Rethinking Sparse Mixture of Experts from a Unified Perspective

Giang Do, Hung Le, Truyen Tran

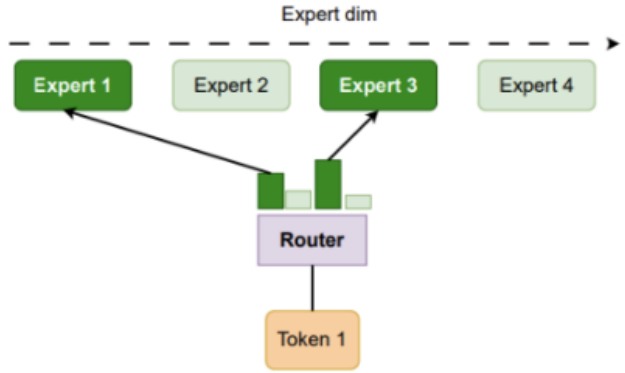
Applied Artificial Intelligence Initiative (A2I2), Deakin University, Victoria, Australia

Content

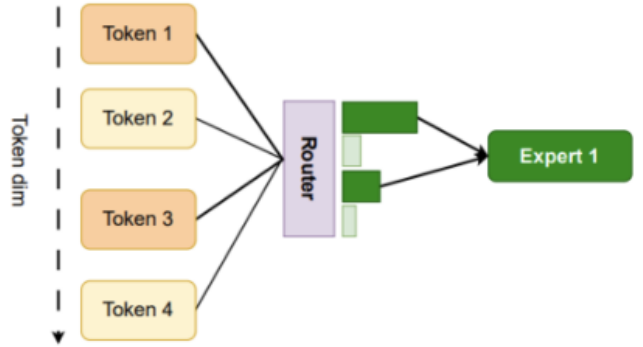
1. Problems

Conventional **Sparse Mixture-of-Experts (SMoE)** routing performs one-dimensional expert selection, either along the *expert dimension* or the *token dimension*:

a) **Token Choice:** each token selects its *top-k* experts
b) **Expert Choice:** each expert selects its *top-k* tokens



(a) Token Choice (TC)

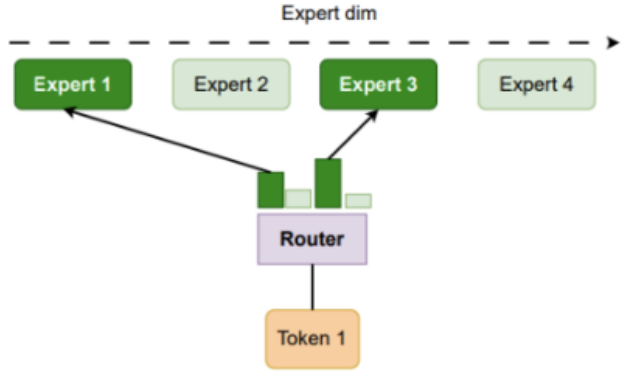


(b) Expert Choice (EC)

1. Problems

Conventional **Sparse Mixture-of-Experts (SMoE)** routing performs one-dimensional expert selection, either along the *expert dimension* or the *token dimension*:

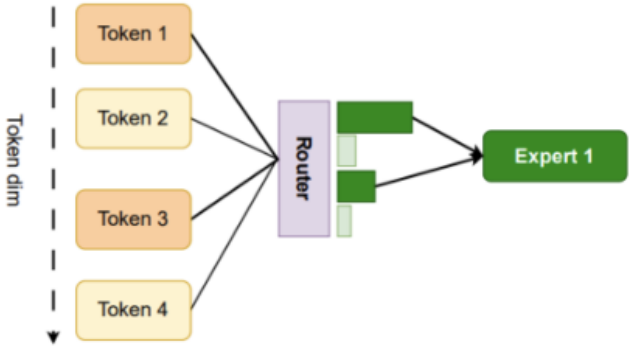
a) **Token Choice:** each token selects its *top-k* experts
b) **Expert Choice:** each expert selects its *top-k* tokens



(a) Token Choice (TC)

Drawbacks of Conventional SMoE Routing:

- **Irrelevant candidate selection**: Both Token Choice and Expert Choice must select a fixed number of experts/tokens, even when some candidates are poorly matched.

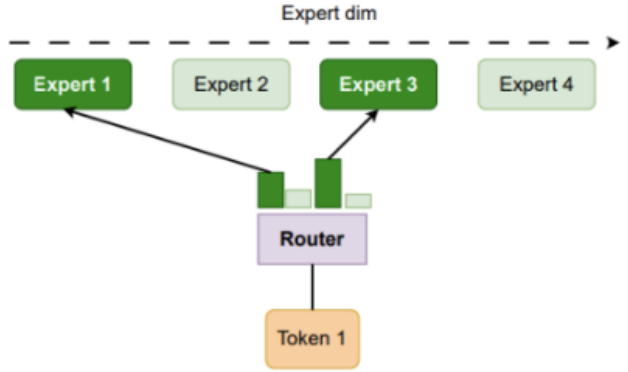


(b) Expert Choice (EC)

