

ProbMoE:

Differentiable Probabilistic Routing
for Mixture-of-Experts

Differentiable Routing

Probabilistic Inference

Sparse MoE

Heng Zhao¹, Zilei Shao², Guy Van Den Broeck², Zhe Zeng¹

1. University of Virginia · 2. University of California, Los Angeles



About Me

- First-Year PhD student- CS@ UVA
- A member of the TAKO Lab advised by Dr. Zhe Zeng.

TAKO@UVA

Principal Investigator



Zhe Zeng

Assistant Professor



The **T**rustworthy **A**I through **K**nowledge and **O**ptimization (**TAKO**) Lab performs research in artificial intelligence (AI) and machine learning (ML) with focus on neurosymbolic AI and probabilistic ML.

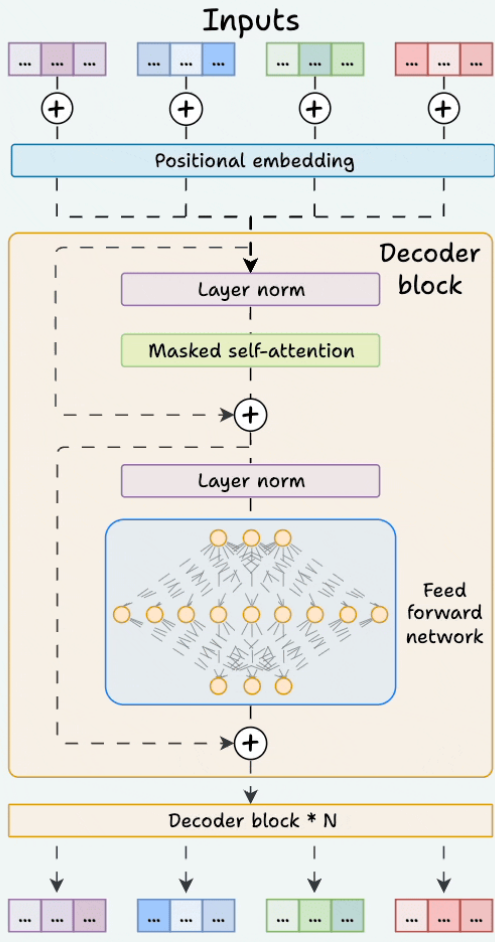


Mixture of Expert (MoE)

