Adaptive Gating in Mixture-of-Experts based Language Models

Jiamin Li[1], Qiang Su[1], Yitao Yang[2], Yimin Jiang, Cong Wang[1], Hong Xu[2]

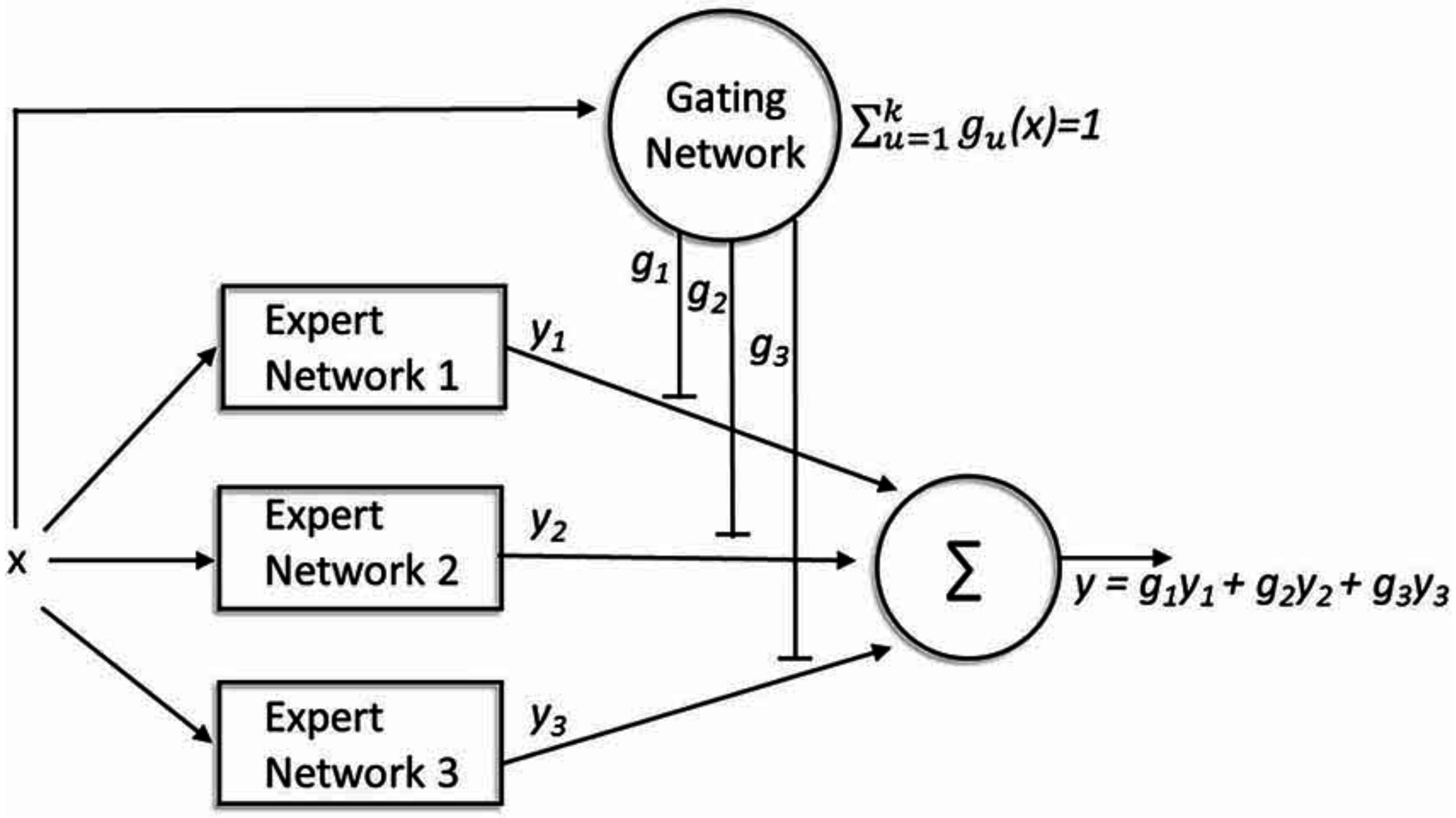
[1]City University of Hong Kong, [2]The Chinese University of Hong Kong

EMNLP 2023

香港城市大學 City University of Hong Kong

香港中文大學 The Chinese University of Hong Kong

Background

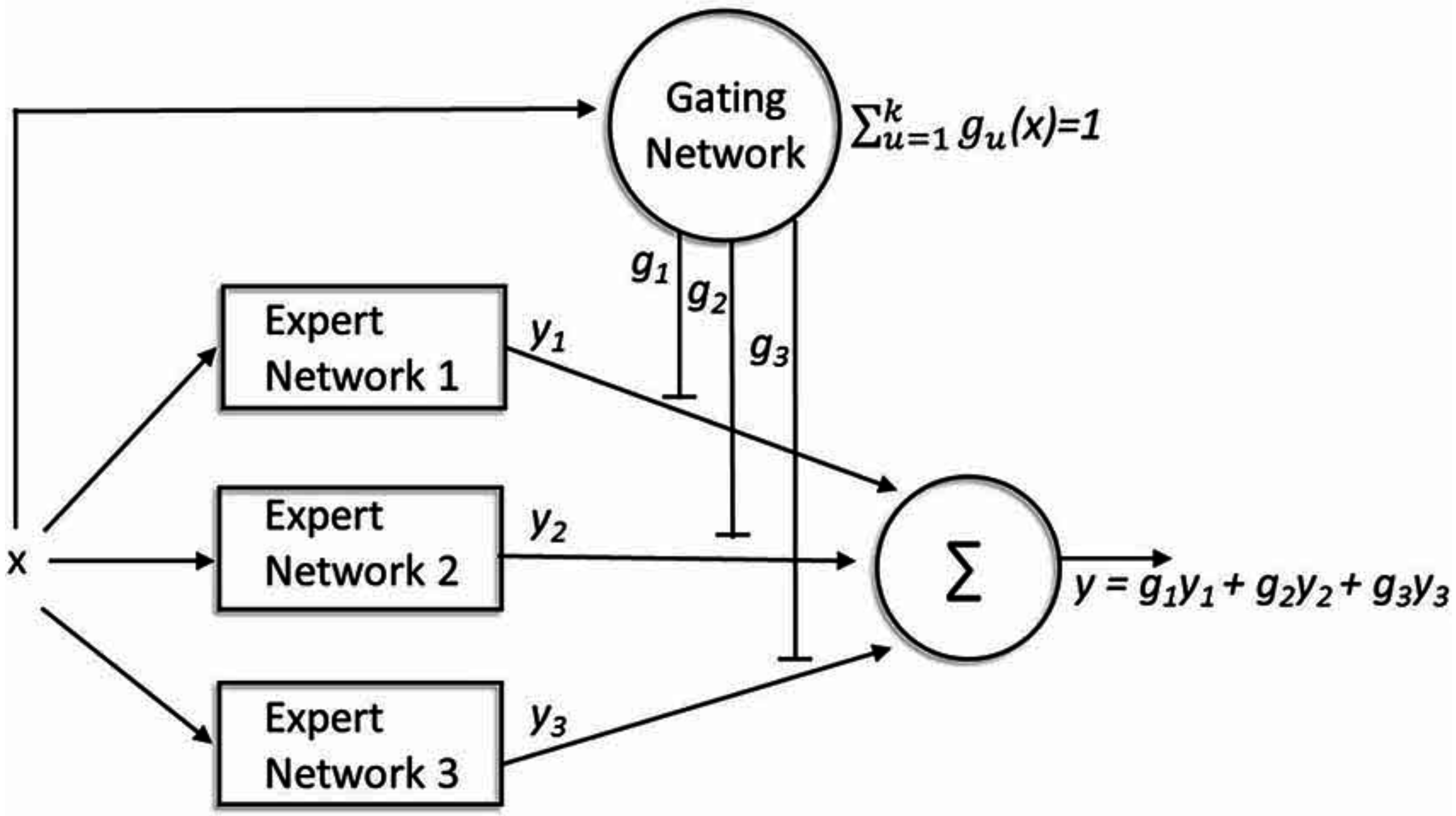


MoE architecture

An ensemble of experts.