1. Problems

Conventional **Sparse Mixture-of-Experts (SMoE)** routing performs one-dimensional expert selection, either along the *expert dimension* or the *token dimension*:

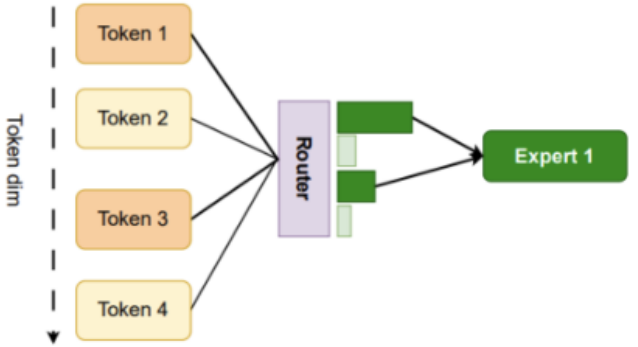
a) **Token Choice:** each token selects its *top-k* experts
b) **Expert Choice:** each expert selects its *top-k* tokens



(a) Token Choice (TC)

Drawbacks of Conventional SMoE Routing:

- **Irrelevant candidate selection**: Both Token Choice and Expert Choice must select a fixed number of experts/tokens, even when some candidates are poorly matched.
- **Limited computational flexibility**: Both methods rely on a fixed integer *top-k* budget, typically (k=4–8), which restricts adaptive computation.



(b) Expert Choice (EC)

2. Unified Sparse Mixture of Experts (USMoE)

Reformulating Expert Selection as Linear Programming

We formulate expert selection as an optimization problem that *maximizes token–expert similarity* under routing constraints:

- **Token Choice constraint**: each token selects a fixed number of experts.

$$\begin{aligned} \text{maximize} \quad & \sum_{i=1}^{T}\sum_{j=1}^{N} S_{ij}x_{ij} \\ \text{subject to} \quad & \sum_{i=1}^{T}\sum_{j=1}^{N} x_{ij} \le c \\ & x_{ij} \in \{0,1\}, \quad \forall i,j. \end{aligned} \tag{2}$$

Token-Choice (TC). TC assumes uniform importance across tokens, assigning each token i to a fixed number of experts $k = \lfloor c/T \rfloor$. This adds the per-token constraint:

$$\sum_{j=1}^{N} x_{ij} = k, \quad \forall i \in 1,\ldots,T. \tag{3}$$

2. Unified Sparse Mixture of Experts (USMoE)

Reformulating Expert Selection as Linear Programming

We formulate expert selection as an optimization problem that *maximizes token–expert similarity* under routing constraints:

- **Token Choice constraint**: each token selects a fixed number of experts.
- **Expert Choice constraint**: each expert selects a fixed number of tokens.

$$\begin{aligned} \text{maximize} \quad & \sum_{i=1}^{T}\sum_{j=1}^{N} S_{ij}x_{ij} \\ \text{subject to} \quad & \sum_{i=1}^{T}\sum_{j=1}^{N} x_{ij} \le c \\ & x_{ij} \in \{0,1\}, \quad \forall i,j. \end{aligned} \tag{2}$$

Token-Choice (TC). TC assumes uniform importance across tokens, assigning each token i to a fixed number of experts $k = \lfloor c/T \rfloor$. This adds the per-token constraint:

$$\sum_{j=1}^{N} x_{ij} = k, \quad \forall i \in 1,\ldots,T. \tag{3}$$

Expert-Choice (EC). Conversely, EC ensures uniform expert utilization by allowing each expert j to select the top $e = \lfloor c/N \rfloor$ tokens. This is modeled by constraining the per-expert capacity:

$$\sum_{i=1}^{T} x_{ij} = e, \quad \forall j \in 1,\ldots,N. \tag{4}$$

2. Unified Sparse Mixture of Experts (USMoE)

Reformulating Expert Selection as Linear Programming

We formulate expert selection as an optimization problem that *maximizes token–expert similarity* under routing constraints:

- **Token Choice constraint**: each token selects a fixed number of experts.
- **Expert Choice constraint**: each expert selects a fixed number of tokens.

➔ **Unified Perspective: Relaxing the constraints of TC and EC enables a more optimal routing solution.**

$$\begin{aligned} \text{maximize} \quad & \sum_{i=1}^{T}\sum_{j=1}^{N} S_{ij}x_{ij} \\ \text{subject to} \quad & \sum_{i=1}^{T}\sum_{j=1}^{N} x_{ij} \le c \\ & x_{ij} \in \{0,1\}, \quad \forall i,j. \end{aligned} \tag{2}$$

