

CSCE 689: Special Topics in Advanced NLP

Lecture 10: Model Efficiency

Kuan-Hao Huang

Fall 2026



Literature Review

- Due: Oct 1
- Page limit: 4-5 pages
- The literature review should cover the 4 suggested papers and at least 10 additional papers published after 2024 (including 2024)
- The review should include:
 - Problem definition and importance of the topic.
 - Background and relevant context from previous works (with additional references, if applicable).
 - A comparative analysis of key methodologies and findings.
 - A critical evaluation of the strengths, limitations, and gaps in the literature.
 - A discussion of open problems and directions for future research.

Course Project – Proposal

- Due: 10/9
- Page limit: 2 pages (excluding references)
- Format: [ACL style](#)
- The proposal should include
 - Introduction to the topic you choose and problem definition
 - Related literature and overview of existing progress and challenges
 - Proposed solutions, novelty, and expected contributions
 - Evaluation metrics
 - Planned implementation details, including dataset, models, codebases, etc.
 - Expected timeline

Course Project

- Sign-up
 - https://docs.google.com/spreadsheets/d/1f02G04qbhnNnsrTylZTFX_MxcNy6xTznYC5Tp8akdg/edit?usp=sharing
- The team and the topic can be **different** from topic study

Team	Topic	Member 1	Email 1	Member 2 (optional)	Email 2 (optional)
Example	Jailbreaking VLMs	First_name Last_name	nlp@tamu.edu	First_name Last_name	nlp@tamu.edu
1					
2					
3					
4					
5					
6					
7					
9					
10					
If you are looking for teammates					
	Name	Email	Potential Topics		
Example	Kuan-Hao Huang	khhuang@tamu.edu	Improving math reasoning for LLMs		

Invited Talks

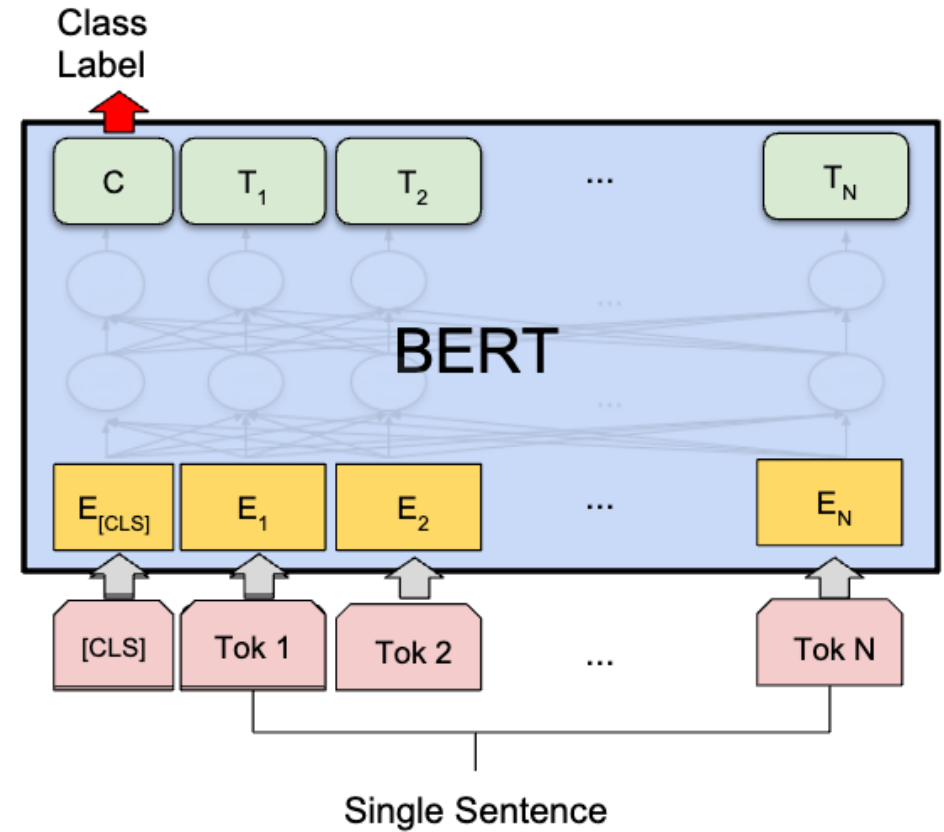
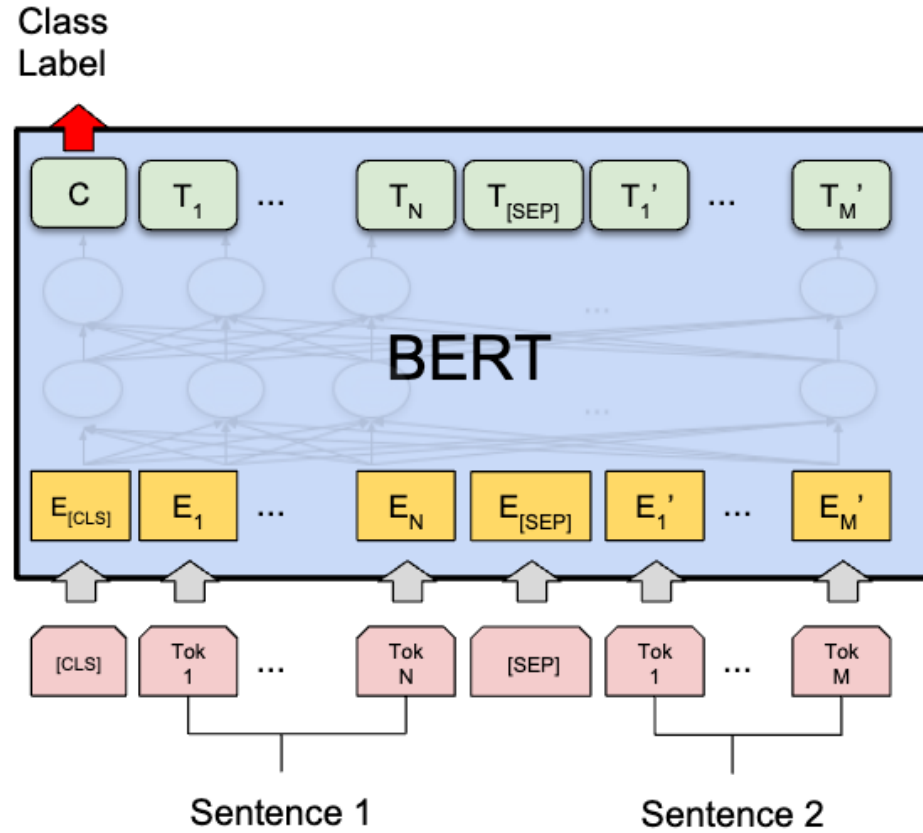
- **Date:** 10/5 online @ Zoom
 - <https://tamu.zoom.us/my/khhuang?pwd=oAdWOKVOCGPAPqDbJnVtktdW2AE6nb.1>
 - **Talk:** Deep Research is Still All About Retrieval
 - **Speaker:** Eugene Yang, Senior Research Scientist at HLTCOE, Johns Hopkins University
- **Date:** 11/23 online @ Zoom
 - <https://tamu.zoom.us/my/khhuang?pwd=oAdWOKVOCGPAPqDbJnVtktdW2AE6nb.1>
 - **Talk:** TBD
 - **Speaker:** Da Yin, Founding Member of Technical Staff at NeoCognition

Talk Abstract

Deep research applications that involve retrieving information and synthesizing comprehensive responses have become increasingly popular. While improvements in LLMs and AI agents have made deep research systems possible, the retrieval component in such systems is still a critical piece that has received less attention. Instead of retrieving relevant but repetitive information from different sources, finding unique pieces of relevant information is a key factor in improving the utility of the final responses. In this talk, I introduce a reproducible evaluation method for information coverage, discuss how information coverage in responses is bottlenecked by upstream retrieval, and present several optimization directions for coverage.

Pre-Trained Models Provide Good Initialization

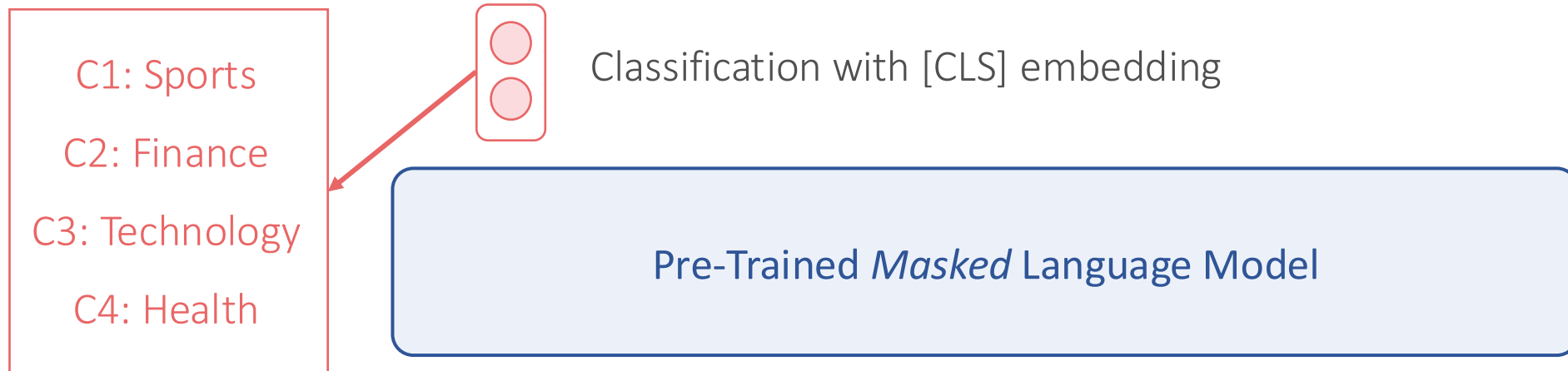
- Pre-training provides a **weight initialization** for continuing fine-tuning