Figure credit to Anatomical and Functional Plasticity in Early Blind ndividuals and the Mixture of Experts Architecture

Background



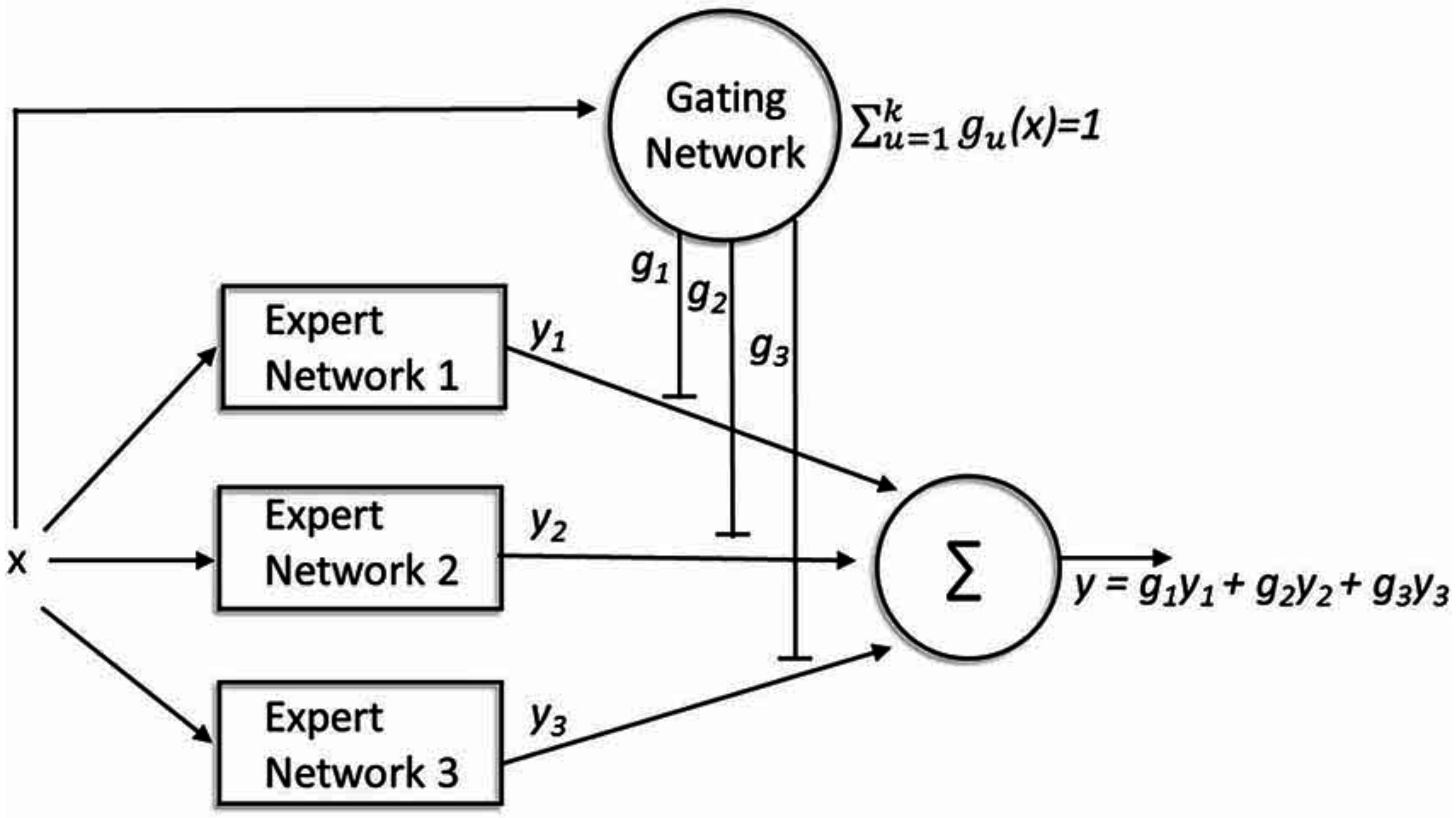
MoE architecture

An ensemble of experts.

- *Sparsely-activated* MoE: each input selects just a few (1 or 2) experts for processing
- Benefit: sub-linear scaling of FLOPS with model size

Figure credit to Anatomical and Functional Plasticity in Early Blind Individuals and the Mixture of Experts Architecture

Background



MoE architecture

An ensemble of experts.

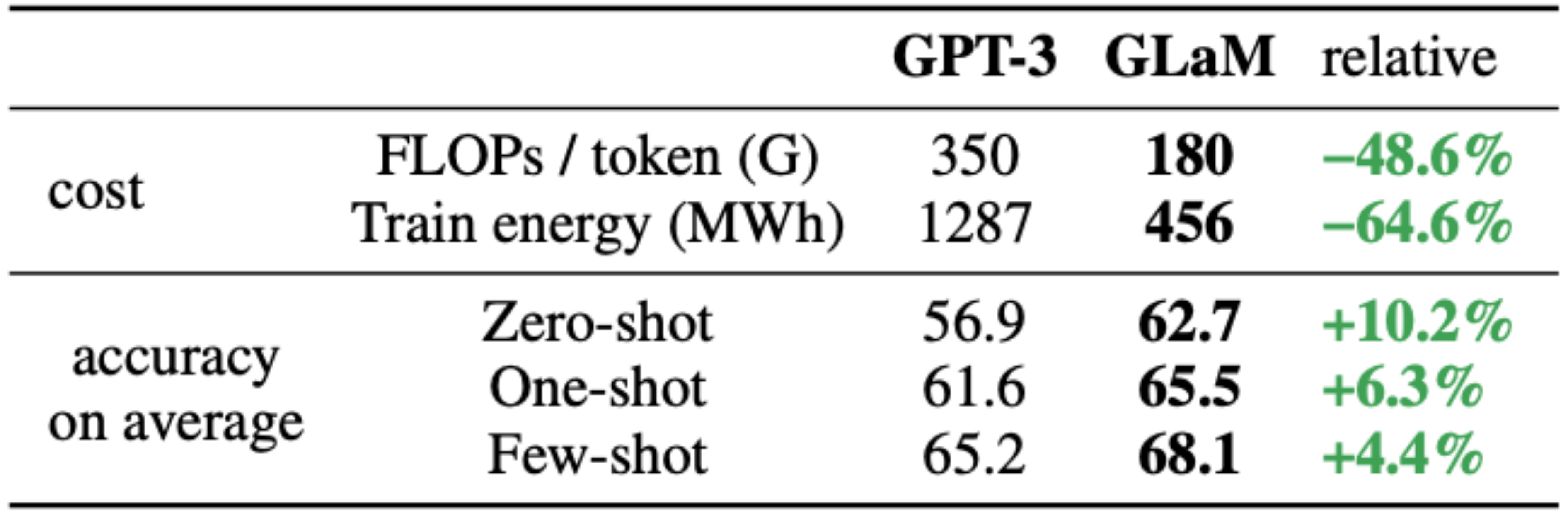
- *Sparsely-activated* MoE: each input selects just a few (1 or 2) experts for processing
- Benefit: sub-linear scaling of FLOPS with model size

Massive model parameters with constant computation cost.

Figure credit to Anatomical and Functional Plasticity in Early Blind Individuals and the Mixture of Experts Architecture

Potential of MoE in Transformer Models

- GLaM by Google
  - GLaM outperforms GPT-3 on 29 tasks



		GPT-3	GLaM	relative
cost	FLOPs / token (G)	350	**180**	−48.6%
	Train energy (MWh)	1287	**456**	−64.6%
accuracy on average	Zero-shot	56.9	**62.7**	+10.2%
	One-shot	61.6	**65.5**	+6.3%
	Few-shot	65.2	**68.1**	+4.4%

- DeepSpeed MoE models
  - Model quality: 6.7B-parameter dense = 1.3B-parameter MoE - 128
  - Training compute reduction of 5x