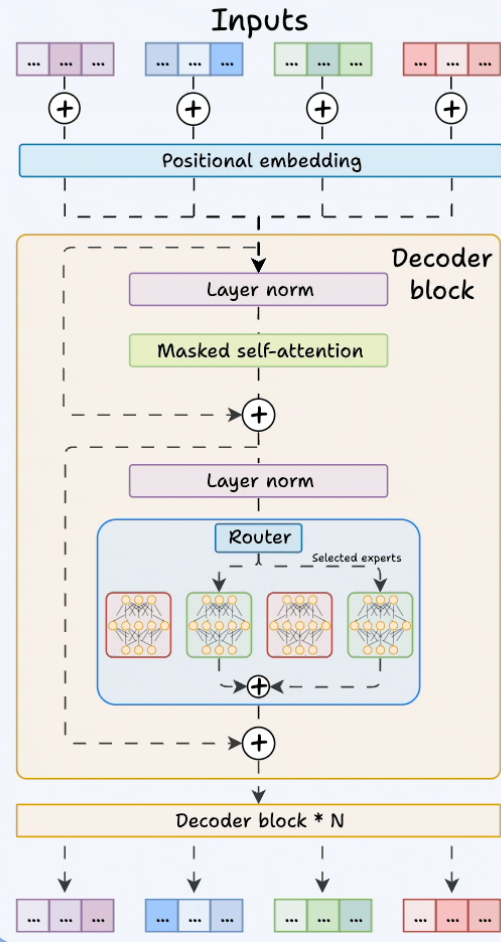
Transformer vs. Mixture of Experts

join.DailyDoseofDS.com

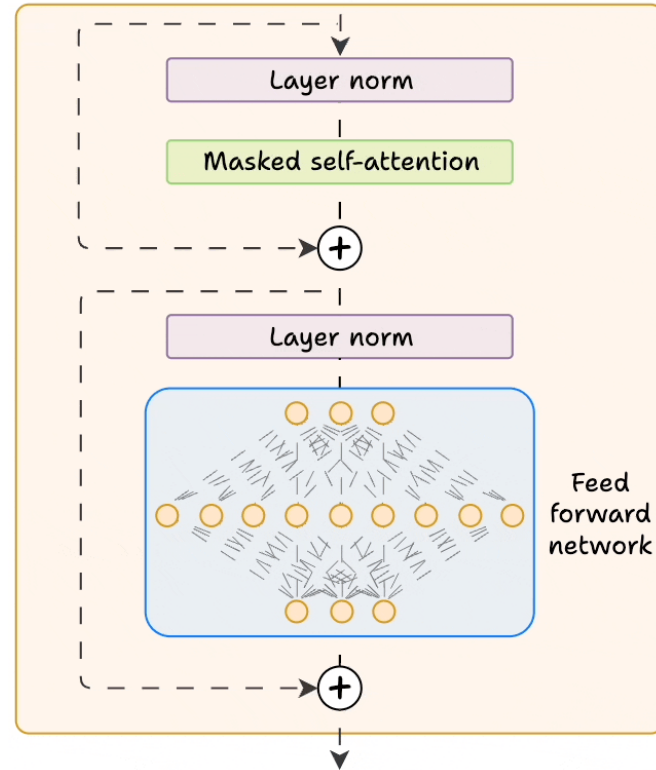
Transformer



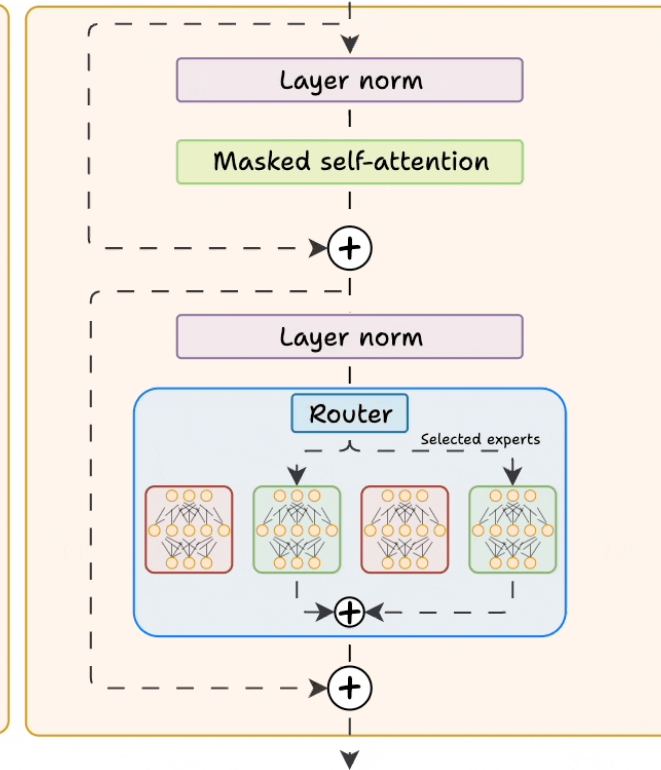
Mixture of Experts



Transformer Decoder Block

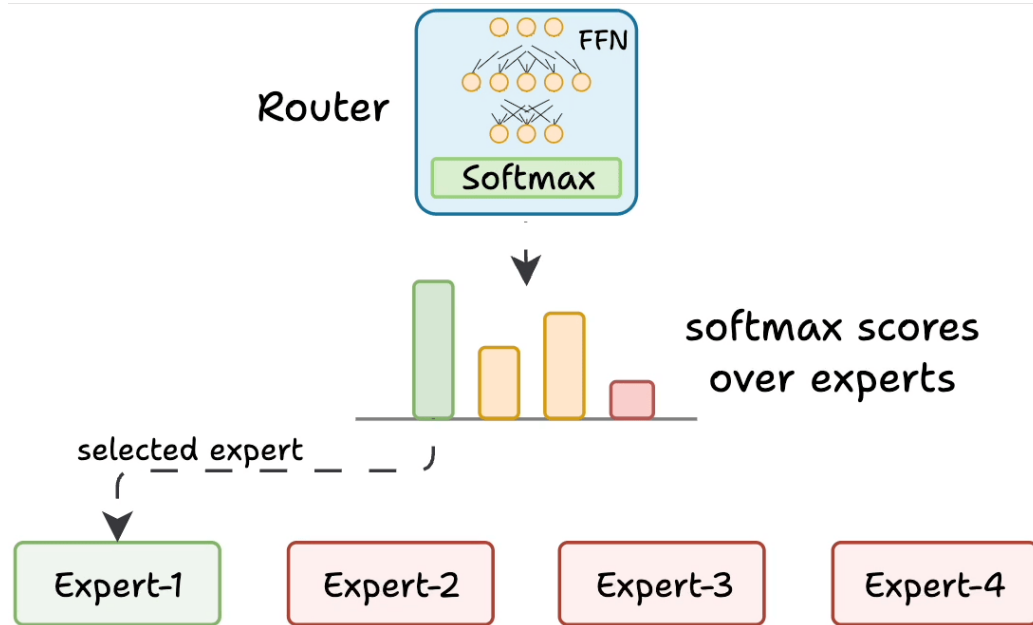


MoE Decoder Block



Source: <https://www.dailydoseofds.com/p/transformer-vs-mixture-of-experts-in-llms/>