Classification with [CLS] Embedding

Topic Classification

The Houston Rockets won an intense overtime game	Sports
Bitcoin hit a new all-time high this week	Finance
Tesla launched a new self-driving software update	Technology
Flu cases are rising in several major cities	Health



[CLS] The Houston Rockets won an intense overtime game

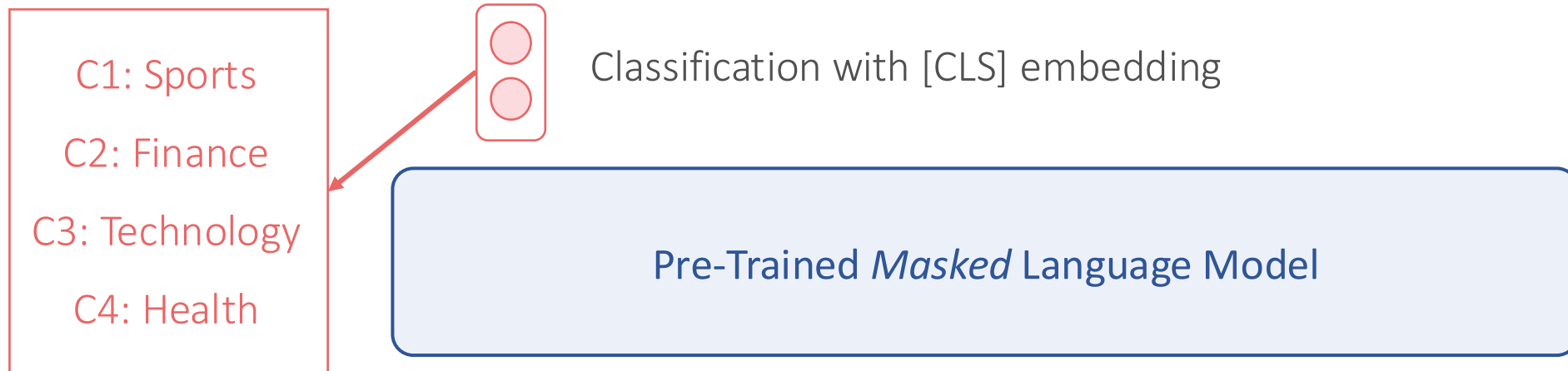
Parameter-Efficient Fine-Tuning (PEFT)

- Motivation
 - Do not fine-tune the whole model
 - Update only a small subset of parameters
 - Keep most parameters frozen
 - Achieve comparable performance to full fine-tuning
- Advantages
 - Lower GPU memory usage
 - Faster training
 - Reduced storage cost
 - Easier deployment across tasks

Classification with [CLS] Embedding

Topic Classification

The Houston Rockets won an intense overtime game	Sports
Bitcoin hit a new all-time high this week	Finance
Tesla launched a new self-driving software update	Technology
Flu cases are rising in several major cities	Health



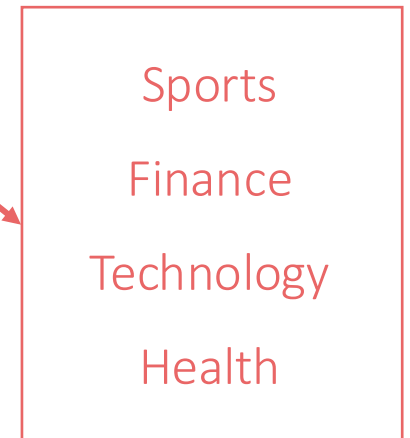
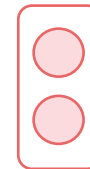
[CLS] The Houston Rockets won an intense overtime game

Classification with [MASK] Embedding

Topic Classification

The Houston Rockets won an intense overtime game	Sports
Bitcoin hit a new all-time high this week	Finance
Tesla launched a new self-driving software update	Technology
Flu cases are rising in several major cities	Health

Classification with [MASK] embedding



Pre-Trained *Masked* Language Model

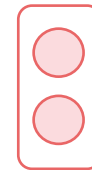
[CLS] The Houston Rockets won an intense overtime game is related to [MASK]

Classification with [MASK] Embedding and Prompt

Topic Classification

The Houston Rockets won an intense overtime game	Sports
Bitcoin hit a new all-time high this week	Finance
Tesla launched a new self-driving software update	Technology
Flu cases are rising in several major cities	Health

Classification with [MASK] embedding



Pre-Trained *Masked* Language Model

Sports
Finance
Technology
Health

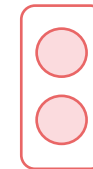
[CLS] The Houston Rockets won an ... overtime game. What is the topic? [MASK]

Classification with [MASK] Embedding and Prompt

Topic Classification

The Houston Rockets won an intense overtime game	Sports
Bitcoin hit a new all-time high this week	Finance
Tesla launched a new self-driving software update	Technology
Flu cases are rising in several major cities	Health

Classification with [MASK] embedding

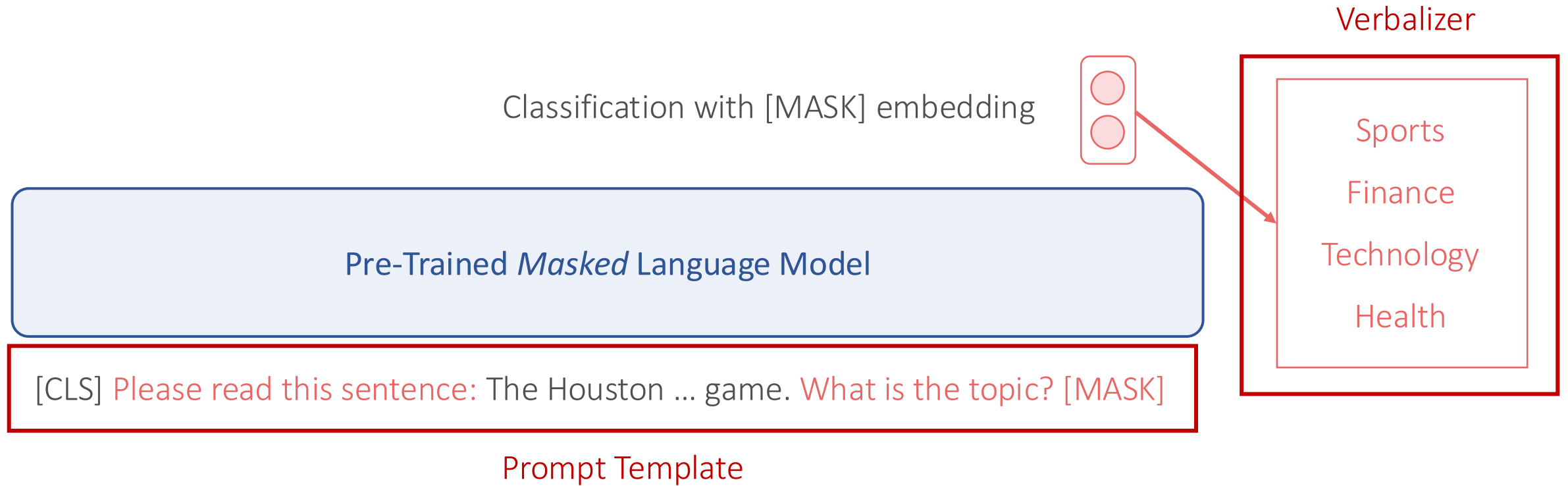


Pre-Trained *Masked* Language Model

Sports
Finance
Technology
Health

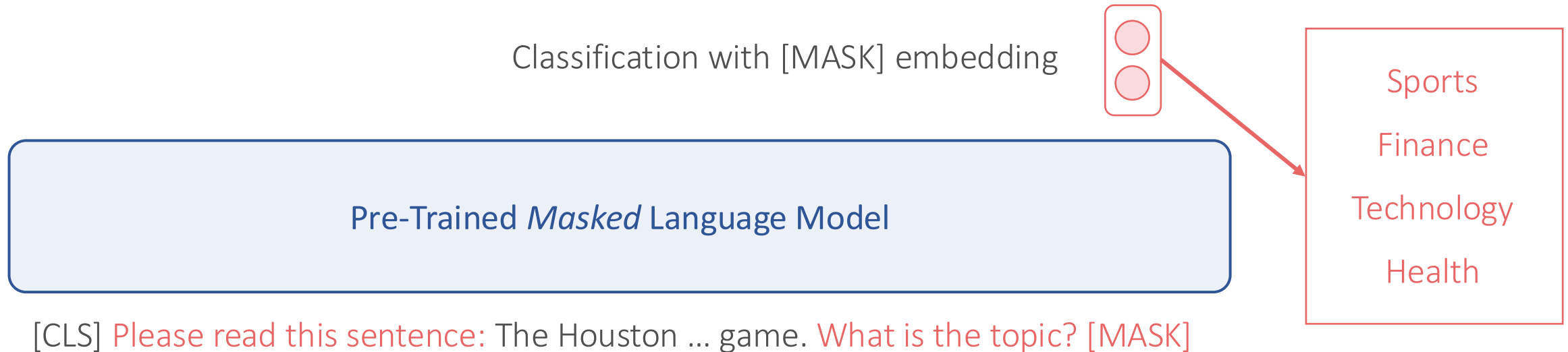
[CLS] Please read this sentence: The Houston ... game. What is the topic? [MASK]