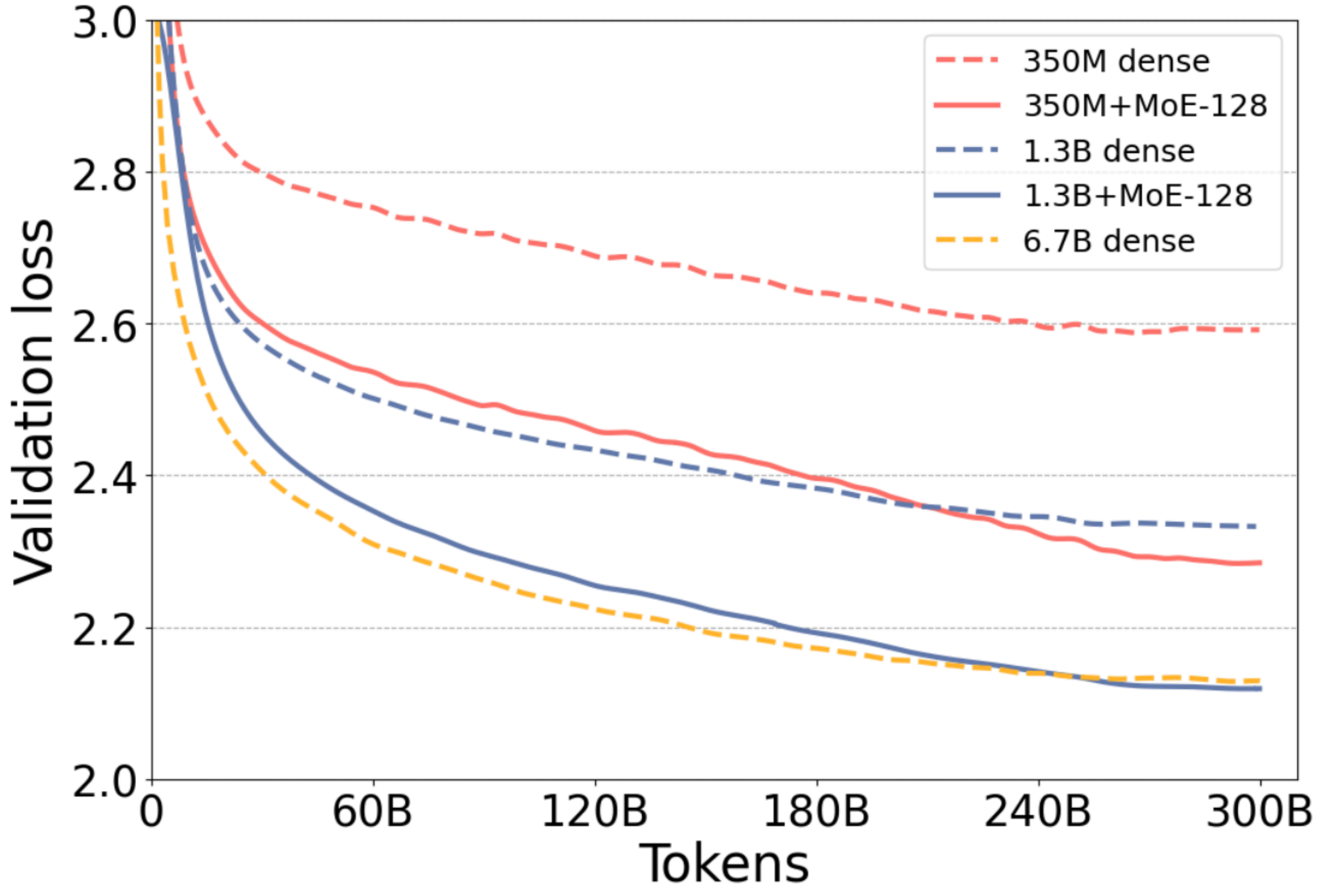


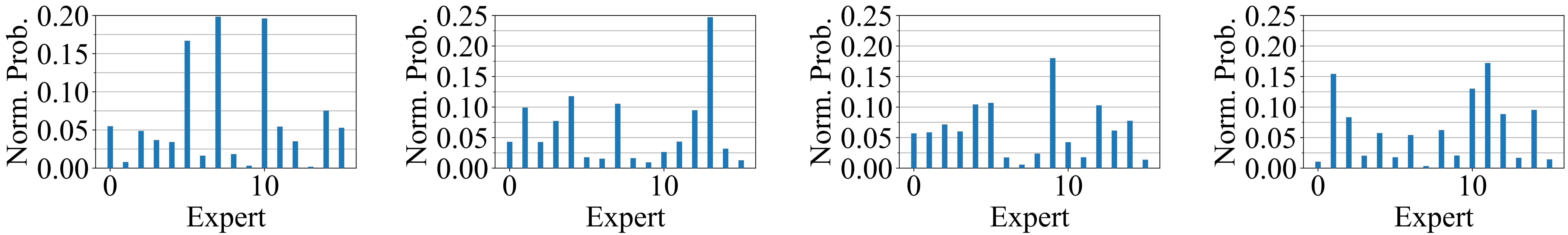
Figure credit to GLaM and DeepSpeed MoE.

Observation & Motivation

- Existing MoE models, adopts a fixed gating policy (i.e. Top-2 gating in training).

Observation & Motivation

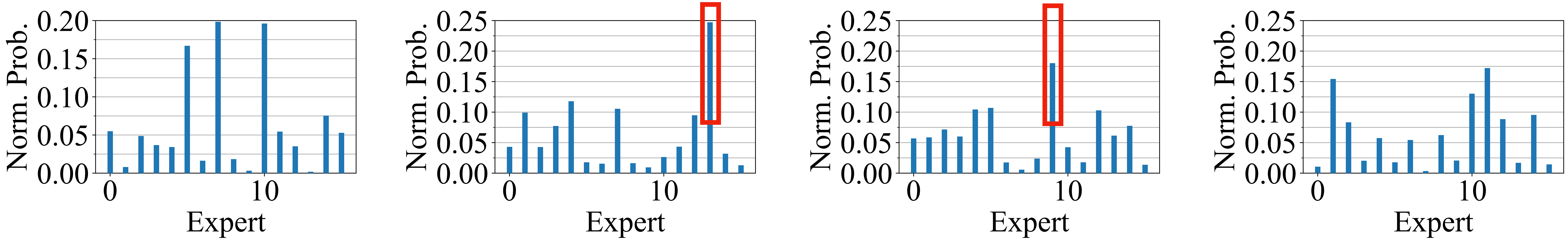
- Existing MoE models, adopts a fixed gating policy (i.e. Top-2 gating in training).



Softmax activations retrieved from MoE gate of four tokens.

Observation & Motivation

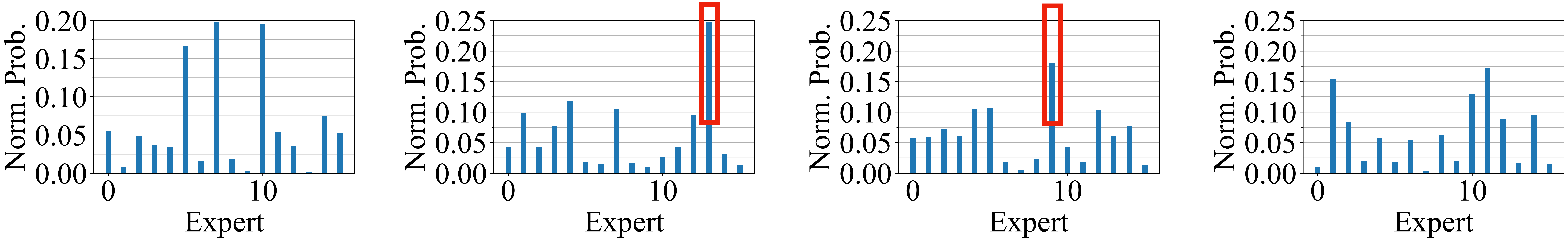
- Existing MoE models, adopts a fixed gating policy (i.e. Top-2 gating in training).



Softmax activations retrieved from MoE gate of four tokens.

Observation & Motivation

- Existing MoE models, adopts a fixed gating policy (i.e. Top-2 gating in training).

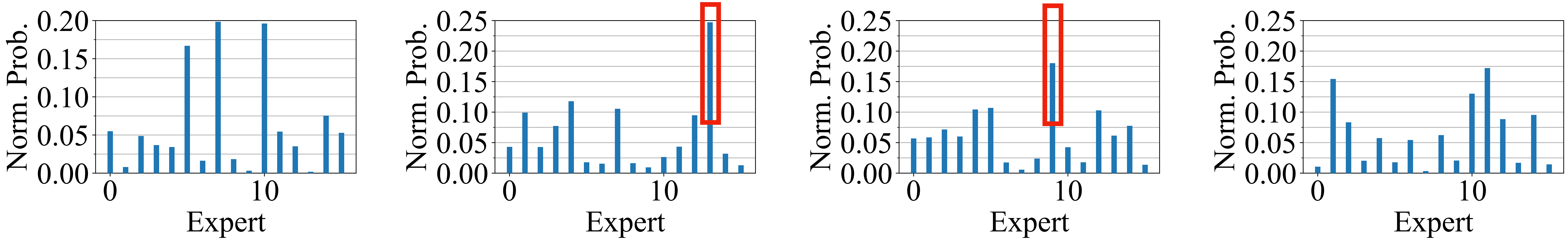


Softmax activations retrieved from MoE gate of four tokens.

Significantly-biased distribution accounts for at least 55% of all the tokens

Observation & Motivation

- Existing MoE models, adopts a fixed gating policy (i.e. Top-2 gating in training).



Softmax activations retrieved from MoE gate of four tokens.

Significantly-biased distribution accounts for at least 55% of all the tokens

- MoE experts specialize in different linguistic aspects.
- Many tokens can be effectively handled by a single expert during the training stage

Adaptive Gating

- Control the number of experts handling each token to reduce training step time

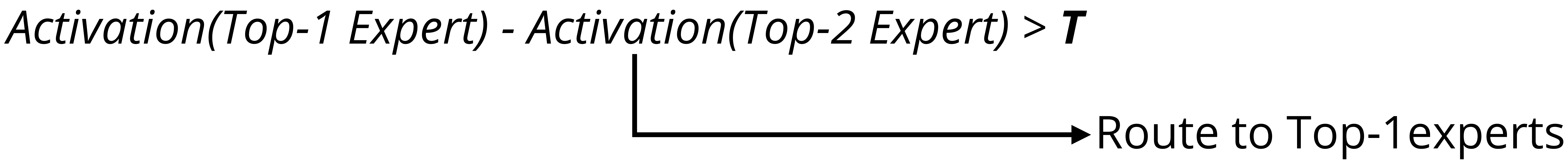
Adaptive Gating

- Control the number of experts handling each token to reduce training step time

Activation(Top-1 Expert) - Activation(Top-2 Expert) > ***T***

Adaptive Gating

- Control the number of experts handling each token to reduce training step time



Activation(Top-1 Expert) - Activation(Top-2 Expert) > ***T***

Route to Top-1experts

Adaptive Gating

- Control the number of experts handling each token to reduce training step time

Activation(Top-1 Expert) - Activation(Top-2 Expert) > ***T***

→ Route to Top-1experts

- Load balancing loss: impose the soft load balancing constraints on the top-1 gating decisions.

