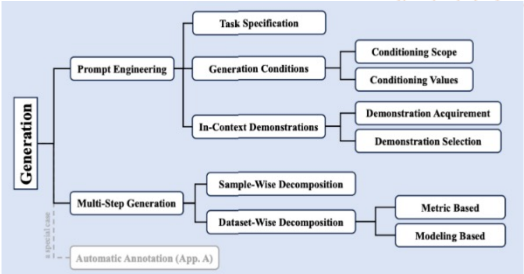
Training data is limited in real-world post-training

Synthetic data are an important supplement

Synthetic data in Hugging Face



Widely utilization of synthetic data[1]



Attention from academic community[2]

[1] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, and et al. The llama 3 herd of models, 2024.

[2] Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. On llms-driven synthetic data generation, curation, and evaluation: A survey, 2024.

Background

- **Synthetic Data Lacks Theoretical Understanding**

  Though it is widely utilized, there is a gap in theoretical analysis

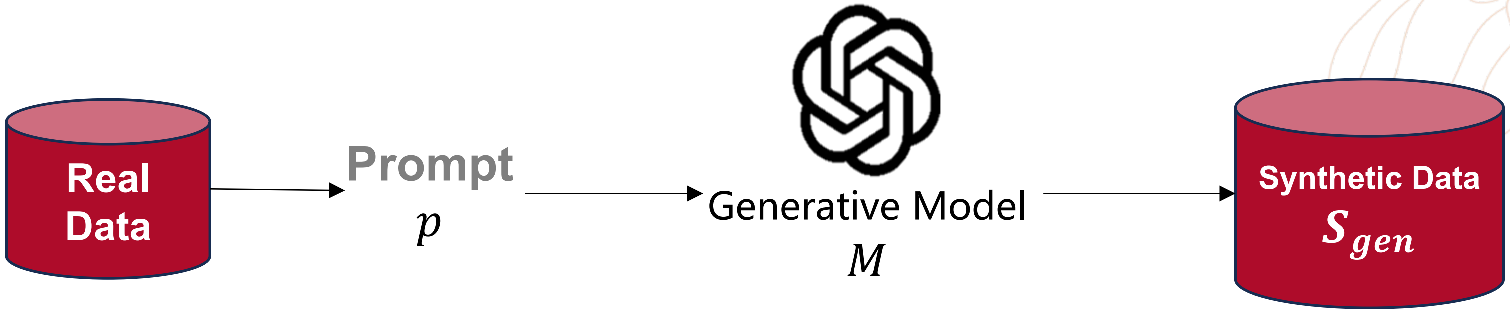
  It is important to provide a formulation

- **In this paper:**
  - 1) We formalize the synthetic data generation
  - 2) We explained the effectiveness of synthetic data in post-training

Formulation of Synthetic Data Generation

• Synthetic Data Generation

A common procedure of synthetic data generation



Synthetic data is generated by a generative model M

The input prompt of M are determined by the real data

e.g.

In code generation, we first obtain human-written code, and obtain similar code data by in-context learning with an LLM.

Formulation of Synthetic Data Generation

- **Synthetic Data Generation**

$$S_{gen} \leftarrow M_p(\mathcal{T}, S_{anchor})$$

➢prompt p can be expressed as the transformation of the anchor data by task $\mathcal{T}$:

