

Foundation of Mixture of Experts in Statistical Machine Learning

Nhat Ho

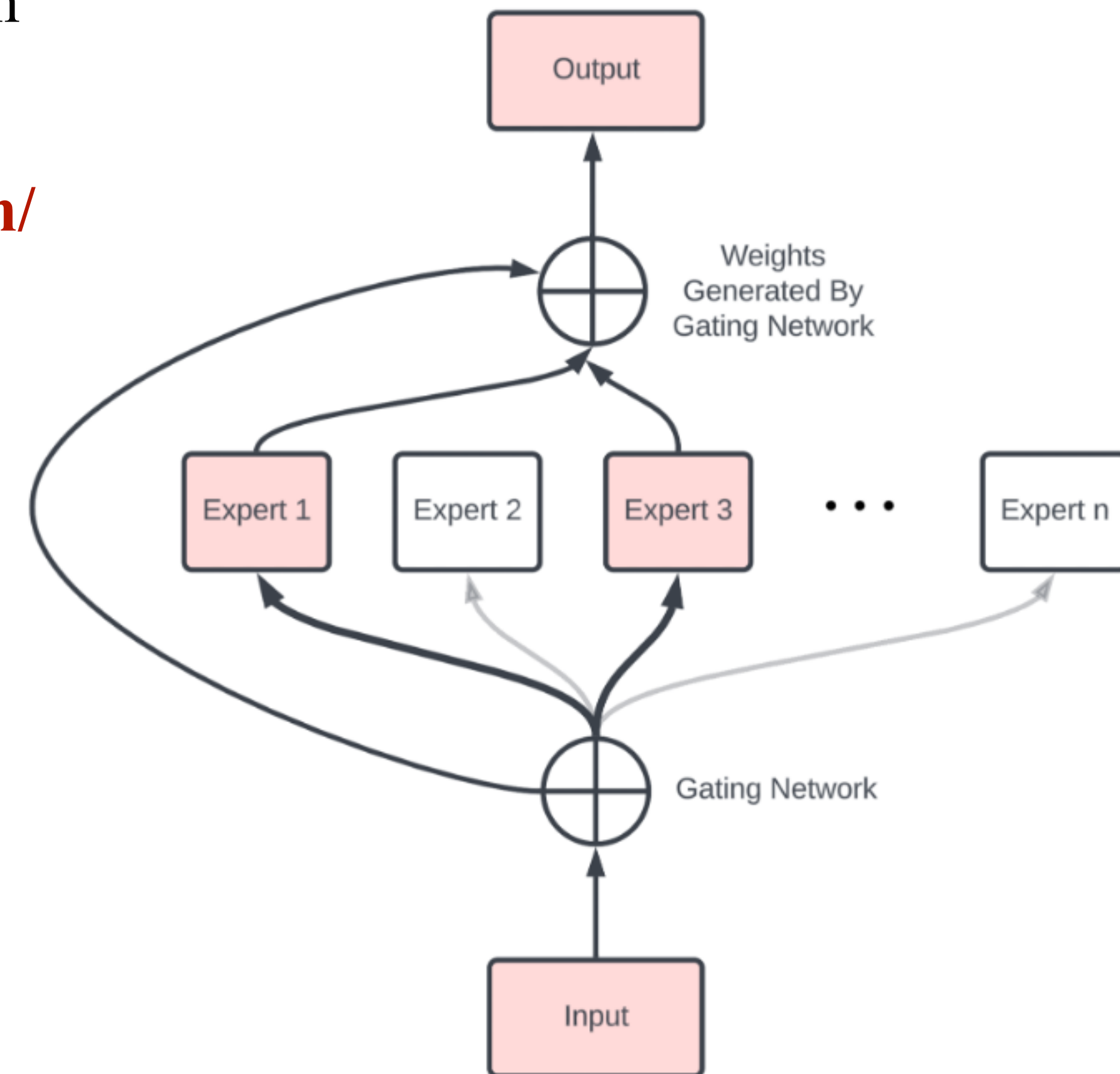
The University of Texas, Austin

Vietnam Institute for Advanced Study in Mathematics (VIASM)

December, 2025

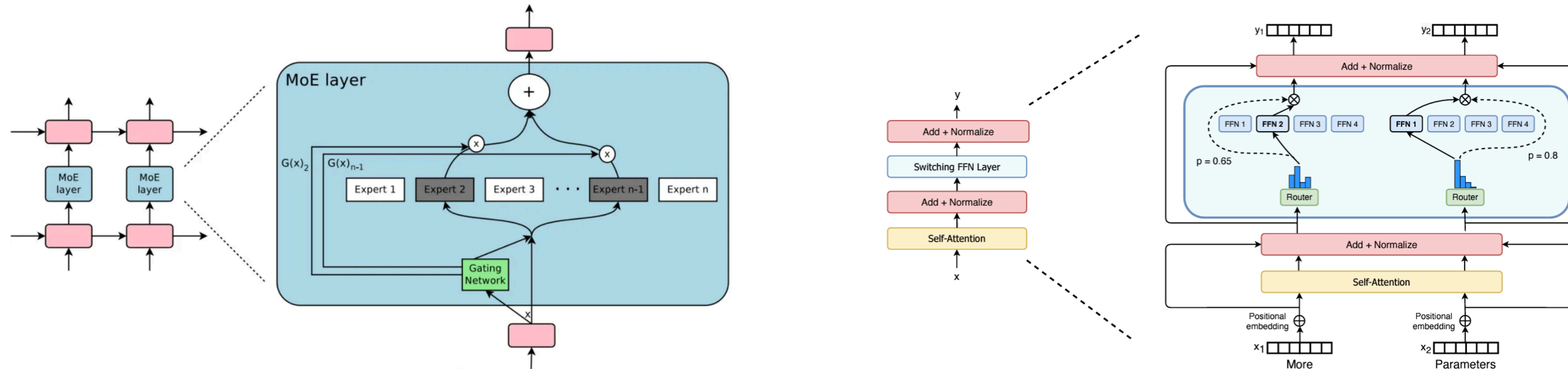
Why Mixture of Experts?

- Mixture of experts was introduced by (Jacobs et al. [1] and Jordan et al. [2])
- Mixture of experts combines multiple **experts** via **gating function/network** to form more complex and accurate models
- In recent years, mixture of experts has been combined into deep learning architectures and complex AI models to improve
 - Scalability (Shazeer et al. [3], Fedus et al. [4], etc.)
 - Multi-modality (Han et al. [6], Han et al. [7], etc.)
 - Safety (Xie et al. [8], etc.)
 - Generalization (aka Performance)



Applications of Mixture of Experts: Brief Introduction

Massive AI Models (e.g., DeepSeek, Mistral, etc.)



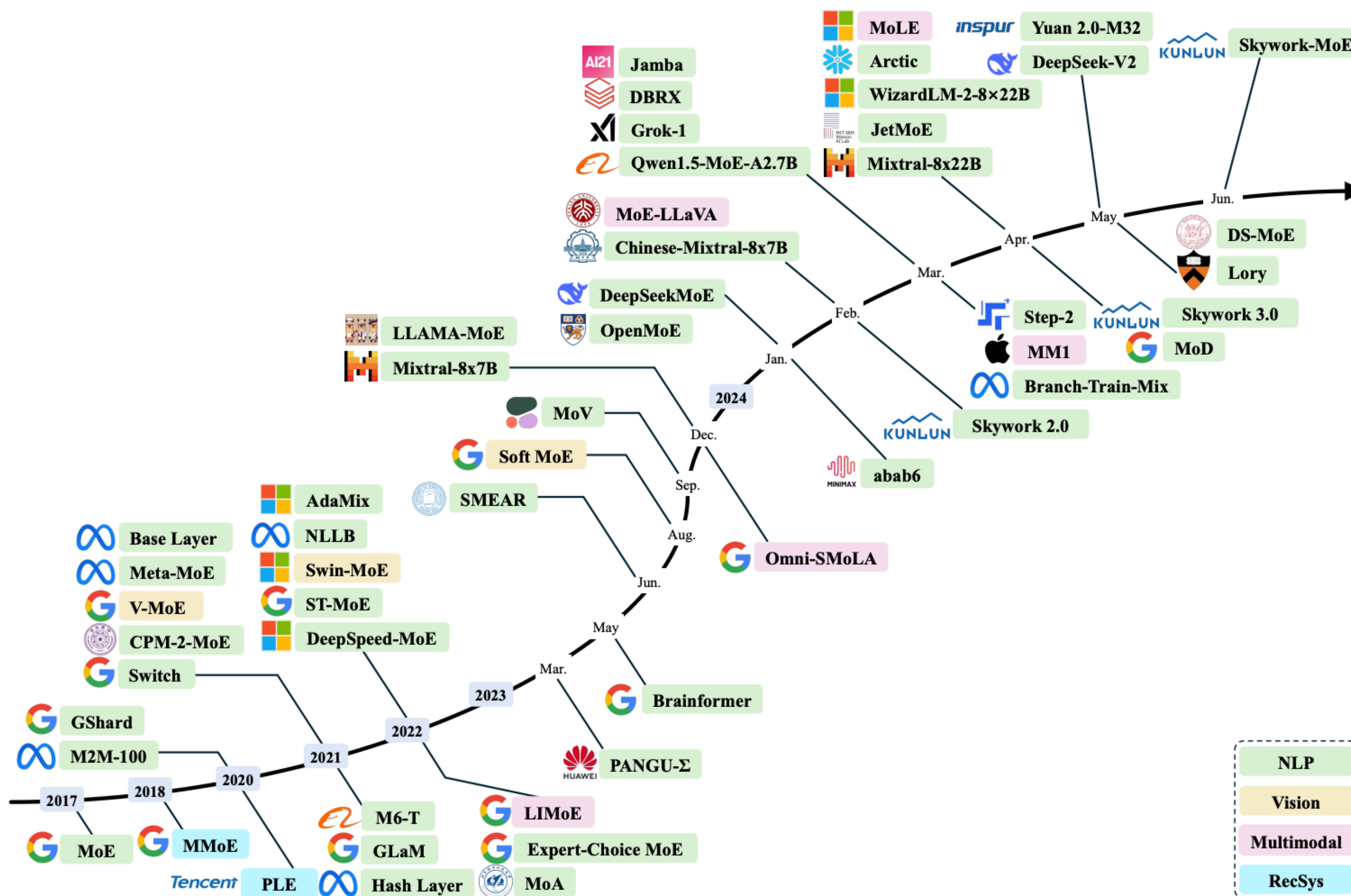
Main Goal: To scale massive AI models to several billions of parameters without increasing computation

Popular approaches:

- Sparse mixture-of-experts layer in massive neural networks (Shazeer et al. [3])
- Switch Transformer (Fedus et al. [4])
- DeepSeek, Mixtral (MistralAI), etc.

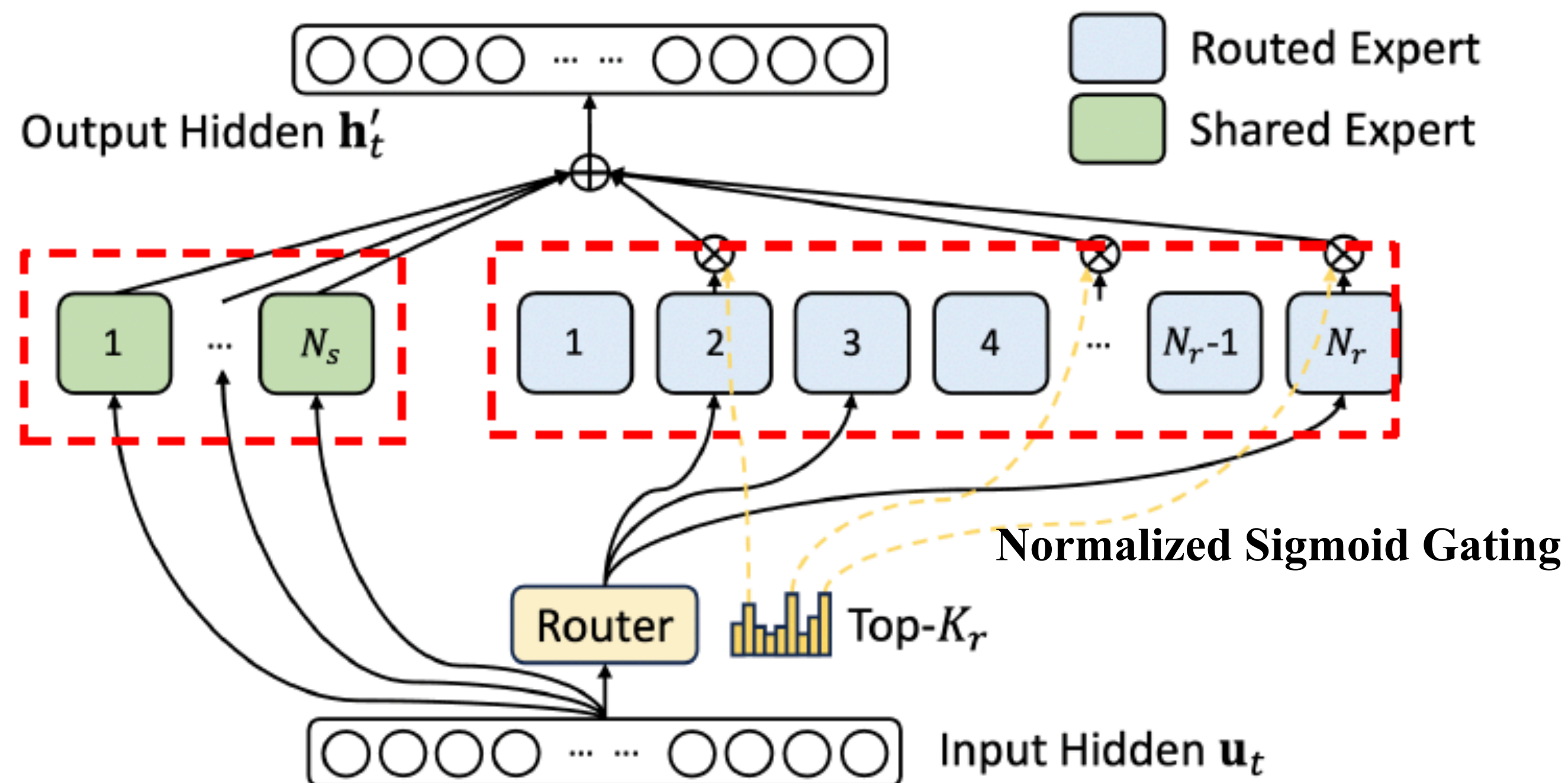
Our approach: We propose competition among experts (Nguyen et al. [5])

Scaling Up Massive AI Models via Sparse MoEs



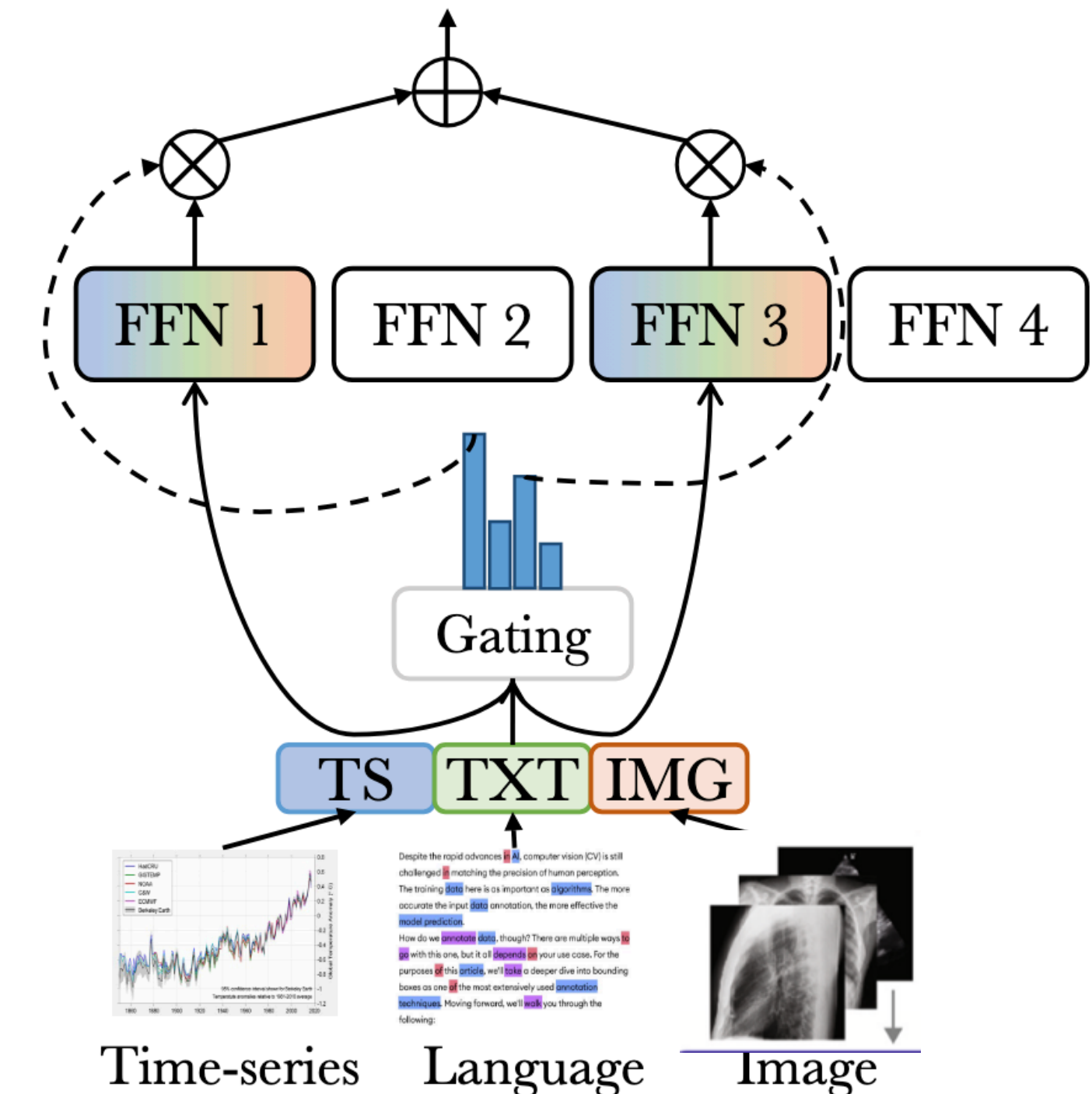
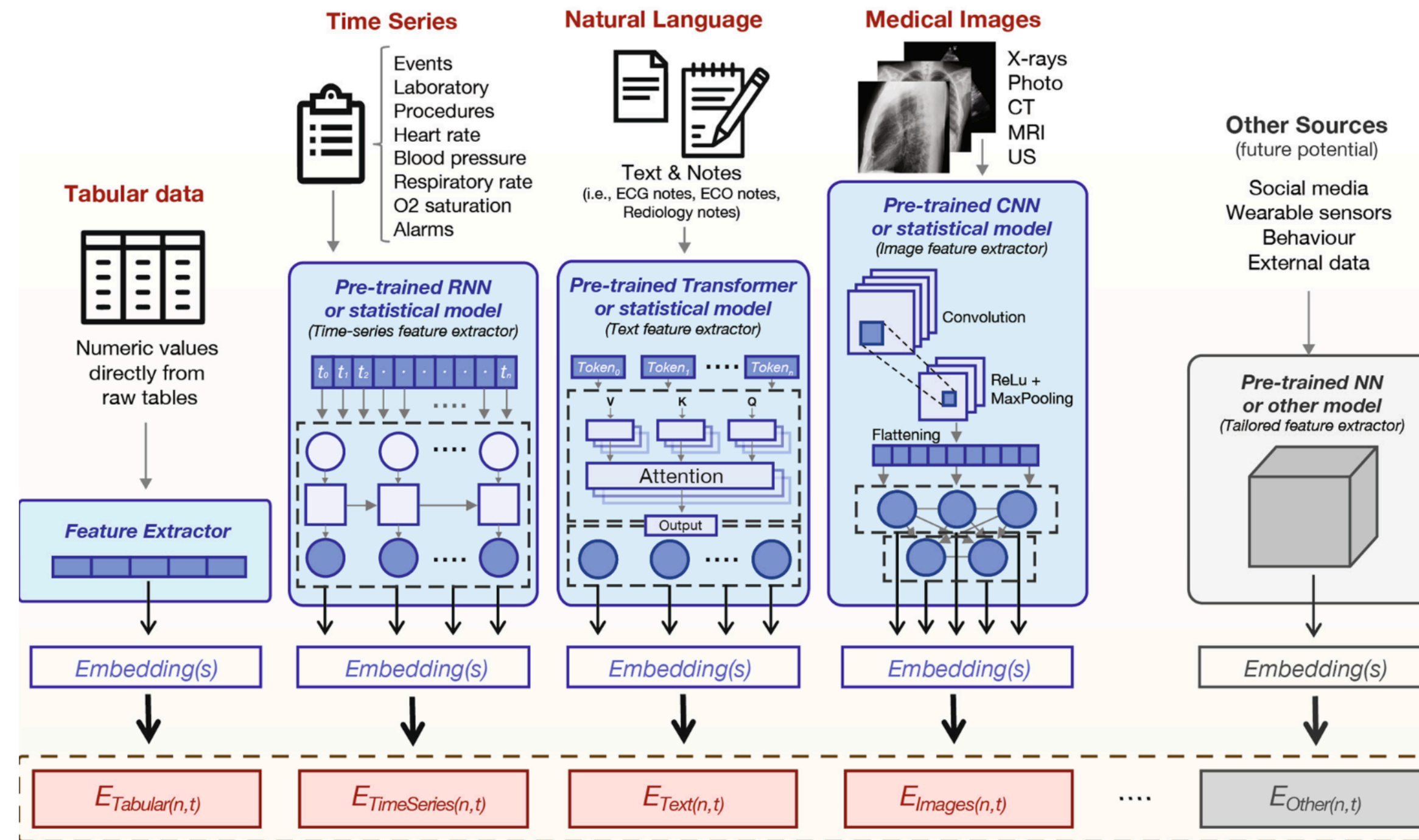
Chronological overview of Sparse MoEs in large-scale AI models (Can et al. [6])

DeepSeek's Sparse MoE



- In ongoing work (Nguyen et al. [7]), we theoretically study the behaviors of shared experts and normalized sigmoid gating in DeepSeek's MoE