$$L_i = E_i \sum_{e \in E} f_e^1 p_e$$

Adaptive Gating

- Control the number of experts handling each token to reduce training step time

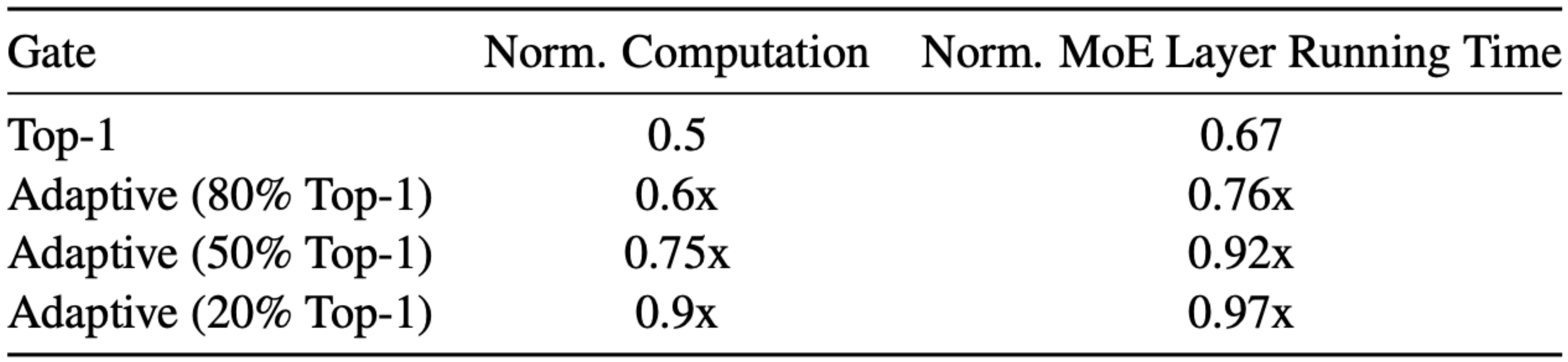
Activation(Top-1 Expert) - Activation(Top-2 Expert) > ***T***

→ Route to Top-1experts

- Load balancing loss: impose the soft load balancing constraints on the top-1 gating decisions.

$$L_i = E_i \sum_{e \in E} f_e^{\boxed{1}} p_e$$

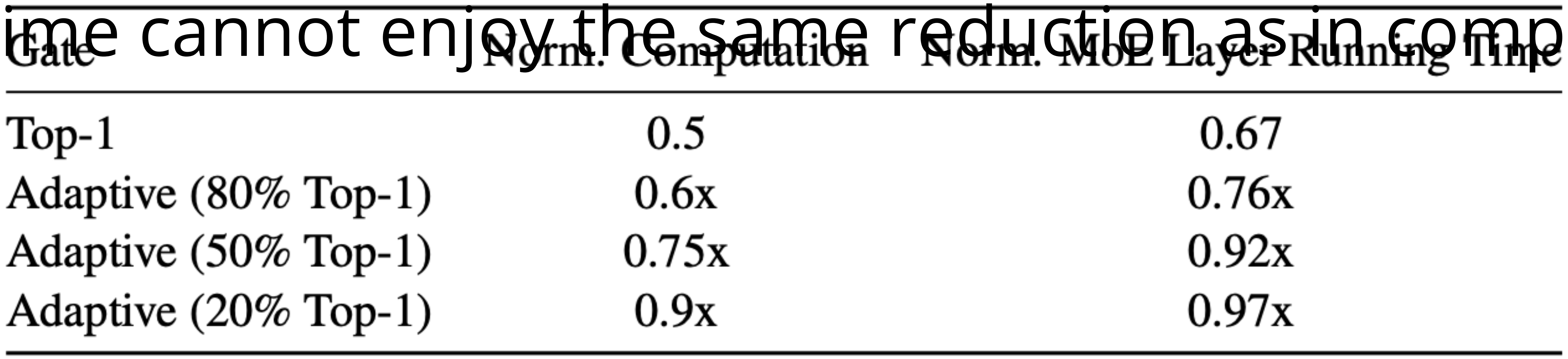
Inefficient Training



Gate	Norm. Computation	Norm. MoE Layer Running Time
Top-1	0.5	0.67
Adaptive (80% Top-1)	0.6x	0.76x
Adaptive (50% Top-1)	0.75x	0.92x
Adaptive (20% Top-1)	0.9x	0.97x

Inefficient Training

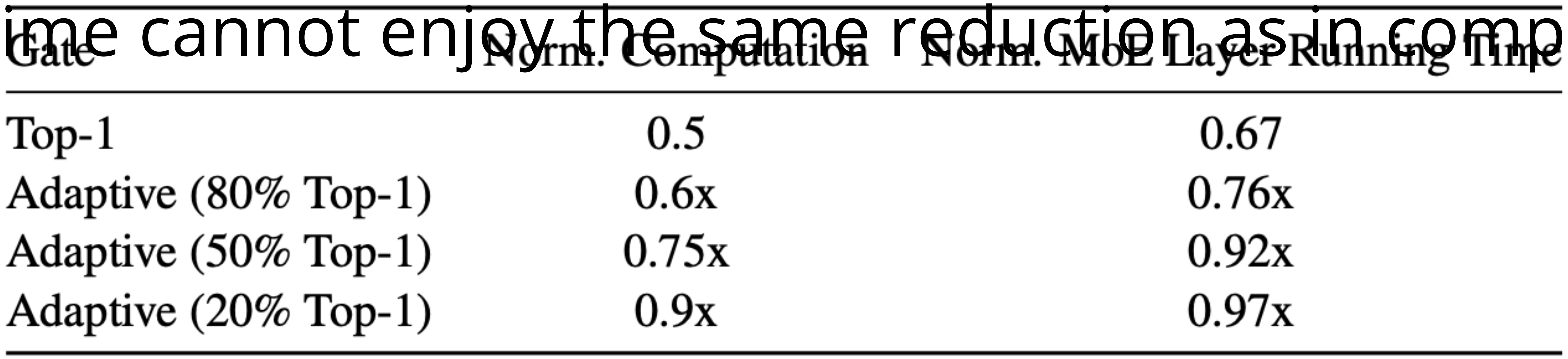
- Training step time cannot enjoy the same reduction as in computation.



Gate	Norm. Computation	Norm. MoE Layer Running Time
Top-1	0.5	0.67
Adaptive (80% Top-1)	0.6x	0.76x
Adaptive (50% Top-1)	0.75x	0.92x
Adaptive (20% Top-1)	0.9x	0.97x

Inefficient Training

- Training step time cannot enjoy the same reduction as in computation.



Gate	Norm. Computation	Norm. MoE Layer Running Time
Top-1	0.5	0.67
Adaptive (80% Top-1)	0.6x	0.76x
Adaptive (50% Top-1)	0.75x	0.92x
Adaptive (20% Top-1)	0.9x	0.97x

Inefficient Training

- Mismatch in the data processing granularity between the MoE experts and the Attention layer.
  - MoE expert -> single tokens
  - Attention layer -> complete sequence
- Training step time cannot enjoy the same reduction as in computation.

Curriculum Learning

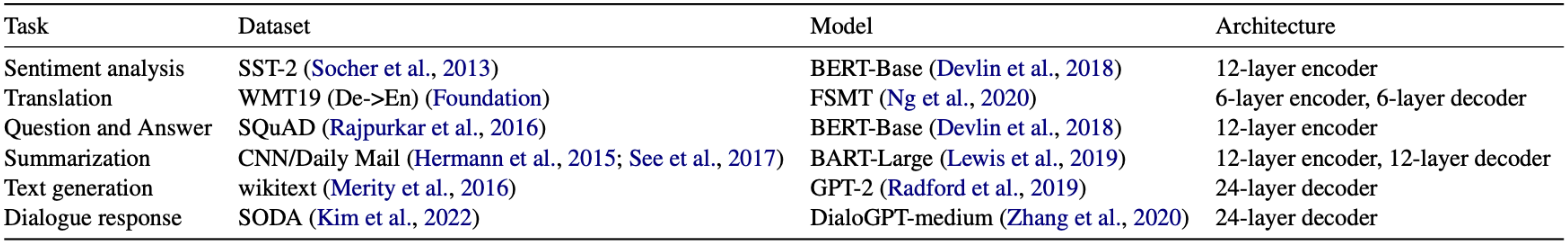
- Mismatch in the data processing granularity between the MoE experts and the Attention layer.
  - MoE expert -> single tokens
  - Attention layer -> complete sequence
- Training step time cannot enjoy the same reduction as in computation.

Curriculum Learning

- Mismatch in the data processing granularity between the MoE experts and the Attention layer.
  - MoE expert -> single tokens
  - Attention layer -> complete sequence
- Training step time cannot enjoy the same reduction as in computation.