Mixture of Expert Model (MoE)



- In most Mixture-of-Experts models, router is usually a small neural network that takes the token representation as input and produces expert logits.

Conventional Mixture-of-Expert (MoE)

- **Treating TopK as fixed**

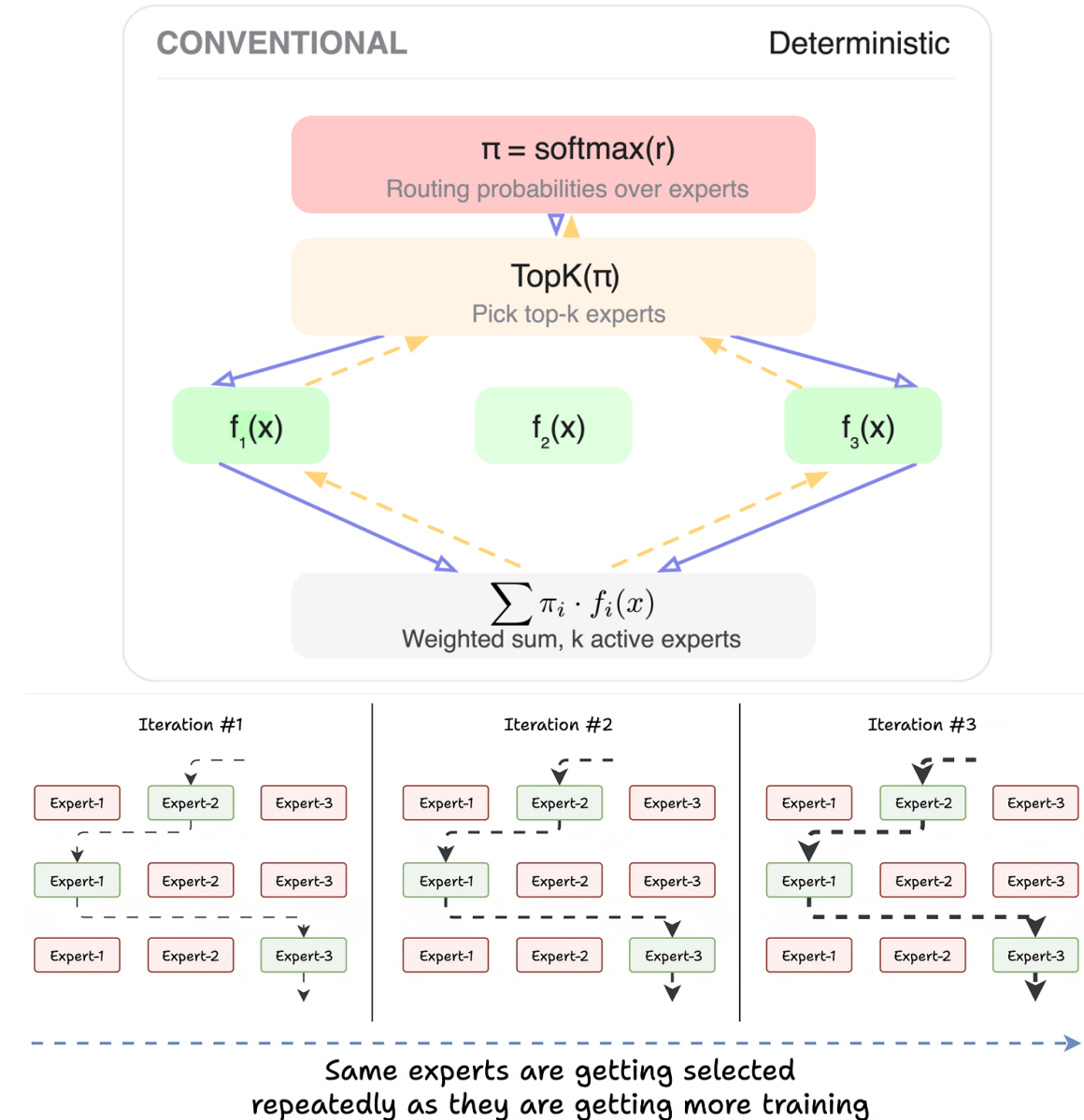
Gradient w.r.t. router logits are computed only through the SoftMax routing weights.