Multimodal Models

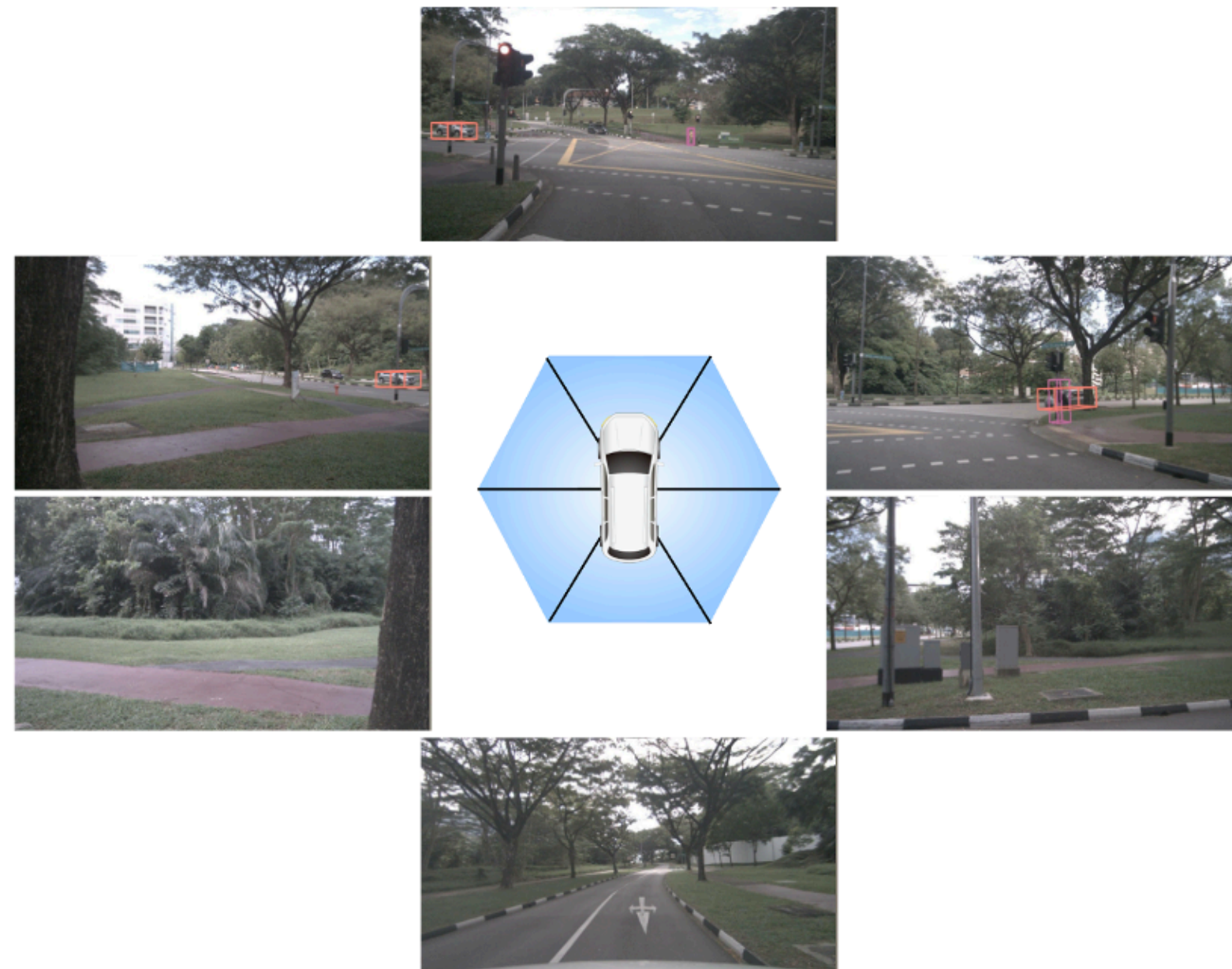


- **Main goal:** To build multimodal models for multimodal data (e.g, healthcare data)
- **Our approach:** We utilize **mixture of experts** (Han et al. [8] and Nguyen et al. [9]) to combine different modalities of data and achieve state-of-the-art results for several multimodal tasks

[8] Xing Han, Huy Nguyen, Carl William Harris, Nhat Ho, Suchi Saria. *FuseMoE: Mixture-of-experts Transformers for fleximodal fusion*. NeurIPS, 2024

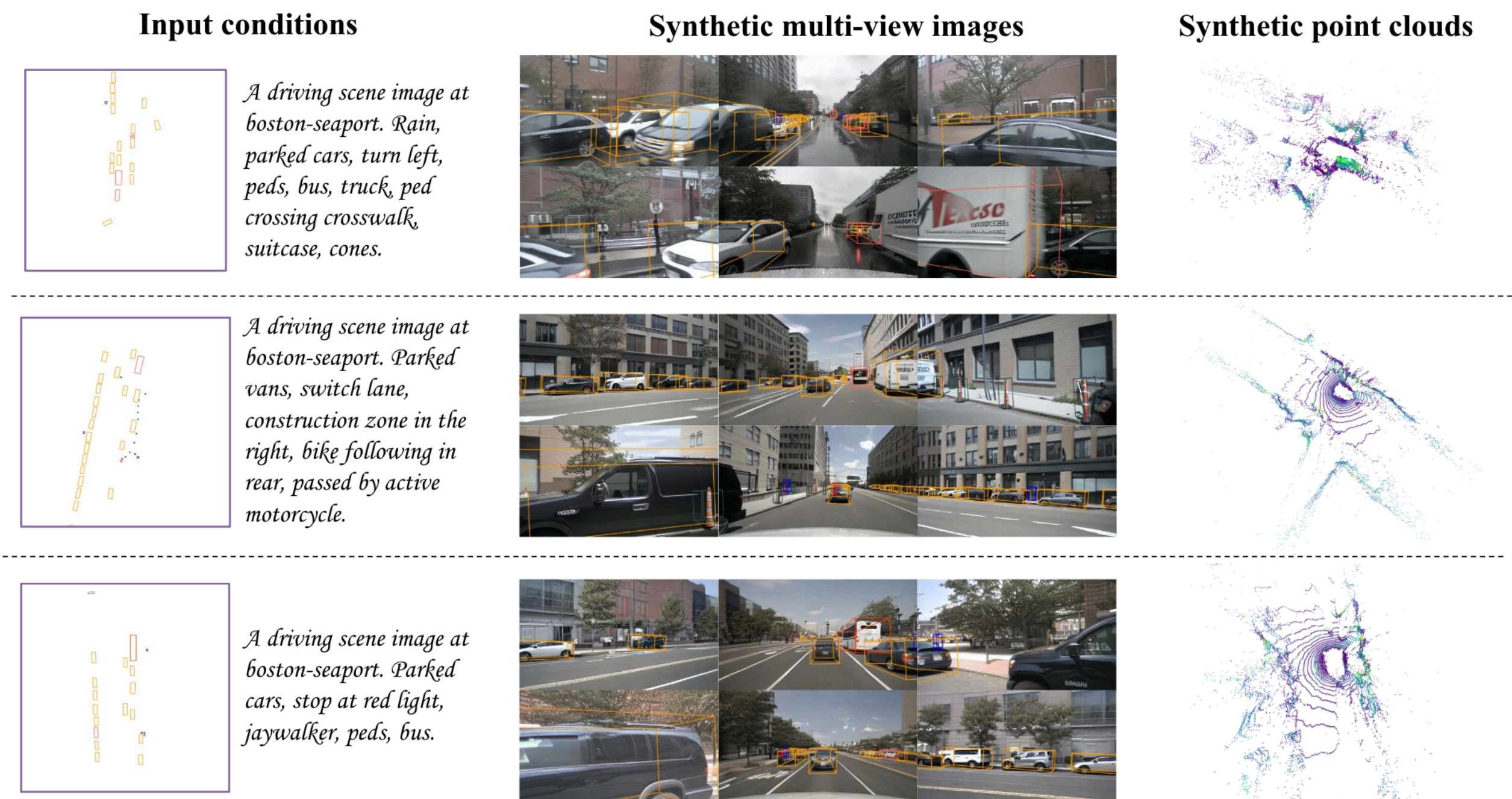
[9] Huy Nguyen, Xing Han, Carl William Harris, Suchi Saria, Nhat Ho. *On expert estimation in hierarchical mixture of experts: Beyond softmax gating functions*. Under review

High-quality Synthetic Data for Autonomous Vehicles



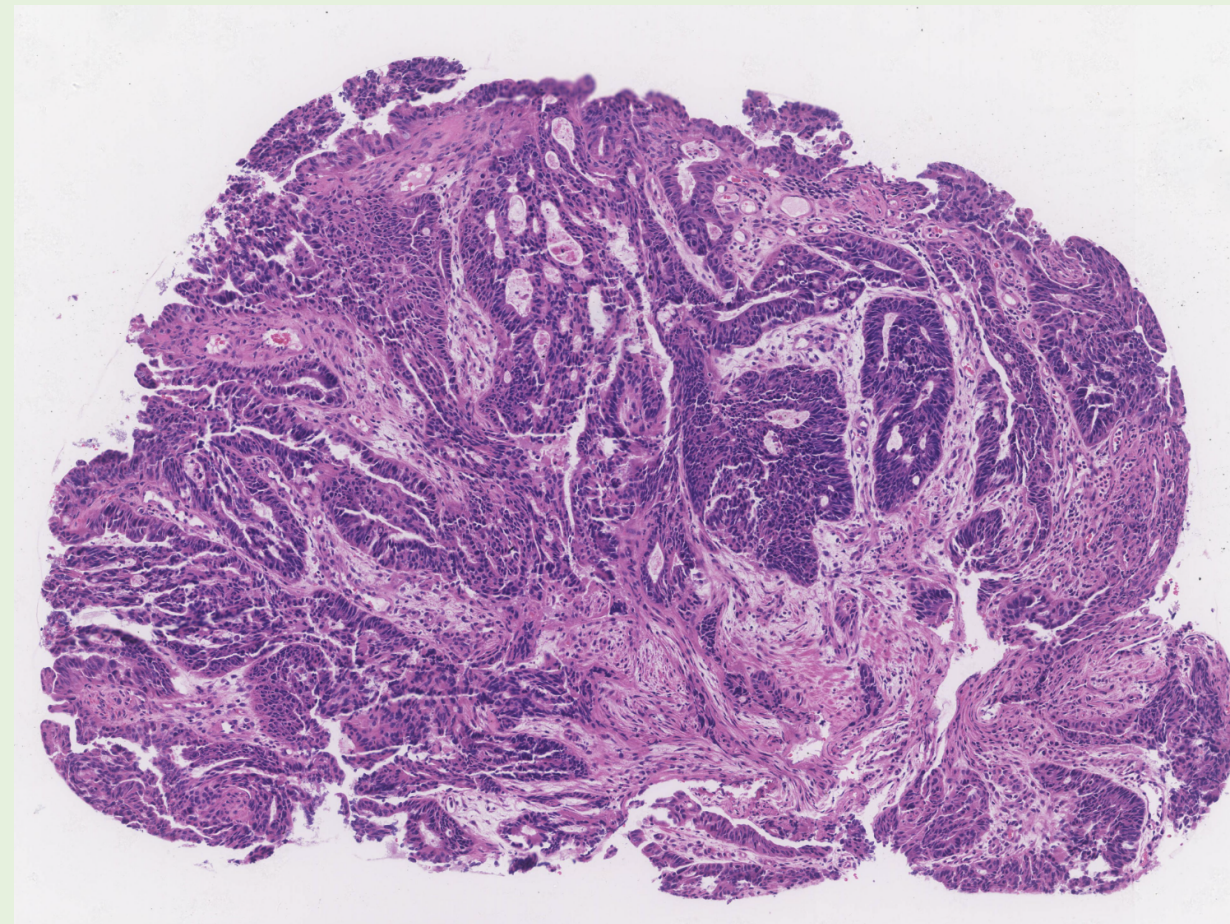
- Autonomous vehicles require a large amount of high-quality multi-view images and point clouds for AI training and evaluation (**Expensive!**)
- Synthetic data are used but need to be consistent to improve safety of vehicles

Our approach: We are working on using **mixture of experts** to scale up our recently proposed diffusion models (Xie et al. [10]) for consistency of generative multi-view images and point clouds

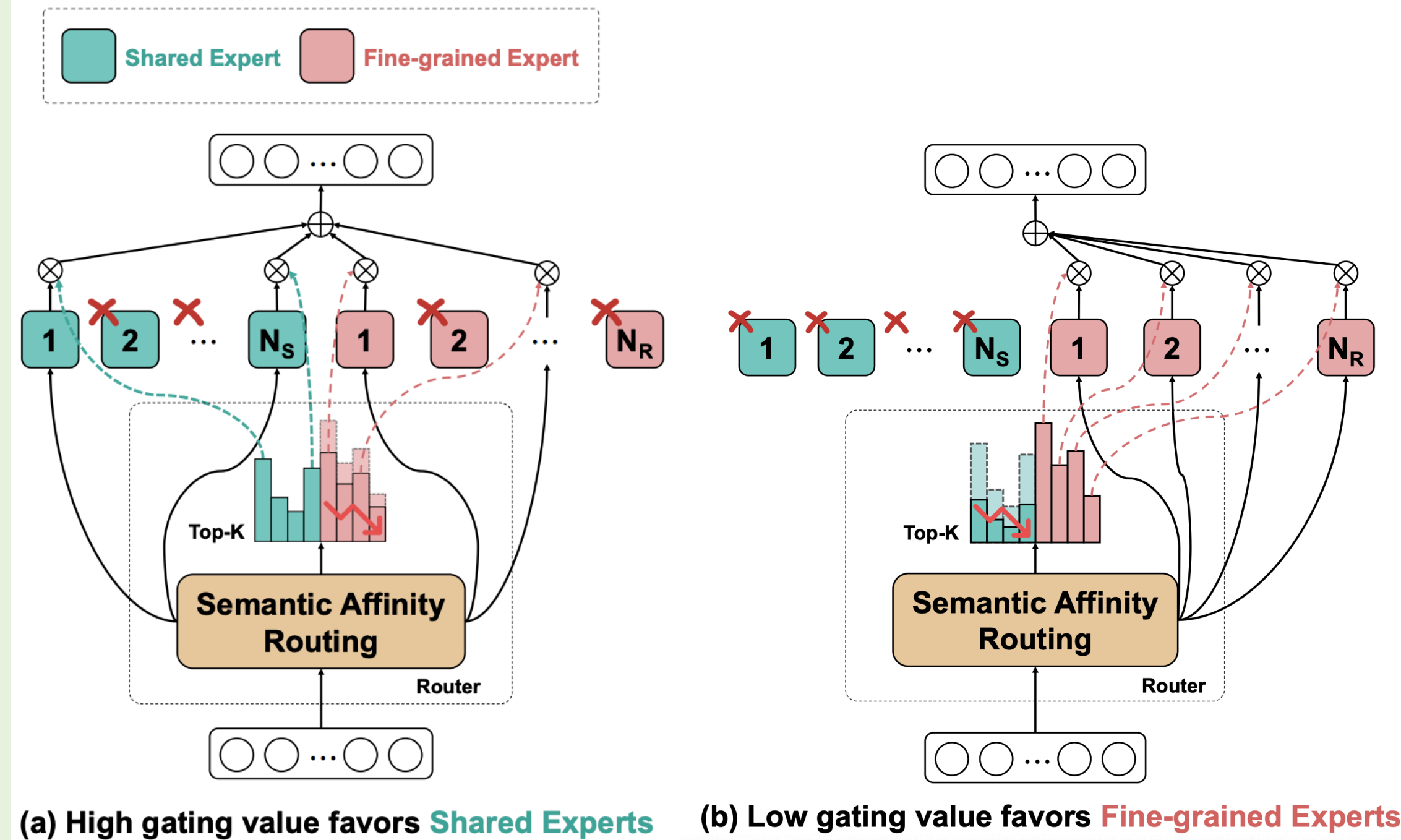


Smarter Healthcare: From Diagnosis to Deployment

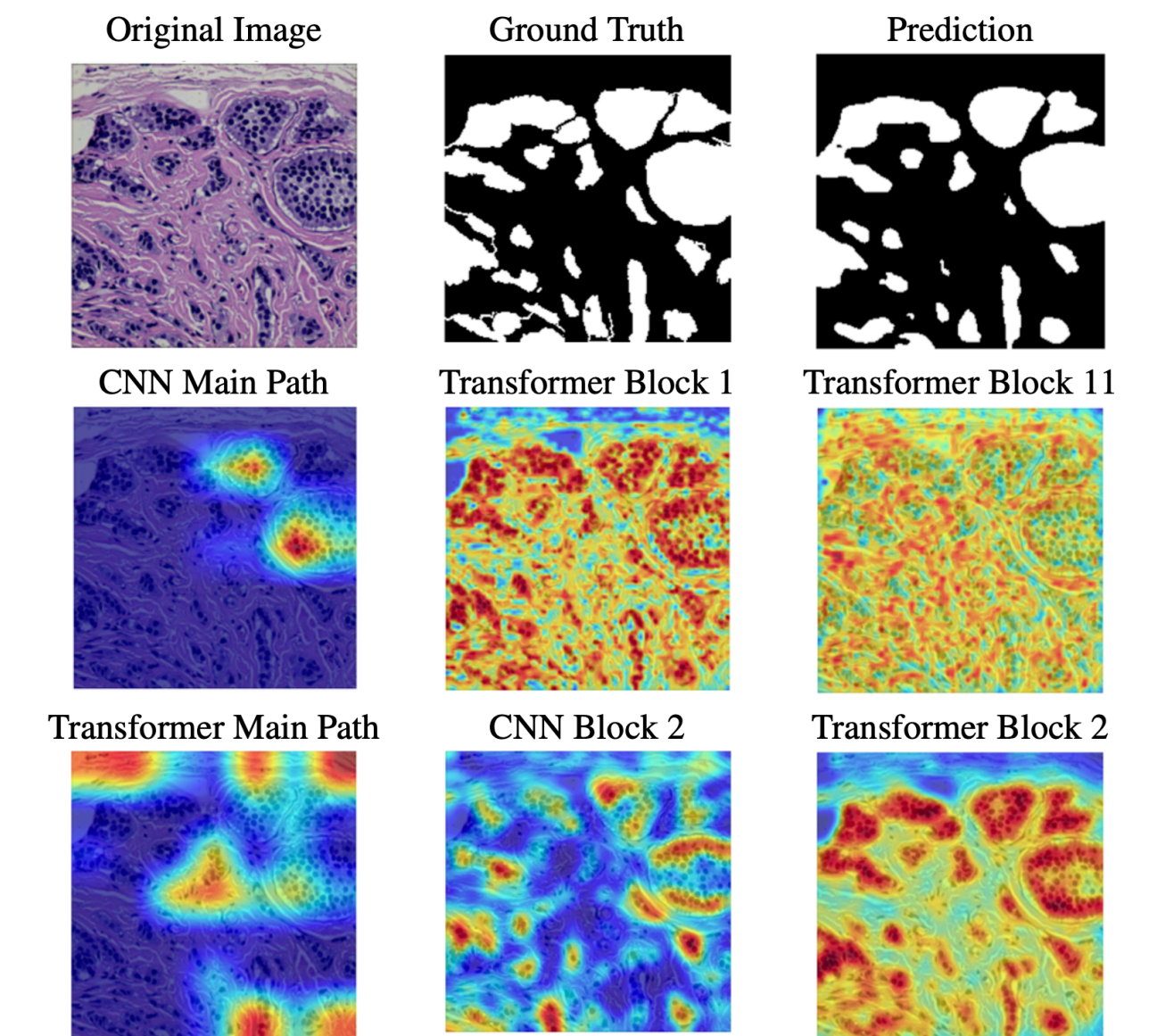
Problems



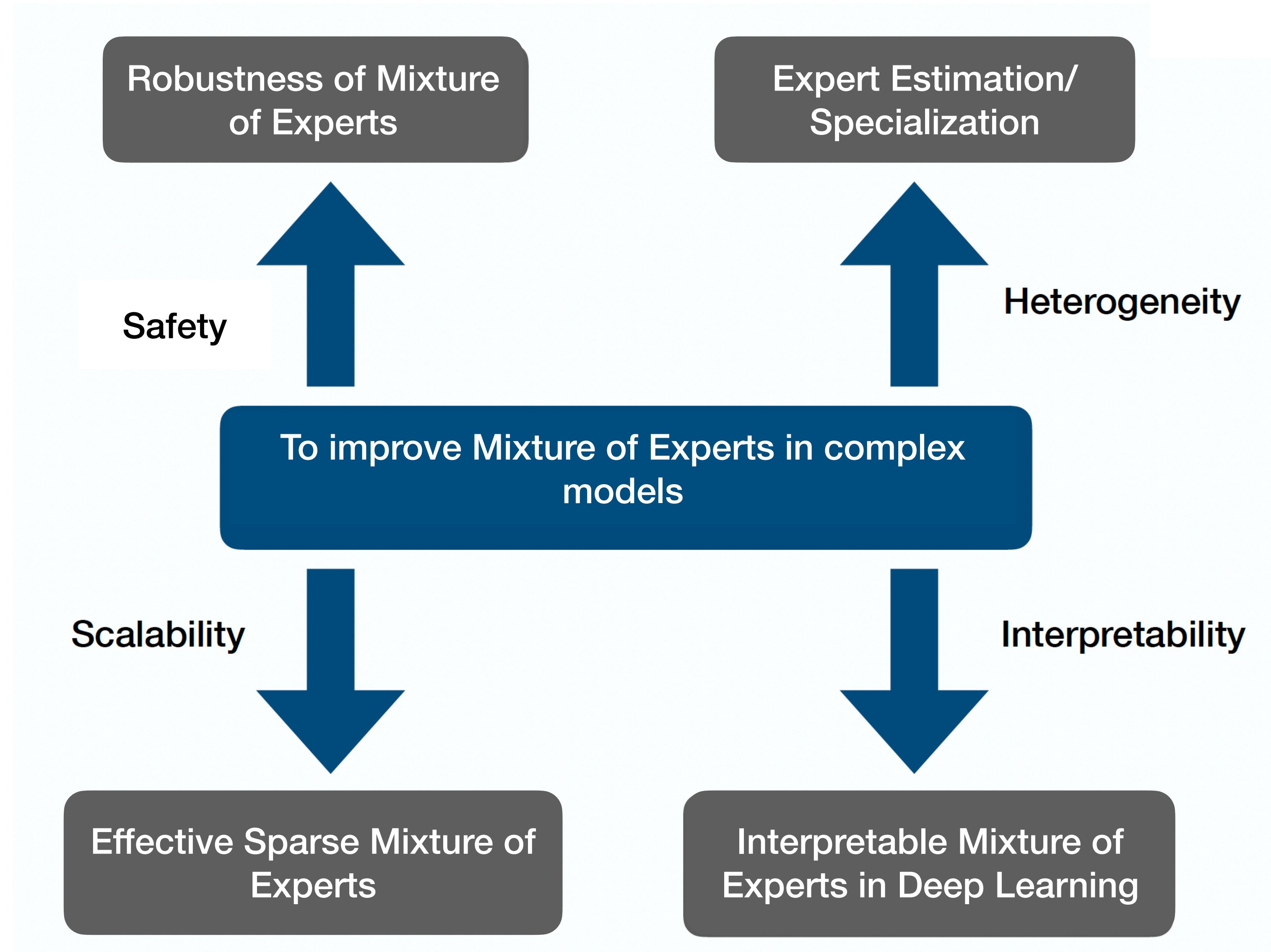
Methodology



Results



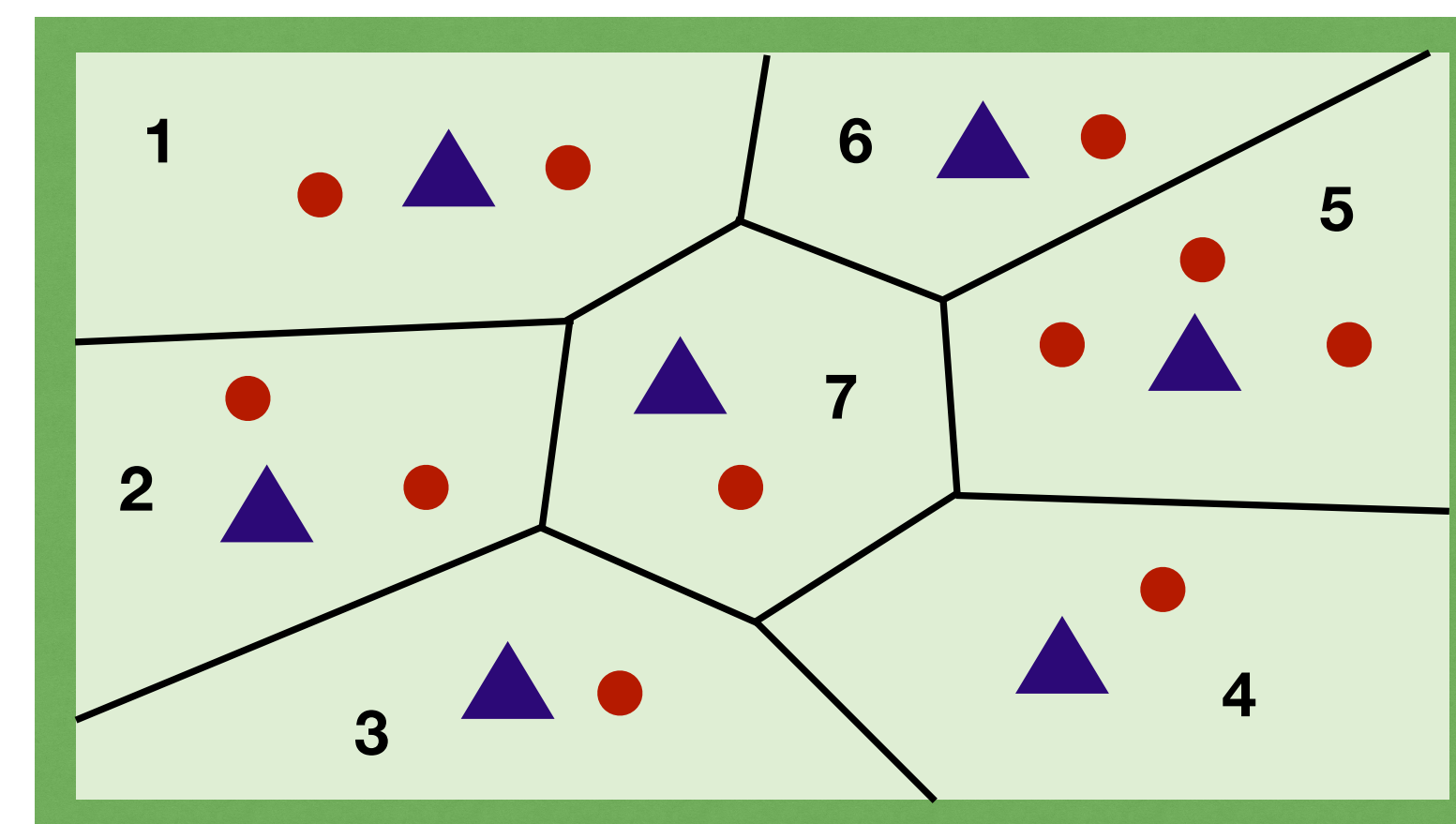
Fundamental Issues



Overview of Our Group Contribution

Theory (Voronoi diagram, algebraic geometry):