- Process *easier* sequences at the initial stages.
- The number of experts required by each token can be an indicator of the token complexity.
- Complexity vector of a sequence: $C_d = [r_0^d, r_1^d, \ldots r_L^d]$

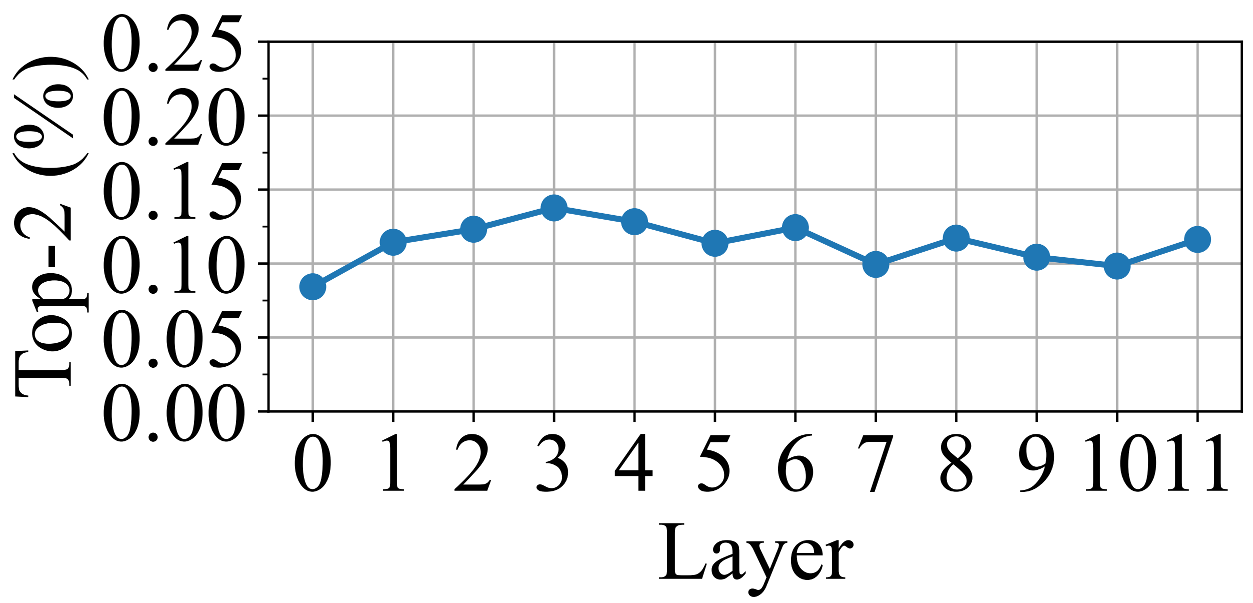
Evaluation



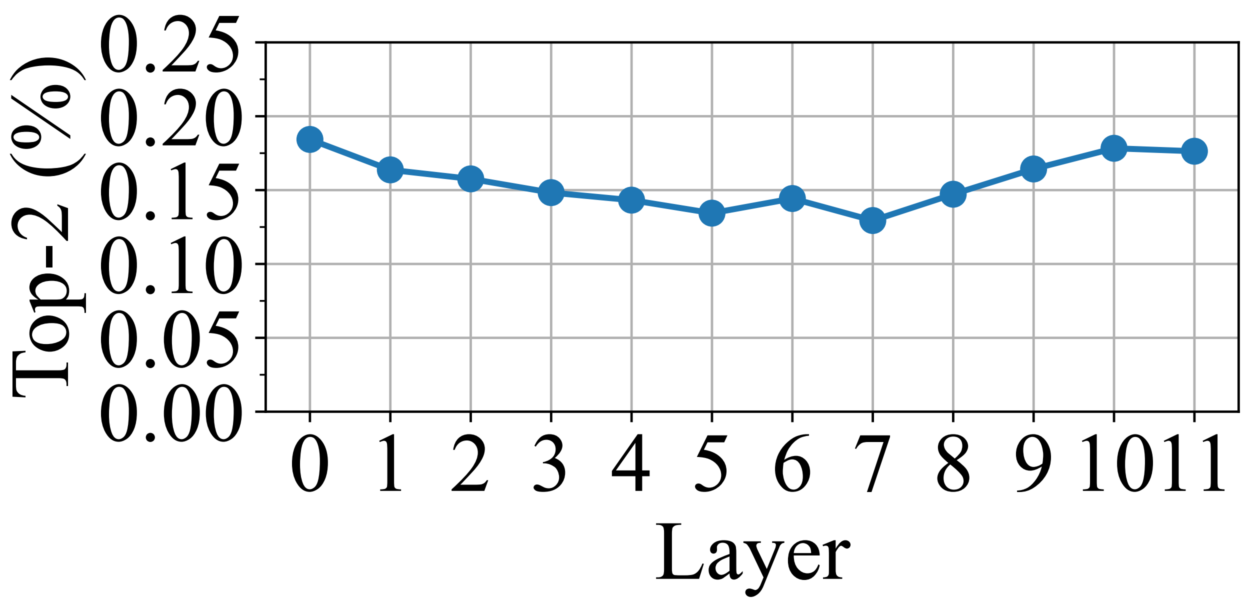
Task	Dataset	Model	Architecture
Sentiment analysis	SST-2 (Socher et al., 2013)	BERT-Base (Devlin et al., 2018)	12-layer encoder
Translation	WMT19 (De->En) (Foundation)	FSMT (Ng et al., 2020)	6-layer encoder, 6-layer decoder
Question and Answer	SQuAD (Rajpurkar et al., 2016)	BERT-Base (Devlin et al., 2018)	12-layer encoder
Summarization	CNN/Daily Mail (Hermann et al., 2015; See et al., 2017)	BART-Large (Lewis et al., 2019)	12-layer encoder, 12-layer decoder
Text generation	wikitext (Merity et al., 2016)	GPT-2 (Radford et al., 2019)	24-layer decoder
Dialogue response	SODA (Kim et al., 2022)	DialoGPT-medium (Zhang et al., 2020)	24-layer decoder

- Testbed
  - 8 A100 GPUs, each with 40 GB memory.
  - Data and expert parallel is used for distributed training.
  - □n terms of hyperparameters and model architecture, we adopt the default configurations established in the existing models