- **Non-differentiable bottleneck**

The resulting learning signal remains tied to TopK decision.

- **Fixed cardinality (k experts per token)**

Each token is always processed by exactly k experts, regardless of complexity.



Source: <https://www.dailydoseofds.com/p/transformer-vs-mixture-of-experts-in-llms/>

ProbMoE

● Probabilistic subset sampling

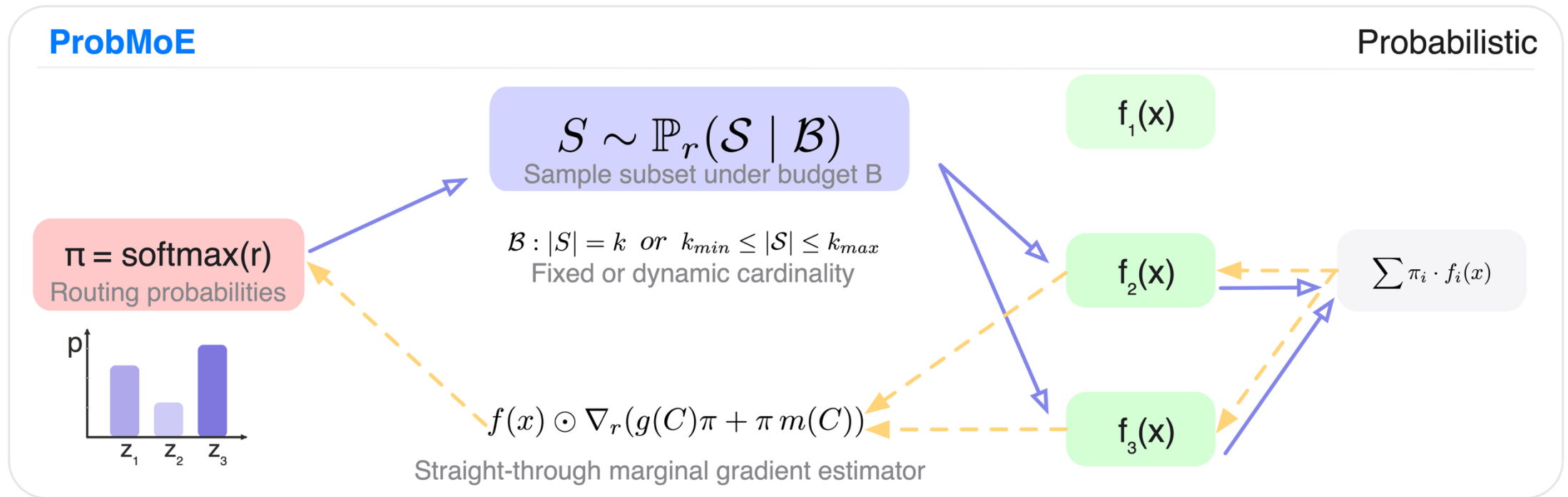
The key idea is that ProbMoE treats MoE routing as a probabilistic and differentiable framework

● Fixed or dynamic cardinality

Supports both exact-k routing and dynamic-k where $|S|$ varies between $k_{\min}$ and $k_{\max}$.

● Differentiable via marginal gradient

Uses straight-through marginal gradient estimator to backpropagate through discrete sampling.



Performance Comparison Across Benchmarks

9.25% ↑

GSM (OLMoE ProbMoE vs Conv.)

16%↑

Law (OLMoE ProbMoE vs Conv.)