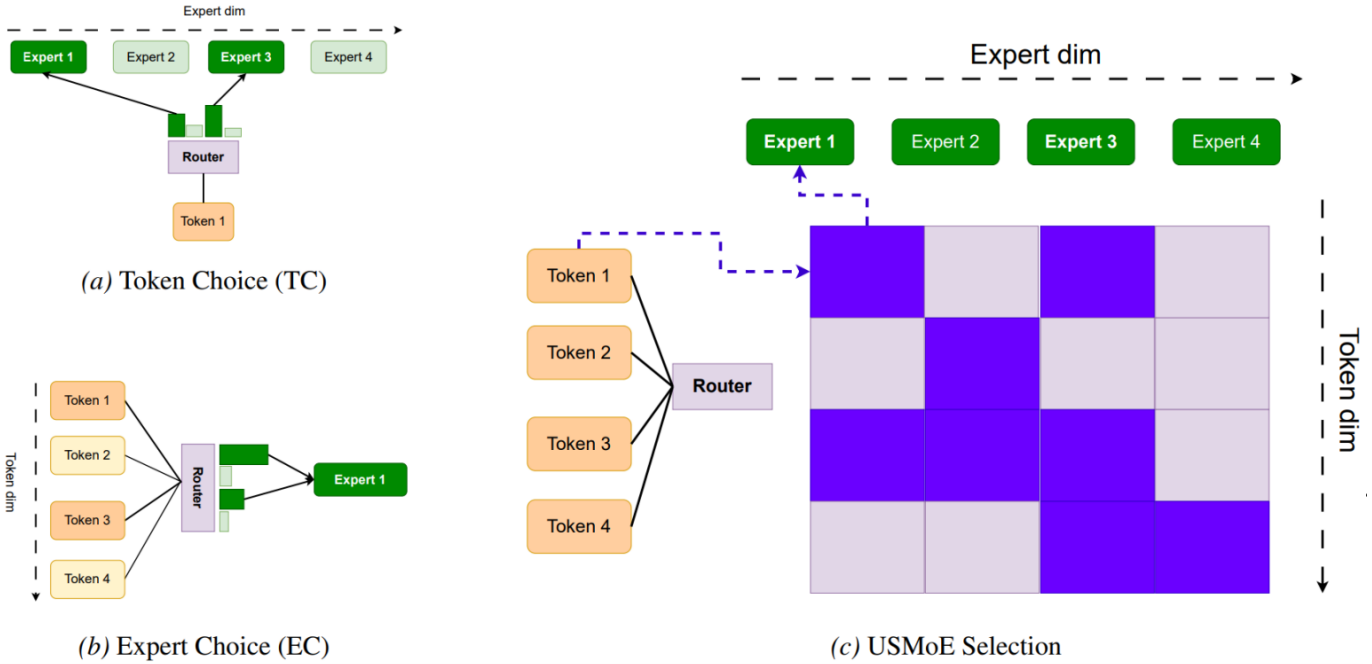
Token-Choice (TC). TC assumes uniform importance across tokens, assigning each token i to a fixed number of experts $k = \lfloor c/T \rfloor$. This adds the per-token constraint:

$$\sum_{j=1}^{N} x_{ij} = k, \quad \forall i \in 1,\ldots,T. \tag{3}$$

Expert-Choice (EC). Conversely, EC ensures uniform expert utilization by allowing each expert j to select the top $e = \lfloor c/N \rfloor$ tokens. This is modeled by constraining the per-expert capacity:

$$\sum_{i=1}^{T} x_{ij} = e, \quad \forall j \in 1,\ldots,N. \tag{4}$$

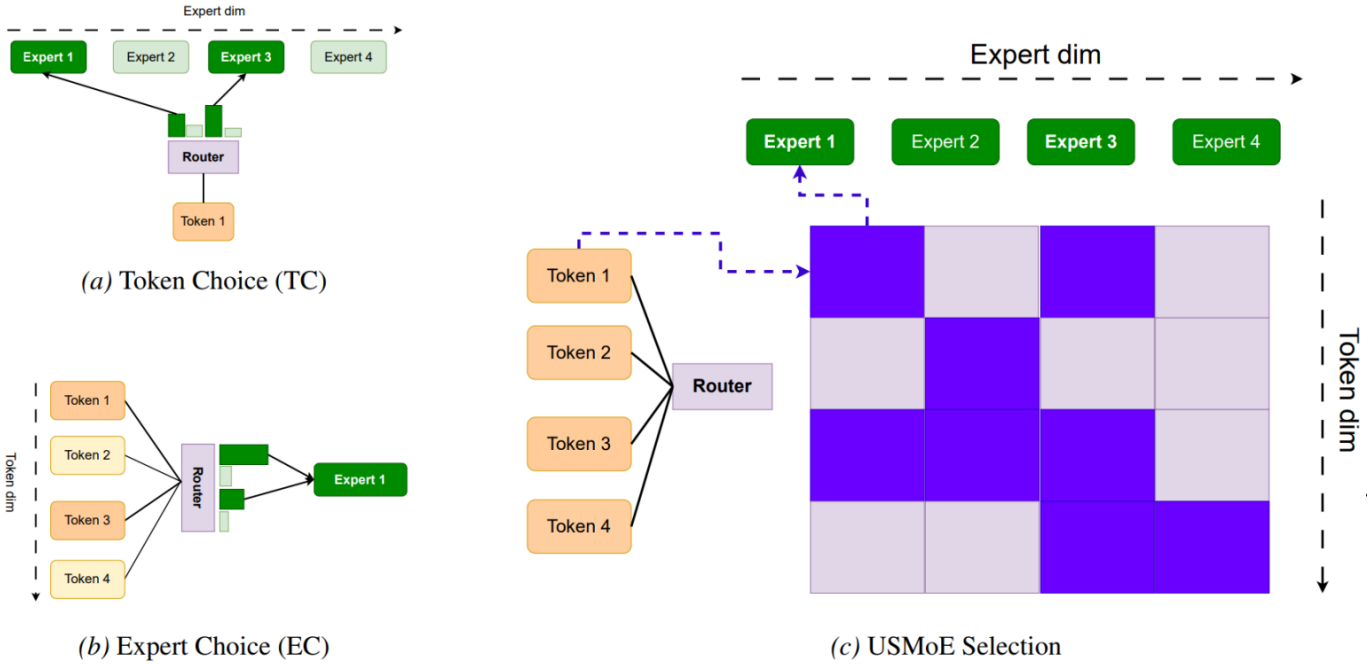
2. Unified Sparse Mixture of Experts (USMoE)