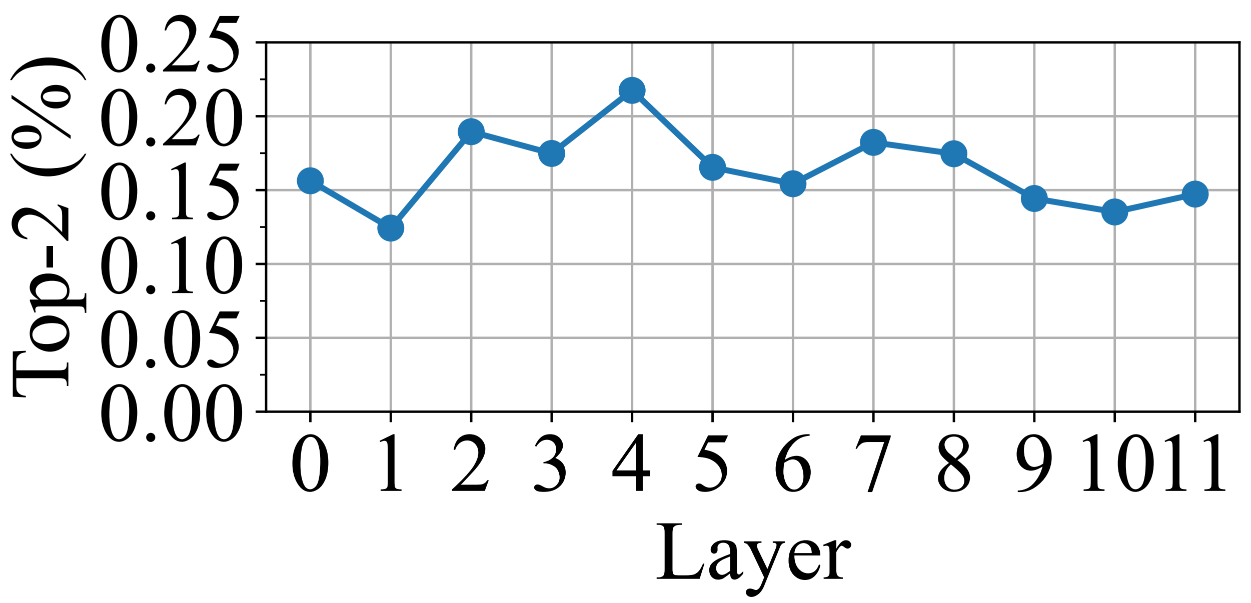
Evaluation



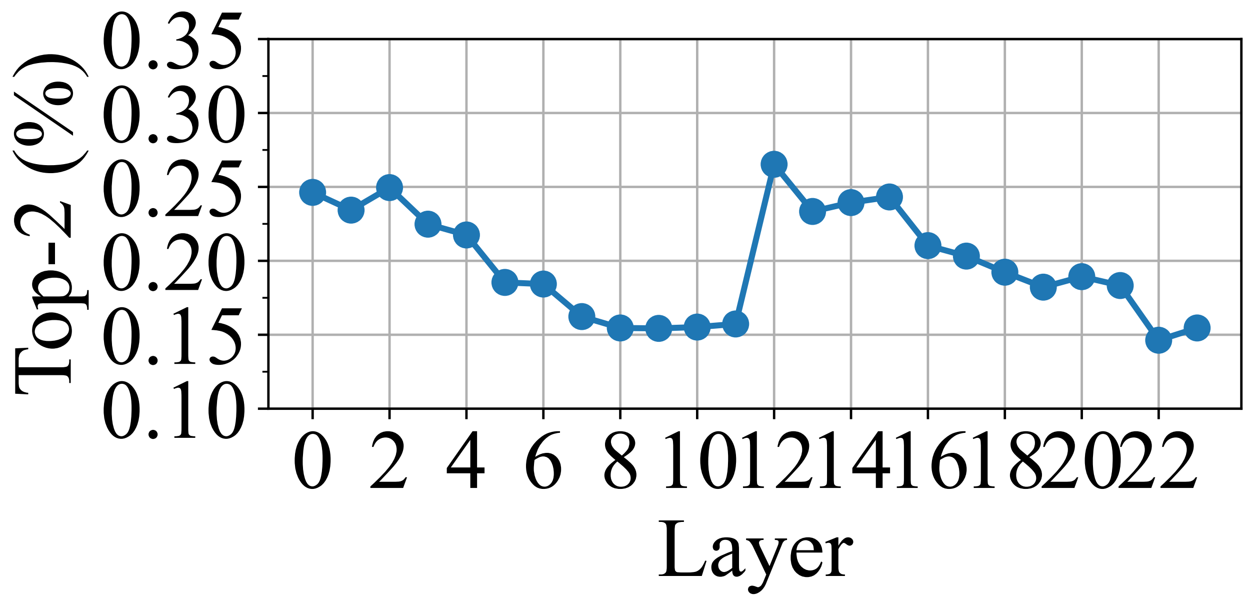
Sentiment Analysis



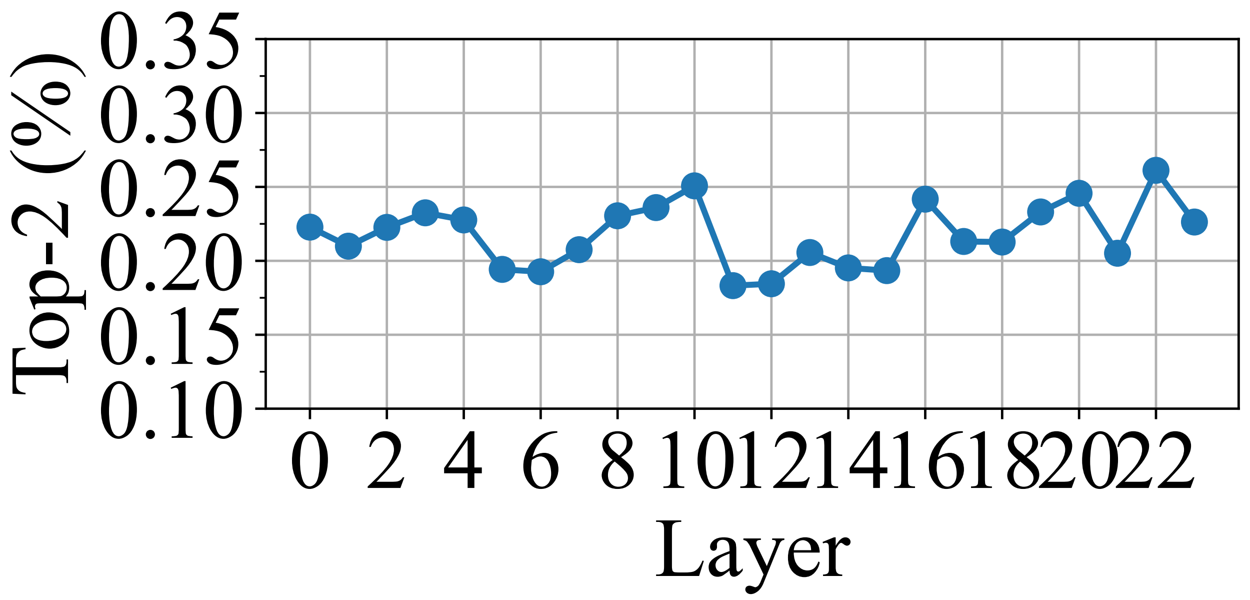
Translation (En->De)