- Understanding heterogeneity and interpretability of (sparse) mixture of experts (Nguyen et al. [12], [13],[14], [15], [16], [17],[18], etc.)
- Explaining the success of Deep Learning Models (Nguyen et al. [19], [20])

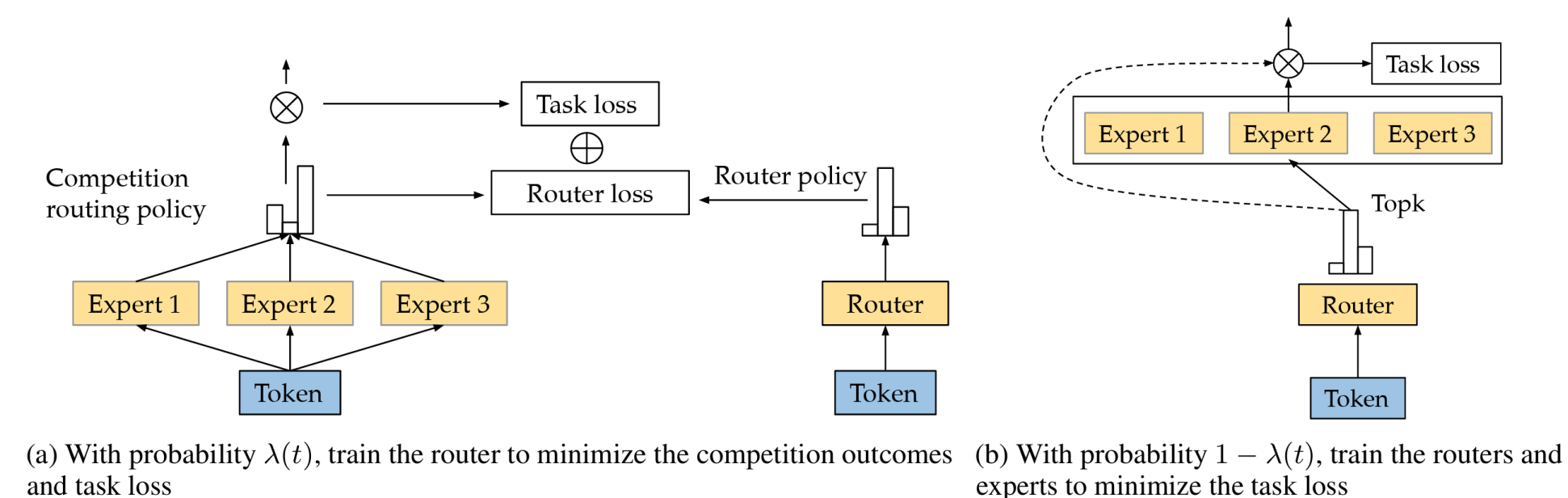


- [12] Huy Nguyen, TrungTin Nguyen, Nhat Ho. *Demystifying softmax gating function in Gaussian mixture of experts*. NeurIPS, 2023
- [13] Huy Nguyen, Pedram Akbarian, Fanqi Yan, Nhat Ho. *Statistical perspective of Top-K sparse softmax gating mixture of experts*. ICLR 2024
- [14] Huy Nguyen, Pedram Akbarian, Fanqi Yan, Nhat Ho. *A general theory for softmax gating multinomial logistic mixture of experts*. ICML 2024
- [15] Huy Nguyen, Pedram Akbarian, Nhat Ho. *Is temperature sample efficient for softmax Gaussian mixture of experts?* ICML 2024
- [16] Huy Nguyen, Nhat Ho†, Alessandro Rinaldo†. *On least square estimation in softmax gating mixture of experts*. ICML 2024
- [17] Huy Nguyen, Nhat Ho†, Alessandro Rinaldo†. *Sigmoid gating is more sample efficient than softmax gating in mixture of experts*. NeurIPS 2024
- [18] Huy Nguyen, Pedram Akbarian, Trang Pham, Trang Nguyen, Shujian Zhang Nhat Ho. *Statistical advantages of perturbing cosine router in sparse mixture of experts*. ICLR 2025
- [19] Huy Nguyen, Thong Doan, Quang Pham, Nghi Bui, Nhat Ho†, Alessandro Rinaldo†. *On DeepSeekMoE: Statistical benefits of shared experts and normalized sigmoid gating*. Under review
- [20] Fanqi Yan, Huy Nguyen, Pedram Akbarian, Nhat Ho†, Alessandro Rinaldo†. *Sigmoid self-attention is better than softmax self-attention: A mixture-of-experts perspective*. Under review

Overview of Our Group Contribution

Method (Efficient, Scalable, and Continual AI):

- Scaling massive AI models using new sparse mixture of experts (Nguyen et al. [5])
- Improving popular parameter efficient fine-tuning methods (PEFTs), such as LoRA, Prompt Tuning, etc. via the perspective of MoEs (Le et al. [21], [22], [27], [28]; Diep et al. [23], [25], [26]; Truong et al. [24])



[21] Minh Le, An Nguyen, Huy Nguyen, Trang Nguyen, Trang Pham, Linh Van Ngo, Nhat Ho. *Mixture of experts meets prompt-based continual learning*. NeurIPS, 2024

[22] Minh Le, Chau Nguyen, Huy Nguyen, Quyen Tran, Trung Le, Nhat Ho. *Revisiting prefix-tuning: Statistical benefits of reparameterization among prompts*. ICLR, 2025

[23] Nghiem Diep, Huy Nguyen, Chau Nguyen, Minh Le, Duy Nguyen, Daniel Sonntag, Mathias Niepert, Nhat Ho. *On zero-initialized attention: Optimal prompt and gating factor estimation*. ICML, 2025

[24] Tuan Truong, Chau Nguyen, Huy Nguyen, Minh Le, Trung Le, Nhat Ho. *RepLoRA: Reparameterizing low-rank adaptation via the perspective of mixture of experts*. ICML, 2025

[25] Nghiem Diep, Dung Le, Tuan Truong, Tan Dinh, Huy Nguyen, Nhat Ho. *HoRA: Cross-head low-rank adaptation with joint hypernetworks*. Under review

[26] Nghiem Tuong Diep, Hien Dang, Tuan Truong, Tan Dinh, Huy Nguyen, Nhat Ho. *DoRAN: Stabilizing weight-decomposed low-rank adaptation via noise injection and auxiliary networks*. Under review

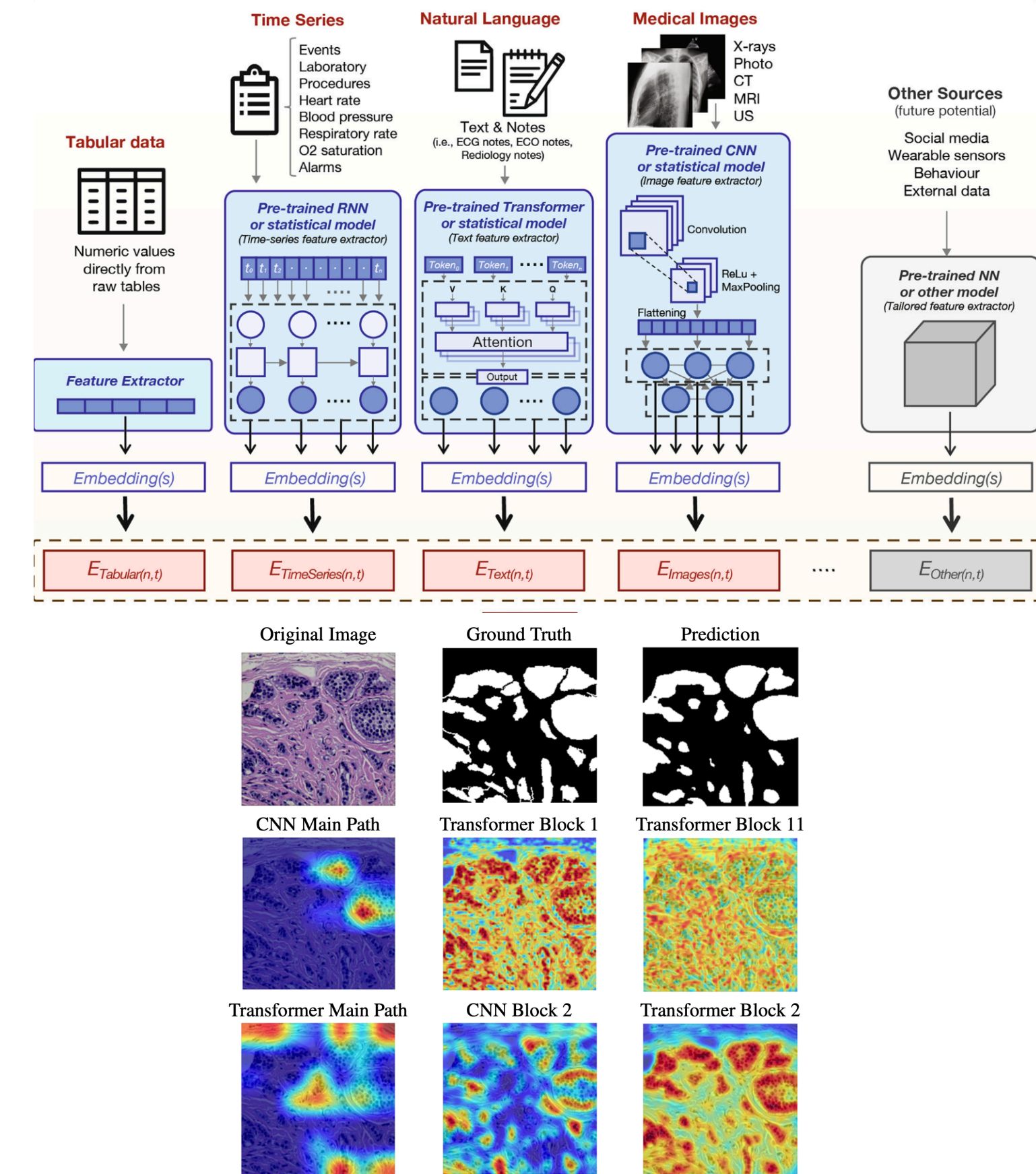
[27] Minh Le, Anh Nguyen, Huy Nguyen, Chau Nguyen, Anh Tran, Nhat Ho. *On the expressiveness of visual prompt experts*. Under review

[28] Minh Le, Bao-Ngoc Dao, Quyen Tran, Huy Nguyen, Anh Nguyen, Nhat Ho. *One-prompt strikes back: Sparse mixture of experts for prompt-based continual learning*. Under review

Overview of Our Group Contribution

Application (Healthcare, Autonomous Vehicle):

- Building large multimodal model for healthcare applications via (hierarchical) mixture of experts (Han et al. [8] and [9], etc.);
- Improving cancer diagnosis (Thai et al. [11], etc.)
- Several other ongoing works



- [8] Xing Han, Huy Nguyen, Carl William Harris, Nhat Ho, Suchi Saria. *FuseMoE: Mixture-of-experts Transformers for fleximodal fusion*. NeurIPS, 2024
- [9] Huy Nguyen, Xing Han, Carl William Harris, Suchi Saria, Nhat Ho. *On expert estimation in hierarchical mixture of experts: Beyond softmax gating functions*. Under review
- [11] Huy Thai, Hoang-Nguyen Vu, Anh-Minh Phan, Quang-Thinh Ly, Tram Dinh, Thi-Ngoc-Truc Nguyen, Nhat Ho. *Shape-adapting gated experts: Dynamic expert routing for colonoscopic lesion segmentation*. Under review

This Minicourse: Understanding Heterogeneity

- **Heterogeneity of Experts:** Mixture of experts are generally over-parameterized
 - Effect of over-parameterization on parameter/ expert estimation has remained poorly understood
- **Understanding heterogeneity has several applications:**
 - Improving **Self-attention in Transformers**, a cornerstone deep learning architecture
 - Boosting the performance of LoRA, a popular parameter-efficient fine-tuning method

This talk:

- **Heterogeneity of Experts:** Establishing the convergence rates of expert estimation
- **Applications of Theories:** Improving self-attention in Transformers and boosting the performance of LoRA

Talk Outline

- **Heterogeneity of Experts in Mixture of Experts**
 - Gaussian Mixture of Experts and Regression Mixture of Experts
 - DeepSeek Mixture of Experts
- **Insights from Theories to Deep Learning Applications**
 - Improving self-attention in Transformer
 - Boosting the performance of LoRA

Key points: Voronoi-based losses are natural and powerful for understanding the heterogeneity of experts

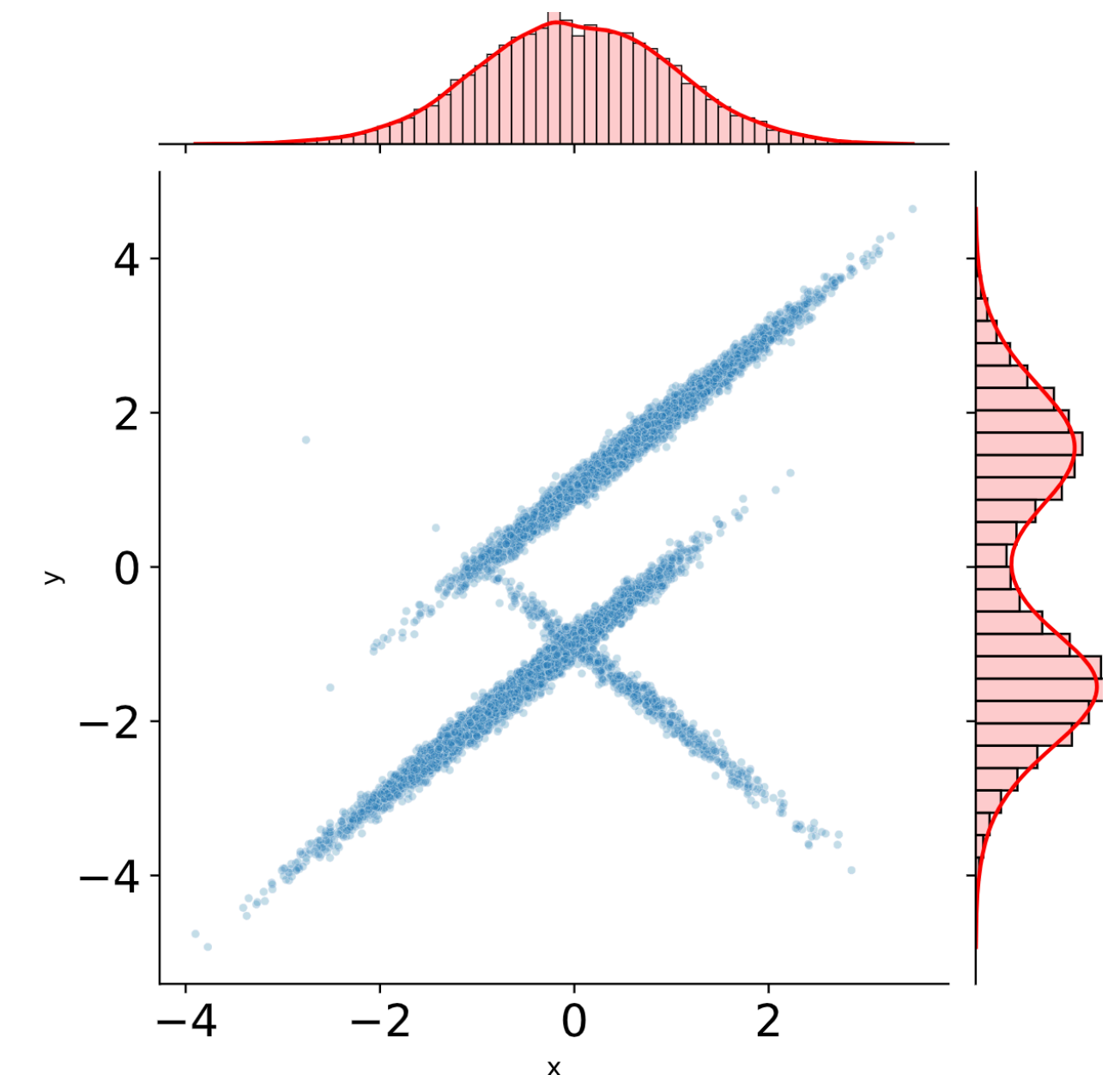
Gaussian Mixture of Experts (MoEs)

- Given a random sample $(Y_1, X_1), \dots, (Y_n, X_n) \in \mathbb{R} \times \mathbb{R}^d$ from the conditional density

$$g_G(Y|X) = \sum_{i=1}^{k_0} \frac{\exp(\beta_{1i}^\top X + \beta_{0i})}{\sum_{j=1}^{k_0} \exp(\beta_{1j}^\top X + \beta_{0j})} \cdot f(Y|a_i^\top X + b_i, \sigma_i) \quad \text{where}$$

$G = \sum_{i=1}^{k_0} \exp(\beta_{0i}) \delta_{(\beta_{1i}, a_i, b_i, \sigma_i)}$ is unknown mixing measure (not necessarily a probability measure)

- The generalization to general experts (deep neural networks) is possible
- Known:** Gaussian family of distributions $\{f(x|\theta, \sigma)\}$
- Unknown:**
 - softmax weights $\{\beta_{1i}\}_{i=1}^{k_0}$ and biases $\{\beta_{0i}\}_{i=1}^{k_0}$
 - Experts parameters $\{a_i, b_i\}_{i=1}^{k_0}$, scales $\{\sigma_i\}_{i=1}^{k_0}$
 - number of experts k_0



Maximum Likelihood Estimation (MLE)

- **Main goal:** To estimate $G = \sum_{i=1}^{k_0} \exp(\beta_{0i}) \delta_{(\beta_{1i}, a_i, b_i, \sigma_i)}$
- **Over-specified/ Over-parameterized setting:** As the number of true experts k_0 is unknown, we use mixture of k experts where k is given and $k > k_0$
- The MLE is then given by:

$$\widehat{G}_n \in \arg \max_{G' \in \mathcal{O}_k} \frac{1}{n} \sum_{i=1}^n \log(g_{G'}(Y_i | X_i))$$

where $\mathcal{O}_k = \{G' = \sum_{i=1}^{k'} \exp(\beta'_{0i}) \delta_{(\beta'_{1i}, a'_i, b'_i, \sigma'_i)} : k' \leq k\}$

Revisiting Identifiability of Gaussian MoEs

- Before studying the convergence of MLE, we need to guarantee the identifiability of Gaussian MoEs

- Identifiability:** For any mixing measures $G' = \sum_{i=1}^{k'} \exp(\beta'_{0i}) \delta_{(\beta'_{1i}, a'_i, b'_i, \sigma'_i)}$,

$$g_G(Y|X) = g_{G'}(Y|X) \text{ for almost surely } (X, Y)$$

if and only if $k' = k_0$ and $G' \equiv G_{t_1, t_2}$ where

$$G_{t_1, t_2} := \sum_{i=1}^{k_0} \exp(\beta_{0i} + t_1) \delta_{(\beta_{1i} + t_2, a_i, b_i, \sigma_i)}, \text{ for some } t_1 \in \mathbb{R} \text{ and } t_2 \in \mathbb{R}^d$$

Identifiability up to translation!

From Density Estimation to Parameter Estimation

- **Density Estimation Rate:** Under the over-specified setting,

$$\mathbb{E}_X[h(g_{\widehat{G}_n}(\cdot | X), g_G(\cdot | X))] = O_P\left(\sqrt{\frac{\log n}{n}}\right)$$

where h stands for the Hellinger distance

- **From Density Estimation to Parameter Estimation:** We aim to establish

$$\mathbb{E}_X[h(g_{\widehat{G}_n}(\cdot | X), g_G(\cdot | X))] \gtrsim D_r(\widehat{G}_n, G)$$

where D_r is some divergence and r is some (vector) power of the parameters

- **High level proof idea:** We use Taylor expansion to determine the value of r

Rate of Parameter Estimation: Challenges

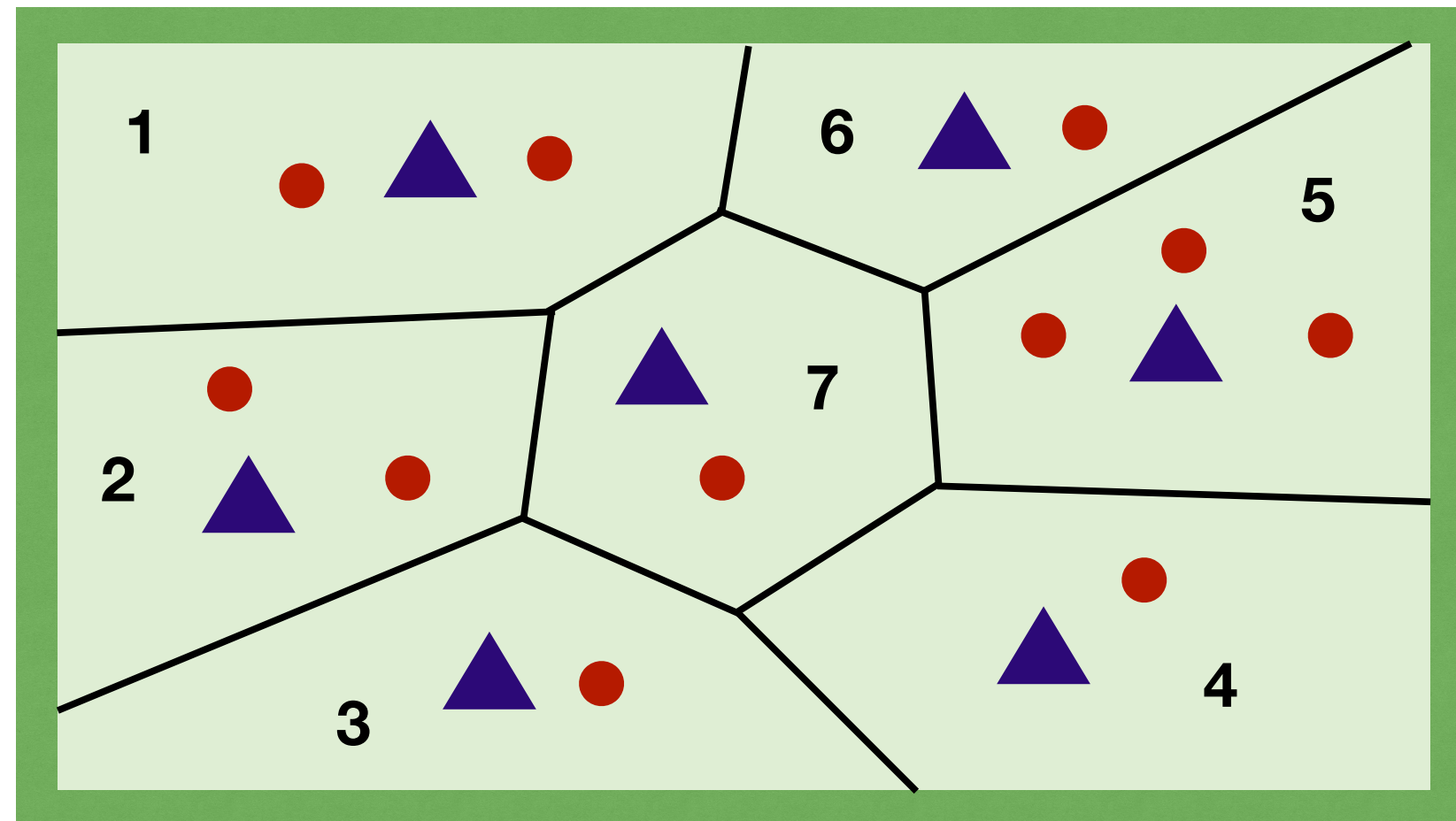
- **Main goal:** To determine the divergence between \widehat{G}_n and G
- **Two Challenges:**
 1. G and \widehat{G}_n are not probability measures
 - **Optimal transport cannot** be used as in standard mixture models (Nguyen [29], Ho and Nguyen [30, 31], etc.)
 2. Complex interaction among parameters

$$\frac{\partial^2 u}{\partial \beta_1 \partial b} = \frac{\partial u}{\partial a}; \quad \frac{\partial^2 u}{\partial b^2} = 2 \frac{\partial u}{\partial \sigma}$$

where $u(Y|X; \beta_1, a, b, \sigma) = \exp(\beta_1^\top X) \cdot f(Y|a^\top X + b, \sigma)$