Experts Selection as a Unified Perspective

- **SMoE**: One-dimensional routing makes it difficult to filter out irrelevant token–expert pairs.
- **USMoE**: Two-dimensional token–expert selection enables globally better matching and suppresses irrelevant pairs.

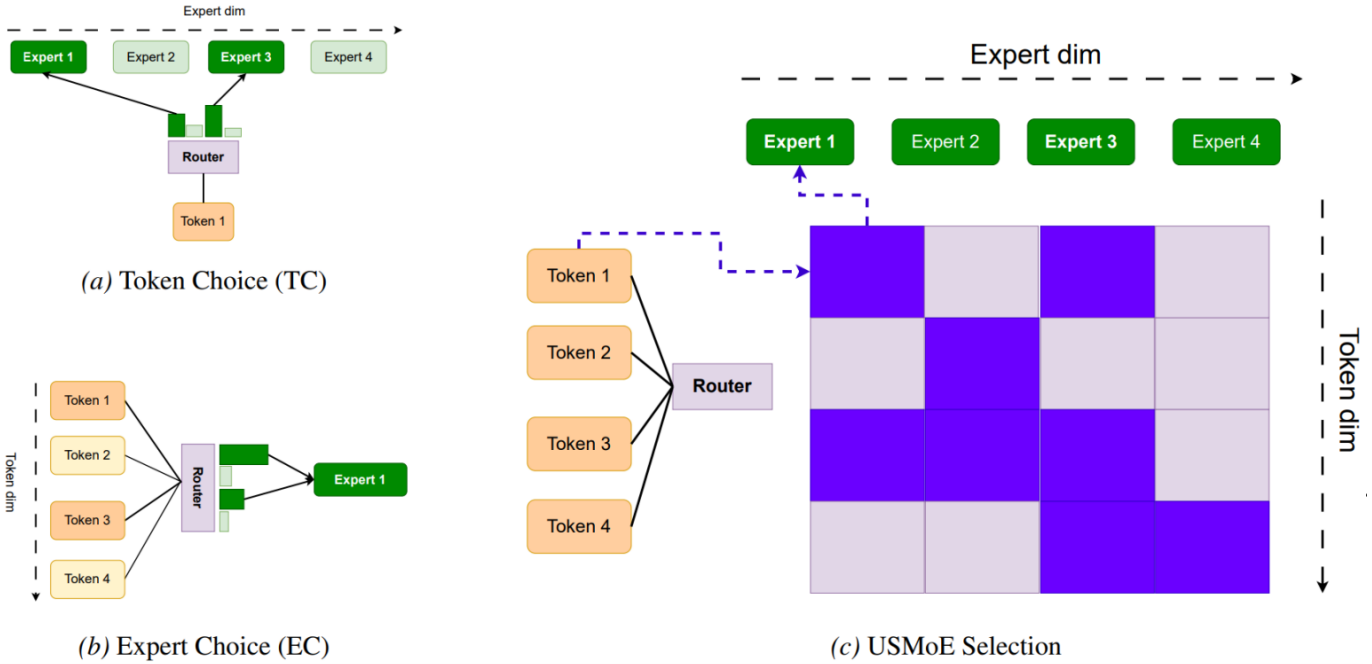
2. Unified Sparse Mixture of Experts (USMoE)



Experts Selection as a Unified Perspective

- **SMoE**: One-dimensional routing makes it difficult to filter out irrelevant token–expert pairs.
- **USMoE**: Two-dimensional token–expert selection enables globally better matching and suppresses irrelevant pairs.
  - ✓ **Avoid missing important** experts/tokens

2. Unified Sparse Mixture of Experts (USMoE)



(a) Token Choice (TC)