Prompt Tuning



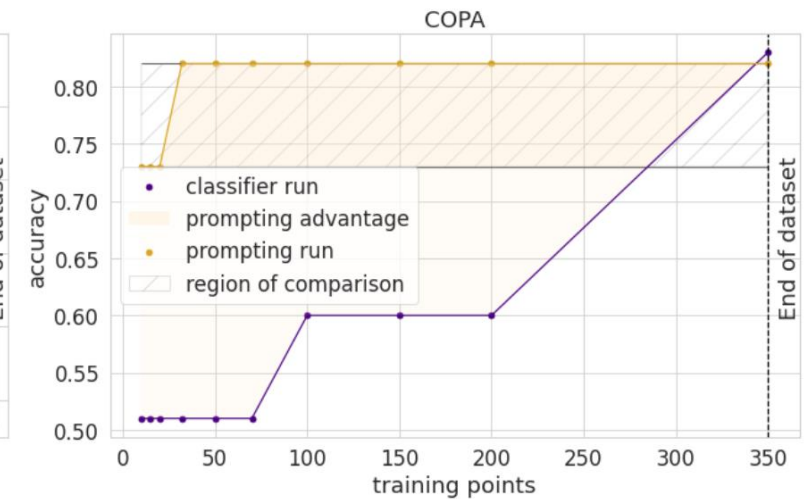
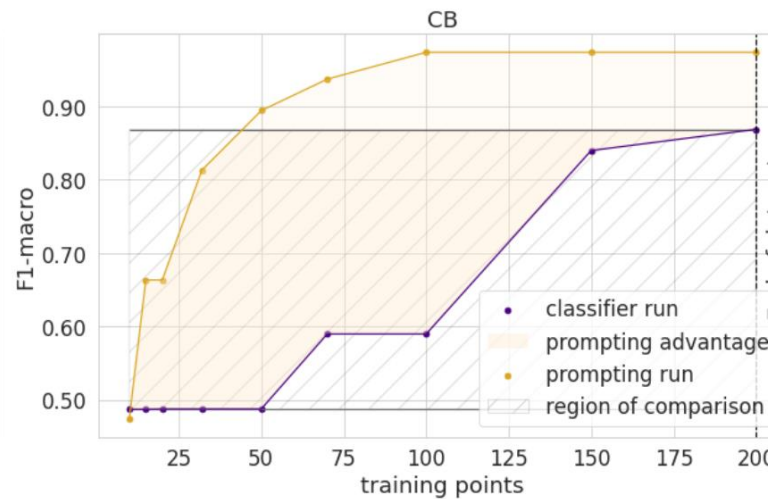
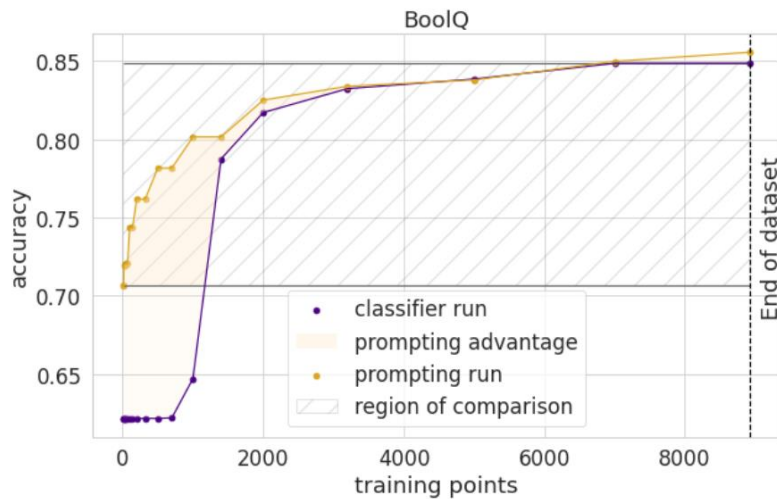
Prompt Tuning

- Better utilize **label semantics** and **pre-trained knowledge**
 - Verbalizer
- Can make **zero-shot** predictions



Prompt Tuning

- Better utilize **label semantics** and **pre-trained knowledge**
 - Verbalizer
- Can make **zero-shot** predictions
- Require **less** training examples

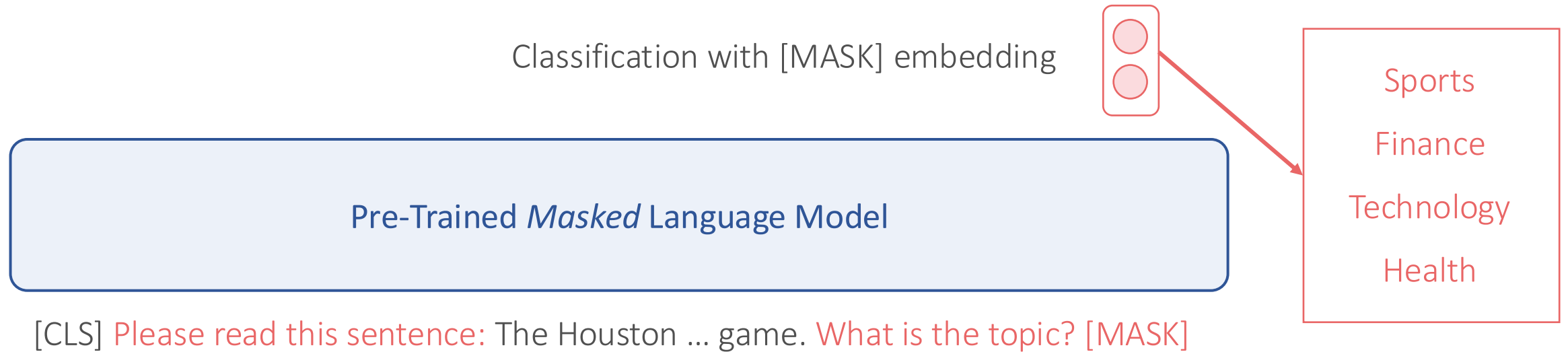


Issues of Discrete/Hard Prompts

- Manually design prompts can be difficult
 - Which one is the best?
- Pre-trained models are sensitive to prompts

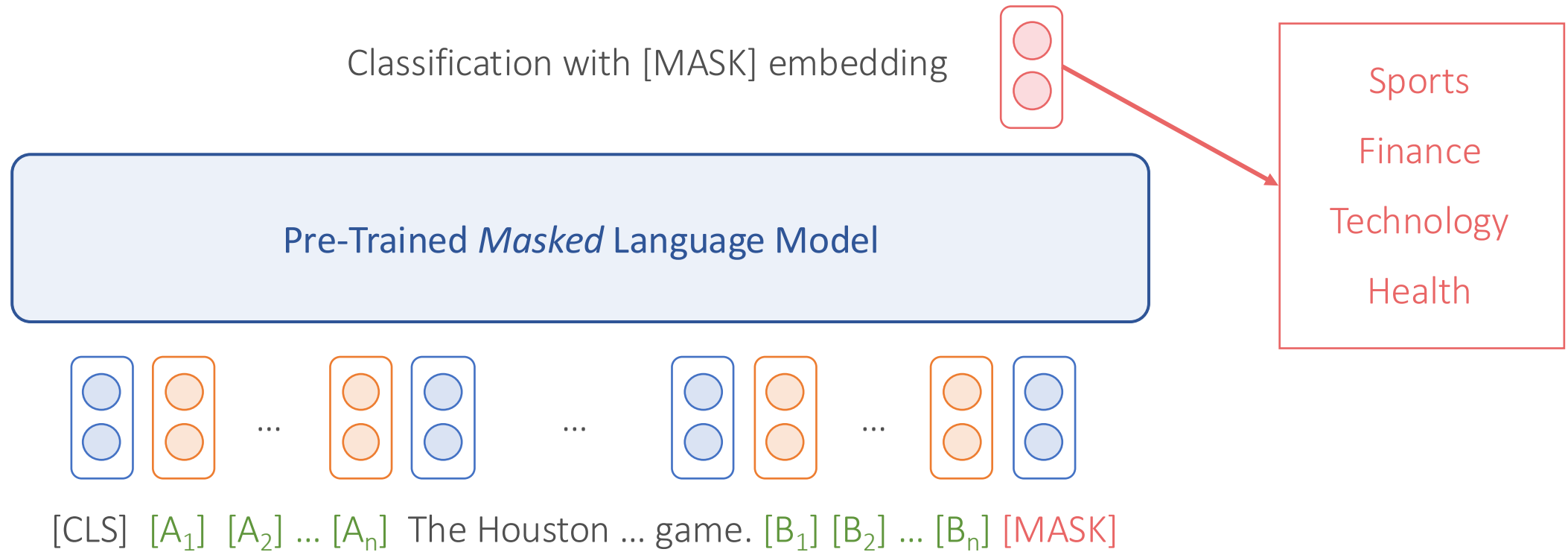
Prompt	P@1
[X] is located in [Y]. (<i>original</i>)	31.29
[X] is located in which country or state? [Y].	19.78
[X] is located in which country? [Y].	31.40
[X] is located in which country? In [Y].	51.08