- Therefore, **different parameters may have different rates**
- **Solution:** We resolve these challenges by developing novel Voronoi losses among mixing measures

Voronoi-based Loss: Definition



Blue triangles: True parameters

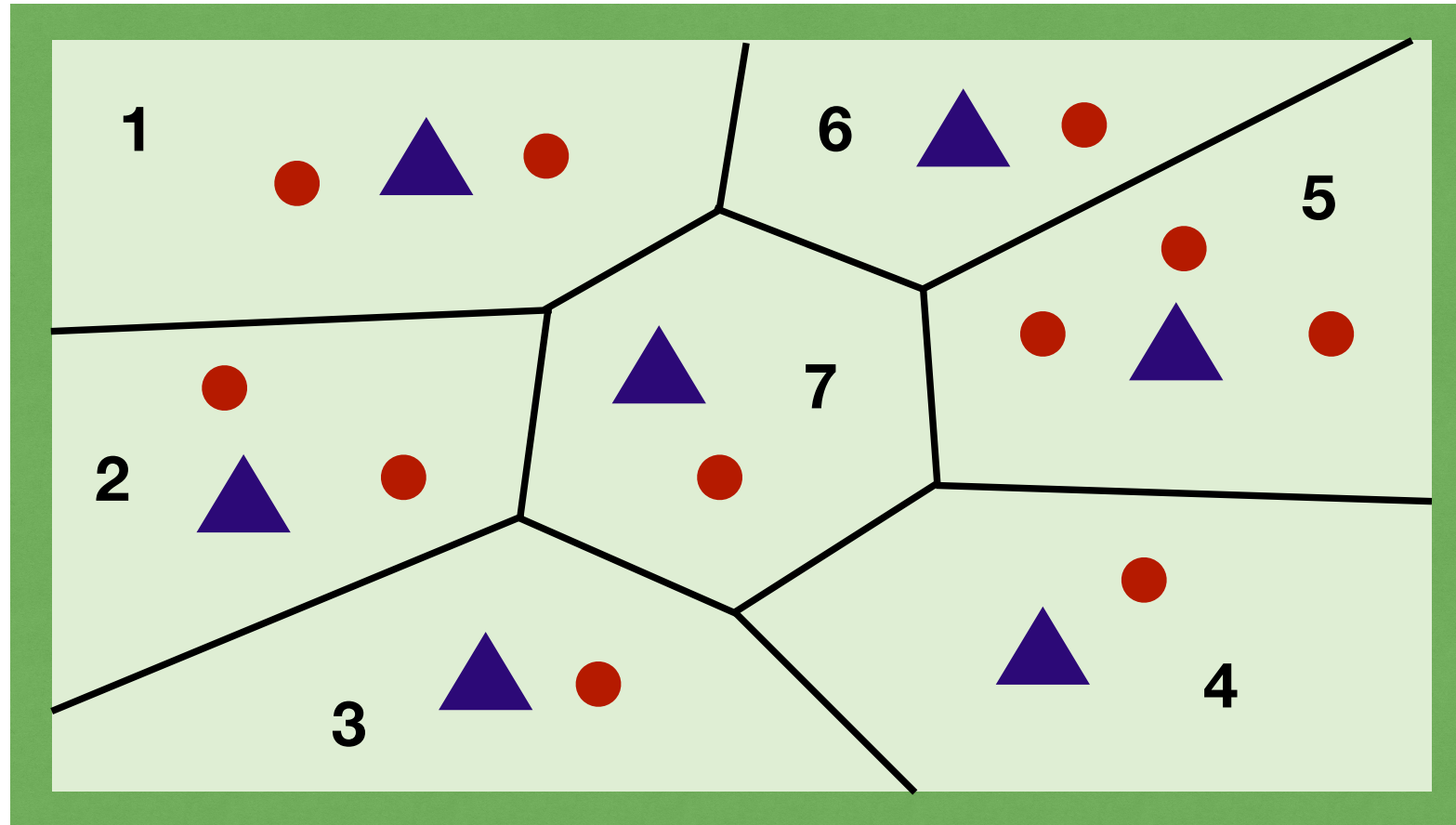
Red points: Parameters from $G' \in \mathcal{O}_k$

- **Voronoi cells:** For any $G' \in \mathcal{O}_k$,

$$\mathcal{A}_j \equiv \mathcal{A}_j(G') := \{i \in \{1, 2, \dots, k\} : \|\omega'_i - \omega_j\| \leq \|\omega'_i - \omega_\ell\|, \forall \ell \neq j\}$$

where $\omega'_i = (\beta'_{1i}, a'_i, b'_i, \sigma'_i)$ and $\omega_j = (\beta_{1j}, a_j, b_j, \sigma_j)$

Voronoi-based Loss: Definition



Blue triangles: True parameters

Red points: Parameters from $G' \in \mathcal{O}_k$

Voronoi loss: For any $r = (r(\mathcal{A}_1), \dots, r(\mathcal{A}_{k_0}))$,

$$\begin{aligned}
 \mathcal{D}_r(G', G) = \inf_{t_1, t_2} & \left\{ \sum_{\substack{j: |\mathcal{A}_j| = 1, \\ i \in \mathcal{A}_j}} \exp(\beta_{0i}) \|(\Delta_{t_2} \beta_{1ij}, \Delta a_{ij}, \Delta b_{ij}, \Delta \sigma_{ij})\| \right. \\
 & + \sum_{\substack{j: |\mathcal{A}_j| > 1, \\ i \in \mathcal{A}_j}} \exp(\beta_{0i}) \left(\|(\Delta_{t_2} \beta_{1ij}, \Delta b_{ij})\|^{r(|\mathcal{A}_j|)} + \|(\Delta a_{ij}, \Delta \sigma_{ij})\|^{r(|\mathcal{A}_j|)/2} \right) + \sum_{j=1}^{k_*} \left| \sum_{i \in \mathcal{A}_j} \exp(\beta'_{0i}) - \exp(\beta_{0j} + t_1) \right| \Big\}
 \end{aligned}$$

where $\Delta_{t_2} \beta_{1ij} = \beta'_{1i} - \beta_{1j} - t_2$, $\Delta a_{ij} = a'_i - a_j$, $\Delta b_{ij} = b'_i - b_j$, and $\Delta \sigma_{ij} = \sigma'_i - \sigma_j$

Parameter Estimation via Voronoi Loss

- **Lower bound:** For any $G' \in \mathcal{O}_k$, we can show that (Nguyen et al. [12])

$$\mathbb{E}_X[h(g_{G'}(\cdot | X), g_G(\cdot | X))] \gtrsim \mathcal{D}_{\bar{r}}(G', G),$$

where $\bar{r} := (\bar{r}(|\mathcal{A}_1|), \dots, \bar{r}(|\mathcal{A}_{k_0}|))$ and $\bar{r}(m)$ is the smallest natural number r such that the following system of polynomial equations does not have any non-trivial solution:

$$\sum_{j=1}^m \sum_{(\alpha_1, \alpha_2, \alpha_3, \alpha_4) \in \mathcal{J}_{\ell_1, \ell_2}} \frac{p_{5j}^2 p_{1j}^{\alpha_1} p_{2j}^{\alpha_2} p_{3j}^{\alpha_3} p_{4j}^{\alpha_4}}{\alpha_1! \alpha_2! \alpha_3! \alpha_4!} = 0,$$

for any $(\ell_1, \ell_2) \in \mathbb{N}^d \times \mathbb{N}$ such that $1 \leq |\ell_1| + \ell_2 \leq r$, where

$$\mathcal{J}_{\ell_1, \ell_2} := \{(\alpha_1, \alpha_2, \alpha_3, \alpha_4) \in \mathbb{N}^d \times \mathbb{N}^d \times \mathbb{N} \times \mathbb{N} : \alpha_1 + \alpha_2 = \ell_1, |\alpha_2| + \alpha_3 + 2\alpha_4 = \ell_2\}$$

Connection to Algebraic Geometry

- $\bar{r}(m)$ is the smallest natural number r such that the following system of polynomial equations does not have any non-trivial solution:

$$\sum_{j=1}^m \sum_{(\alpha_1, \alpha_2, \alpha_3, \alpha_4) \in \mathcal{F}_{\ell_1, \ell_2}} \frac{p_{5j}^2 p_{1j}^{\alpha_1} p_{2j}^{\alpha_2} p_{3j}^{\alpha_3} p_{4j}^{\alpha_4}}{\alpha_1! \alpha_2! \alpha_3! \alpha_4!} = 0,$$

for any $(\ell_1, \ell_2) \in \mathbb{N}^d \times \mathbb{N}$ such that $1 \leq |\ell_1| + \ell_2 \leq r$, where

$$\mathcal{F}_{\ell_1, \ell_2} := \{(\alpha_1, \alpha_2, \alpha_3, \alpha_4) \in \mathbb{N}^d \times \mathbb{N}^d \times \mathbb{N} \times \mathbb{N} : \alpha_1 + \alpha_2 = \ell_1, |\alpha_2| + \alpha_3 + 2\alpha_4 = \ell_2\}$$

- **Some values of $\bar{r}(m)$:** When $m = 1$, $\bar{r}(m) = 4$;

When $m = 2$, $\bar{r}(m) = 6$;

- It is challenging to determine exact value of $\bar{r}(m)$ for general over-specified setting

Rate of Parameter Estimation

- Recall that for any $G' \in \mathcal{O}_k$, $\mathbb{E}_X[h(g_{G'}(\cdot | X), g_G(\cdot | X))] \gtrsim \mathcal{D}_{\bar{r}}(G', G)$
- Since $\mathbb{E}_X[h(g_{\hat{G}_n}(\cdot | X), g_G(\cdot | X))] = O_P(\sqrt{\log n/n})$, we directly have

$$\mathcal{D}_{\bar{r}}(\hat{G}_n, G) = O_P(\sqrt{\log(n)/n})$$

- Implication:**
 - Rates of softmax weights and expert biases are respectively $n^{-1/2\bar{r}(|\mathcal{A}_j|)}$ and $n^{-1/2}$ for those in Voronoi cells $|\mathcal{A}_j| > 1$ and $|\mathcal{A}_j| = 1$
 - Rates of expert weights and variances are respectively $n^{-1/\bar{r}(|\mathcal{A}_j|)}$ and $n^{-1/2}$ for those in Voronoi cells $|\mathcal{A}_j| > 1$ and $|\mathcal{A}_j| = 1$
 - Rates of softmax biases are $n^{-1/2}$

Remark on the Rates

- Recall that we have

$$\mathcal{D}_{\bar{r}}(\widehat{G}_n, G) = \mathcal{O}_P(\sqrt{\log(n)/n})$$

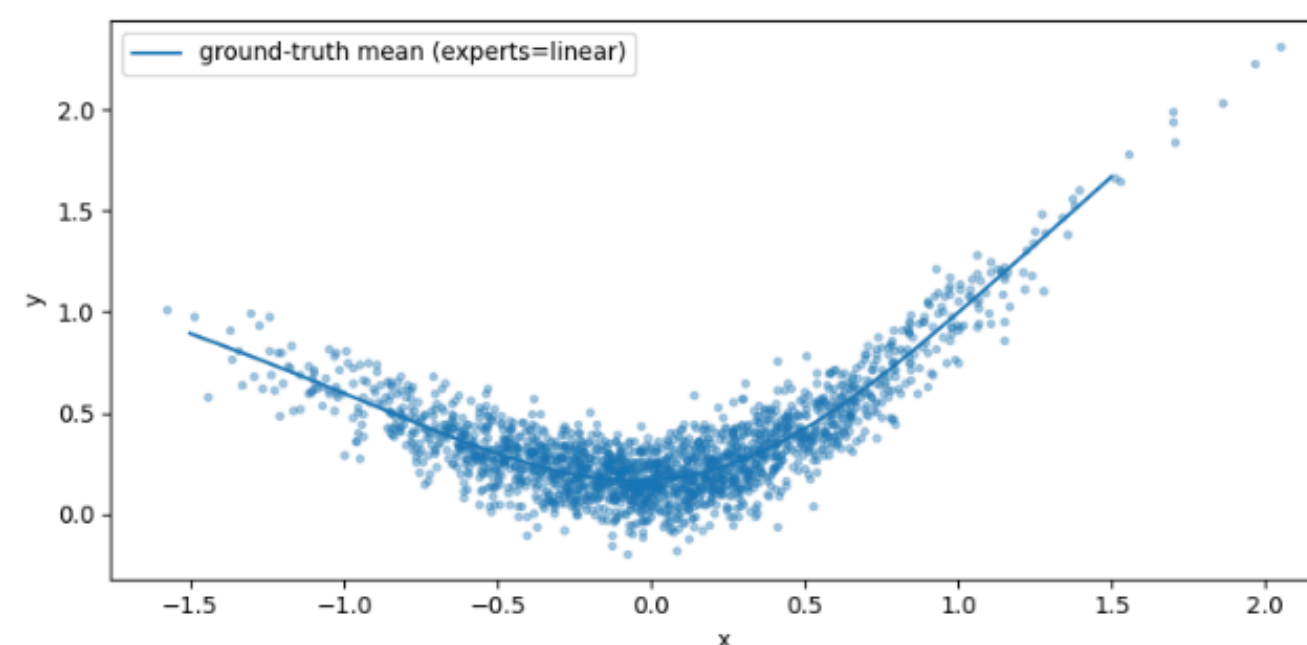
- Unfortunately, the order \bar{r} is not optimal
- Current progress to sharpen the rate of MLE and extend Gaussian MoEs beyond linear experts:
 - We develop comprehensive condition on expert functions (including deep neural networks) to have much faster rates on estimating all the parameters and expert functions

Beyond the Gaussian MoEs: Regression MoE

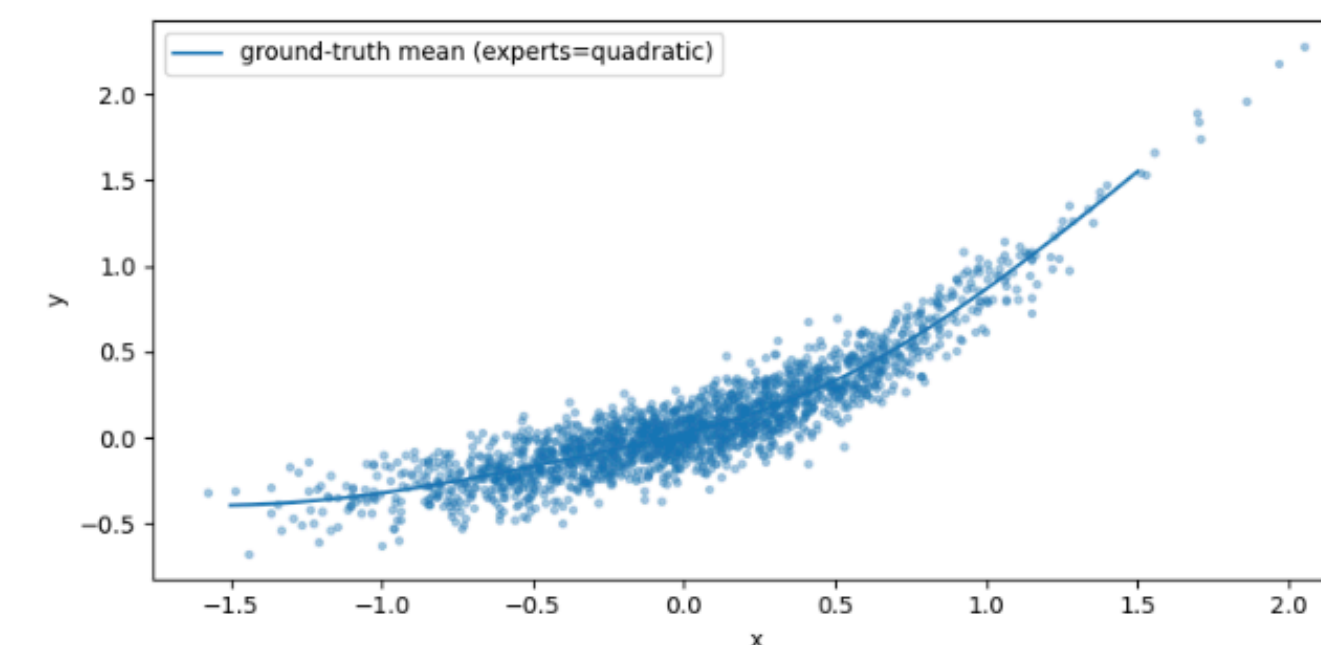
- **Regression setting of MoE (Nguyen et al. [16]):** Data $(Y_1, X_1), \dots, (Y_n, X_n) \in \mathbb{R} \times \mathbb{R}^d$ are generated from

$$Y_i = \sum_{j=1}^{k_0} \frac{\exp(\beta_{1j}^\top X + \beta_{0j})}{\sum_{j=1}^{k_0} \exp(\beta_{1j}^\top X + \beta_{0j})} \cdot h(X, \eta_j) + \epsilon_i$$

where $\epsilon_1, \dots, \epsilon_n$ are i.i.d. Gaussian noises



(a) Linear experts



(b) Quadratic experts

- **Main goal:** To estimate $G = \sum_{i=1}^{k_0} \exp(\beta_{0i}) \delta_{(\beta_{1i}, \eta_i)}$

Regression MoE: Least-square loss

- **Over-specified/ Over-parameterized setting:** As the number of true experts k_0 is unknown, we use mixture of k experts where k is given and $k > k_0$
- The least-square loss is then given by:

$$\widehat{G}_n \in \arg \max_{G' \in \mathcal{O}_k} \frac{1}{n} \sum_{i=1}^n (Y_i - p_{G'}(X_i))^2$$

where $\mathcal{O}_k = \{G' = \sum_{i=1}^{k'} \exp(\beta'_{0i}) \delta_{(\beta'_{1i}, \eta'_i)} : k' \leq k\},$

$$p_{G'}(X) = \sum_{i=1}^{k_0} \frac{\exp((\beta'_{1i})^\top X + \beta'_{0i})}{\sum_{j=1}^{k_0} \exp((\beta'_{1j})^\top X + \beta'_{0j})} \cdot h(X, \eta'_i)$$

Strongly Identifiable Experts

Definition (Strong Identifiability): An expert function $h(\cdot, \eta)$ is strongly identifiable if

$$\left\{ x^\nu \cdot \frac{\partial^{|\rho|} h}{\partial \eta^\rho}(x, \eta_j) : j \in [k], \nu \in \mathbb{N}^d, \rho \in \mathbb{N}^q, 0 \leq |\nu| + |\rho| \leq 2 \right\} \quad \text{are linearly independent}$$

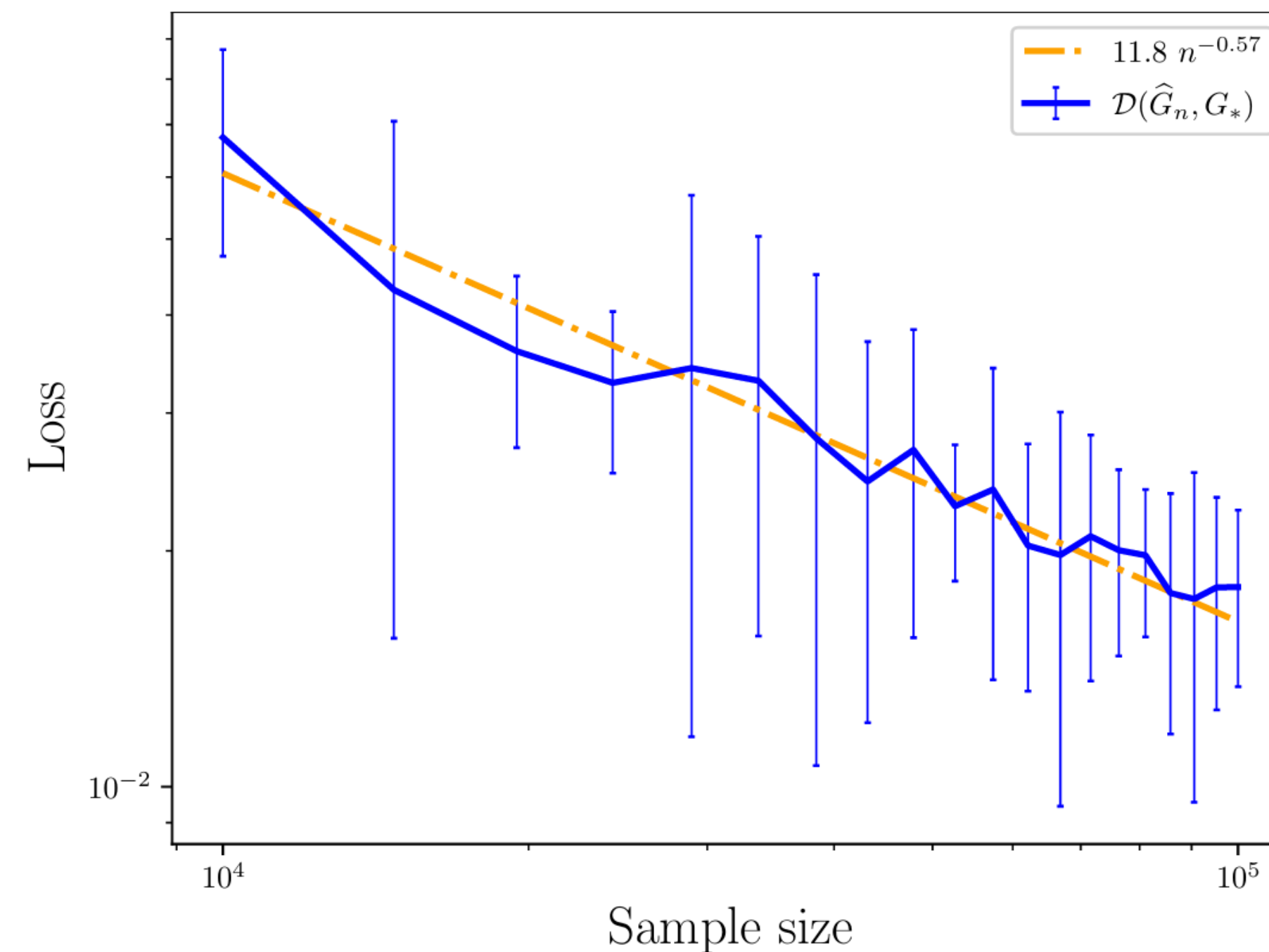
Example: The strong identifiability holds for feed-forward networks with activations like sigmoid, tanh, and GeLU and non-linear transformed input, i.e.,

$$h(x, a, b) = \sigma\left(a \frac{x}{\|x\|} + b\right)$$

Convergence Rates: Strong Identifiability

Informal Theorem: When the experts are strongly identifiable,

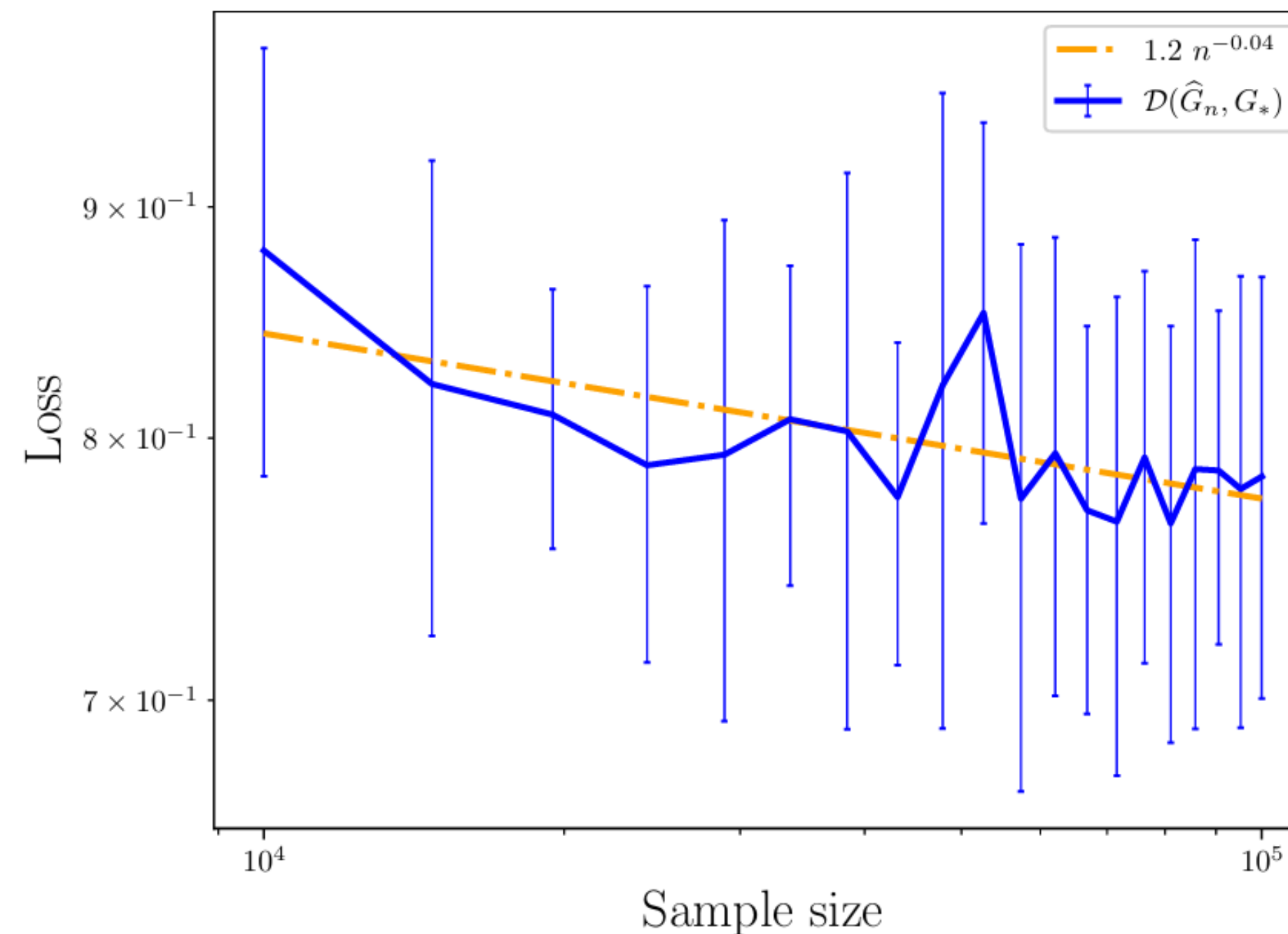
- Rates of softmax weights and expert parameters are respectively $n^{-1/4}$ and $n^{-1/2}$ for those in Voronoi cells $|\mathcal{A}_j| > 1$ and $|\mathcal{A}_j| = 1$
- Rates of softmax biases are $n^{-1/2}$