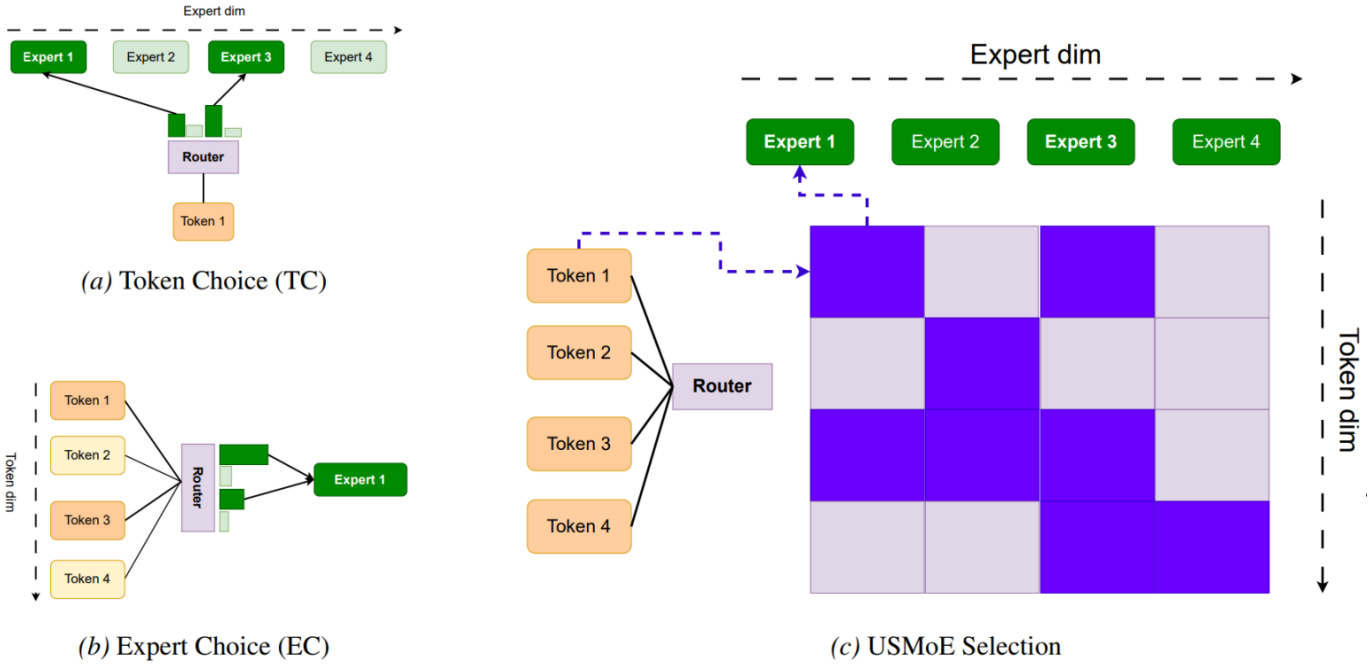
(b) Expert Choice (EC)

(c) USMoE Selection

Experts Selection as a Unified Perspective

- **SMoE**: One-dimensional routing makes it difficult to filter out irrelevant token–expert pairs.
- **USMoE**: Two-dimensional token–expert selection enables globally better matching and suppresses irrelevant pairs.
  - ✓ **Avoid missing important** experts/tokens
  - ✓ **Avoid selecting irrelevant** experts/tokens

2. Unified Sparse Mixture of Experts (USMoE)



(a) Token Choice (TC)

(b) Expert Choice (EC)

(c) USMoE Selection

Experts Selection as a Unified Perspective

- **SMoE**: One-dimensional routing makes it difficult to filter out irrelevant token–expert pairs.
- **USMoE**: Two-dimensional token–expert selection enables globally better matching and suppresses irrelevant pairs.
  - ✓ **Avoid missing important** experts/tokens
  - ✓ **Avoid selecting irrelevant** experts/tokens
  - ✓ **Enable dynamic expert** selection with integer or non-negative fractional top-(k) budgets

2. Unified Sparse Mixture of Experts (USMoE)

Router Logits as a Unified Score

- **USMoE**: Introduces a **unified score** that integrates *token-centric* and *expert-centric* signals for more balanced and informed routing decisions.

Algorithm 1 Unified Sparse Mixture-of-Experts (USMoE) Layer

Input: Input tensor $X \in \mathbb{R}^{B \times L \times D}$, router weights $W_r \in \mathbb{R}^{D \times N}$, experts $\{E_i\}_{i=1}^{N}$, controlling factor $\alpha \in [0, 1]$, score functions S_t (TC) and S_e (EC)
Output: Output tensor $Y \in \mathbb{R}^{B \times L \times D}$
$S \leftarrow XW_r$ {Compute routing logits via dot product}
$U \leftarrow (1 - \alpha) \cdot S_t + \alpha \cdot S_e$ {Compute unified scores}
$U_{flat} \leftarrow \text{reshape}(U, [B, L \cdot N])$ {Flatten for global or sequence-level selection}
$V_{top}, I_{top} \leftarrow \text{TopK}(U_{flat}, k = n)$ {Select top-n expert-token pairs}
$Y \leftarrow \text{SparseDispatcher}(X, \{E_i\}, V_{top}, I_{top})$ {Execute sparse MoE computation}
return Y

3. Theoretical Results

Key Theoretical Results of USMoE:

- **Proposition 3.3:** USMoE achieves the *optimal solution* by maximizing total token–expert similarity under a *global routing budget*.
- **Lemma 3.5:** USMoE provides a natural *filtering effect*: clean token–expert pairs with high affinity scores crowd out noisy pairs amplified by TC/EC routing.
- **Lemma 3.7:** USMoE *mitigates representation collapse*, leading to stronger routing behavior than conventional SMoE methods, as discussed in Chi et al. (2022).