Hard Prompt Tuning



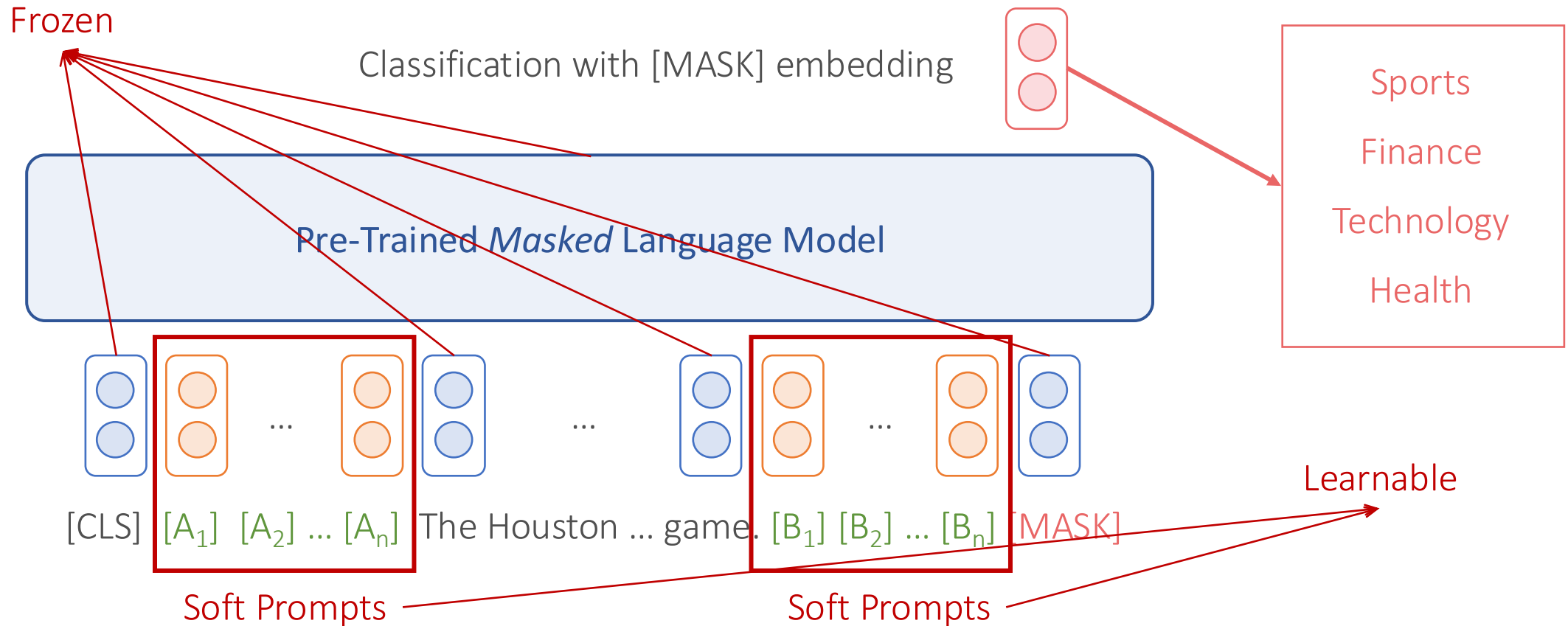
Soft Prompt Tuning

- Let model learn good prompts by itself



Soft Prompt Tuning

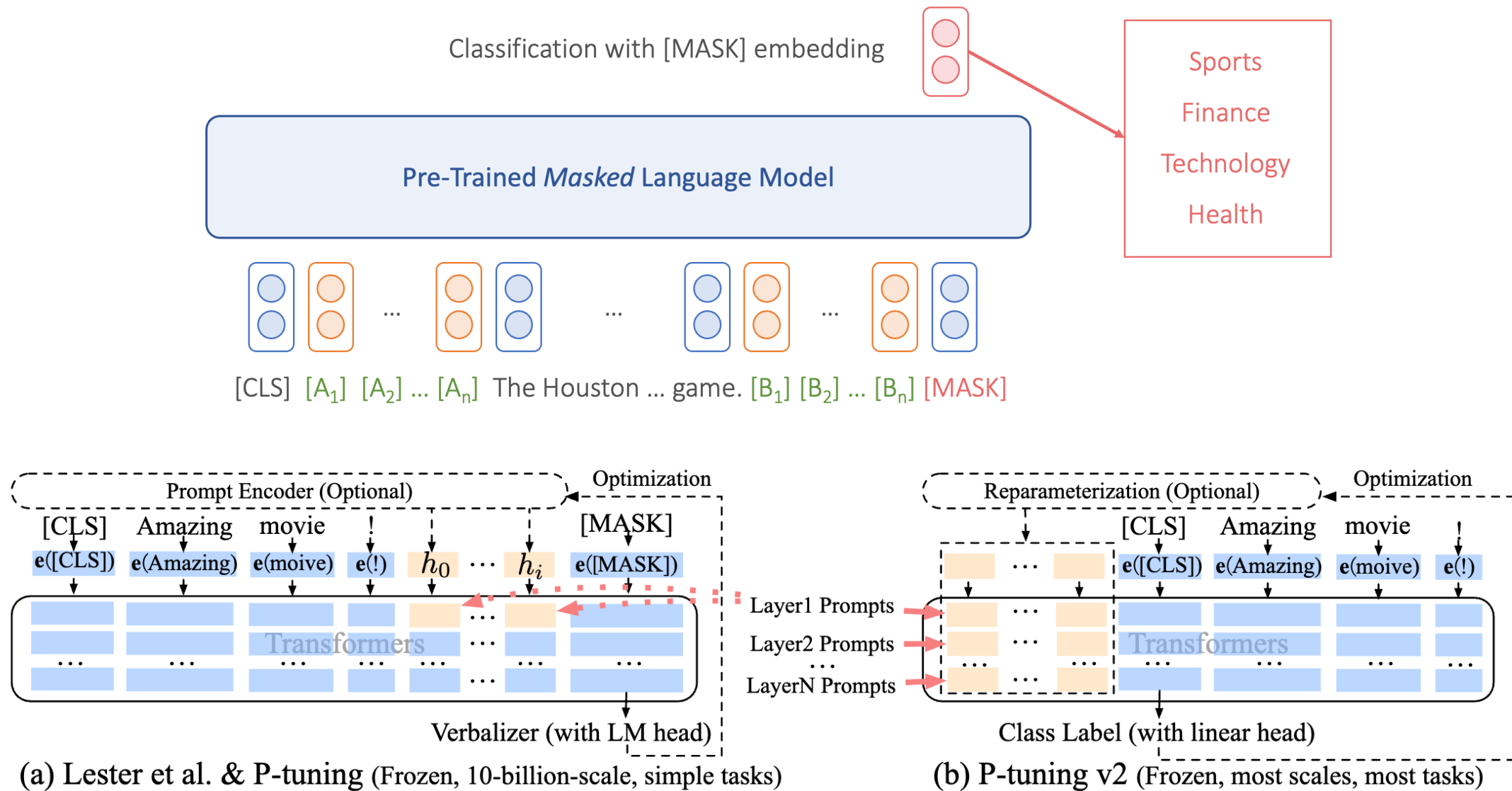
- Let model learn good prompts by itself



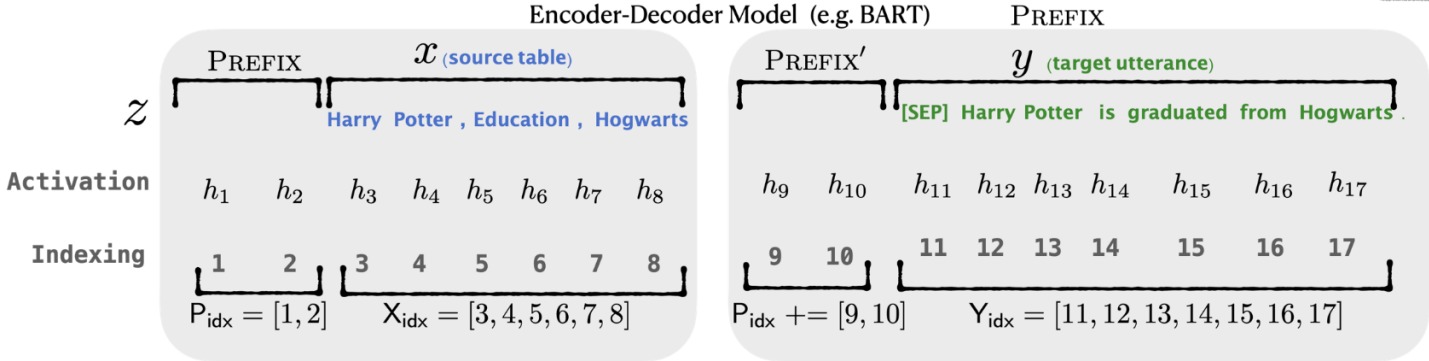
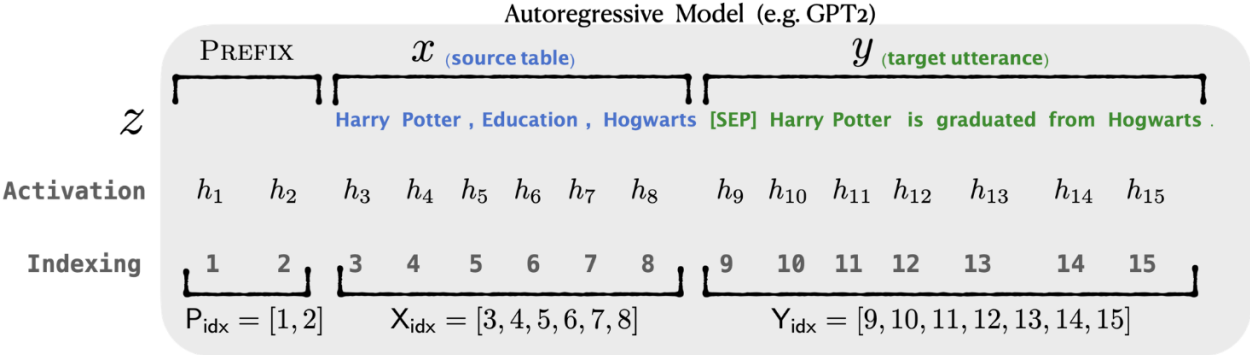
Soft Prompt Tuning

Prompt	$\mathcal{D}_{dev}$ Acc.
Does [PRE] agree with [HYP]? [MASK].	57.16
Does [HYP] agree with [PRE]? [MASK].	51.38
Premise: [PRE] Hypothesis: [HYP] Answer: [MASK].	68.59
[PRE] question: [HYP]. true or false? answer: [MASK].	70.15
P-tuning	76.45

From Prompt Tuning to Prefix Tuning



Prefix Tuning for Generation



Summarization Example

Article: Scientists at University College London discovered people tend to think that their hands are wider and their fingers are shorter than they truly are.They say the confusion may lie in the way the brain receives information from different parts of the body.Distorted perception may dominate in some people, leading to body image problems ... [ignoring 308 words] could be very motivating for people with eating disorders to know that there was a biological explanation for their experiences, rather than feeling it was their fault."

Summary: The brain naturally distorts body image – a finding which could explain eating disorders like anorexia, say experts.