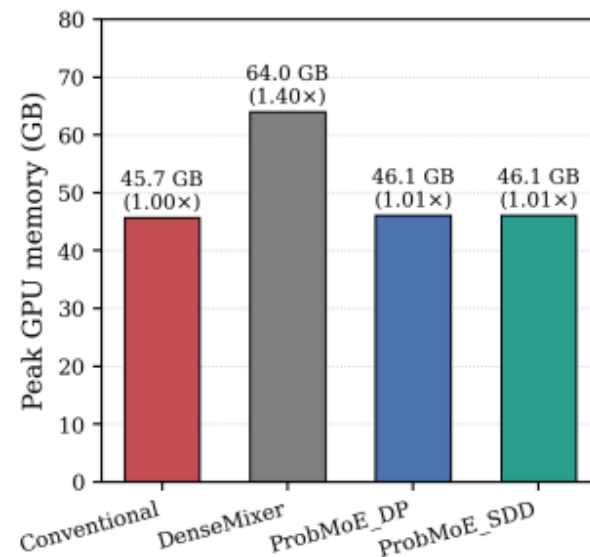
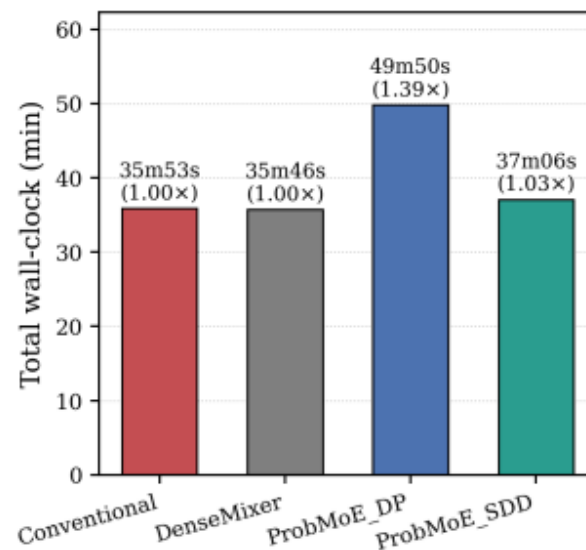
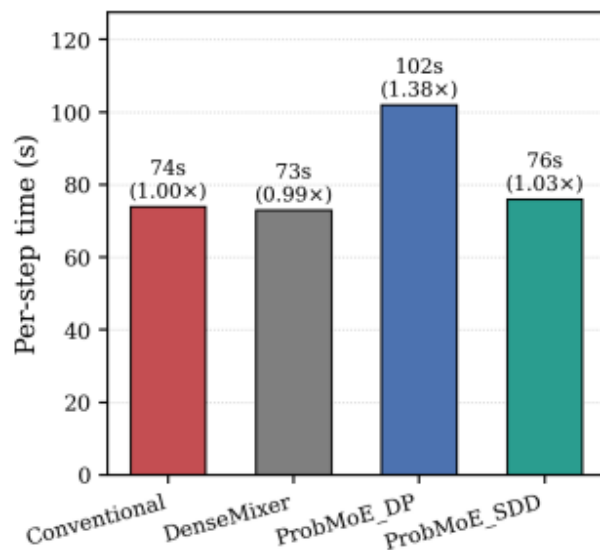
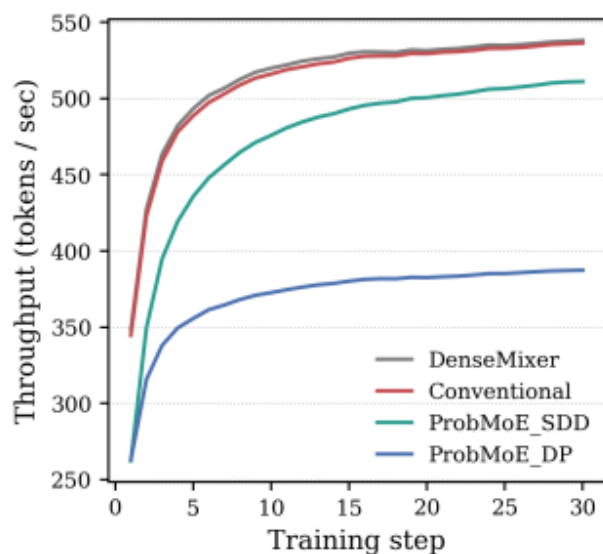
13.85%↑

Summary (Qwen ProbMoE vs Conv.)

ProbMoE achieves top-1 or top-2 accuracy on 5 of 7 benchmarks across both backbones

Model	GSM	Law	Translation	MBPP	Summary	MMLU Overall	MMLU Stem
OLMoE Backbone (Top- <i>k</i> = 8)							
Base Model	15.85	5.70	11.09	19.80	7.40	53.40	44.91
Frozen Router	44.88	22.50	28.29	17.80	36.05	53.97	46.18
Conventional	45.94	25.00	27.56	23.20	33.70	54.04	46.46
DenseMixer	47.00	27.90	30.32	24.40	37.50	53.95	46.18
ProbMoE	50.19	29.00	31.63	22.80	39.29	53.69	45.54
Qwen backbone (Top- <i>k</i> = 4)							
Base Model	38.69	18.20	16.53	38.84	28.29	60.85	52.27
Frozen Router	53.37	33.01	32.75	35.20	38.29	61.08	53.19
Conventional	53.30	29.50	30.00	32.80	39.00	61.03	53.16
DefaultMoE	51.00	31.20	24.00	33.20	38.40	—	—
SparseMixer	1.30	3.40	3.50	0.00	2.10	—	—
ReMoE	46.30	25.50	16.99	33.00	25.80	—	—
DenseMixer	54.97	30.75	33.75	34.00	41.00	61.03	52.87
ProbMoE	53.29	34.40	39.23	35.00	44.40	61.05	53.82

Wall-clock Report



● In terms of peak GPU memory, ProbMoE DP and ProbMoE SDD closely match Conventional fine-tuning, whereas DenseMixer requires substantially more memory than any other method.