(b) Over-specified setting

Weakly Identifiable Experts - Ridge Experts

- (**Weak identifiability**) When the strong identifiability condition fails to hold, e.g., the experts take the form of ridge experts, namely, $\sigma(a^\top x + b)$ and σ is a polynomial function
- The rates of estimating experts and experts parameters can be as slow as **exponential (in minimax sense)**



Gaussian MoE versus Regression MoE

Model Type	Parameters a_j^*		Parameters b_j^*		Experts $(a_j^*)^\top x + b_j^*$	
	$j : \mathcal{A}_j^n = 1$	$j : \mathcal{A}_j^n > 1$	$j : \mathcal{A}_j^n = 1$	$j : \mathcal{A}_j^n > 1$	$j : \mathcal{A}_j^n = 1$	$j : \mathcal{A}_j^n > 1$
Probabilistic	$\mathcal{O}_P(n^{-1/2})$	$\mathcal{O}_P(n^{-1/\bar{r}_j})$	$\mathcal{O}_P(n^{-1/2})$	$\mathcal{O}_P(n^{-1/2\bar{r}_j})$	$\mathcal{O}_P(n^{-1/2})$	$\mathcal{O}_P(n^{-1/2\bar{r}_j})$
Deterministic	Slower than $\mathcal{O}_P(n^{-1/2r}), \forall r \geq 1$					

Summary of the rates under Probabilistic MoE (Gaussian MoE) versus
Deterministic MoE (Regression MoE)

Application of Theories: Improving Self-attention in Transformers

[18] Pedram Akbarian, Huy Nguyen, Xing Han, Nhat Ho. *Quadratic Gating Functions in Mixture of Experts: A Statistical Insight*. Under review

Background on Transformers

- Transformers were proposed by (Vaswani et al. [29]) in 2017
- The driving force behind the success of Transformers is the **self-attention mechanism**
- For each position, the self-attention calculates a weighted average of the feature representations of all other positions

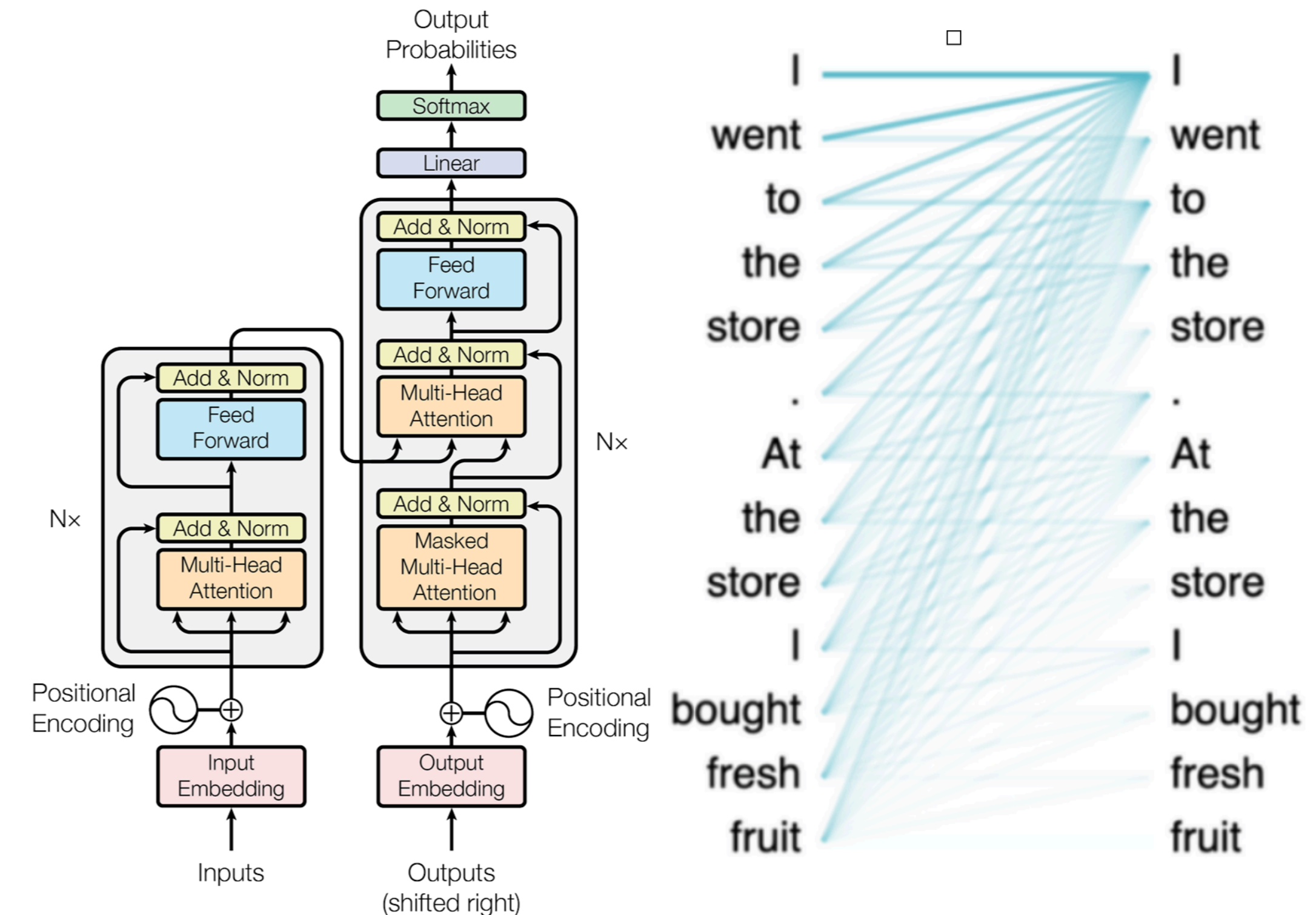


Figure 1: The Transformer - model architecture.

Background on Transformers

For each $Y \in \mathbb{R}^{N \times d}$, the ℓ -layer of Transformers takes the following form:

$$T_{\ell}(Y) = f_{\ell}(A_{\ell}(Y) + Y)$$

where f_{ℓ} is two-layer feedforward neural networks;

the **self-attention function**

$$A_{\ell}(Y) = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{d}}\right)V;$$

Query matrix $Q = YW_Q^{\top}$; **Key matrix** $K = YW_K^{\top}$; and **Value matrix** $V = YW_V^{\top}$

Embedding matrices $W_Q \in \mathbb{R}^{m \times d}$, $W_K \in \mathbb{R}^{m \times d}$, $W_V \in \mathbb{R}^{d \times d}$

We drop index ℓ
in query, key, and value
matrices for simplicity

Interpreting Transformers as Mixture of Experts

Claim: Each row of the self-attention is a mixture of **linear experts** with **quadratic gating function**

- Indeed, we can verify that

Quadratic gating function

Linear experts

$$A_i = \sum_{j=1}^N \left[\frac{\exp(\mathbf{Y}^\top B_{i,j} \mathbf{Y})}{\sum_{j'=1}^N \exp(\mathbf{Y}^\top B_{i,j'} \mathbf{Y})} \right] \cdot \mathbf{Y}^\top R_j$$

where A_i stands for the i -th row of the self-attention matrix ($1 \leq i \leq N$);

\mathbf{Y} is the vectorization of input $Y \in \mathbb{R}^{N \times d}$; $B_{i,j}$ and R_j are some functions of embedding matrices W_Q, W_K, W_V

Hardness of Learning the Self-Attention

- In Transformers, we learn the embedding matrices W_Q , W_K , W_V in the self-attention
 - Efficiently estimating these matrices implies good performance of Transformers
- By viewing the self-attention as mixture of experts, it is equivalent to learn the **heterogeneity of experts**

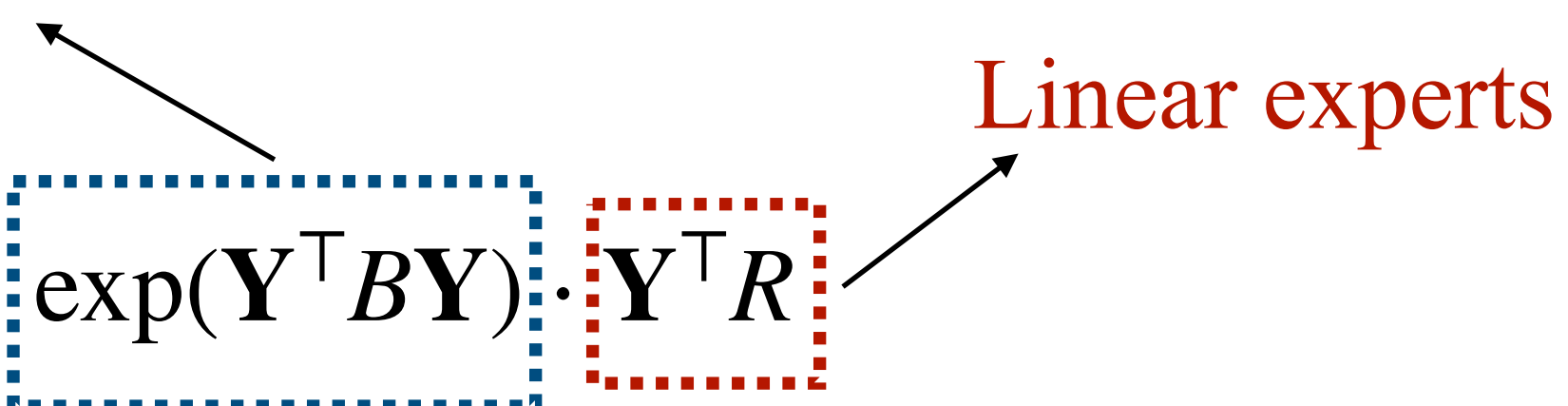
Informal Result (Pedram et al. [30]):

We need **exponential number of data** to guarantee **good estimation of linear experts** in quadratic gating mixture of experts

Hardness of Learning the Self-Attention

- Main reason: **Linear Experts**

Quadratic gating function

- We define $u(\mathbf{Y}, B, R) := \exp(\mathbf{Y}^\top B \mathbf{Y}) \cdot \mathbf{Y}^\top R$
- 

- **Fact:** $u(\mathbf{Y}, B, R)$ and its first order derivative with respect to R are *linearly dependent*

In high level, that linear dependence leads to **flat** landscape of loss function of Transformers
 (Optimization and Learning can become difficult!!!)

Transformers with Non-linear Values

- Proposed solution (Pedram et al. [30]): **Non-linear Experts**

Quadratic gating function

Non-linear experts

- Idea: $\bar{u}(\mathbf{Y}, B, R) := \boxed{\exp(\mathbf{Y}^\top B \mathbf{Y})} \cdot \boxed{\sigma(\mathbf{Y}^\top R)}$

where σ is a non-linear activation function

- New self-attention - Active-Attention:** $\bar{A}(Y) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right) \sigma(V)$
 (Computationally comparable to the original self-attention)

[30] Pedram Akbarian, Huy Nguyen, Xing Han, Nhat Ho. *Quadratic Gating Functions in Mixture of Experts: A Statistical Insight*. Under review

Transformers with Non-linear Values

Quadratic gating function

Non-linear experts

- Idea: $\bar{u}(\mathbf{Y}, B, R) := \boxed{\exp(\mathbf{Y}^\top B \mathbf{Y})} \cdot \boxed{\sigma(\mathbf{Y}^\top R)}$

where σ is a non-linear activation function

- Fact:** With sufficiently non-linear function σ (e.g., GELU, Tanh, etc.), $\bar{u}(\mathbf{Y}, B, R)$ and its first order derivatives are *linearly independent*
- That linear independence leads to **sharper** landscape of loss function (**Optimization and Learning become easier**)

Transformers with Non-linear Values

Quadratic gating function

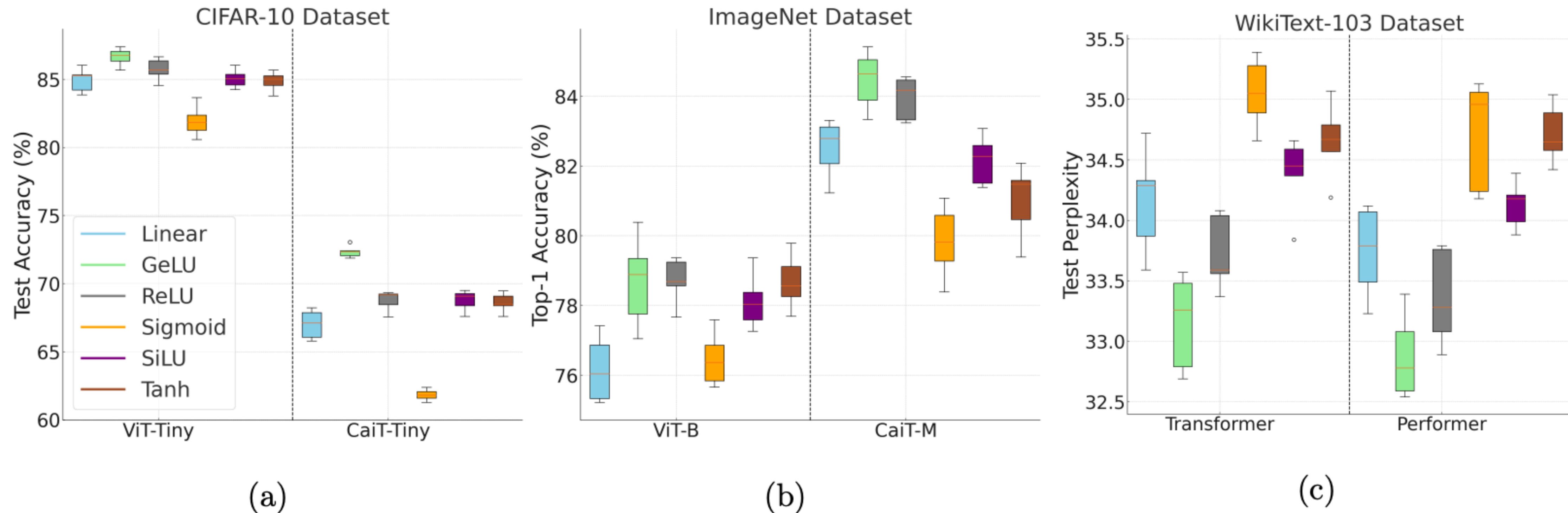
Non-linear experts

- Idea: $\bar{u}(\mathbf{Y}, B, R) := \boxed{\exp(\mathbf{Y}^\top B \mathbf{Y})} \cdot \boxed{\sigma(\mathbf{Y}^\top R)}$

where σ is a non-linear activation function

Informal Result: Much less data are required to estimate the new self-attention (**Optimal Sample Efficiency**).

Transformers with Non-linear Values



Active-Attention versus Standard Attention on three benchmark datasets (a) CIFAR-10, (b) ImageNet, and (c) WikiText-103.

Transformers with Non-linear Values

Model \ Dataset		Weather	Traffic	Electricity	Illness	ETTh1	ETTh2	ETTm1	ETTm2
PatchTST	<i>Linear</i>	<i>0.197</i>	<i>0.383</i>	<i>0.152</i>	<i>1.474</i>	<i>0.414</i>	<i>0.338</i>	<i>0.331</i>	<i>0.220</i>
	GELU	0.195	0.382	0.149	<u>1.520</u>	0.413	0.337	0.332	0.221
	ReLU	0.196	<u>0.380</u>	0.150	1.551	0.413	0.336	0.331	0.218
	Sigmoid	<u>0.192</u>	0.386	<u>0.146</u>	1.613	<u>0.411</u>	0.325	<u>0.328</u>	<u>0.216</u>
	SiLU	0.196	0.381	0.149	1.559	0.413	0.337	0.333	0.221
	Tanh	0.187	0.375	0.141	1.447	0.410	<u>0.329</u>	0.325	0.212
Transformer	<i>Linear</i>	<i>0.835</i>	<i>0.748</i>	<i>0.296</i>	<i>4.832</i>	<i>1.328</i>	<i>1.152</i>	<i>1.138</i>	<i>1.389</i>
	GELU	<u>0.804</u>	0.726	0.302	4.129	<u>1.269</u>	1.134	1.134	1.353
	ReLU	0.839	0.735	<u>0.272</u>	4.224	1.314	<u>1.102</u>	1.116	1.382
	Sigmoid	0.811	0.714	0.278	4.972	1.285	1.086	1.132	<u>1.357</u>
	SiLU	0.823	0.756	0.293	4.535	1.334	1.157	1.153	1.379
	Tanh	0.797	<u>0.721</u>	0.269	<u>4.216</u>	1.255	1.114	<u>1.125</u>	1.364

Active-Attention versus Standard Attention on time-series forecasting tasks.

Application of Theories: Improving LoRA via Reparameterization

[24] Tuan Truong, Chau Nguyen, Huy Nguyen, Minh Le, Trung Le, Nhat Ho. *RepLoRA: Reparameterizing low-rank adaptation via the perspective of mixture of experts*. ICML, 2025

[25] Nghiem Diep, Dung Le, Tuan Truong, Tan Dinh, Huy Nguyen, Nhat Ho. *HoRA: Cross-head low-rank adaptation with joint hypernetworks*. Under review

[26] Nghiem Tuong Diep, Hien Dang, Tuan Truong, Tan Dinh, Huy Nguyen, Nhat Ho. *DoRAN: Stabilizing weight-decomposed low-rank adaptation via noise injection and auxiliary networks*. Under review

Background on LoRA

- LoRA was proposed by (Hu et al. [19]) in 2021 as a parameter-efficient fine-tuning (PEFT) method
- Given a pre-trained weight matrix $W_0 \in \mathbb{R}^{m \times n}$, the output is

$$Y = W_0 X + B A X$$

where $B \in \mathbb{R}^{m \times r}$, $A \in \mathbb{R}^{r \times n}$ with $r \ll \min\{m, n\}$

- During training, W_0 is fixed and A, B are updated

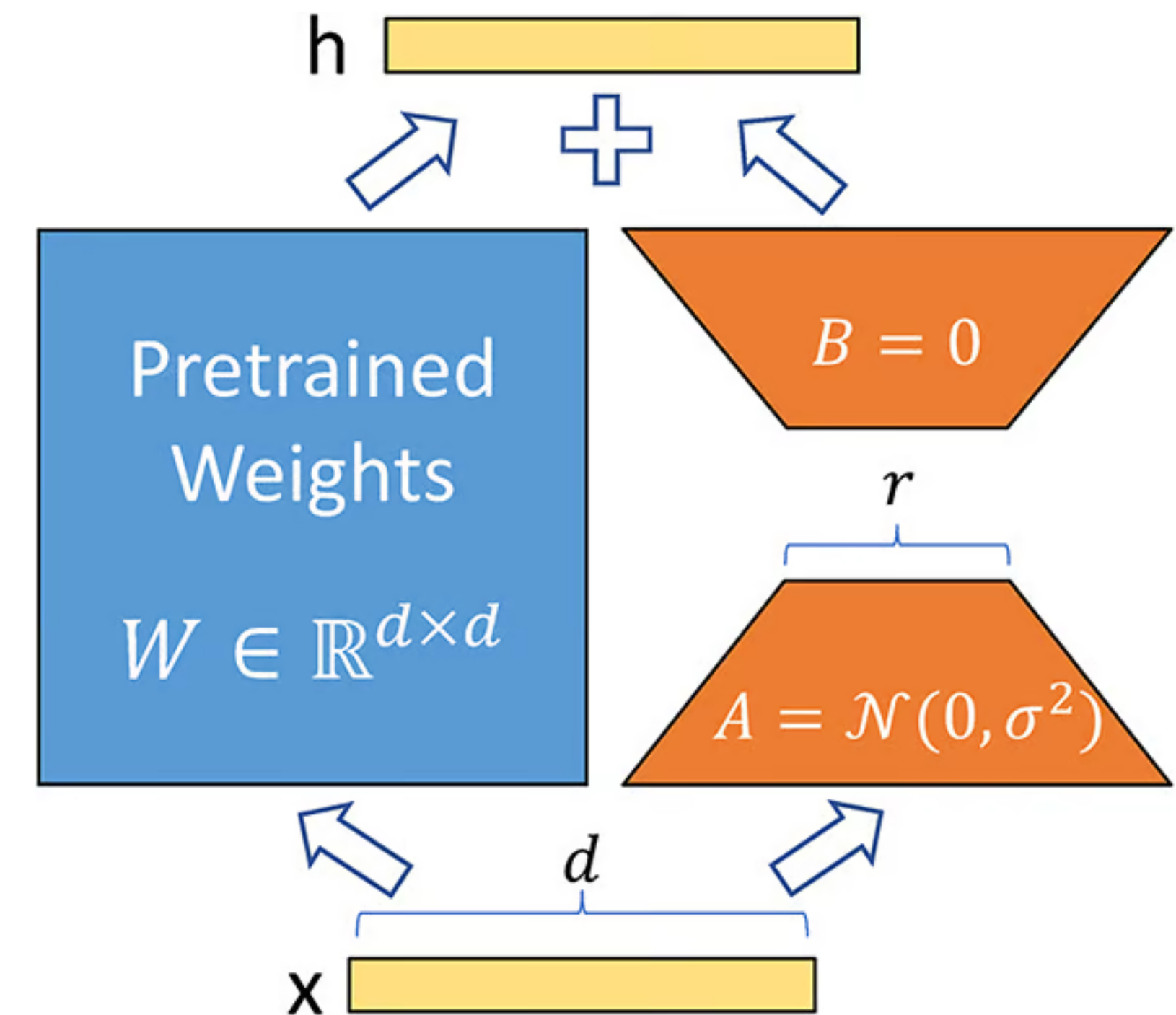


Illustration of LoRA (Hu et al. [16])

LoRA on Pretrained Transformers

For each $X \in \mathbb{R}^{N \times d}$, the self-attention of pre-trained Transformers with LoRA takes the following form:

$$f_{\text{LoRA}}(X) = \text{softmax} \left(\frac{(XW_Q + XB_Q A_Q)W_K^\top X^\top}{\sqrt{d}} \right) (XW_V + XB_V A_V)$$

Where W_Q, W_K, W_V are pretrained matrices;

A_Q, B_Q, A_V, B_V are updated matrices

LoRA as Mixture of Experts

Claim: Each row of the self-attention with LoRA is a mixture of **linear experts** with **quadratic gating function**

- Indeed, we can verify that

$$M_i = \sum_{j=1}^N \frac{\exp(\mathbf{X}^\top C_{i,j} \mathbf{X} + \mathbf{X}^\top B_Q A_Q D_j \mathbf{X})}{\sum_{j'=1}^N \exp(\mathbf{X}^\top C_{i,j'} \mathbf{X} + \mathbf{X}^\top B_Q A_Q D_{j'} \mathbf{X})} \cdot (W_V + B_V A_V)^\top R_j \mathbf{X}$$

where M_i stands for the i -th row of the self-attention matrix with LoRA ($1 \leq i \leq N$);

\mathbf{X} is the vectorization of input $Y \in \mathbb{R}^{N \times d}$;

$C_{i,j}$, D_j , and R_j are some fixed matrices;

Without Reparameterization among Matrices

- We would like to estimate the low-rank matrices B_Q, A_Q, B_V, A_V

Informal Result (Truong et al. [20]):

We need **exponential number of data** to estimate the low-rank matrices B_Q, A_Q, B_V, A_V

- **Main reason:** The low-rank matrices B_Q, A_Q and B_V, A_V are estimated **separately**
 - The MoE and its first order derivative with respect to these matrices are linearly dependent
 - Flat landscape and hard estimation

[20] Tuan Truong*, Chau Nguyen*, Huy Nguyen*, Minh Le, Trung Le, Nhat Ho. *RepLoRA: Reparameterizing low-rank adaptation via the perspective of mixture of experts*. ICML, 2025

RepLoRA: Reparameterization among Matrices

- Reparameterization (Shared structures) among low-rank matrices: We consider


$$A_Q = A_V = \varphi_1(A) \quad \text{and} \quad B_Q = B_V = \varphi_2(B)$$

Informal Result (Truong et al. [20]): With sufficient conditions on the functions φ_1, φ_2 , much less data are required to estimate the matrices A and B (**Optimal Sample Efficiency**).