Table-to-text Example

Table: name[Clowns] customer-rating[1 out of 5] eatType[coffee shop] food[Chinese] area[riverside] near[Clare Hall]

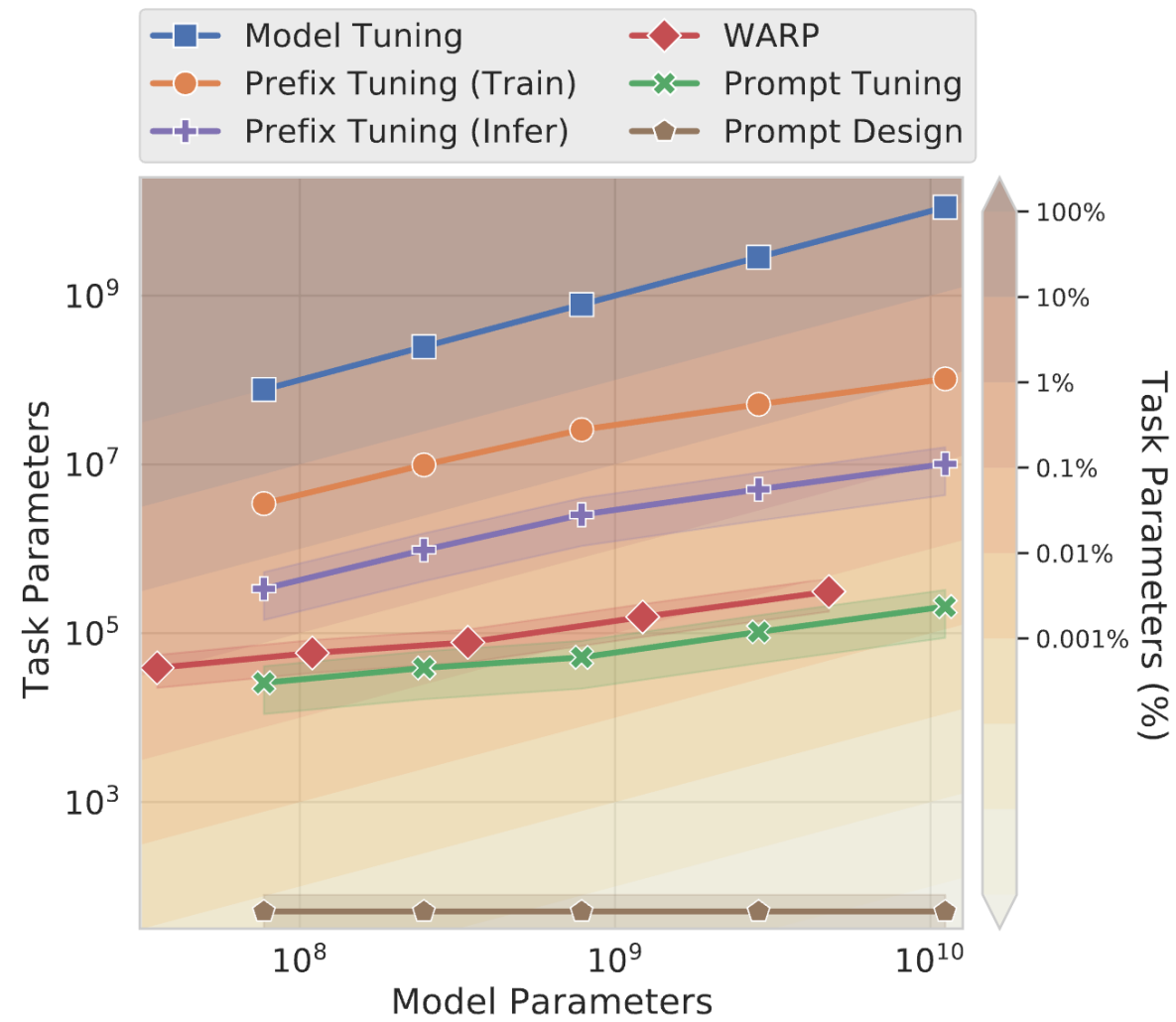
Textual Description: Clowns is a coffee shop in the riverside area near Clare Hall that has a rating 1 out of 5 . They serve Chinese food .

Prefix Tuning

	#Size	BoolQ			CB			COPA			MultiRC (F1a)		
		FT	PT	PT-2	FT	PT	PT-2	FT	PT	PT-2	FT	PT	PT-2
BERT _{large}	335M	77.7	67.2	<u>75.8</u>	94.6	80.4	94.6	<u>69.0</u>	55.0	73.0	<u>70.5</u>	59.6	70.6
RoBERTa _{large}	355M	86.9	62.3	<u>84.8</u>	<u>98.2</u>	71.4	100	94.0	63.0	<u>93.0</u>	85.7	59.9	<u>82.5</u>
GLM _{xlarge}	2B	88.3	79.7	<u>87.0</u>	96.4	<u>76.4</u>	96.4	93.0	<u>92.0</u>	91.0	<u>84.1</u>	77.5	84.4
GLM _{xxlarge}	10B	<u>88.7</u>	88.8	88.8	98.7	<u>98.2</u>	96.4	98.0	98.0	98.0	88.1	<u>86.1</u>	88.1

	#Size	ReCoRD (F1)			RTE			WiC			WSC		
		FT	PT	PT-2	FT	PT	PT-2	FT	PT	PT-2	FT	PT	PT-2
BERT _{large}	335M	<u>70.6</u>	44.2	72.8	<u>70.4</u>	53.5	78.3	<u>74.9</u>	63.0	75.1	68.3	64.4	68.3
RoBERTa _{large}	355M	<u>89.0</u>	46.3	89.3	<u>86.6</u>	58.8	89.5	75.6	56.9	<u>73.4</u>	<u>63.5</u>	64.4	<u>63.5</u>
GLM _{xlarge}	2B	<u>91.8</u>	82.7	91.9	90.3	<u>85.6</u>	90.3	74.1	71.0	<u>72.0</u>	95.2	87.5	<u>92.3</u>
GLM _{xxlarge}	10B	94.4	87.8	<u>92.5</u>	93.1	<u>89.9</u>	93.1	75.7	71.8	<u>74.0</u>	95.2	<u>94.2</u>	93.3

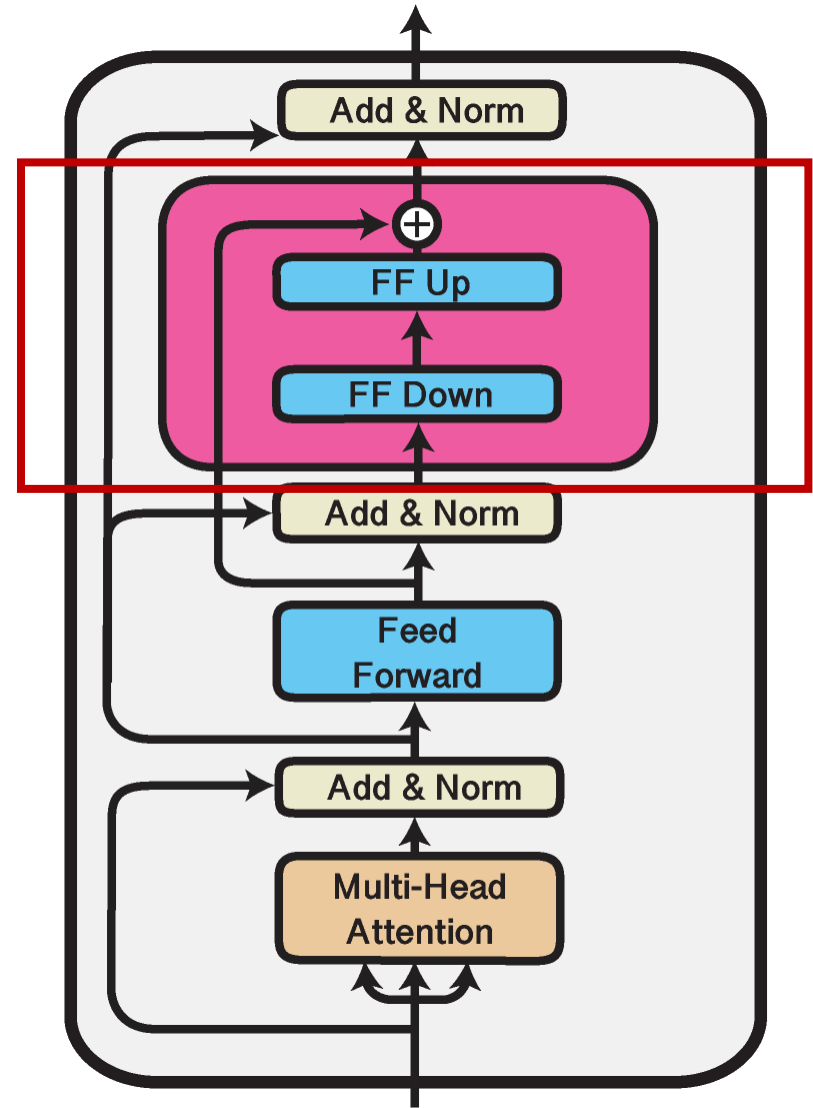
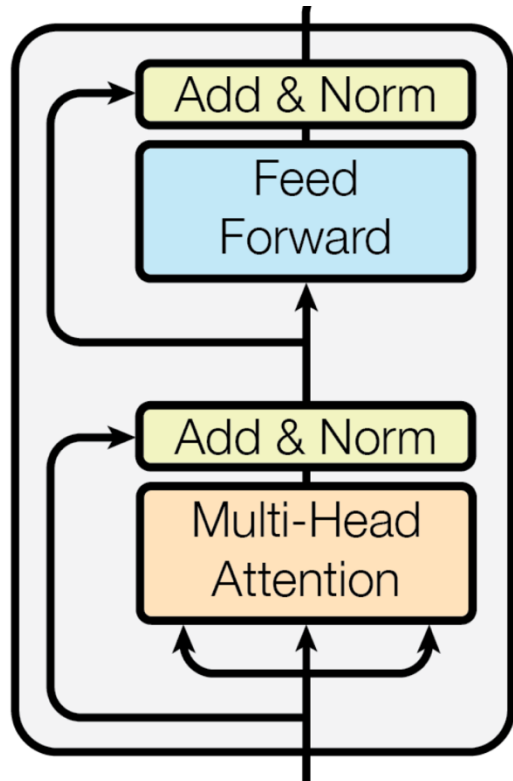
Prefix Tuning – Parameter-Efficient