Gradient variance comparison

Model	Aggregated Bias	Aggregated Variance	Aggregated Error
ProbMoE	0.0702 ± 0.0426	0.0076 ± 0.0110	0.0967 ± 0.0659
DenseMixer	0.1052 ± 0.0839	0.0201 ± 0.0427	0.1349 ± 0.0995

- marginal-based estimator provides a more accurate and stable approximation of the true gradient

Expert Diversity

- **ProbMoE activates more diverse experts**

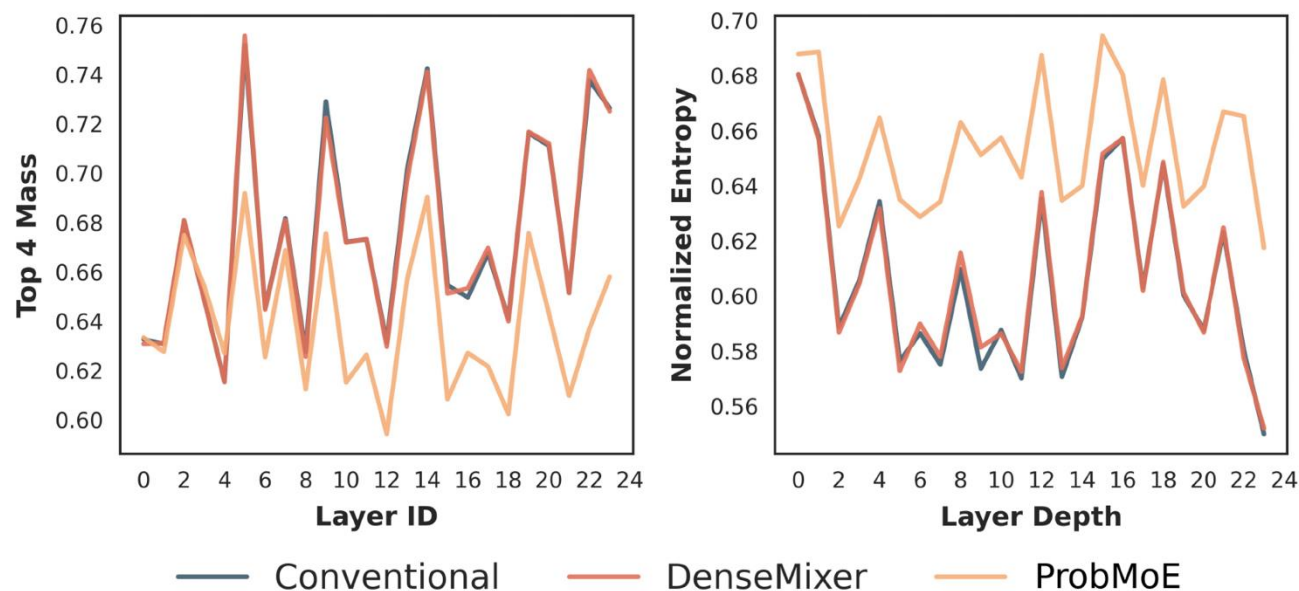
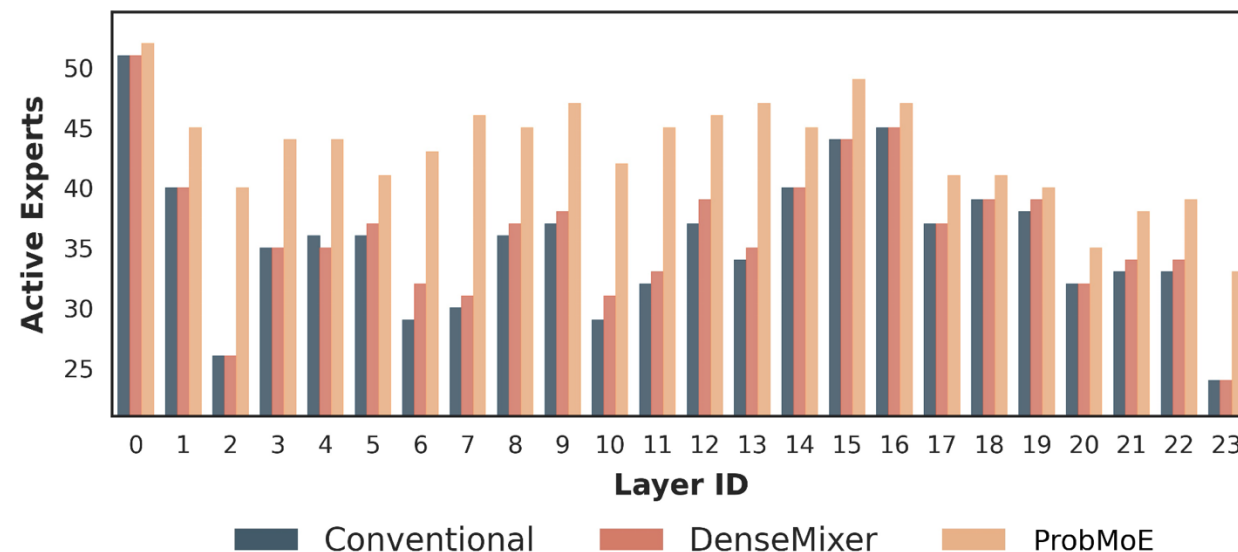
Exact-k routing activates more distinct experts per layer vs. Conventional and DenseMixer on Qwen.