4. Experiment Results

- **Training-free Evaluation:** USMoE improves reasoning LLMs over TC/EC without retraining.

Benchmark	Clean Setting				Corrupt Setting			
	TC		EC	USMoE	TC		EC	USMoE
	Original	MoEE			Original	MoEE		
Qwen3-30B-A3B-Thinking								
ARC-C	0.584 ± .014	0.275 ± .013	0.470 ± .015	**0.600** ± .014	0.244 ± .013	0.279 ± .013	0.353 ± .014	**0.383** ± .014
ARC-E	0.819 ± .008	0.241 ± .009	0.670 ± .010	**0.822** ± .008	0.341 ± .010	0.237 ± .009	0.442 ± .010	**0.558** ± .010
BoolQ	0.866 ± .006	0.378 ± .008	0.785 ± .007	**0.870** ± .006	0.556 ± .009	0.378 ± .008	0.627 ± .008	**0.727** ± .008
OBQA	0.424 ± .022	0.292 ± .020	0.378 ± .022	**0.450** ± .022	0.294 ± .020	0.300 ± .021	0.322 ± .021	**0.344** ± .020
PIQA	0.812 ± .009	0.517 ± .012	0.649 ± .011	**0.842** ± .009	0.520 ± .012	0.530 ± .012	0.554 ± .012	**0.619** ± .011
WinoGrande	0.729 ± .012	0.505 ± .014	0.596 ± .014	**0.739** ± .012	0.490 ± .014	0.522 ± .014	0.513 ± .014	**0.560** ± .014
Average	0.706 ± .012	0.368 ± .013	0.591 ± .013	**0.721** ± .012	0.408 ± .013	0.374 ± .013	0.469 ± .013	**0.532** ± .013
Qwen3-30B-A3B-Instruct								
ARC-C	0.631 ± .014	0.270 ± .013	0.512 ± .015	**0.646** ± .014	0.253 ± .013	0.275 ± .013	0.354 ± .014	**0.433** ± .014
ARC-E	0.838 ± .008	0.239 ± .009	0.702 ± .009	**0.846** ± .007	0.361 ± .010	0.239 ± .009	0.483 ± .010	**0.609** ± .010
BoolQ	0.886 ± .006	0.378 ± .008	0.853 ± .006	**0.894** ± .005	0.600 ± .009	0.378 ± .008	0.747 ± .008	**0.808** ± .007
OBQA	0.454 ± .022	0.292 ± .020	0.374 ± .022	**0.460** ± .022	0.300 ± .021	0.302 ± .021	0.326 ± .021	**0.336** ± .021
PIQA	0.805 ± .009	0.521 ± .012	0.646 ± .011	**0.810** ± .009	0.516 ± .012	0.530 ± .012	0.568 ± .012	**0.605** ± .011
WinoGrande	0.733 ± .012	0.497 ± .014	0.583 ± .014	**0.736** ± .012	0.494 ± .014	0.518 ± .014	0.513 ± .014	**0.545** ± .014
Average	0.724 ± .012	0.366 ± .013	0.612 ± .013	**0.732** ± .012	0.421 ± .013	0.374 ± .013	0.499 ± .013	**0.556** ± .013

Table 1. Performance comparison of various methods for Qwen3-30B-A3B-Thinking and Qwen3-30B-A3B-Instruct using 0-shot evaluation across six reasoning benchmarks. We evaluate both clean and corrupt settings, grouping methods by selection strategy: **TC** (Token Choice), **EC** (Expert Choice), and our proposed **USMoE**. The best results are highlighted in **bold**.

4. Experiment Results

- **Training-free Evaluation:** USMoE significantly outperforms standard routing baselines across a range of MTEB tasks without retraining.

Model	Task	Router	TC	EC	MoEE	USMoE
OLMoE-1B-7B	Classification	43.1	57.7	56.2	51.7	**61.4**
	Clustering	16.2	24.8	26.9	23.2	**32.1**
	PairClassification	53.5	62.0	58.9	66.0	**68.9**
	Reranking	41.7	51.3	51.0	53.2	**55.1**
	STS	49.4	63.5	44.2	67.8	**71.1**
	Summarization	25.6	28.9	29.7	30.4	**30.5**
	Average	38.3	48.0	44.5	48.7	**53.2**
Qwen1.5-MoE-A2.7B	Classification	48.8	58.0	35.2	54.0	**59.7**
	Clustering	14.3	34.2	29.2	30.1	**37.5**
	PairClassification	51.9	60.5	56.0	60.3	**66.6**
	Reranking	41.0	46.6	45.0	51.1	**56.8**
	STS	48.3	50.1	50.0	64.3	**69.0**
	Summarization	27.0	23.0	21.9	27.3	**31.0**
	Average	38.6	45.4	39.6	47.9	**53.4**
DeepSeekMoE-16B	Classification	48.6	56.4	55.4	53.0	**60.4**
	Clustering	17.8	29.0	20.3	28.5	**32.8**
	PairClassification	57.4	59.8	53.8	63.3	**67.9**
	Reranking	43.8	45.7	40.9	50.6	**52.4**
	STS	52.8	49.0	37.1	63.4	**68.1**
	Summarization	29.1	24.4	25.7	29.2	**30.7**
	Average	41.6	44.0	38.9	48.0	**52.0**

Table 2. Performance comparison of USMoE, Token Choice (TC), Expert Choice (EC), and MoEE across across MTEB Tasks with **PromptEOL (Jiang et al., 2024b)**. The best result for each row is highlighted in **bold**.