RepLoRA: Reparameterization among Matrices

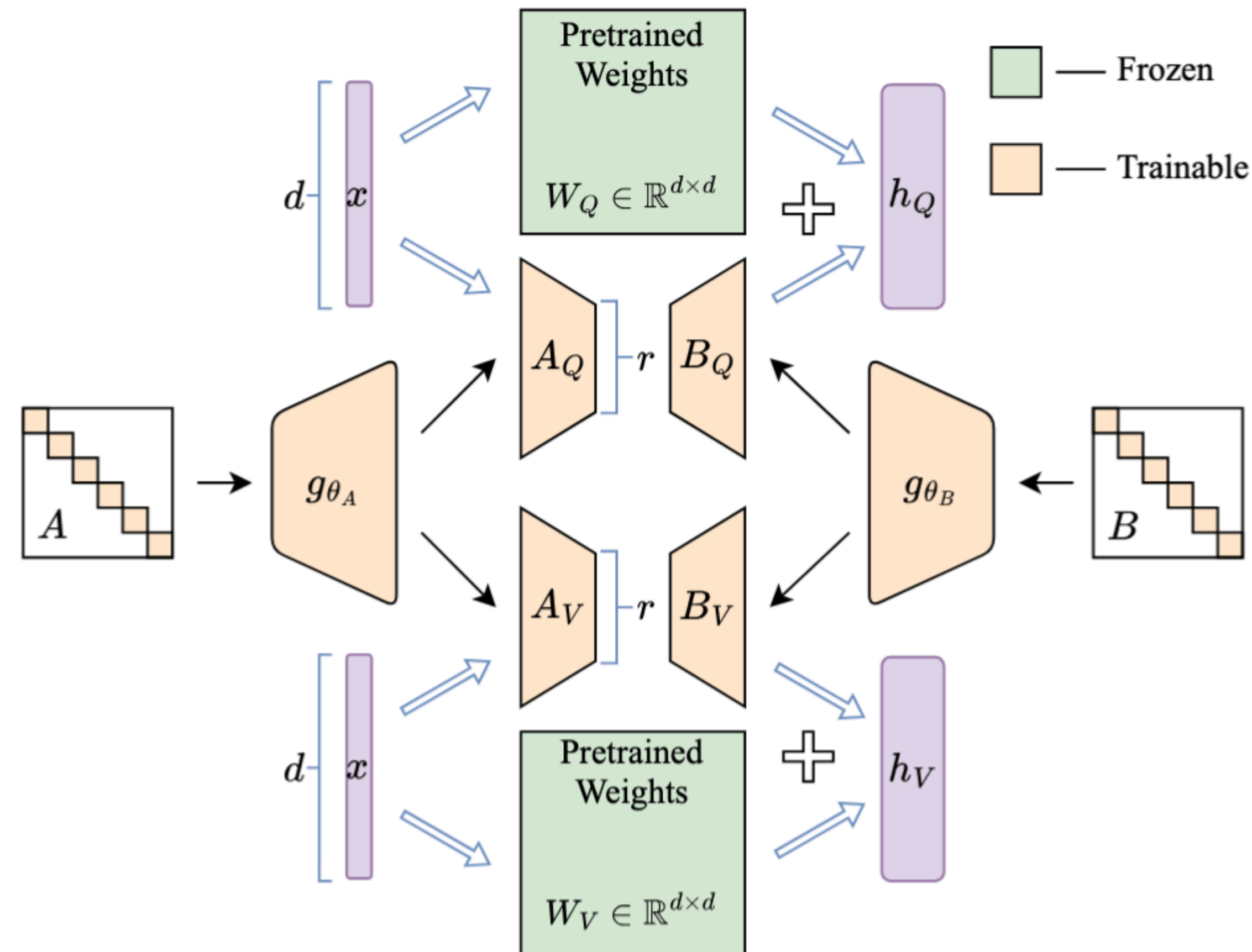
Quadratic gating function

Non-linear experts

- 
Idea: $\tilde{u}(\mathbf{X}, A, B) := \exp(\mathbf{X}^\top \varphi_2(B) \varphi_1(A) \mathbf{X}) \cdot \varphi_2(B) \varphi_1(A) \mathbf{X}$
- With sufficient conditions on φ_1 and φ_2 , $\tilde{u}(\mathbf{X}, A, B)$ and its first order derivative are linearly independent
- It leads to faster rates of estimating the matrices A and B

RepLoRA: Experiments

- We consider two-layer MLPs for the reparametrization among low-rank matrices in RepLoRA
- The shared matrices are chosen to be diagonal matrices



RepLoRA: Experiments

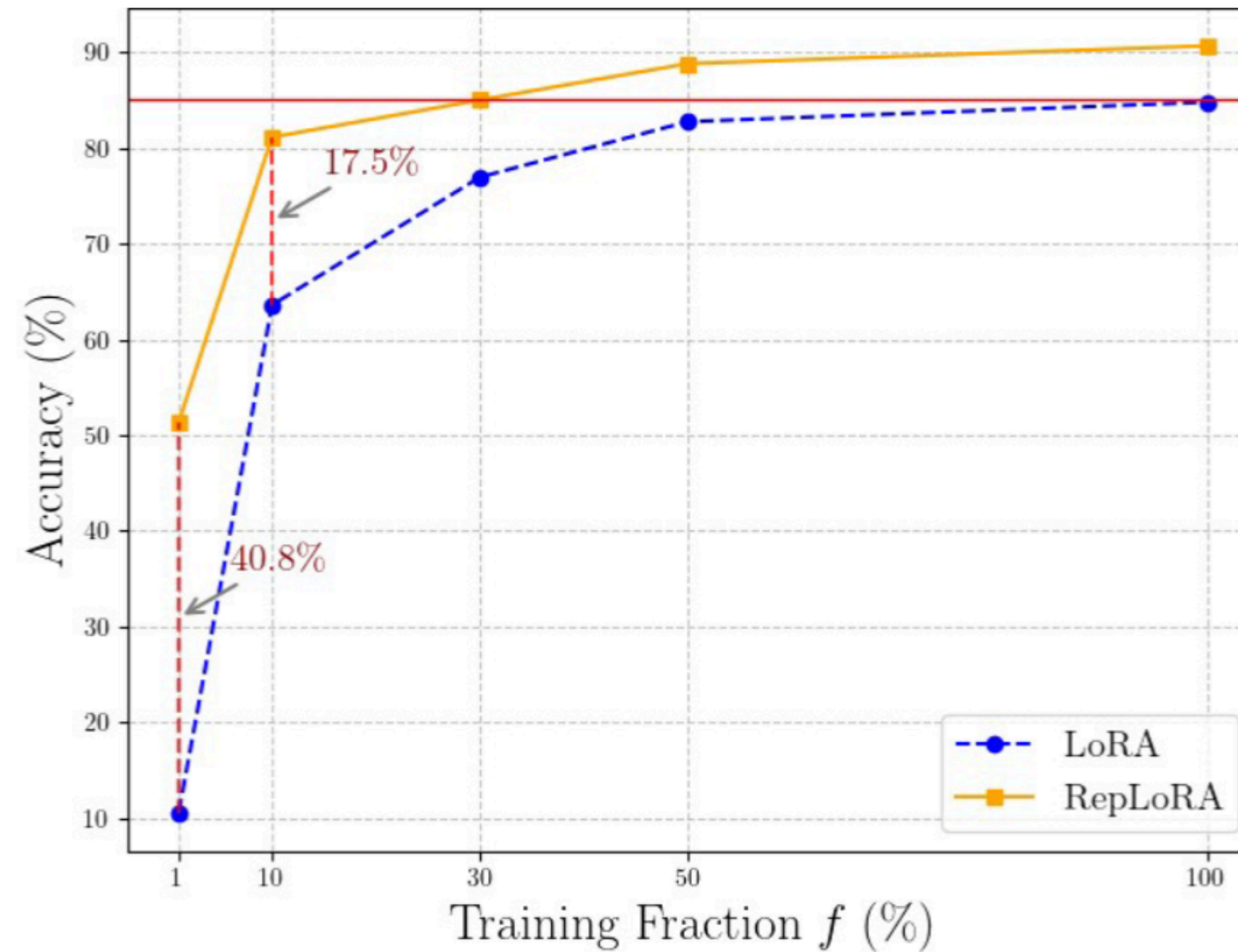


Figure 2: Sample Efficiency on FGVC Datasets. RepLoRA not only outperforms LoRA consistently but also achieves LoRA performance on a full dataset with only $f = 30\%$ training fraction.

RepLoRA: Experiments

Table 1: Top-1 Accuracy and PPT on commonsense datasets. The accuracies are reported with LLaMA-7B and LLaMA-13B.

Model	Method	#Params (%)	BoolQ	PIQA	SIQA	HellaSwag	WinoGrande	ARC-e	ARC-c	OBQA	AVG	PPT
ChatGPT	-	-	73.1	85.4	68.5	78.5	66.1	89.8	79.9	74.8	77.0	-
LLaMA-7B	Prefix	0.11	64.3	76.8	73.9	42.1	72.1	72.9	54	60.6	64.6	0.83
	LoRA	0.83	67.2	79.4	76.6	78.3	78.4	77.1	61.5	74.2	74.1	1.70
	Adapter	0.99	63.0	79.2	76.3	67.9	75.7	74.5	57.1	72.4	70.8	1.74
	RepLoRA	1.01	71.8	84.1	79.3	85.2	83.3	82.4	66.2	81.2	79.1	1.96
LLaMA-13B	Prefix	0.03	65.3	75.4	72.1	55.2	68.6	79.5	62.9	68	68.4	0.79
	LoRA	0.67	71.7	82.4	79.6	90.4	83.6	83.1	68.5	82.1	80.2	2.15
	Adapter	0.80	71.8	83.0	79.2	88.1	82.4	82.5	67.3	81.8	79.5	1.80
	RepLoRA	0.99	73.1	85.2	84.7	91.1	85.9	84.7	73.4	85.6	82.9	2.60

RepLoRA: Experiments

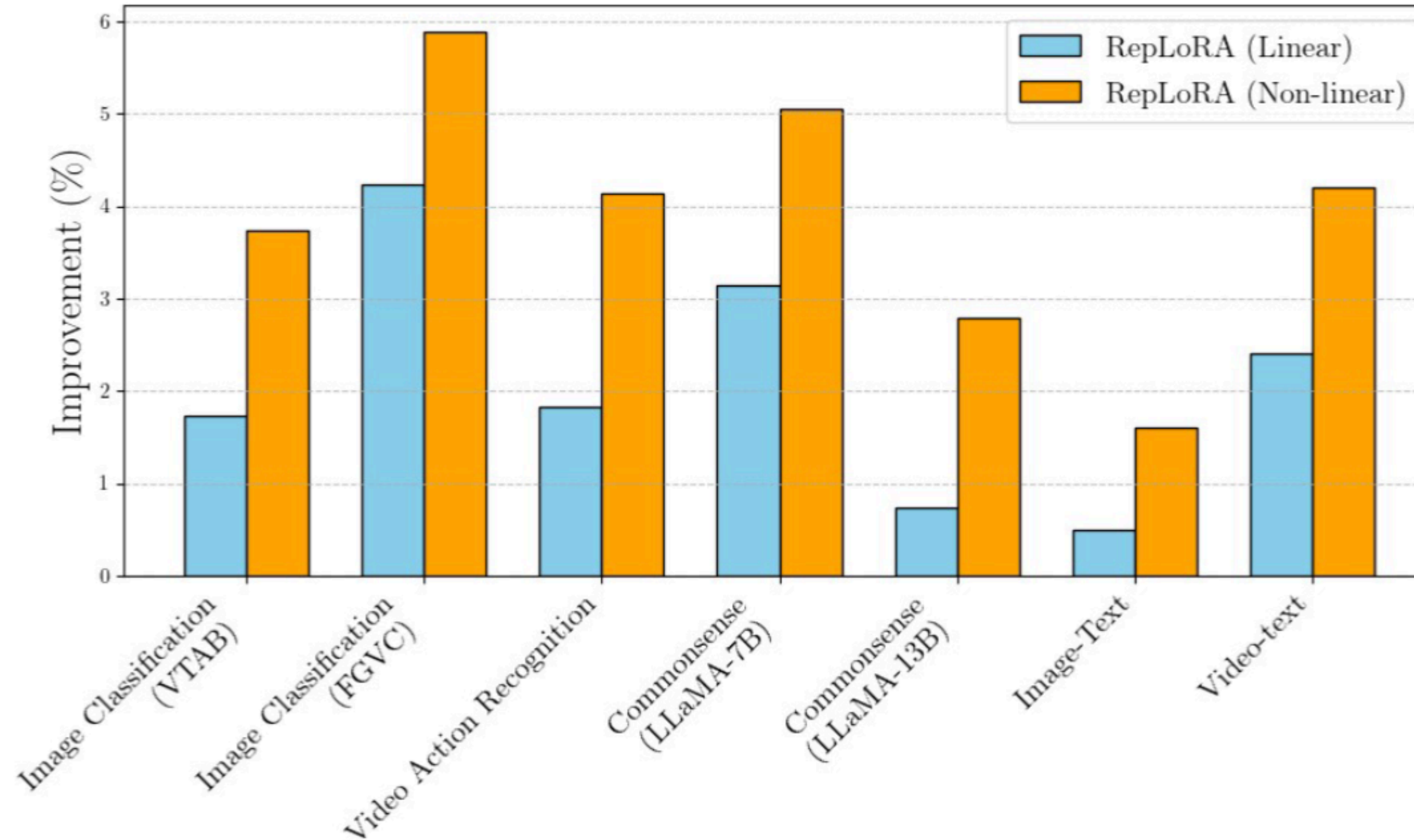
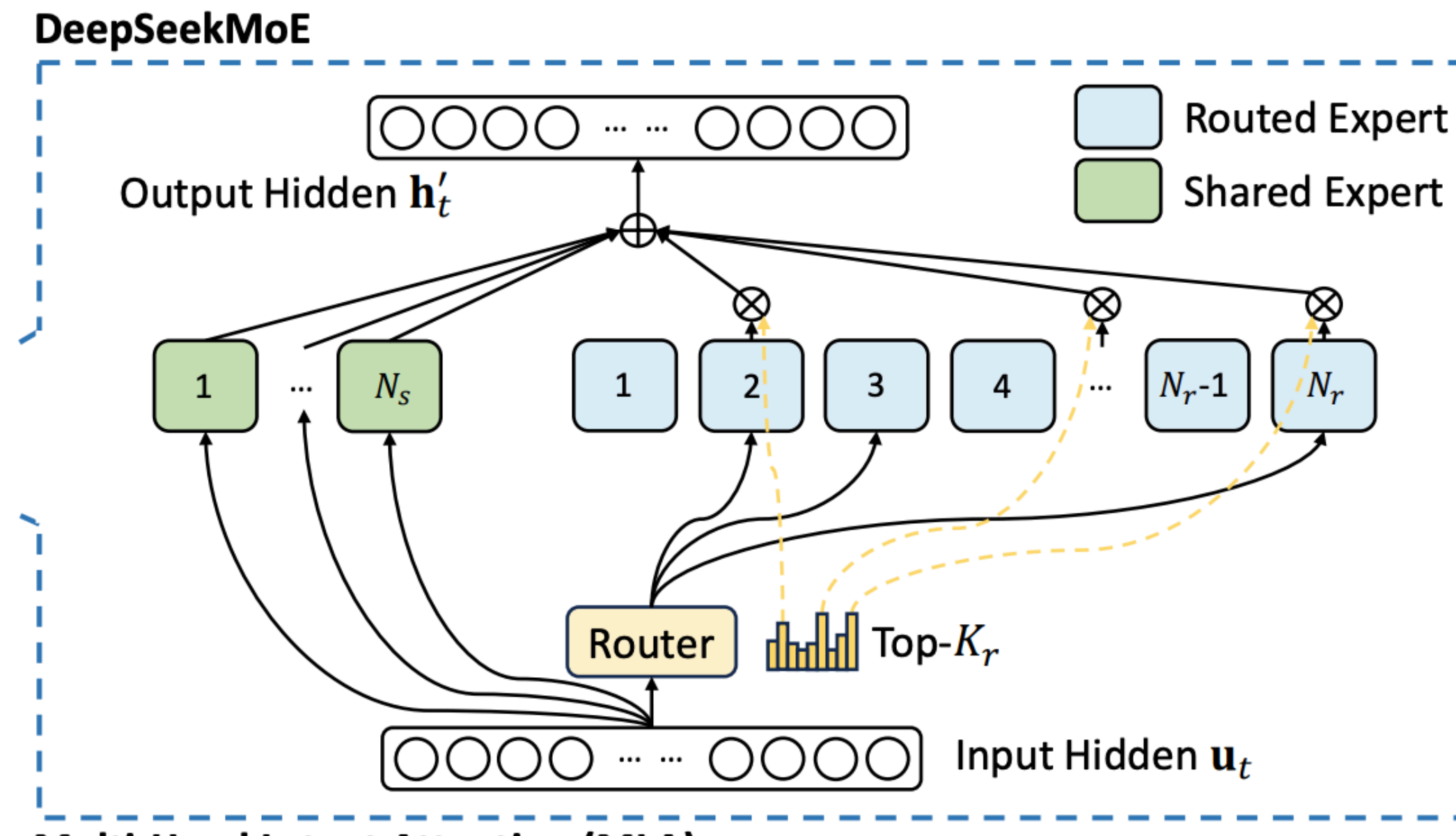


Figure 3: Performance improvements over LoRA. RepLoRA outperforms LoRA across all domains, with non-linear reparameterization substantially surpassing its linear counterpart.

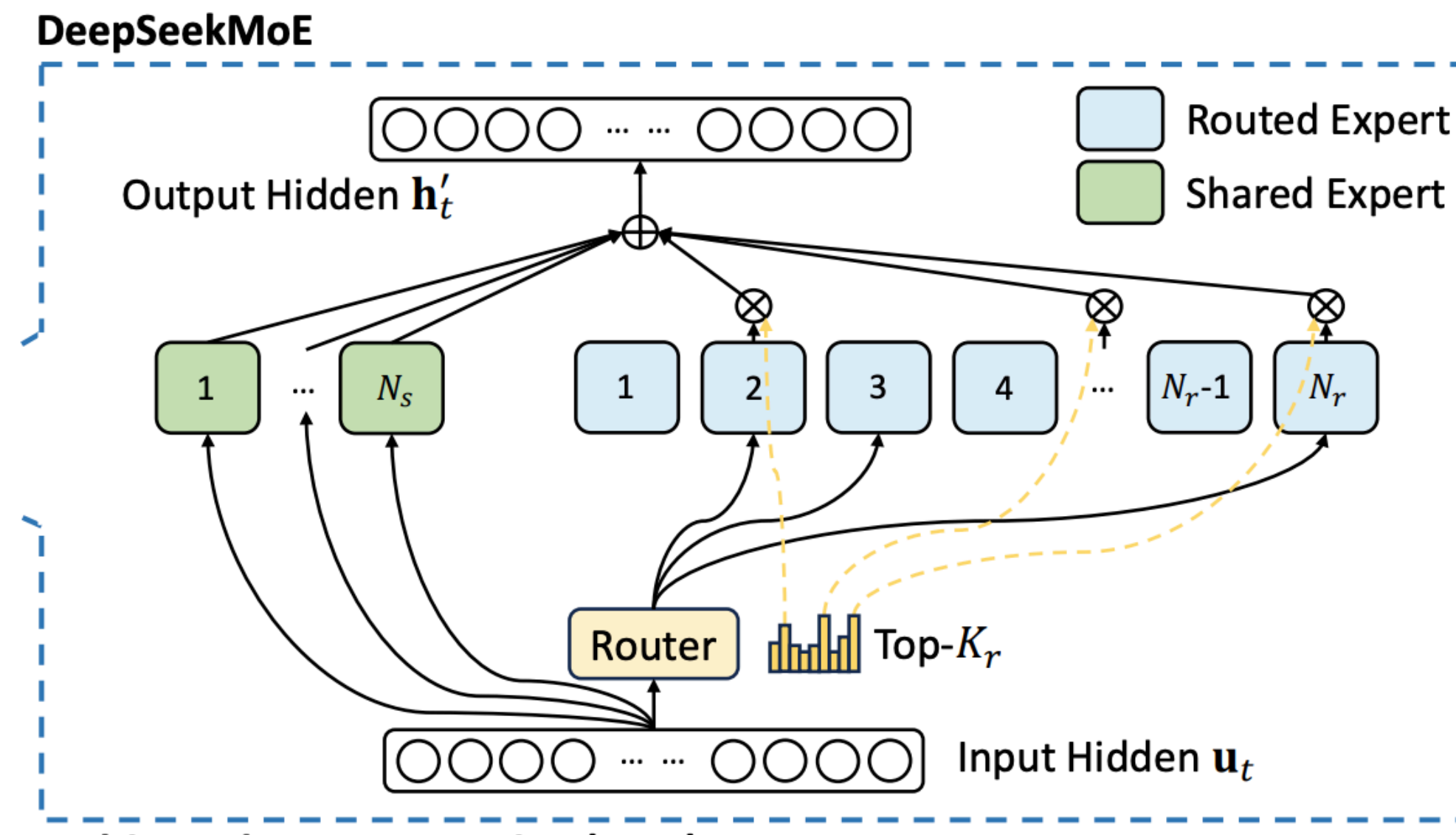
On DeepSeekMoE: Statistical Benefits of Shared Experts and Normalized Sigmoid Gating