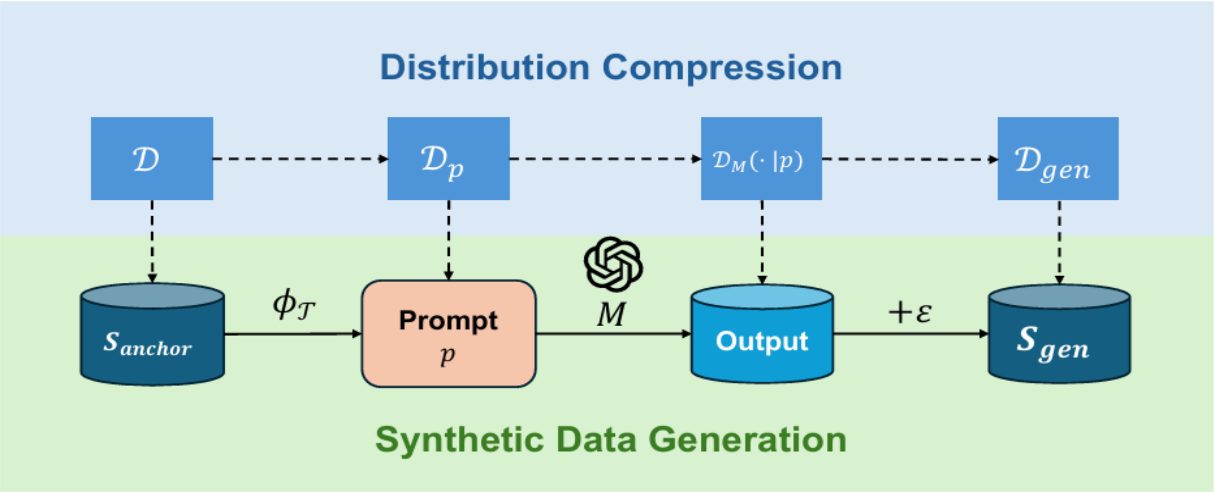
$$p = \phi_{\mathcal{T}}(S_{anchor})$$

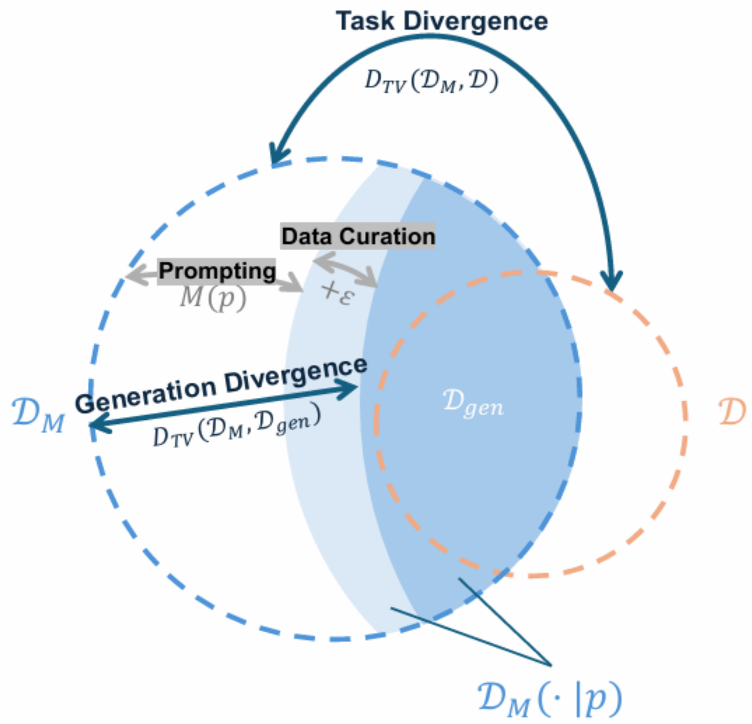
➢Synthetic data is the output of M on p

$$S_{gen} = M(p) + \epsilon$$

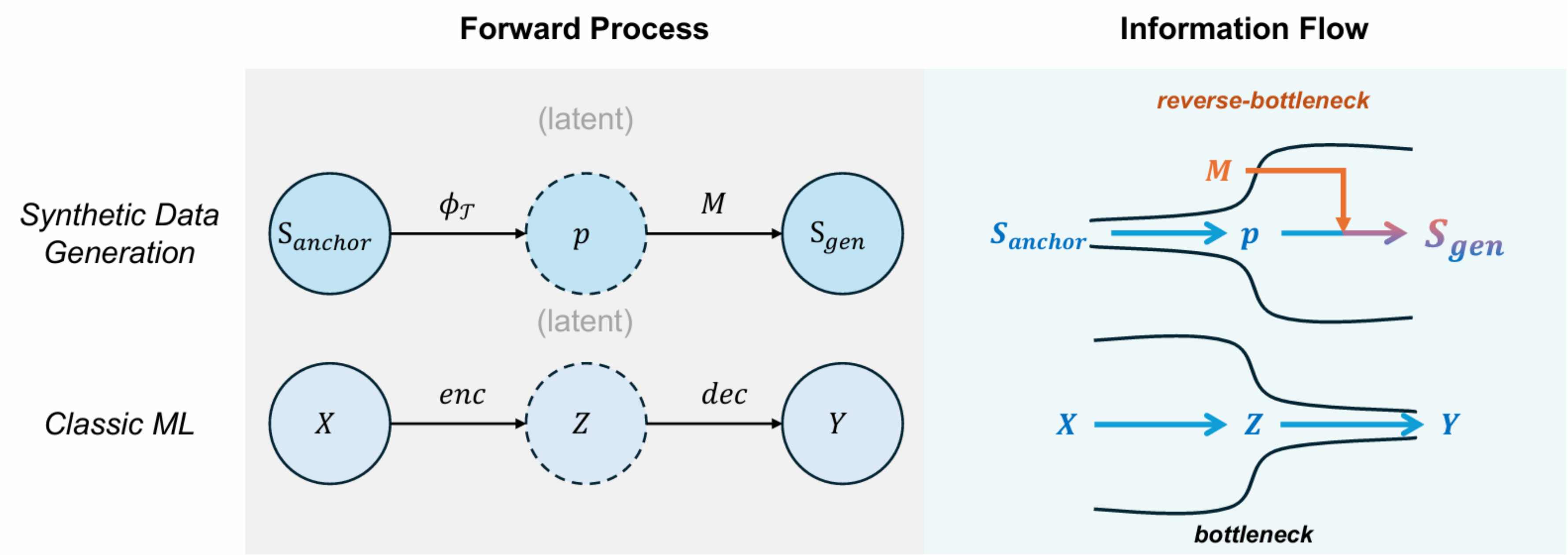
- **Synthetic Data Generation – Distribution Shift**