4. Experiment Results

Training from Scratch:

- **Language Models:** USMoE consistently improves over TC and EC across Enwik8, Text8, WikiText-103, and LM1B
- **Vision**: USMoE achieves the highest mean accuracy with lower variance across diverse vision datasets.

Setting	Dataset	USMoE $k=2$	USMoE $k=1.5$	TC $k=2$	EC $k=2$
Original	Enwik8 (↓)	**1.18**	1.19	1.20	**1.18**
	Text8 (↓)	**1.20**	1.28	1.29	1.24
	WikiText-103 (↓)	**29.20**	30.67	30.16	29.83
	lm1b (↓)	**56.90**	57.55	58.00	58.60
Corrupt	Enwik8 (↓)	**1.75**	1.78	1.77	1.76
	Text8 (↓)	**1.83**	1.95	1.86	1.89
	WikiText-103 (↓)	**38.45**	40.39	39.31	39.28
	lm1b (↓)	**68.43**	70.47	79.73	71.75

Table 3. Performance comparison of USMoE, Token Choice (TC), and Expert Choice (EC) across multiple datasets for Transformer-XL(20M) (Dai et al., 2019), with BPC on the Enwik8 and Text8 test sets, and perplexity on the WikiText-103 and One Billion Word test sets. Lower values are better, with the best results highlighted in **bold**.

Size	Dataset	USMoE	TC	EC	SoftMoE
10M	Cifar10	**89.6**$_{\pm 0.3}$	88.7$_{\pm 0.2}$	88.9$_{\pm 0.3}$	85.6$_{\pm 0.3}$
	Cifar100	**66.6**$_{\pm 0.5}$	65.4$_{\pm 0.5}$	65.7$_{\pm 0.4}$	61.4$_{\pm 0.3}$
	STL-10	**66.7**$_{\pm 0.4}$	66.4$_{\pm 0.1}$	66.1$_{\pm 0.4}$	65.4$_{\pm 0.2}$
	SVHN	**95.6**$_{\pm 0.1}$	95.1$_{\pm 0.1}$	95.0$_{\pm 0.1}$	94.8$_{\pm 0.1}$
	ImageNet-1K	**60.2**$_{\pm 0.1}$	56.6$_{\pm 0.5}$	56.2$_{\pm 0.4}$	46.8$_{\pm 0.6}$
110M	Cifar10	**91.5**$_{\pm 0.5}$	85.7$_{\pm 0.8}$	87.4$_{\pm 0.7}$	80.3$_{\pm 1.0}$
	Cifar100	**67.3**$_{\pm 0.5}$	55.5$_{\pm 2.8}$	66.2$_{\pm 0.9}$	42.9$_{\pm 1.4}$
	STL-10	**66.2**$_{\pm 0.1}$	64.4$_{\pm 0.2}$	65.5$_{\pm 0.4}$	63.9$_{\pm 1.2}$
	SVHN	**96.1**$_{\pm 0.1}$	94.5$_{\pm 0.4}$	93.2$_{\pm 0.2}$	93.5$_{\pm 0.1}$
	ImageNet-1K	**73.5**$_{\pm 0.4}$	72.0$_{\pm 0.4}$	70.9$_{\pm 0.5}$	71.2$_{\pm 0.3}$
Avg.	All	**77.3**$_{\pm 0.3}$	74.4$_{\pm 1.5}$	75.5$_{\pm 1.2}$	70.6$_{\pm 1.5}$

Table 4. Accuracy of VIT-MoE evaluated on vision classification tasks. Each method is evaluated 3 times, reporting the mean and standard deviation. Higher is better, the best results are in bold.

4. Experiment Results

More Flexible Compute:

- **Fractional Top-k Routing**: USMoE supports *non-integer top-k* budgets. With k=3.5, USMoE already outperforms Token Choice with (k=8), offering a new way to reduce computation without sacrificing performance.