DeepSeekMoE



- In response to these issues, **DeepSeekMoE** proposes
 - (i) Employing **shared expert strategy**
 - (ii) Replacing softmax gating with **normalized sigmoid gating**

Shared Expert Strategy



- **Shared experts:** are always activated to learn **general knowledge**
 - **Routed experts:** a few of them are activated per input to learn **specific knowledge**
- **Improve expert specialization**

Gaussian MoE

- Given a random sample $(X_1, Y_1), \dots, (X_n, Y_n) \in \mathbb{R}^d \times \mathbb{R}$ from the conditional density

$$f_{G_1, G_2}(y | x) := \frac{1}{2} \sum_{i=1}^{k_1^*} \omega_i \pi(Y | h_1(x, \kappa_i), \tau_i) + \frac{1}{2} \sum_{i=1}^{k_2^*} \frac{\exp((\beta_{1i})^\top x + \beta_{0i})}{\sum_{j=1}^{k_2^*} \exp((\beta_{1j})^\top x + \beta_{0j})} \pi(y | h_2(x, \eta_i), \nu_i) .$$

where $G_1 := \sum_{i=1}^{k_1^*} \omega_i \delta_{(\kappa_i, \tau_i)}$ and $G_2 := \sum_{i=1}^{k_2^*} \exp(\beta_{0i}) \delta_{(\beta_{1i}, \eta_i, \nu_i)}$ are unknown mixing measures (not necessarily a probability measure)

- Known:** Gaussian family of distributions $\{\pi(\cdot | \mu, \nu)\}$ with mean μ and variance ν
- Unknown:**
 - Constant weights $\{\omega_i\}_{i=1}^{k_1^*}$ softmax weights $\{\beta_{1i}\}_{i=1}^{k_2^*}$ and biases $\{\beta_{0i}\}_{i=1}^{k_2^*}$
 - Shared expert parameters $\{\kappa_i\}_{i=1}^{k_1^*}$, routed expert parameters $\{\eta_i\}_{i=1}^{k_2^*}$, variances $\{\tau_i\}_{i=1}^{k_1^*}$, $\{\nu_i\}_{i=1}^{k_2^*}$
 - The numbers of shared experts k_1^* and routed experts k_2^*

Maximum Likelihood Estimation (MLE)

- **Main goal:** To estimate $G_1 := \sum_{i=1}^{k_1^*} \omega_i \delta_{(\kappa_i, \tau_i)}$ and $G_2 := \sum_{i=1}^{k_2^*} \exp(\beta_{0i}) \delta_{(\beta_{1i}, \eta_i, \nu_i)}$
- **Over-specified/ Over-parameterized setting:** As the numbers of true experts k_1^* and k_2^* unknown, we use mixture of k_1 shared experts and k_2 where k_1, k_2 are given and $k_1 > k_1^*, k_2 > k_2^*$
- The MLE is then given by:

$$(\hat{G}_1^n, \hat{G}_2^n) \in \arg \max_{(G'_1, G'_2) \in \mathcal{O}_{k_1, k_2}} \frac{1}{n} \sum_{i=1}^n \log(f_{G'_1, G'_2}(Y_i | X_i))$$

$$\text{where } \mathcal{O}_{k_1, k_2} = \left\{ G'_1 = \sum_{i=1}^{k'_1} \omega'_i \delta_{(\kappa'_i, \tau'_i)} : k'_1 \leq k_1 \right\} \times \left\{ G'_2 = \sum_{i=1}^{k'_2} \exp(\beta'_{0i}) \delta_{(\beta'_{1i}, \eta'_i, \nu'_i)} : k'_2 \leq k_2 \right\}$$

From Density Estimation to Parameter Estimation

- **Density Estimation Rate:** Under the over-specified setting,

$$\mathbb{E}_X[d_H(f_{\widehat{G}_1^n}, \widehat{G}_2^n(\cdot | X), f_{G_1, G_2}(\cdot | X))] = O_P\left(\sqrt{\frac{\log n}{n}}\right)$$

where d_H stands for the Hellinger distance

- **From Density Estimation to Parameter Estimation:** We aim to establish

$$\mathbb{E}_X[d_H(f_{\widehat{G}_1^n}, \widehat{G}_2^n(\cdot | X), f_{G_1, G_2}(\cdot | X))] \gtrsim D(\widehat{G}_n, G)$$

where D is some divergence or loss

- **High level proof idea:** We use Taylor expansion to decompose the density discrepancy

Rate of Parameter Estimation: Challenges

- **Main goal:** To determine the divergence between $(\widehat{G}_1^n, \widehat{G}_2^n)$ and (G_1, G_2)
- **Two Challenges:**
 1. G_2 and \widehat{G}_2^n are not probability measures
 2. Complex interaction among parameters occurs when experts are of linear forms, e.g., $h_2(x, (\eta_1, \eta_0)) := \eta_1^\top x + \eta_0$.

$$\frac{\partial^2 u}{\partial \beta_1 \partial \eta_0} = \frac{\partial u}{\partial \eta_1}; \quad \frac{\partial^2 u}{\partial \eta_0^2} = 2 \frac{\partial u}{\partial \nu}$$

where $u(Y|X; \beta_1, \eta_1, \eta_0, \nu) = \exp(\beta_1^\top X) \cdot \pi(Y|\eta_1^\top x + \eta_0, \nu)$

- Therefore, **different parameters may have different rates**
- **Solution 1:** We resolve challenge 1 by developing novel Voronoi losses among mixing measures
- **Solution 2:** We resolve challenge 2 by establishing a strong identifiability condition on the expert functions.

Solution 2: Strongly Identifiable Experts

Definition 1 (Strong Identifiability). We say that expert functions $x \mapsto h_1(x, \kappa)$ and $x \mapsto h_2(x, \eta)$ are strongly identifiable if they are twice differentiable w.r.t κ and η , respectively, and if for any $k_1, k_2 \geq 1$ and distinct parameters $\kappa_1, \dots, \kappa_{k_1}$ and $\eta_1, \dots, \eta_{k_2}$, each of the sets

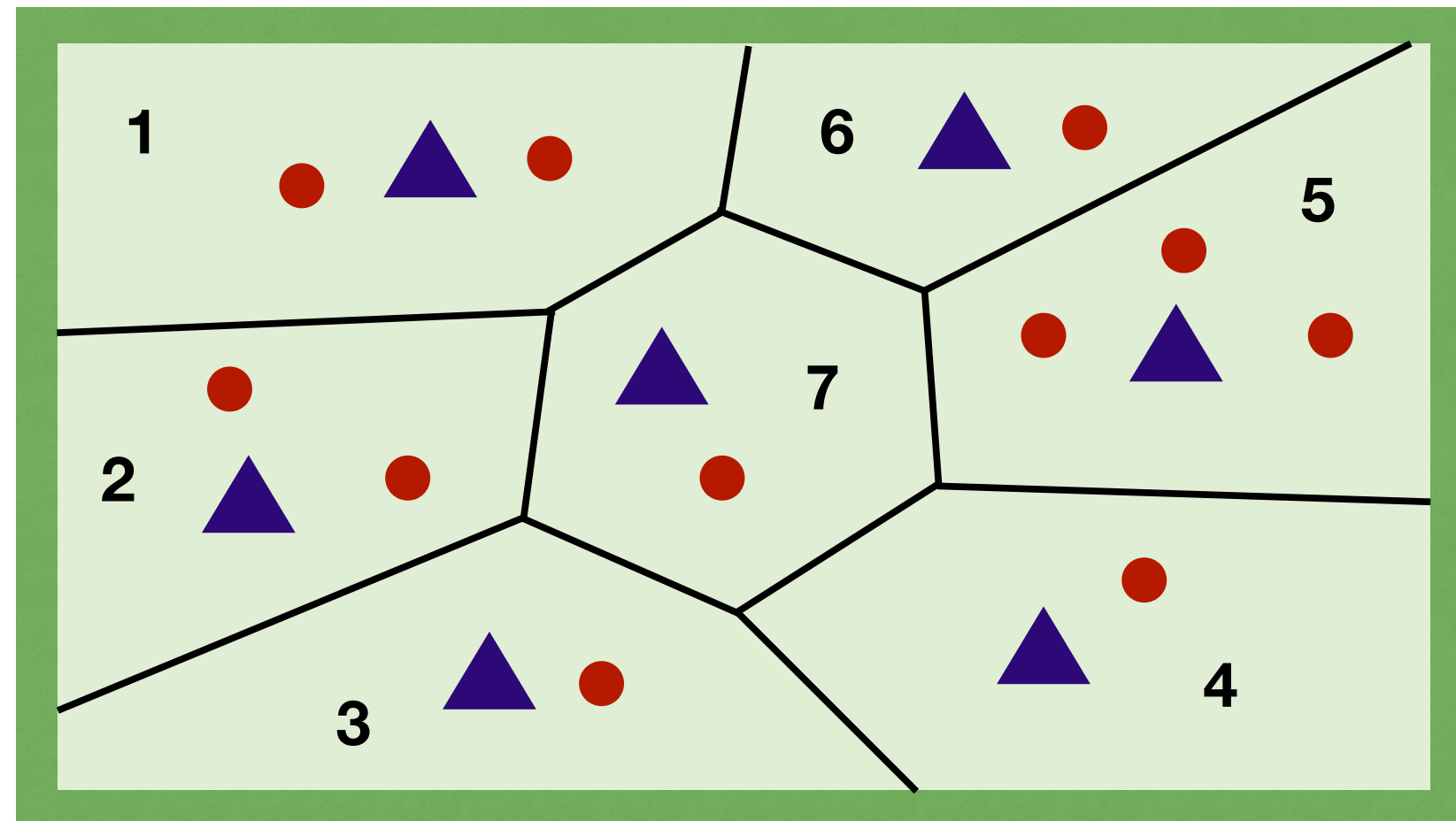
$$\left\{ \frac{\partial h_1}{\partial \kappa^{(u_1)}}(x, \kappa_i) : i \in [k_1], u_1 \in [d_1] \right\}, \left\{ \frac{\partial h_1}{\partial \kappa^{(u_1)}}(x, \kappa_i) \frac{\partial h_1}{\partial \kappa^{(v_1)}}(x, \kappa_i), 1 : i \in [k_1], u_1, v_1 \in [d_1] \right\},$$

$$\left\{ \frac{\partial h_2}{\partial \eta^{(u_2)}}(x, \eta_j), \frac{\partial^2 h_2}{\partial \eta^{(u_2)} \partial \eta^{(v_2)}}(x, \eta_j), x^{(u)} \frac{\partial h_2}{\partial \eta^{(v_2)}}(x, \eta_j) : j \in [k_2], u_2, v_2 \in [d_2], u \in [d] \right\}$$

consists of linearly independent functions (in x).

Examples. Two-layer FFNs $h_1(x, (\kappa_2, \kappa_1, \kappa_0)) := \kappa_2 \text{ReLU}(\kappa_1^\top x + \kappa_0)$ and $h_2(x, (\eta_2, \eta_1)) := \eta_2 \text{GELU}(\eta_1^\top x)$ are strongly identifiable. The same claim holds when replacing the ReLU function with other activation functions such as sigmoid and tanh. On the other hand, linear experts $h_1(x, (\kappa_1, \kappa_0)) := \kappa_1^\top x + \kappa_0$ and $h_2(x, (\eta_1, \eta_0)) := \eta_1^\top x + \eta_0$ fail to satisfy the strong identifiability condition because $\frac{\partial h_1}{\partial \kappa_0} \frac{\partial h_1}{\partial \kappa_0} = 1$ and $\frac{\partial h_2}{\partial \eta_1} = x \frac{\partial h_2}{\partial \eta_0}$ for all x .

Solution 1: Voronoi-based Loss



Blue triangles: True parameters

Red points: Parameters from $G'_1 \in \mathcal{O}_{k_1}, G'_2 \in \mathcal{O}_{k_2}$

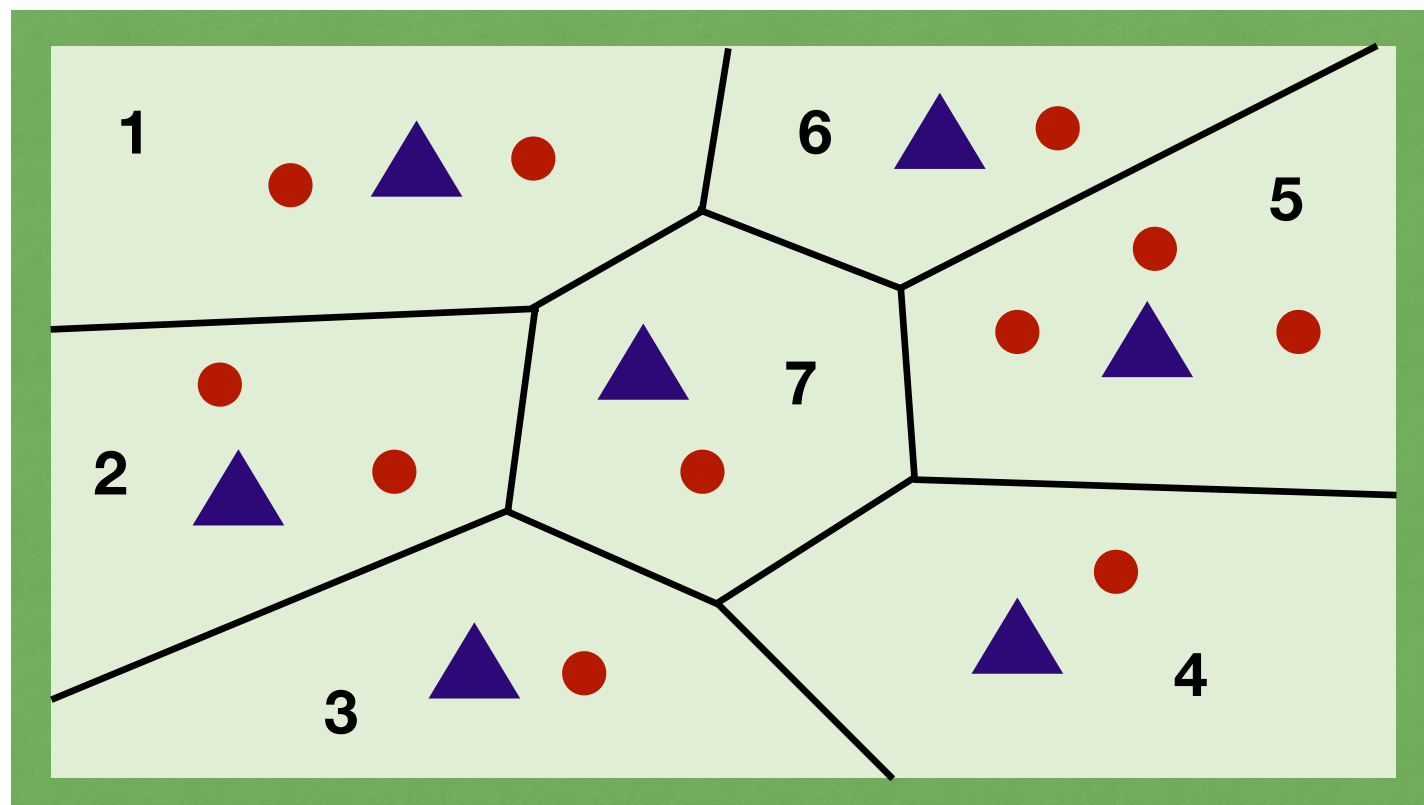
- **Voronoi cells:** For any $G'_1 \in \mathcal{O}_{k_1}$,

$$\mathcal{V}_{1,j} \equiv \mathcal{V}_{1,j}(G') := \{i \in \{1, 2, \dots, k_1\} : \|\xi'_i - \xi_j\| \leq \|\xi'_i - \xi_\ell\|, \forall \ell \neq j\}$$

$$\mathcal{V}_{2,j} \equiv \mathcal{V}_{2,j}(G') := \{i \in \{1, 2, \dots, k_2\} : \|\zeta'_i - \zeta_j\| \leq \|\zeta'_i - \zeta_\ell\|, \forall \ell \neq j\}$$

where $\xi'_i = (\kappa'_i, \tau'_i)$, $\xi_j = (\kappa_j, \tau_j)$ and $\zeta'_i = (\beta'_{1i}, \eta'_i, \nu'_i)$, $\zeta_j = (\beta_{1j}, \eta_j, \nu_j)$.

Voronoi-based Loss: For Strongly Identifiable Experts



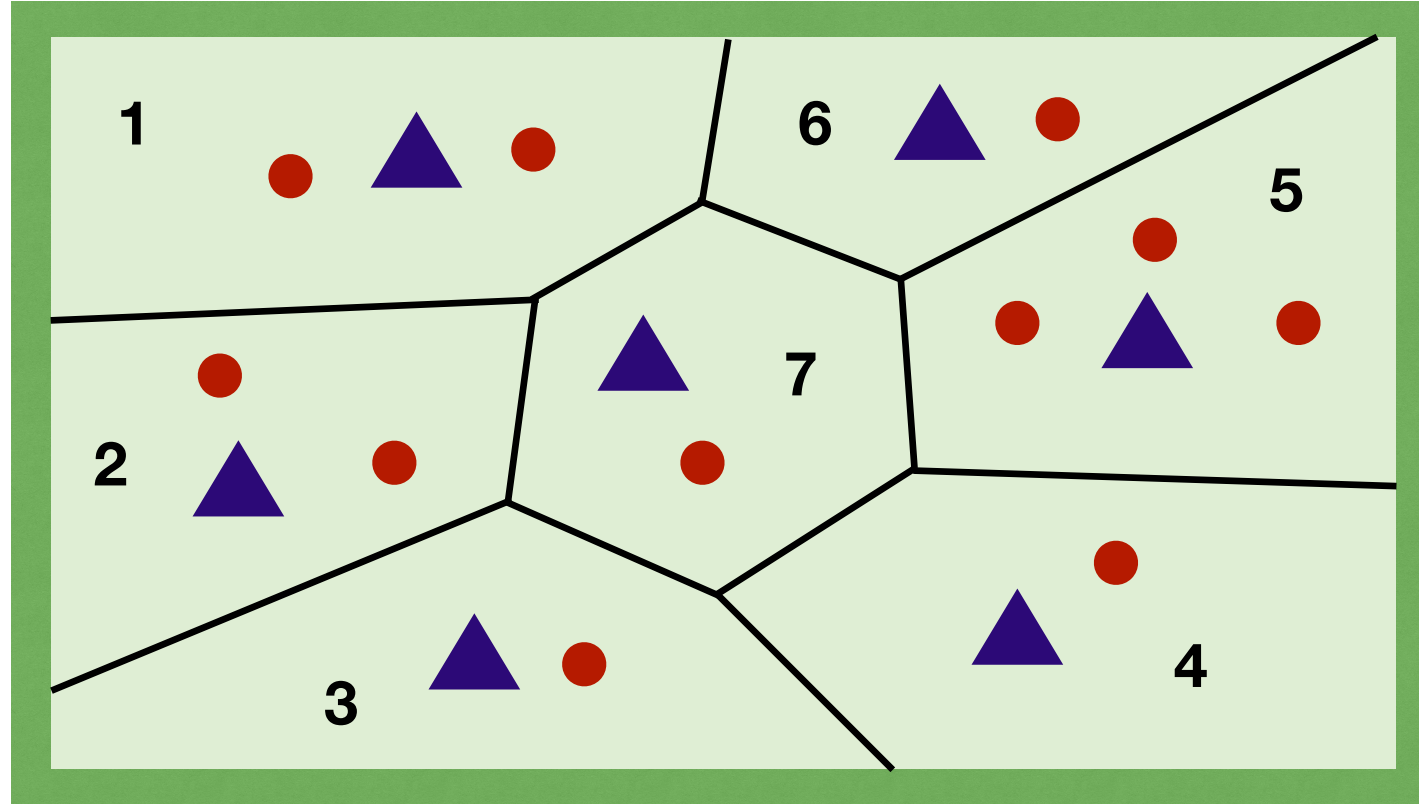
Blue triangles: True parameters

Red points: Parameters from $G' \in \mathcal{O}_k$

$$\begin{aligned} \mathcal{D}_1((G_1, G_2), (G_1^*, G_2^*)) &:= \sum_{j=1}^{k_1^*} \left| \sum_{i \in \mathcal{V}_{1,j}} \omega_i - \omega_j^* \right| + \sum_{j=1}^{k_2^*} \left| \sum_{i \in \mathcal{V}_{2,j}} \exp(\beta_{0i}) - \exp(\beta_{0j}^*) \right| \\ &+ \sum_{\substack{j \in [k_1^*], \\ |\mathcal{V}_{1,j}|=1}} \sum_{i \in \mathcal{V}_{1,j}} \omega_i (\|\Delta \kappa_{ij}\| + |\Delta \tau_{ij}|) + \sum_{\substack{j \in [k_2^*], \\ |\mathcal{V}_{2,j}|=1}} \sum_{i \in \mathcal{V}_{2,j}} \exp(\beta_{0i}) (\|\Delta \beta_{1ij}\| + \|\Delta \eta_{ij}\| + |\Delta \nu_{ij}|) \\ &+ \sum_{\substack{j \in [k_1^*], \\ |\mathcal{V}_{1,j}|>1}} \sum_{i \in \mathcal{V}_{1,j}} \omega_i (\|\Delta \kappa_{ij}\|^2 + |\Delta \tau_{ij}|^2) + \sum_{\substack{j \in [k_2^*], \\ |\mathcal{V}_{2,j}|>1}} \sum_{i \in \mathcal{V}_{2,j}} \exp(\beta_{0i}) (\|\Delta \beta_{1ij}\|^2 + \|\Delta \eta_{ij}\|^2 + |\Delta \nu_{ij}|^2), \end{aligned}$$

where we denote $\Delta p_{ij} := p_i' - p_j$.

Voronoi-based Loss: For Linear Experts



Blue triangles: True parameters

Red points: Parameters from $G' \in \mathcal{O}_k$

$$\begin{aligned}
 \mathcal{D}_2((G_1, G_2), (G_1^*, G_2^*)) &:= \sum_{j=1}^{k_1^*} \left| \sum_{i \in \mathcal{V}_{1,j}} \omega_i - \omega_j^* \right| + \sum_{j=1}^{k_2^*} \left| \sum_{i \in \mathcal{V}_{2,j}} \exp(\beta_{0i}) - \exp(\beta_{0j}^*) \right| \\
 &+ \sum_{\substack{j \in [k_1^*], \\ |\mathcal{V}_{1,j}|=1}} \sum_{i \in \mathcal{V}_{1,j}} \omega_i (\|\Delta \kappa_{1ij}\| + |\Delta \kappa_{0ij}| + |\Delta \tau_{ij}|) + \sum_{\substack{j \in [k_1^*], \\ |\mathcal{V}_{1,j}|>1}} \sum_{i \in \mathcal{V}_{1,j}} \omega_i (\|\Delta \kappa_{ij}\|^2 + |\Delta \kappa_{0ij}|^{r_{1,j}} + |\Delta \tau_{ij}|^{r_{1,j}/2}) \\
 &+ \sum_{j \in [k_2^*]: |\mathcal{V}_{2,j}|=1} \sum_{i \in \mathcal{V}_{2,j}} \exp(\beta_{0i}) (\|\Delta \beta_{1ij}\| + \|\Delta \eta_{1ij}\| + |\Delta \eta_{0ij}| + |\Delta \nu_{ij}|) \\
 &+ \sum_{j \in [k_2^*]: |\mathcal{V}_{2,j}|>1} \sum_{i \in \mathcal{V}_{2,j}} \exp(\beta_{0i}) (\|\Delta \beta_{1ij}\|^{r_{2,j}} + \|\Delta \eta_{1ij}\|^{r_{2,j}/2} + |\Delta \eta_{0ij}|^{r_{2,j}} + |\Delta \nu_{ij}|^{r_{2,j}/2}), \quad (6)
 \end{aligned}$$

where we denote $\Delta p_{ij} := p_i' - p_j$.