Parameter-Efficient Fine-Tuning (PEFT)

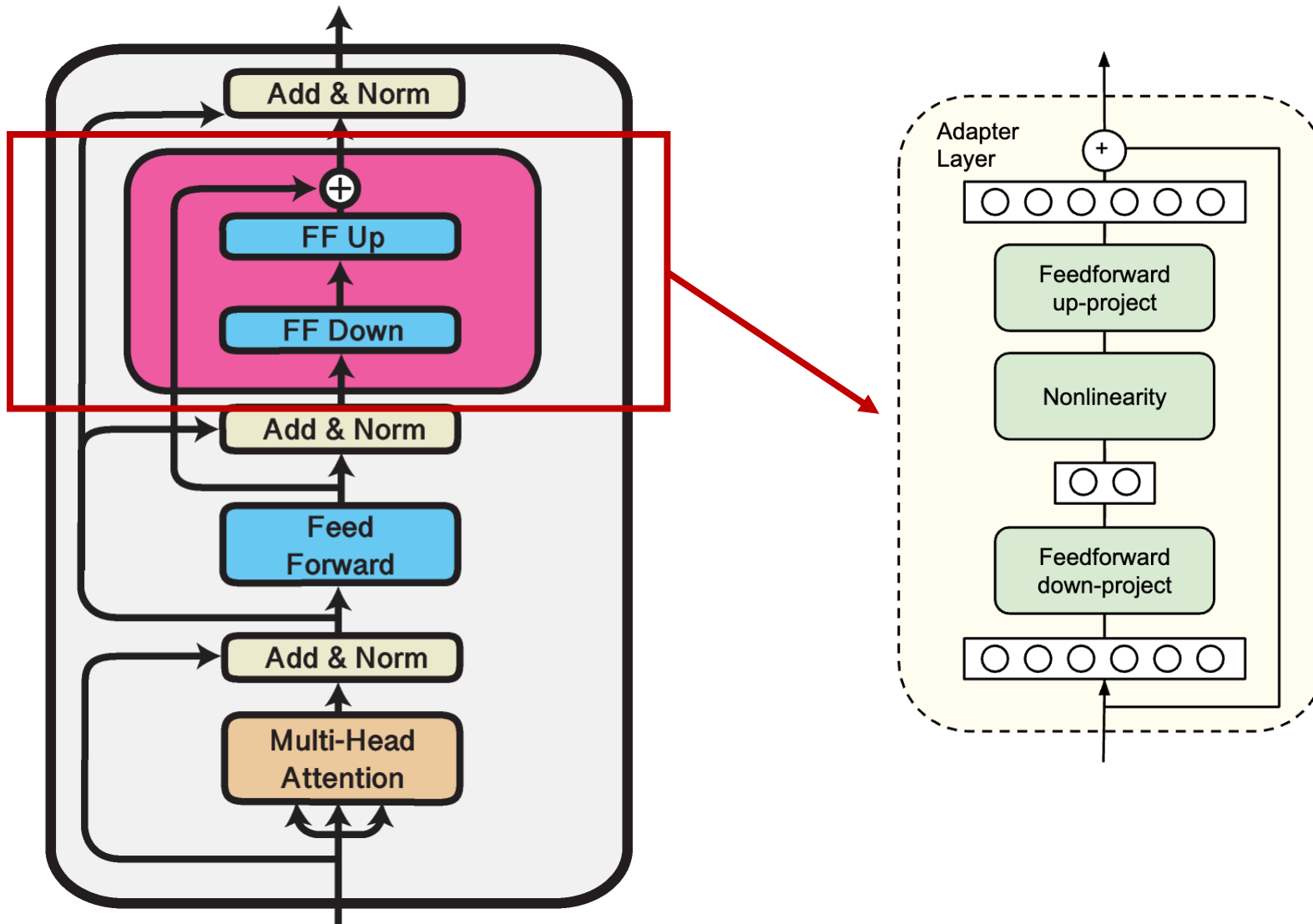
- Motivation
 - Do not fine-tune the whole model
 - Update only a small subset of parameters
 - Keep most parameters frozen
 - Achieve comparable performance to full fine-tuning
- Advantages
 - Lower GPU memory usage
 - Faster training
 - Reduced storage cost
 - Easier deployment across tasks

Adapter

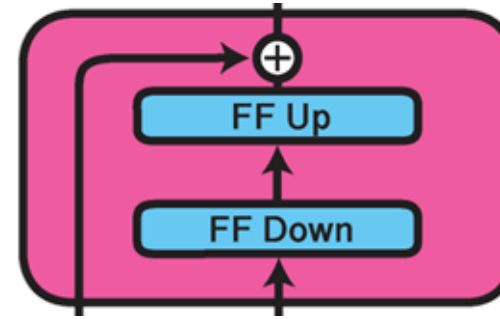
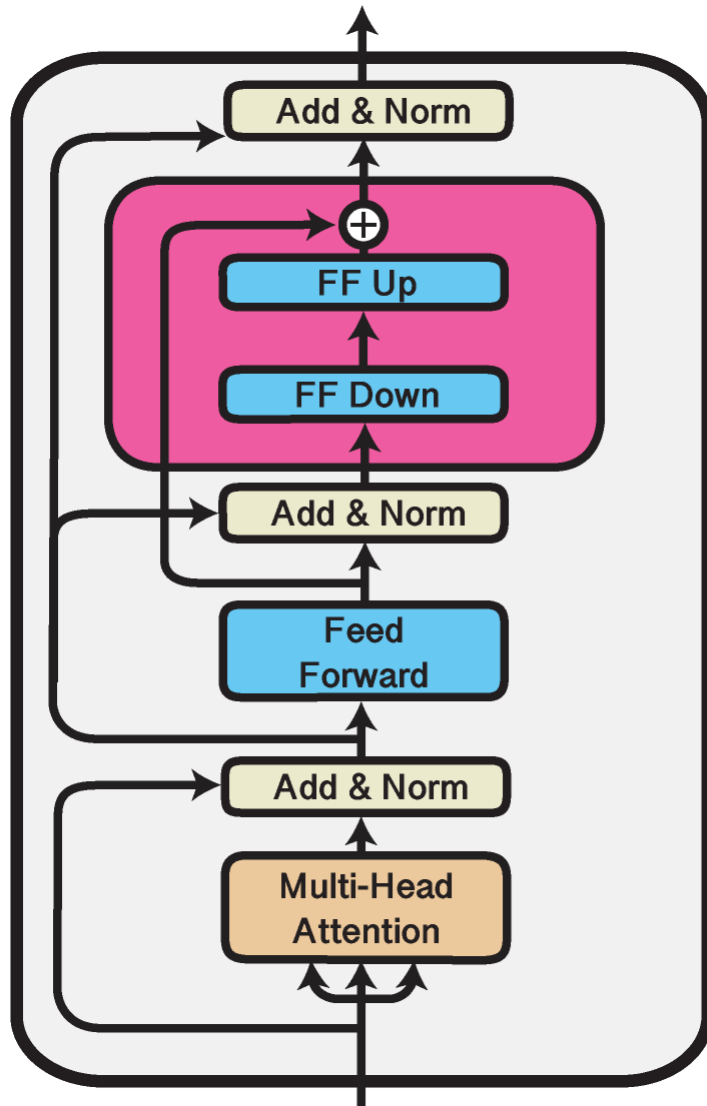