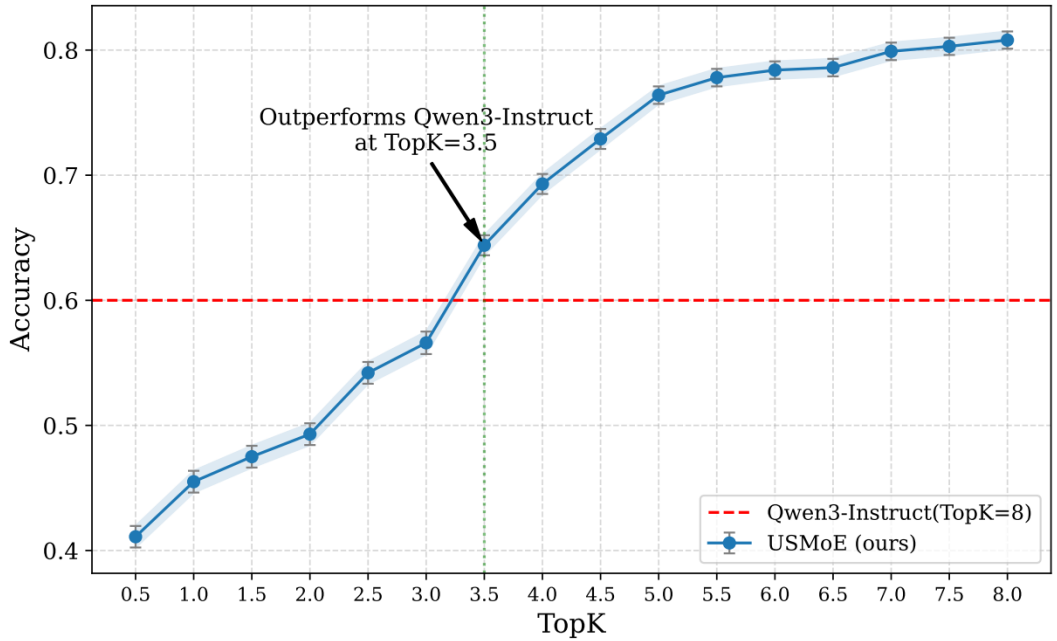


Figure 4. Demonstration of the flexibility of USMoE in supporting both fractional and non-fractional Top-K routing, compared with Token Choice, on Qwen3-30B-A3B-Instruct (abbreviated as Qwen3-Instruct) evaluated on the BoolQ dataset under corrupted settings. USMoE outperforms the original Qwen3-30B-A3B-Instruct starting from Top-$K = 3.5$, enabling reduced computational cost while maintaining competitive performance. Best viewed in color.

4. Experiment Results

More Explanatory:

- **Improved Explainability:** USMoE provides clearer interpretability by mapping clusters of tokens to groups of experts that share similar semantic meaning.

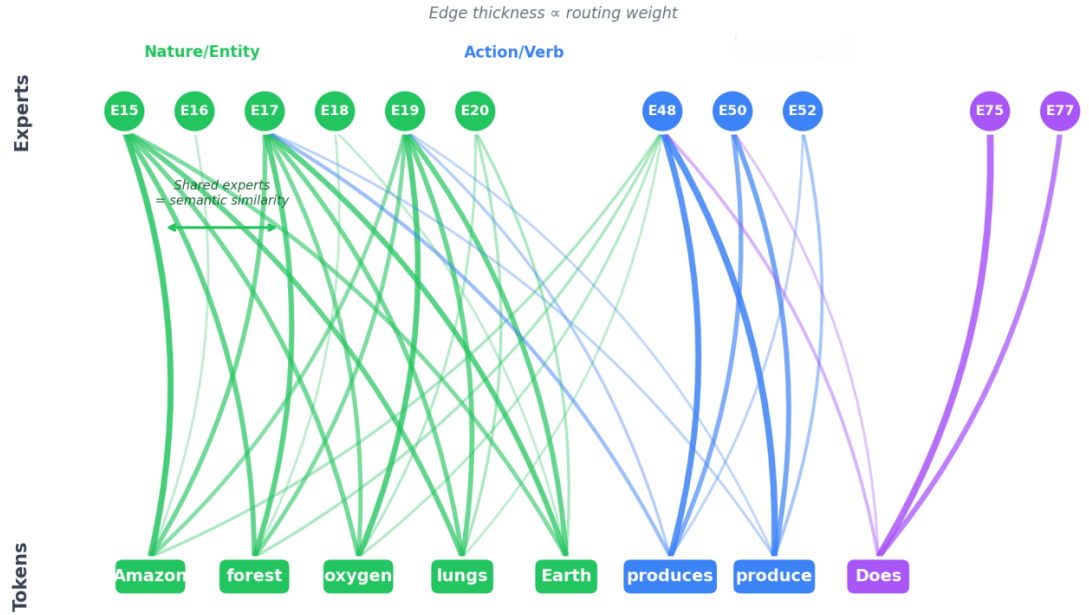


Figure 5. Explainability comparison between USMoE and the conventional view on the BoolQ dataset using Qwen3-30B-A3B-Instruct. Under USMoE, tokens routed to the same experts exhibit coherent semantic characteristics, while experts that attend to similar token sets tend to correspond to interpretable sub-domains (e.g., "Nature" experts spanning Expert 15 to Expert 20). This unified perspective reveals structured token–expert relationships that are less apparent under conventional routing formulations. Best viewed in color.

5. Conclusion

- **Conventional TC/EC routing is suboptimal:** Fixed expert selection limits routing quality and reduces flexibility in controlling model computation.
- **USMoE improves expert selection:** USMoE provides stronger routing from both theoretical and empirical perspectives.
- **USMoE enables flexible computation:** Expert selection can be dynamically adjusted under a constrained computation budget.
- **Unified perspective for MoE:** Viewing expert selection jointly over tokens and experts opens a new direction for understanding and improving Mixture-of-Experts models.

End