- **Less concentrated expert assignment**

Top-4 mass is lower for ProbMoE: routing probability is spread more evenly across experts.

- **Higher normalized entropy**

Routing distributions are more uniform — a healthier, more balanced use of model capacity.



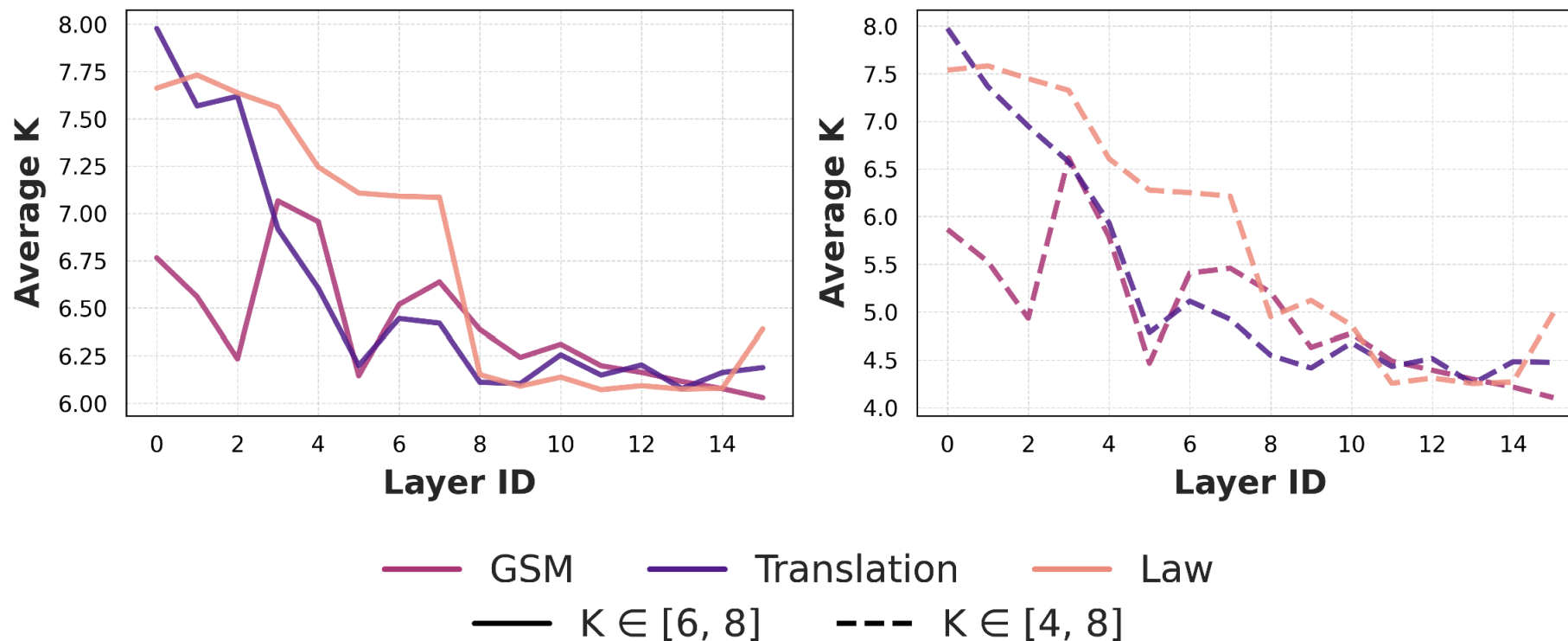
Dynamic-k vs Exact-k Routing Performance