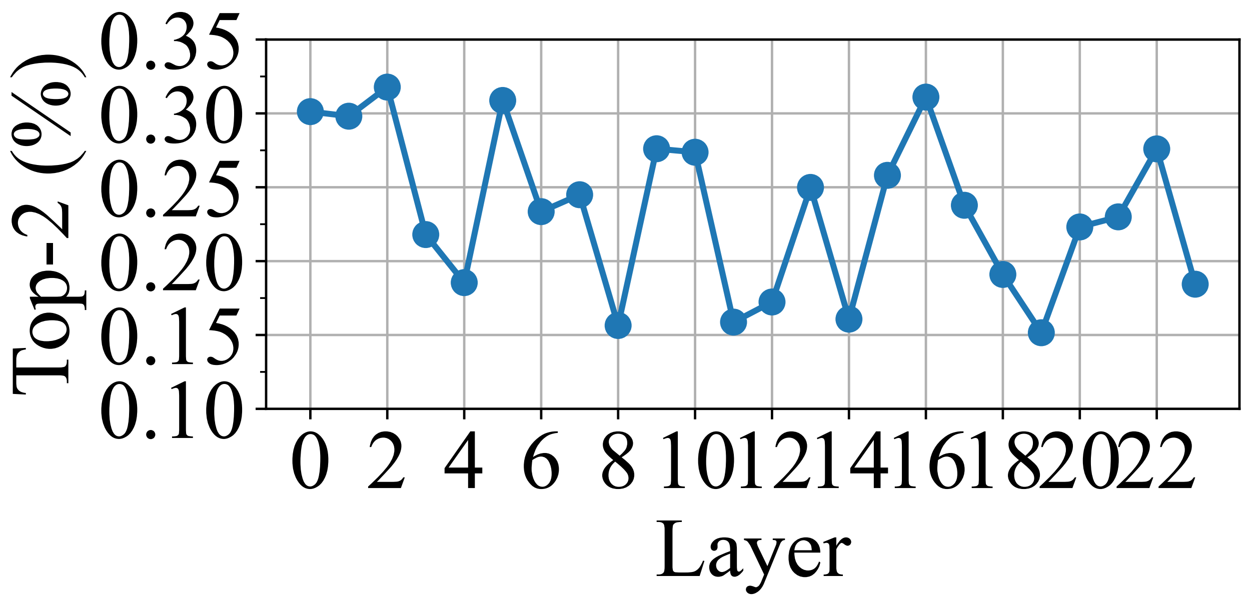
Q&A



Summarisation

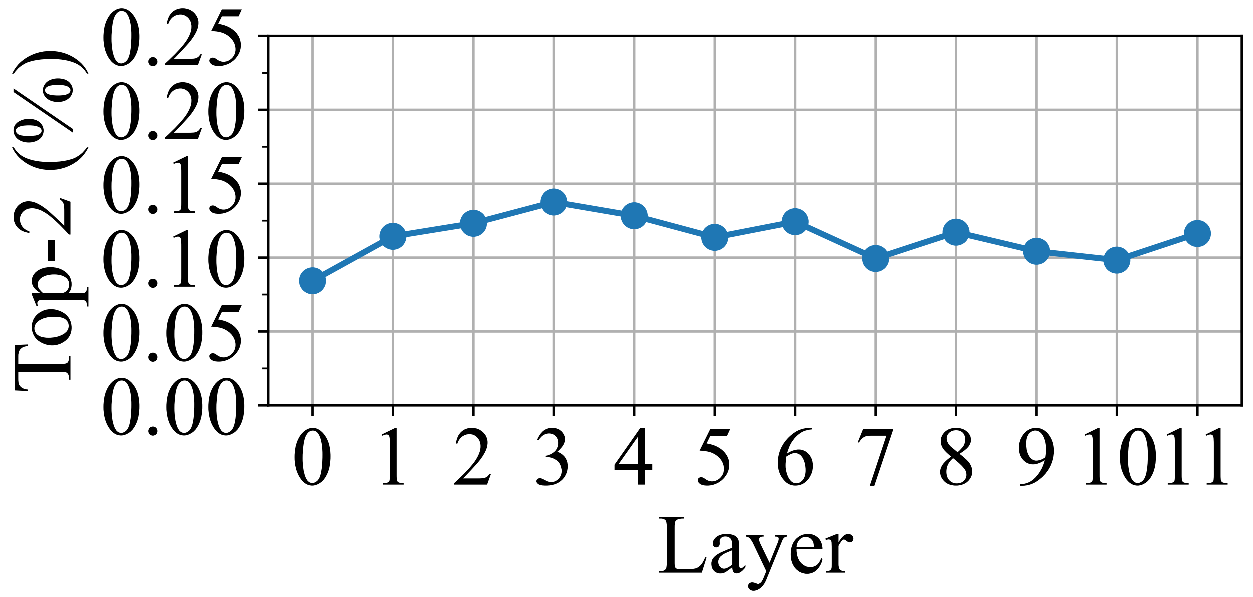


Text Generation

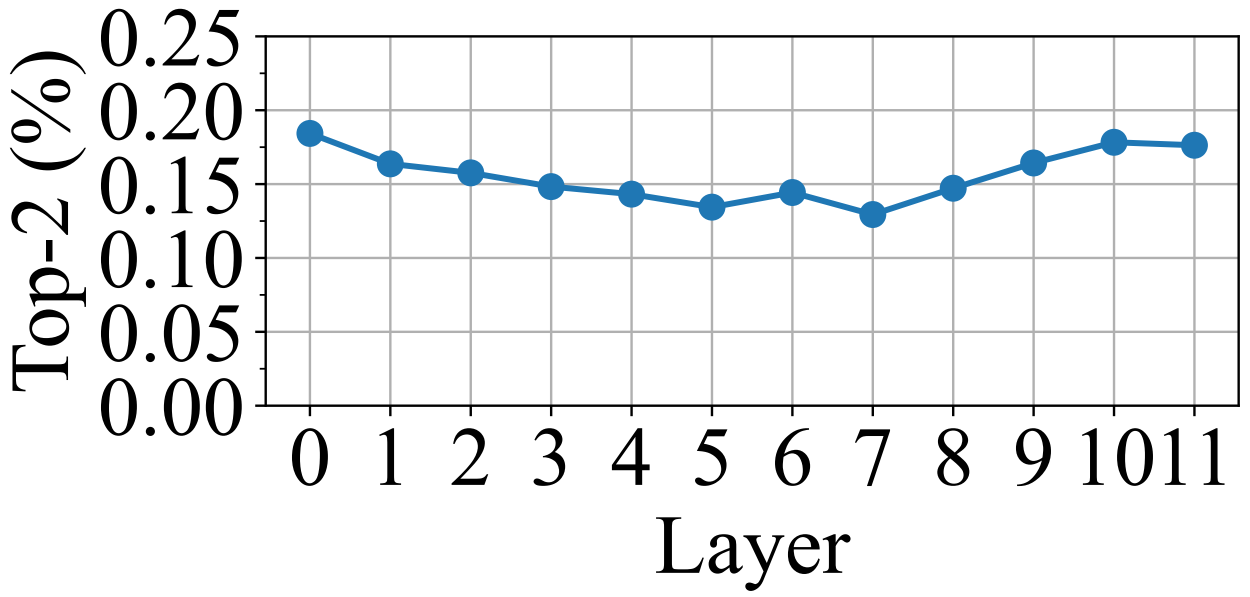


Dialogue Response

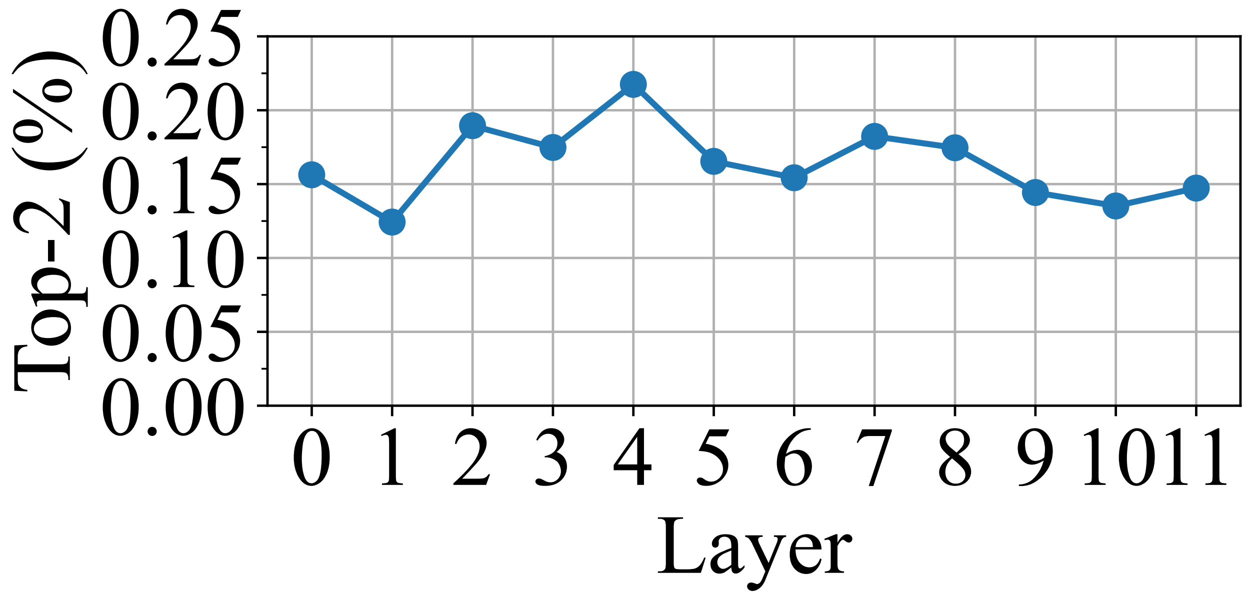
Evaluation



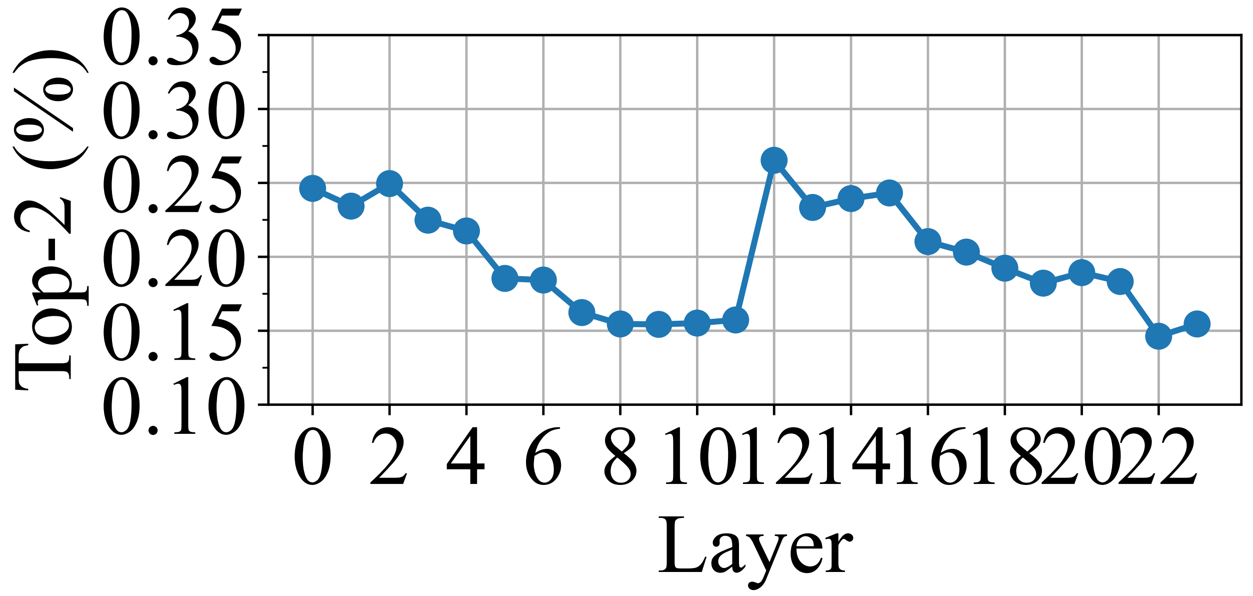
Sentiment Analysis



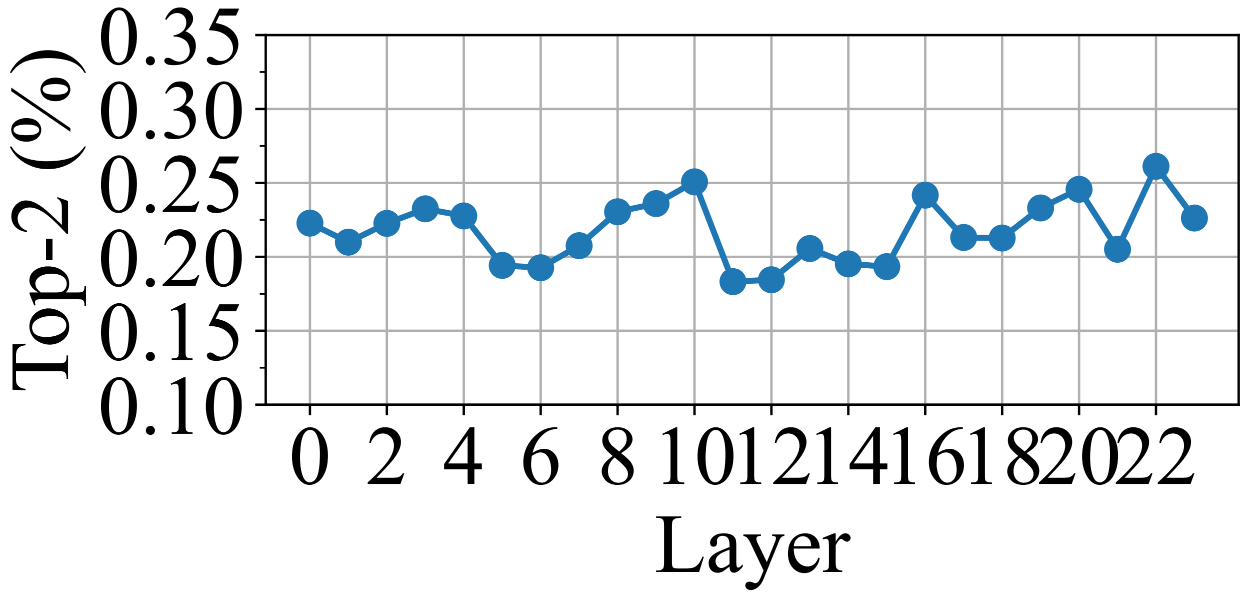
Translation (En->De)