Adapter

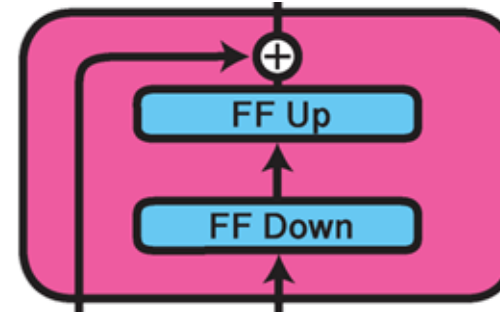
Adapter



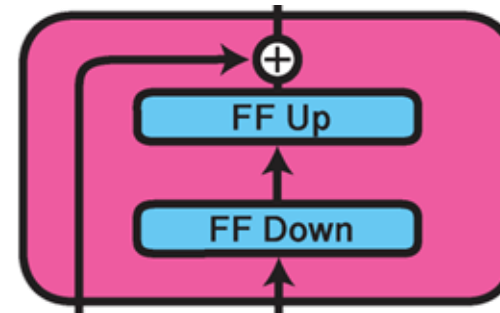
Adapter



Task 1

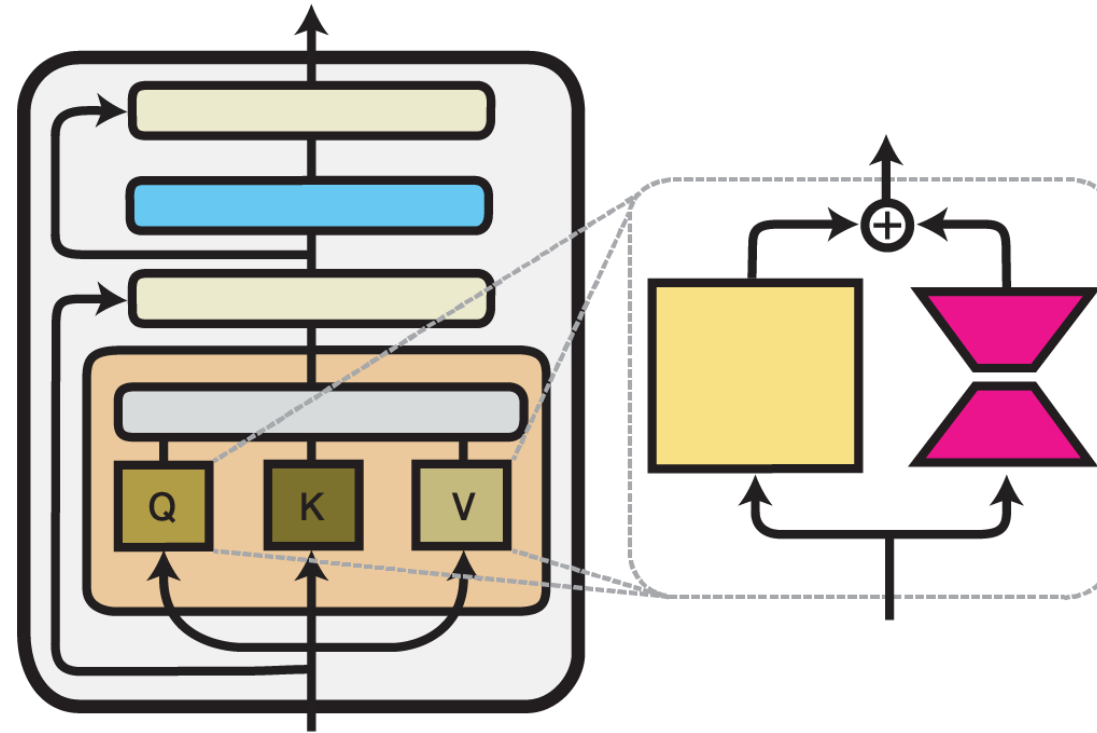


Task 2

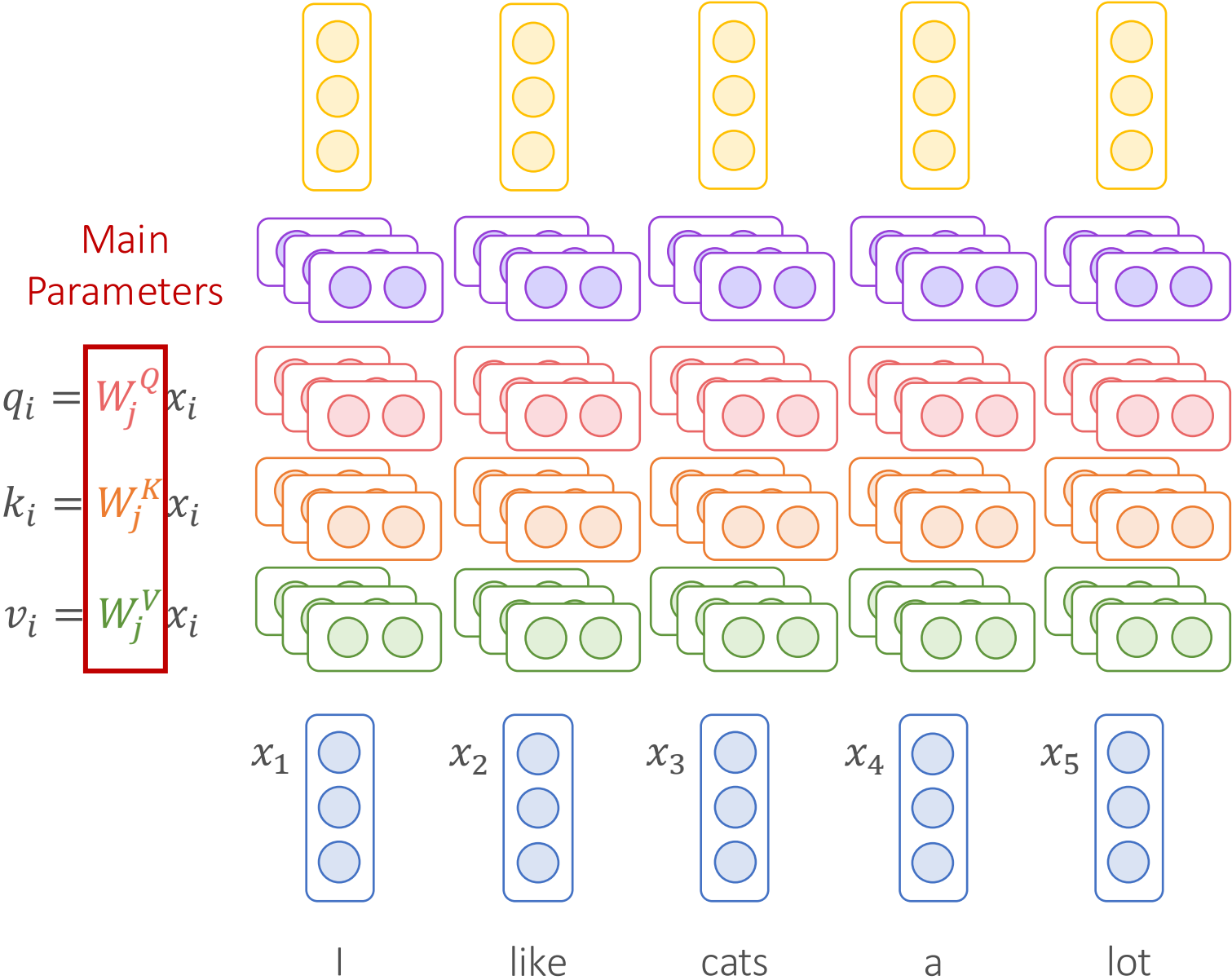


Task 3

Low-Rank Adaptation (LoRA)



Low-Rank Adaptation (LoRA)



Low-Rank Adaptation (LoRA)