Dataset	OLMoE	Qwen1.5
GSM	−1.82 (80.00%)	−4.29 (75.00%)
Law	−0.04 (84.50%)	+2.70 (75.00%)
Translation	+0.36 (82.00%)	+3.22 (75.00%)

- **Accuracy V.S. efficiency trade-off**

ProbMoE Dynamic-k achieves performance comparable to ProbMoE Exact-k while activating fewer experts.

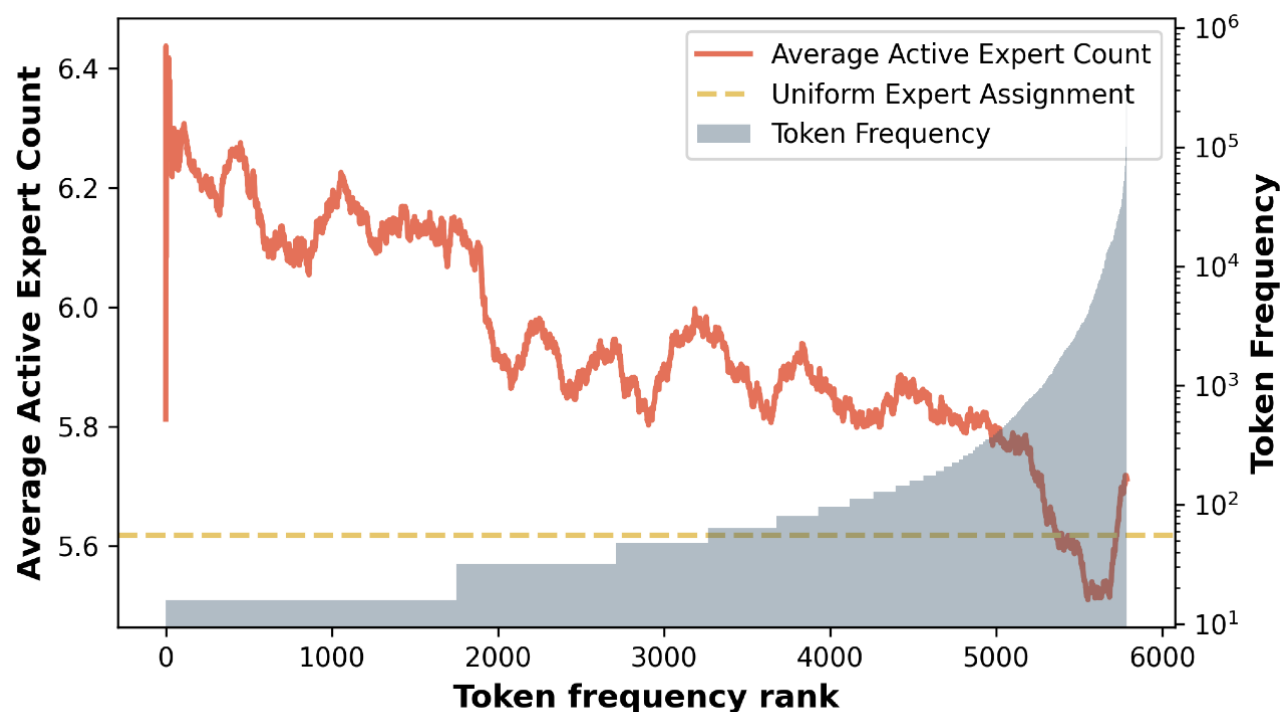
Dynamic-k routing behavior



- **Different tasks uses different**

ProbMoE allocates systematically different amounts of expert capacity across tasks depending on their difficulty

Token Frequency vs Expert Activation



- **Rare tokens activate more experts**

ProbMoE Dynamic-k routes rarer tokens through more experts, allocating more capacity to hard cases.

- **Frequent tokens use fewer experts**

Common tokens require less computation — the model dynamically routes them through fewer experts.

Semantic-Aware Expert Allocation

Avg. Active Experts	Examples
Top 20%	\n, :, ?, ons, eli
Bottom 20%	16, 5, downward, Meat, reduce

- **Semantics drives expert allocation**

Ambiguous tokens activates more experts while concrete words uses fewer.

THANKS!