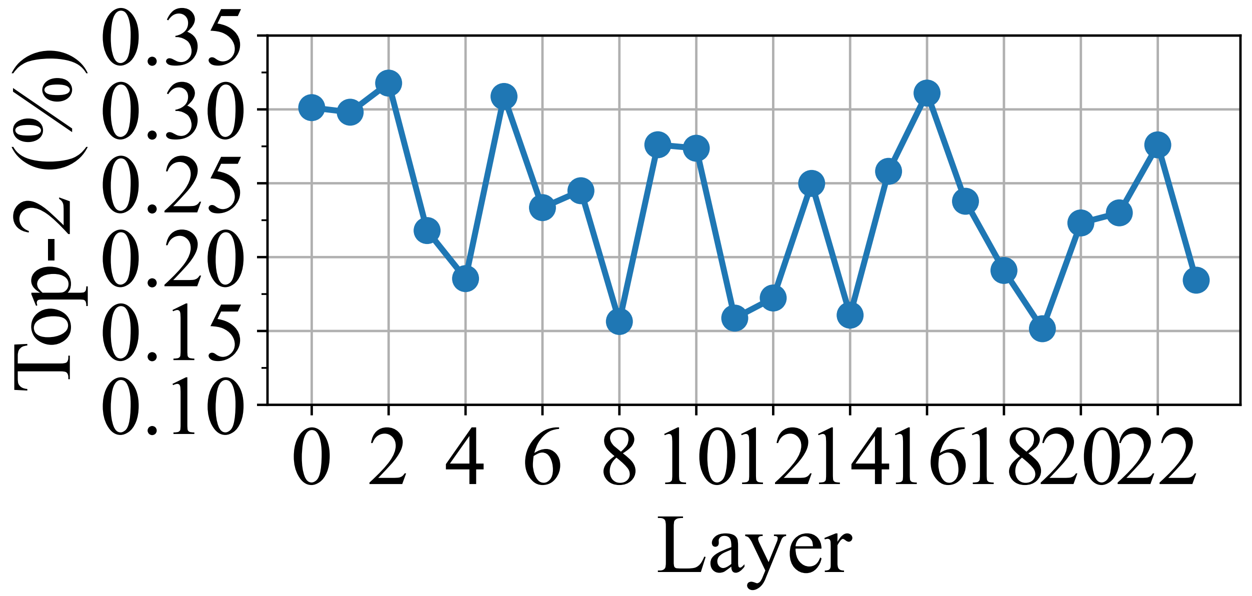
Q&A



Summarisation



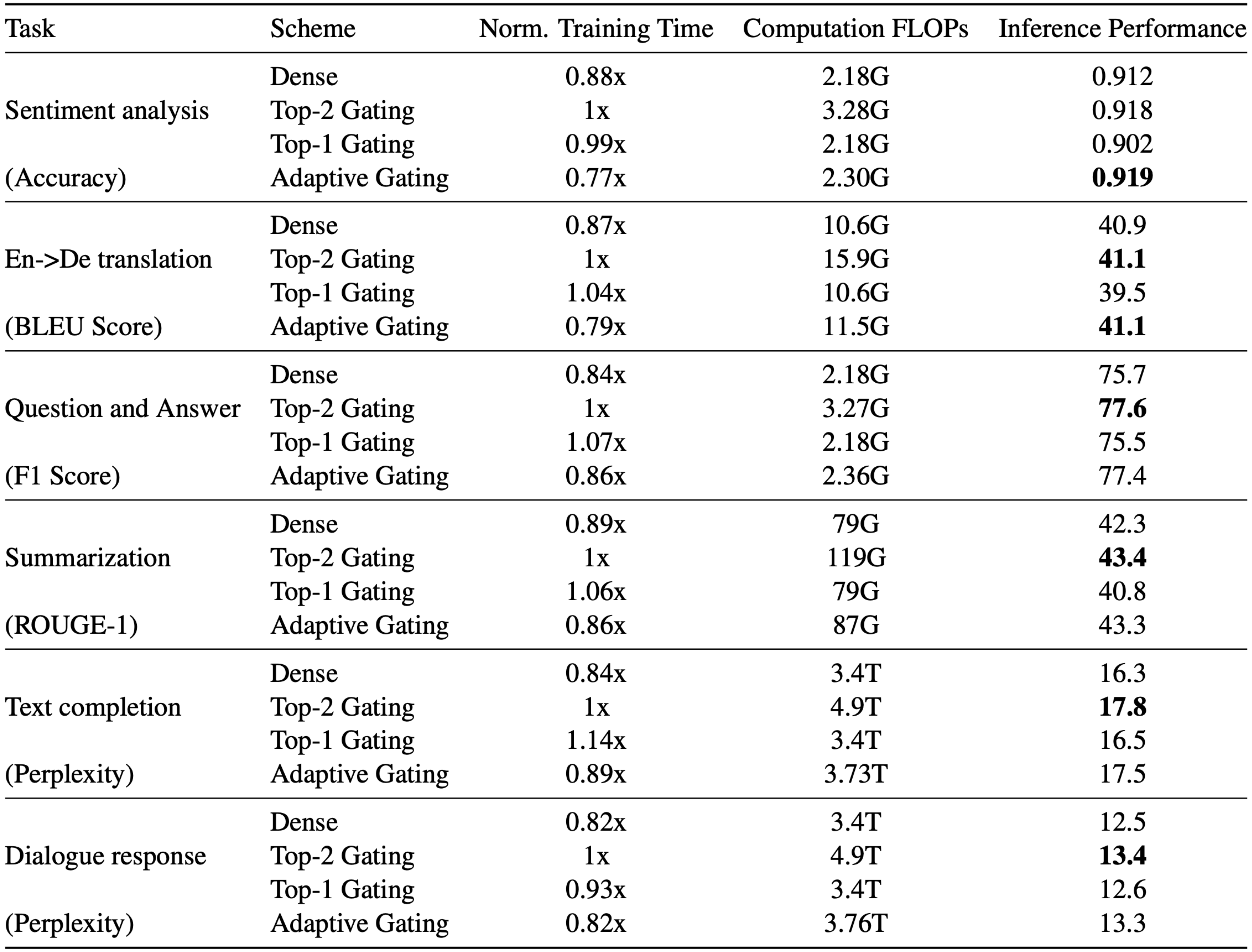
Text Generation



Dialogue Response

About ~25% of the tokens are routed to two experts.

Evaluation



Task	Scheme	Norm. Training Time	Computation FLOPs	Inference Performance
	Dense	0.88x	2.18G	0.912
Sentiment analysis	Top-2 Gating	1x	3.28G	0.918
	Top-1 Gating	0.99x	2.18G	0.902
(Accuracy)	Adaptive Gating	0.77x	2.30G	**0.919**
	Dense	0.87x	10.6G	40.9
En->De translation	Top-2 Gating	1x	15.9G	**41.1**
	Top-1 Gating	1.04x	10.6G	39.5
(BLEU Score)	Adaptive Gating	0.79x	11.5G	**41.1**
	Dense	0.84x	2.18G	75.7
Question and Answer	Top-2 Gating	1x	3.27G	**77.6**
	Top-1 Gating	1.07x	2.18G	75.5
(F1 Score)	Adaptive Gating	0.86x	2.36G	77.4
	Dense	0.89x	79G	42.3
Summarization	Top-2 Gating	1x	119G	**43.4**
	Top-1 Gating	1.06x	79G	40.8
(ROUGE-1)	Adaptive Gating	0.86x	87G	43.3
	Dense	0.84x	3.4T	16.3
Text completion	Top-2 Gating	1x	4.9T	**17.8**
	Top-1 Gating	1.14x	3.4T	16.5
(Perplexity)	Adaptive Gating	0.89x	3.73T	17.5
	Dense	0.82x	3.4T	12.5
Dialogue response	Top-2 Gating	1x	4.9T	**13.4**
	Top-1 Gating	0.93x	3.4T	12.6
(Perplexity)	Adaptive Gating	0.82x	3.76T	13.3

Adaptive gating reduces at most 22.5% training time while maintaining inference quality.