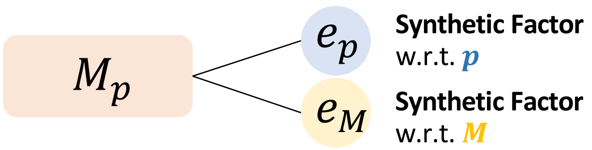
  The generation process can be regarded as a distribution shift





- **Reverse-Bottleneck**





Information Gain

$$\Delta I = \mathbf{H}\big(M(p)\big) - \mathbf{I}\Big(h(e_p), M(p)\Big)$$

Notations

$\mathbf{H}(\cdot)$ *denotes the information entropy.*
$\mathbf{I}(\cdot,\cdot)$ *denotes the mutual information.*

- ## Generalization Error

Theorem 4.7. *(Synthetic data post-training upper bound.) For the same condition as lemma* $\boxed{4.6}$ *and a synthetic data generation process described above, the generalization error of the model* π *post-trained on the synthetic data can be bounded as:*

$$\mathbb{E}(\mathrm{Err}(\pi^{S_{gen}})) \leq C \underbrace{\left(D_{\mathrm{TV}}(\mathcal{D}, \mathcal{D}_M) + D_{\mathrm{TV}}(\mathcal{D}_M, \mathcal{D}_{gen})\right)}_{\text{Distributions' Divergence}} + \underbrace{\exp\left(-\frac{L}{2}\log\frac{1}{\eta}\right)\sqrt{\frac{2\sigma^2 \boxed{[-\Delta I} + B_{syn} + H(e_M) + \delta_{\epsilon,p}]}{n}}}_{\text{Generalization Error w.r.t. synthetic data}}. \tag{7}$$

➢ The upper bound is controlled by $-\Delta I$.When more information gain is introduced, $\pi^{S_{gen}}$ will obtain better generalization capability.

- **The Generalization Gain of Synthetic Data**

Definition 4.9. *(Generalization Gain via Mutual Information, GGMI.) GGMI is defined as the difference between the mutual information terms in the two generalization upper bounds:*

$$\mathrm{GGMI} = I(S_{anchor}, W^{'}) - I(S_{gen}, W). \tag{9}$$

Generalization Gain

Theorem 4.10. *(Upper bound of GGMI.) Given the synthetic data generation above, W' is parameterized by training with S_{anchor}, and W is parameterized by training with S_{gen}, the GGMI can be bounded as follows:*

$$\mathrm{GGMI} \leq \boxed{\Delta I} - (\alpha+1)H(S_{anchor}|W) + \boxed{2\Delta H} + H(S_{gen}|W) + \epsilon_{W,p}, \tag{10}$$

where $\Delta H = H\left(S_{anchor}\right) - H\left(S_{gen}\right)$, $\epsilon_{W,p} = H(S_{anchor}|W) - H(S_{anchor}|M(p))$, *it is assumed that* $H(S_{anchor}|W') = \alpha H(S_{anchor}|W)$, $\alpha \geq 0$.

Diversity

Faithfulness

The benefits of synthetic data are presented in two aspects: **Diversity** and **Faithfulness**, corresponding to ΔI and ΔH

中国人民大學
RENMIN UNIVERSITY OF CHINA

高瓴人工智能学院
Gaoling School of Artificial Intelligence

THANKS

THANK YOU