Parameter Estimation via Voronoi Loss

- **Lower bound:** For any $G' \in \mathcal{O}_k$, we can show that

$$\mathbb{E}_X[d_H(f_{G'_1, G'_2}(\cdot | X), f_{G_1, G_2}(\cdot | X))] \gtrsim \mathcal{D}_2((G'_1, G'_2), (G_1, G_2)),$$

where $\bar{r}_2 := (\bar{r}(|\mathcal{V}_{2,j}|), \dots, \bar{r}(|\mathcal{V}_{2,k_2}|))$ and $\bar{r}(m)$ is the smallest natural number r such that the following system of polynomial equations does not have any non-trivial solution:

$$\sum_{j=1}^m \sum_{(\alpha_1, \alpha_2, \alpha_3, \alpha_4) \in \mathcal{J}_{\ell_1, \ell_2}} \frac{p_{5j}^2 p_{1j}^{\alpha_1} p_{2j}^{\alpha_2} p_{3j}^{\alpha_3} p_{4j}^{\alpha_4}}{\alpha_1! \alpha_2! \alpha_3! \alpha_4!} = 0,$$

for any $(\ell_1, \ell_2) \in \mathbb{N}^d \times \mathbb{N}$ such that $1 \leq |\ell_1| + \ell_2 \leq r$, where

$$\mathcal{J}_{\ell_1, \ell_2} := \{(\alpha_1, \alpha_2, \alpha_3, \alpha_4) \in \mathbb{N}^d \times \mathbb{N}^d \times \mathbb{N} \times \mathbb{N} : \alpha_1 + \alpha_2 = \ell_1, |\alpha_2| + \alpha_3 + 2\alpha_4 = \ell_2\}$$

Connection to Algebraic Geometry

- $\bar{r}(m)$ is the smallest natural number r such that the following system of polynomial equations does not have any non-trivial solution:

$$\sum_{j=1}^m \sum_{(\alpha_1, \alpha_2, \alpha_3, \alpha_4) \in \mathcal{F}_{\ell_1, \ell_2}} \frac{p_{5j}^2 p_{1j}^{\alpha_1} p_{2j}^{\alpha_2} p_{3j}^{\alpha_3} p_{4j}^{\alpha_4}}{\alpha_1! \alpha_2! \alpha_3! \alpha_4!} = 0,$$

for any $(\ell_1, \ell_2) \in \mathbb{N}^d \times \mathbb{N}$ such that $1 \leq |\ell_1| + \ell_2 \leq r$, where

$$\mathcal{F}_{\ell_1, \ell_2} := \{(\alpha_1, \alpha_2, \alpha_3, \alpha_4) \in \mathbb{N}^d \times \mathbb{N}^d \times \mathbb{N} \times \mathbb{N} : \alpha_1 + \alpha_2 = \ell_1, |\alpha_2| + \alpha_3 + 2\alpha_4 = \ell_2\}$$

- **Some values of $\bar{r}(m)$:** When $m = 1$, $\bar{r}(m) = 4$;

When $m = 2$, $\bar{r}(m) = 6$;

- It is challenging to determine exact value of $\bar{r}(m)$ for general over-specified setting

Sample Efficiency of Shared Expert Strategy

DeepSeek-V2's MoE	ReLU FFN Experts	Linear Experts
Shared Experts	$\tilde{\mathcal{O}}_P(n^{-1/4})$	$\tilde{\mathcal{O}}_P(n^{-1/4})$
Routed Experts	$\tilde{\mathcal{O}}_P(n^{-1/4})$	$\tilde{\mathcal{O}}_P(n^{-1/r_2(\mathcal{V}_{2,j})})$

- **Without Shared Expert Strategy:** one may need $\mathcal{O}(n^{-1/12})$ many data points to achieve good estimation of experts
- **With Shared Expert Strategy:** one only needs $\mathcal{O}(n^{-1/4})$ many data points to guarantee good estimation of experts
 - **Shared expert strategy improves the model sample efficiency**

Normalized Sigmoid Gating

- DeepSeek-V3's MoE replaces the **Softmax gating** with the **Normalized sigmoid gating**

$$y = \sum_{i=1}^N \frac{\exp(\omega_i^\top x + \beta_i)}{\sum_{j=1}^N \exp(\omega_j^\top x + \beta_j)} \cdot E(x, \eta_i) \longrightarrow y = \sum_{i=1}^N \frac{\sigma(\omega_i^\top x + \beta_i)}{\sum_{j=1}^N \sigma(\omega_j^\top x + \beta_j)} \cdot E(x, \eta_i)$$



- **Expert utilization:** **Softmax gating** causes **low expert utilization**, i.e., a few experts are activated more often than others. **Normalized sigmoid gating** improves **expert utilization**

Gaussian MoE

- Given a random sample $(X_1, Y_1), \dots, (X_n, Y_n) \in \mathbb{R}^d \times \mathbb{R}$ from the conditional density

$$g_{G_1, G_2}(y | x) := \frac{1}{2} \sum_{i=1}^{k_1^*} \omega_i \pi(y | h_1(x, \kappa_i), \tau_i) + \frac{1}{2} \sum_{i=1}^{k_2^*} \frac{\sigma((\beta_{1i})^\top x + \beta_{0i})}{\sum_{j=1}^{k_2^*} \sigma((\beta_{1j})^\top x + \beta_{0j})} \cdot \pi(y | h_2(x, \eta_i), \nu_i),$$

where $z \in \mathbb{R} \mapsto \sigma(z) := \frac{1}{1 + \exp(-z)}$ is the sigmoid function.

Normalized Sigmoid Gating

	ReLU FFN Experts	Linear Experts
DeepSeek-V2's MoE	$\tilde{\mathcal{O}}_P(n^{-1/4})$	$\tilde{\mathcal{O}}_P(n^{-1/r_2(\mathcal{V}_{2,j})})$
DeepSeek-V3's MoE	$\tilde{\mathcal{O}}_P(n^{-1/2})$	$\tilde{\mathcal{O}}_P(n^{-1/2})$

- **When using the softmax gating**, the rates for estimating experts are significantly slow as they depend on the number of fitted experts
 - **When using the normalized sigmoid gating**, the expert estimation rates remains constant at the order of $\mathcal{O}(n^{-1/2})$
- **Normalized sigmoid gating improves the model sample efficiency**

Language Modeling

- Experiments on language modeling using Switch Transformer in two scales: small (**158M parameters on 6.5B tokens**) and large (**679M parameters on 26.2B tokens**)
- Models are configured with 66 total experts, utilizing **top-8 expert routing** v.s. a **top-6 plus 2 shared experts** routing scheme

Table 2: Performance comparisons of different Sparse Mixture of Experts (SMoE) models on subsets of the SlimPajama dataset using a small-scale model with 158M parameters and large-scale model with 679M parameters. (SMoE-SG refers to SMoE Sigmoid Gating). PPL indicates the perplexity score.

	Small Models (158M)				Large Models (679M)			
	SMoE	DeepSeek-V3	DeepSeek-V2	SMoE-SG	SMoE	DeepSeek-V3	DeepSeek-V2	SMoE-SG
PPL ↓	13.63	13.42	<u>13.49</u>	13.61	9.51	<u>9.49</u>	9.52	9.46
LAMBADA	25.27%	25.49%	25.29%	25.43%	37.13%	36.88%	37.11%	37.56%
BLiMP	77.71%	77.20%	77.37%	77.38%	80.47%	81.28%	80.98%	81.08%
CBT	84.18%	84.40%	84.33%	84.23%	89.83%	89.65%	89.93%	89.57%
HellaSwag	29.43%	29.38%	29.38%	29.13%	37.49%	37.32%	37.14%	37.52%
PIQA	57.94%	59.14%	60.17%	58.92%	64.36%	65.72%	64.36%	64.91%
ARC-Challenge	21.20%	21.63%	20.52%	21.37%	23.09%	23.95%	24.21%	23.09%
RACE	30.11%	30.60%	31.02%	31.05%	33.03%	33.12%	33.17%	32.68%
SIQA	35.62%	35.57%	34.90%	34.90%	37.41%	38.59%	36.95%	37.67%
CommonSenseQA	24.65%	25.47%	24.98%	24.90%	26.54%	28.09%	27.35%	28.50%
Average	42.90%	43.21%	<u>43.11%</u>	43.04%	47.71%	48.29%	47.91%	<u>48.06%</u>

Language Modeling

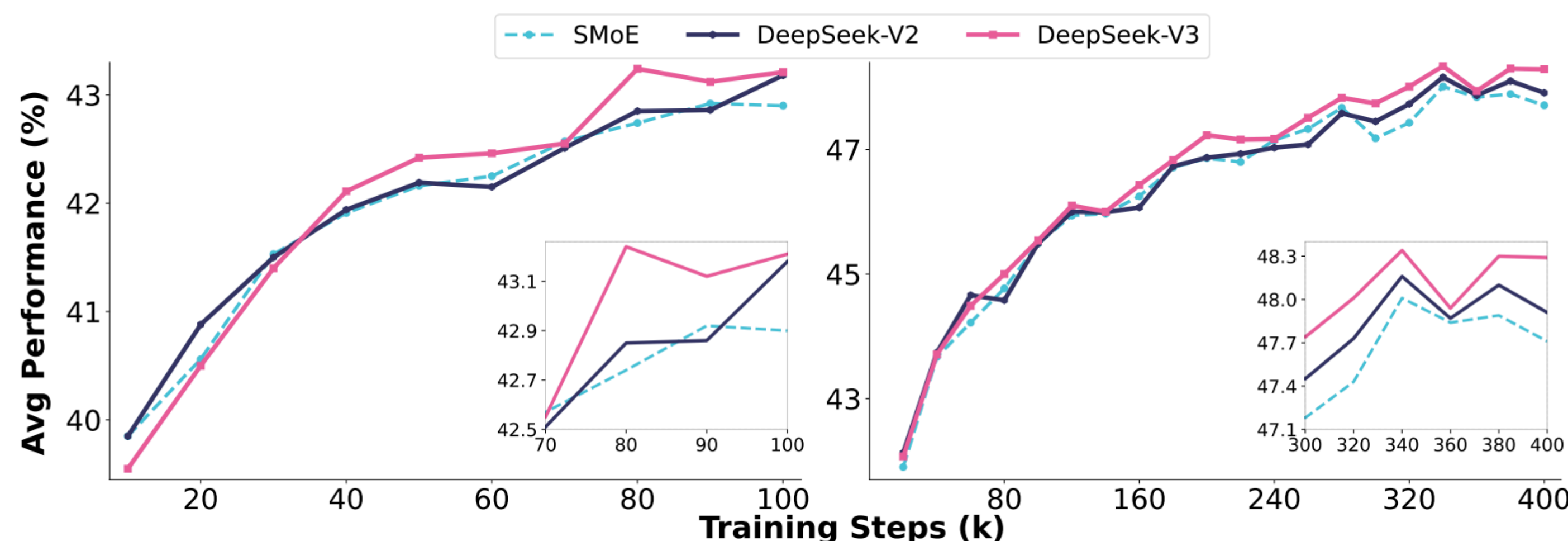


Figure 1: Average performance (%) over training steps in language modeling tasks. **Left:** Model with 158M parameters; **Right:** Model with 679M parameters.

- DeepSeek-V3 and DeepSeek-V2 consistently reach the final performance of Vanilla SMOE **using only 70-80% of the total training steps**
- DeepSeek-V3 demonstrates **marginal improvements** over DeepSeek-V2 in both **convergence speed and final task performance**

→ These results highlight the **efficiency gains** introduced by the shared expert and normalized sigmoid gating mechanisms

Vision-Language Modeling

- We conduct experiments on the visual instruction tuning tasks using the popular **LLaVA architecture**. Models are configured with **66 total experts**, utilizing **top-8 expert routing** v.s. a **top-6 plus 2 shared experts** routing scheme
- To compare different SMoE algorithms, we use **a subset of the LLaVA 1.5 dataset** (332K samples and 287M tokens) to train the models in the Visual Instruction Tuning (VIT) stage

Table 3: Vision-language model performance across benchmarks. (SMoE-SG refers to SMoE Sigmoid Gating)

	AI2D	MMStar	POPE	Science QA	TextVQA	GQA	MME-RW -Lite	MMMU Pro-S	OCR Bench	Average
SMoE	64.90%	41.66%	85.67%	81.61%	40.92%	60.19%	31.79%	25.61%	30.90%	51.47%
DeepSeek-V3	65.45%	41.40%	85.44%	81.94%	40.69%	60.01%	32.20%	26.01%	32.60%	51.75%
DeepSeek-V2	64.70%	41.55%	85.80%	82.20%	40.51%	60.15%	31.11%	25.72%	31.00%	51.41%
SMoE-SR	64.64%	41.51%	85.87%	82.17%	40.54%	60.07%	31.68%	25.95%	31.00%	<u>51.49%</u>

Vision-Language Modeling

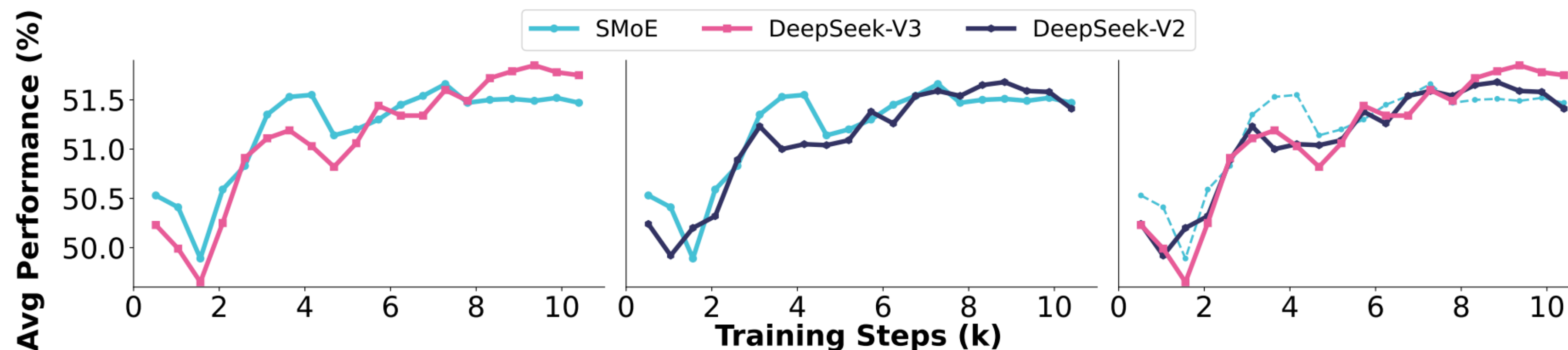


Figure 2: Average performance (%) over training steps on vision-language pretraining tasks. **Left:** Vanilla SMOE vs. DeepSeek-V3; **Center:** Vanilla SMOE vs. DeepSeek-V2; **Right:** DeepSeek-V2 vs. DeepSeek-V3.

- DeepSeek variants exhibit **faster and more stable convergence** compared to Vanilla SMOE
- Both DeepSeek-V2 and DeepSeek-V3 demonstrate **accelerated convergence during the final stages of training**
 - Both shared expert strategy and normalized sigmoid routing significantly contribute to **faster learning in vision-language pre-training**

Router Saturation

- Router saturation quantifies **the proportion of overlapping activated experts** between the final checkpoint and an intermediary checkpoint at time t
- **A higher router saturation** value indicates **stronger alignment in expert selection**, signifying that the router's decisions become increasingly consistent with its final configuration

Router Saturation

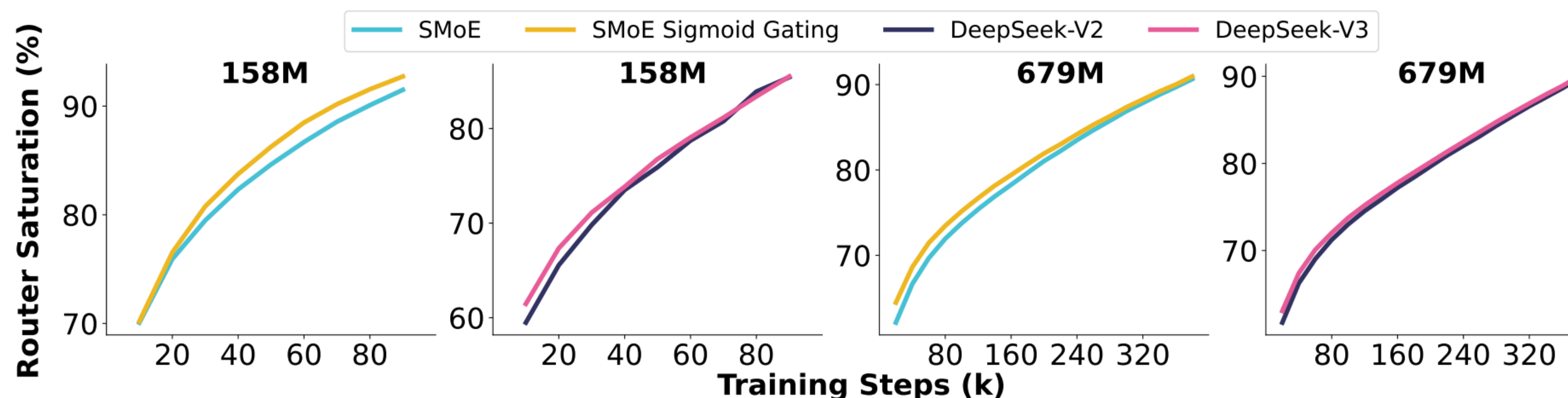


Figure 3: Evolution of router saturation (averaged across all layers) during training for language-modeling tasks with 158 M (left) and 679 M (right) parameter models. We compute saturation by comparing the routing to the top-8 experts with SMOE and SMOE Sigmoid Gating, and the top-6 experts with DeepSeek variants.

- SMOE Sigmoid Gating exhibits consistently steeper saturation curves compared to Vanilla SMOE, reflecting more rapid convergence in expert selection
- A similar pattern is observed in the comparison between DeepSeek-V3 and DeepSeek-V2 under the shared expert configuration

→ These findings highlight the effectiveness of **normalized sigmoid gating** in **accelerating router convergence, potentially reducing the training time required for convergence.**

Router Change Rate

- Router change rate quantifies **the proportion of expert activation decisions that change between consecutive checkpoints**
- **A lower router change rate** implies **greater consistency in routing decisions** over time, reflecting a more stable training process

Router Change Rate

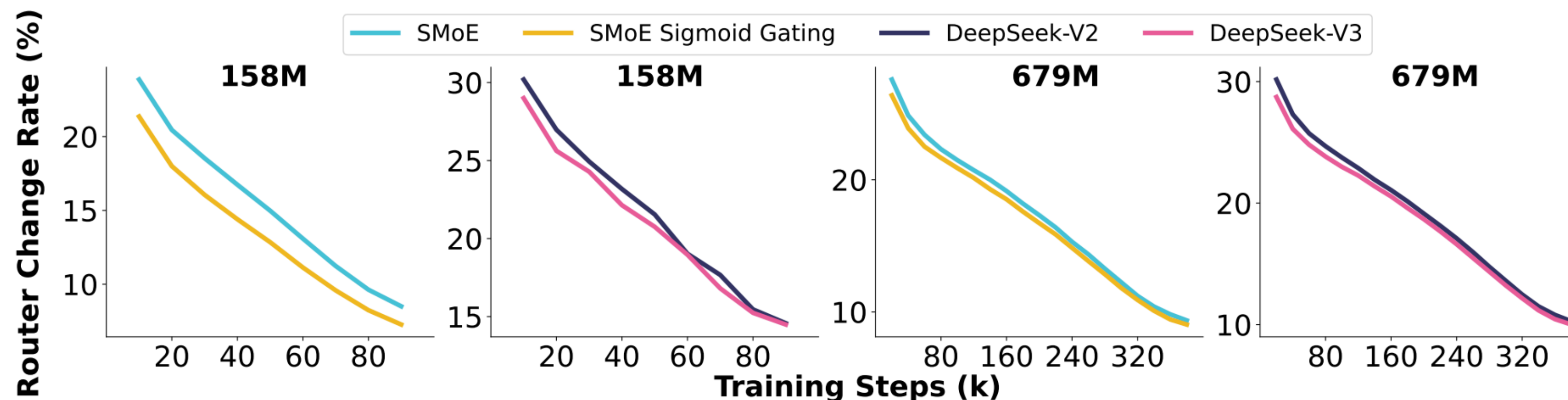


Figure 4: Router Change Rate (averaged across all layers) during training for language-modeling tasks with 158 M (left) and 679 M (right) parameter models. We compute router change rate by comparing the routing to the top-8 experts with SMoE and SMoE Sigmoid Gating, and the top-6 experts with DeepSeek variants.

- **Employing normalized sigmoid gating** have significantly **lower change rates** in both non-shared and shared expert settings
- These findings underscore the efficiency of normalized sigmoid gating in **stabilizing routing decisions throughout training**

Expert Utilization

- To quantify expert utilization, we apply Jain's Fairness Index to the router's resource allocation across n experts

$$J(R) = J(r_1, r_2, \dots, r_n) = \frac{(\sum_{i=1}^n r_i)^2}{n \sum_{i=1}^n r_i^2},$$

where $r_i \geq 0$ represent the proportion of input tokens assigned to expert i

- $J(R) = 1$ indicates **perfectly uniform expert usage**, while $J(R) = 1/n$ signifies **complete imbalance, with only one active expert**

Expert Utilization

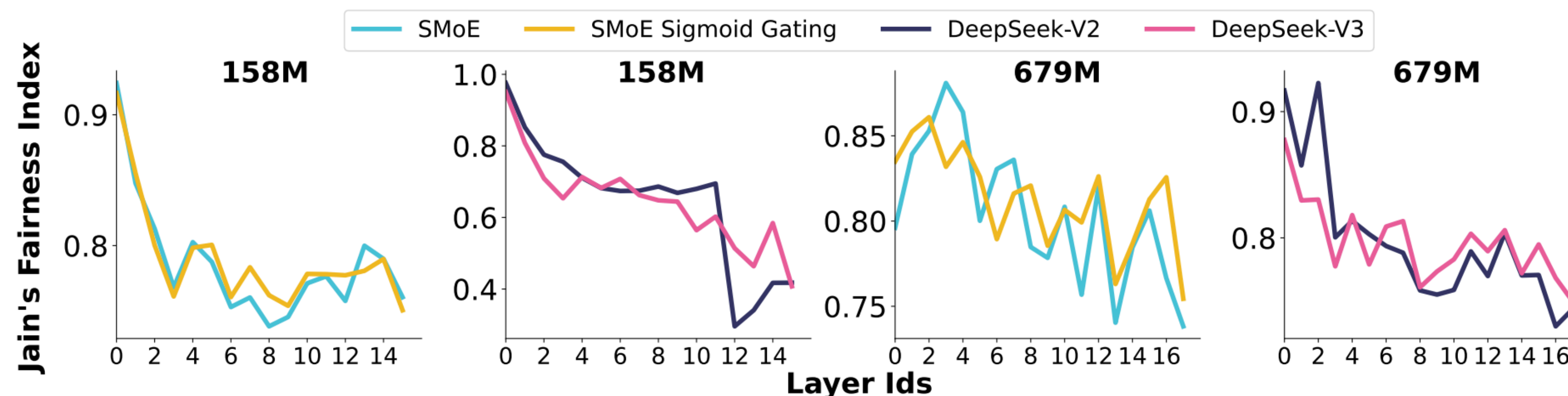


Figure 13: Jain's Fairness Index across MoE layers for language-modeling tasks with 158 M (left) and 679 M (right) parameter models.

- Fairness in expert utilization **is highest in the initial layers** and **declines in subsequent layers**, suggesting that earlier layers facilitate broader expert utilization
- **Models employing normalized sigmoid gating** (SMoE-SG and DeepSeek-V3) maintain a higher fairness index, especially in the later layers, **indicating better expert utilization**

Takeaways

- The shared expert strategy and the normalized sigmoid gating in DeepSeekMoE improves the model sample efficiency
- These two ingredients also yield substantial gains in router convergence, routing stability, and expert utilization

Takeaways

- The shared expert strategy and the normalized sigmoid gating in DeepSeekMoE improves the model sample efficiency
- These two ingredients also yield substantial gains in router convergence, routing stability, and expert utilization

Discussion

There are several on-going directions:

1. **Statistical perspective of parameter fine-tuning techniques (PEFT) via mixture of experts**
2. **Bayesian mixture of experts:**
 - Uncertainty quantification has remained non-trivial
 - Bayesian mixture of experts offers good solution to these problems (Ongoing works)
3. Model selection in mixture of experts:

Thank You!