Before fine-tuning

$$q_i = W_j^Q x_i$$

$$k_i = W_j^K x_i$$

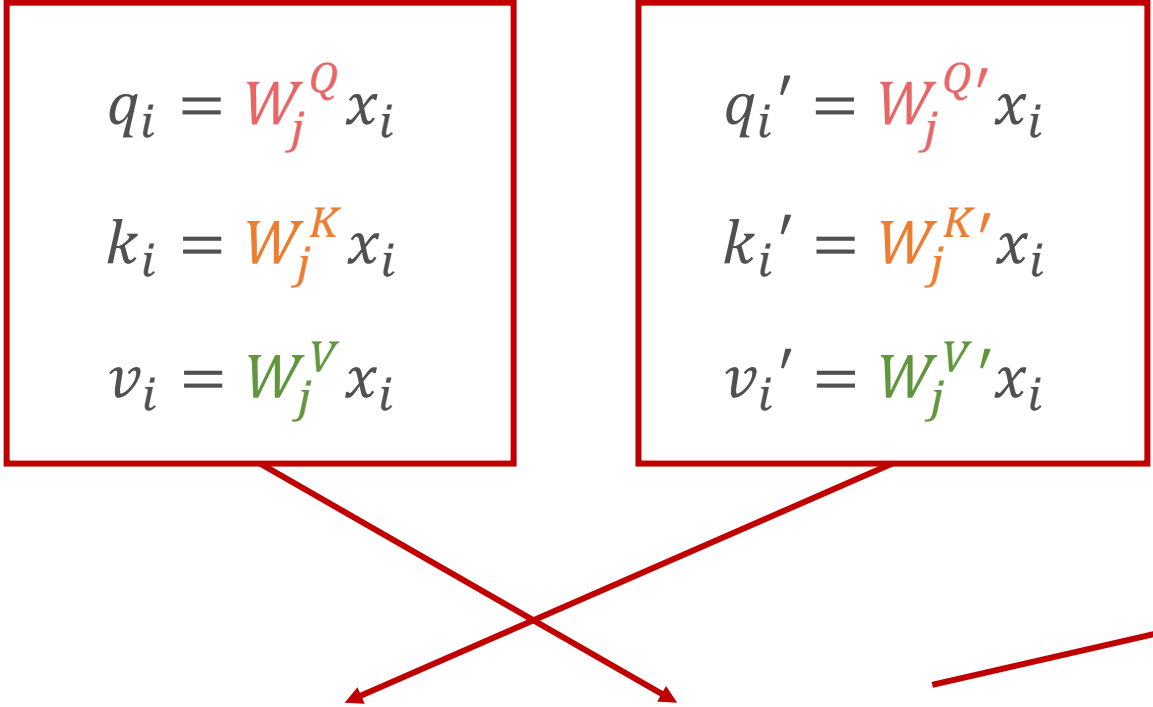
$$v_i = W_j^V x_i$$

After fine-tuning

$$q_i' = W_j^{Q'} x_i$$

$$k_i' = W_j^{K'} x_i$$

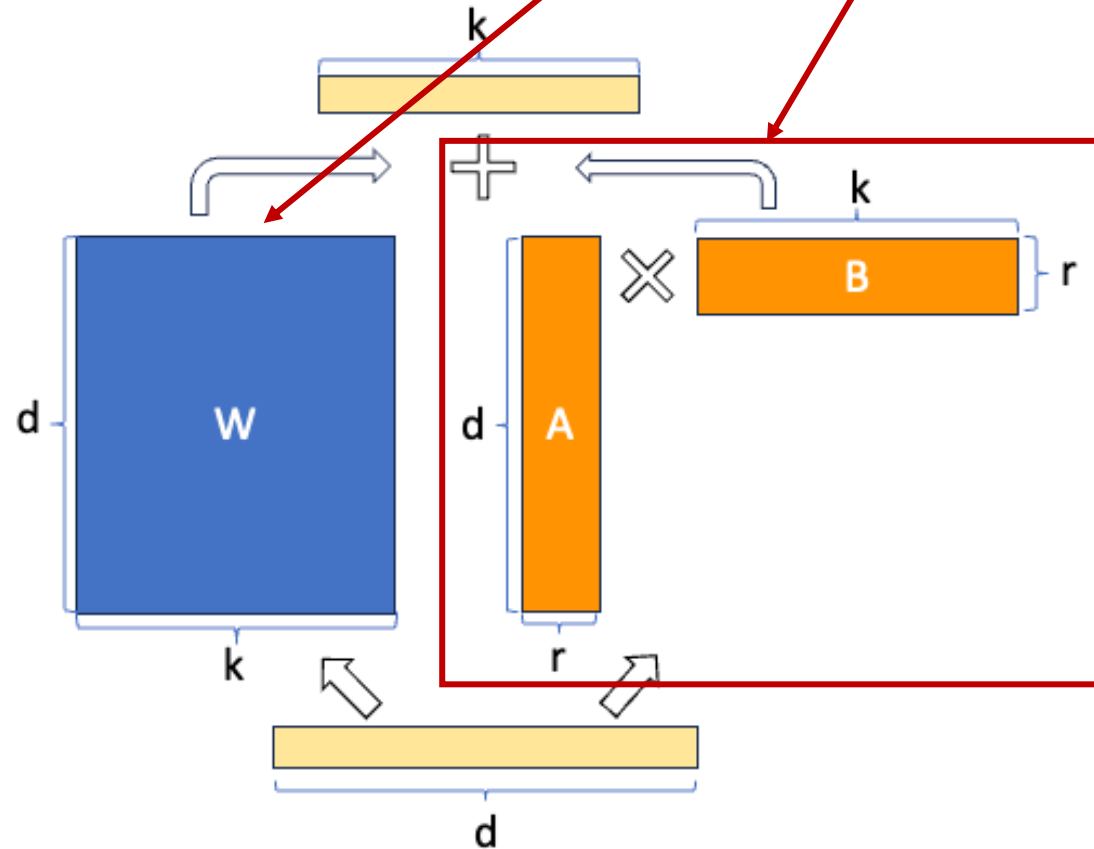
$$v_i' = W_j^{V'} x_i$$


$$h = W_{new} x = W_{old} x + W_{\Delta} x$$

Learnable
Parameters

Low-Rank Adaptation (LoRA)

$$h = W_{new}x = W_{old}x + W_{\Delta}x$$



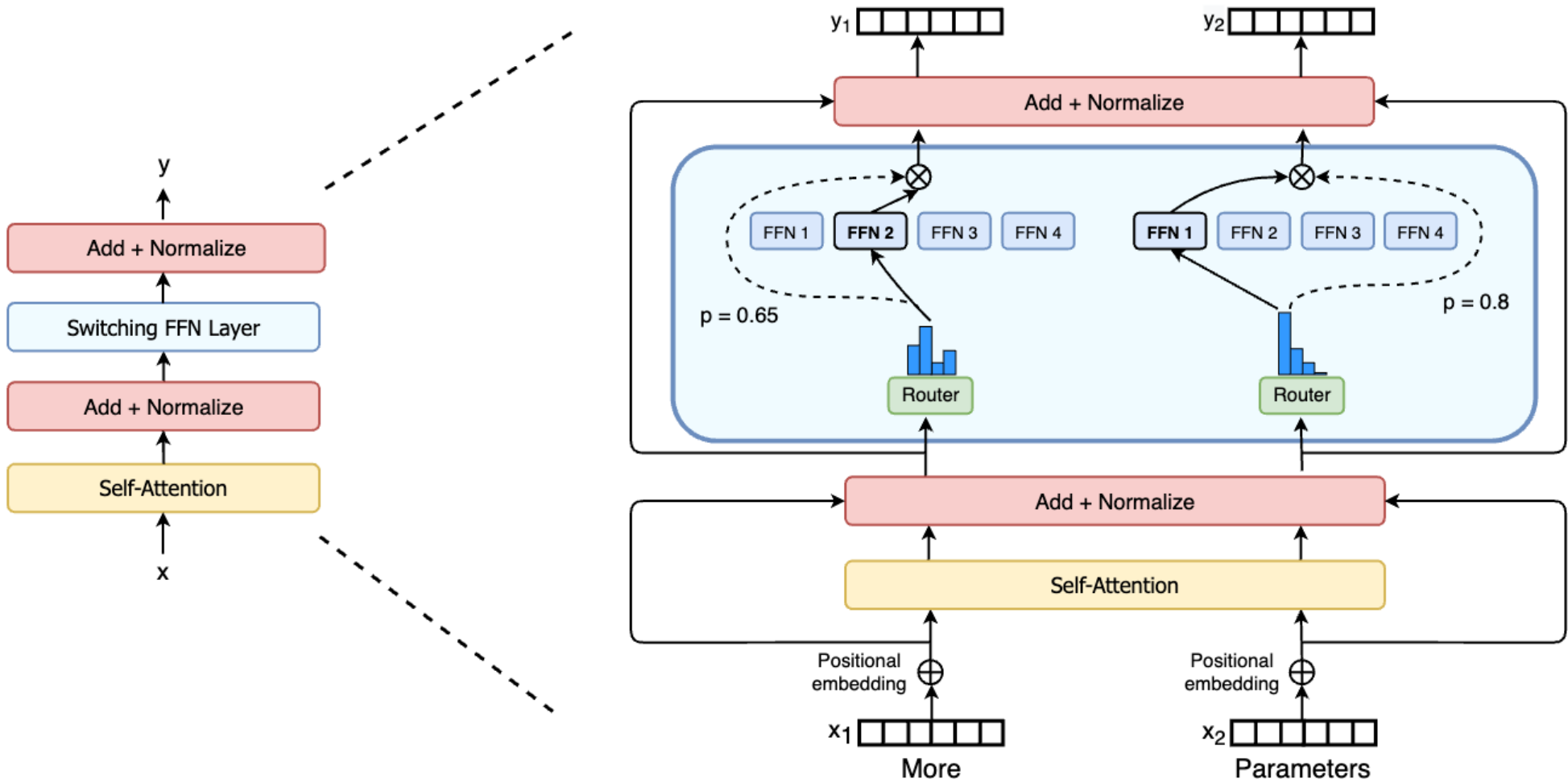
LoRA: Low-Rank Adaptation

Model & Method	# Trainable Parameters	MNLI	SST-2	MRPC	CoLA	QNLI	QQP	RTE	STS-B	Avg.
RoB _{base} (FT)*	125.0M	87.6	94.8	90.2	63.6	92.8	91.9	78.7	91.2	86.4
RoB _{base} (BitFit)*	0.1M	84.7	93.7	92.7	62.0	91.8	84.0	81.5	90.8	85.2
RoB _{base} (Adpt ^D)*	0.3M	87.1 $\pm$.0	94.2 $\pm$.1	88.5 $\pm$ 1.1	60.8 $\pm$.4	93.1 $\pm$.1	90.2 $\pm$.0	71.5 $\pm$ 2.7	89.7 $\pm$.3	84.4
RoB _{base} (Adpt ^D)*	0.9M	87.3 $\pm$.1	94.7 $\pm$.3	88.4 $\pm$.1	62.6 $\pm$.9	93.0 $\pm$.2	90.6 $\pm$.0	75.9 $\pm$ 2.2	90.3 $\pm$.1	85.4
RoB _{base} (LoRA)	0.3M	87.5 $\pm$.3	95.1$\pm$.2	89.7 $\pm$.7	63.4 $\pm$ 1.2	93.3$\pm$.3	90.8 $\pm$.1	86.6$\pm$.7	91.5$\pm$.2	87.2
RoB _{large} (FT)*	355.0M	90.2	96.4	90.9	68.0	94.7	92.2	86.6	92.4	88.9
RoB _{large} (LoRA)	0.8M	90.6$\pm$.2	96.2 $\pm$.5	90.9$\pm$1.2	68.2$\pm$1.9	94.9$\pm$.3	91.6 $\pm$.1	87.4$\pm$2.5	92.6$\pm$.2	89.0
RoB _{large} (Adpt ^P)†	3.0M	90.2 $\pm$.3	96.1 $\pm$.3	90.2 $\pm$.7	68.3$\pm$1.0	94.8$\pm$.2	91.9$\pm$.1	83.8 $\pm$ 2.9	92.1 $\pm$.7	88.4
RoB _{large} (Adpt ^P)†	0.8M	90.5$\pm$.3	96.6$\pm$.2	89.7 $\pm$ 1.2	67.8 $\pm$ 2.5	94.8$\pm$.3	91.7 $\pm$.2	80.1 $\pm$ 2.9	91.9 $\pm$.4	87.9
RoB _{large} (Adpt ^H)†	6.0M	89.9 $\pm$.5	96.2 $\pm$.3	88.7 $\pm$ 2.9	66.5 $\pm$ 4.4	94.7 $\pm$.2	92.1 $\pm$.1	83.4 $\pm$ 1.1	91.0 $\pm$ 1.7	87.8
RoB _{large} (Adpt ^H)†	0.8M	90.3 $\pm$.3	96.3 $\pm$.5	87.7 $\pm$ 1.7	66.3 $\pm$ 2.0	94.7 $\pm$.2	91.5 $\pm$.1	72.9 $\pm$ 2.9	91.5 $\pm$.5	86.4
RoB _{large} (LoRA)†	0.8M	90.6$\pm$.2	96.2 $\pm$.5	90.2$\pm$1.0	68.2 $\pm$ 1.9	94.8$\pm$.3	91.6 $\pm$.2	85.2$\pm$1.1	92.3$\pm$.5	88.6
DeB _{XXL} (FT)*	1500.0M	91.8	97.2	92.0	72.0	96.0	92.7	93.9	92.9	91.1
DeB _{XXL} (LoRA)	4.7M	91.9$\pm$.2	96.9 $\pm$.2	92.6$\pm$.6	72.4$\pm$1.1	96.0$\pm$.1	92.9$\pm$.1	94.9$\pm$.4	93.0$\pm$.2	91.3

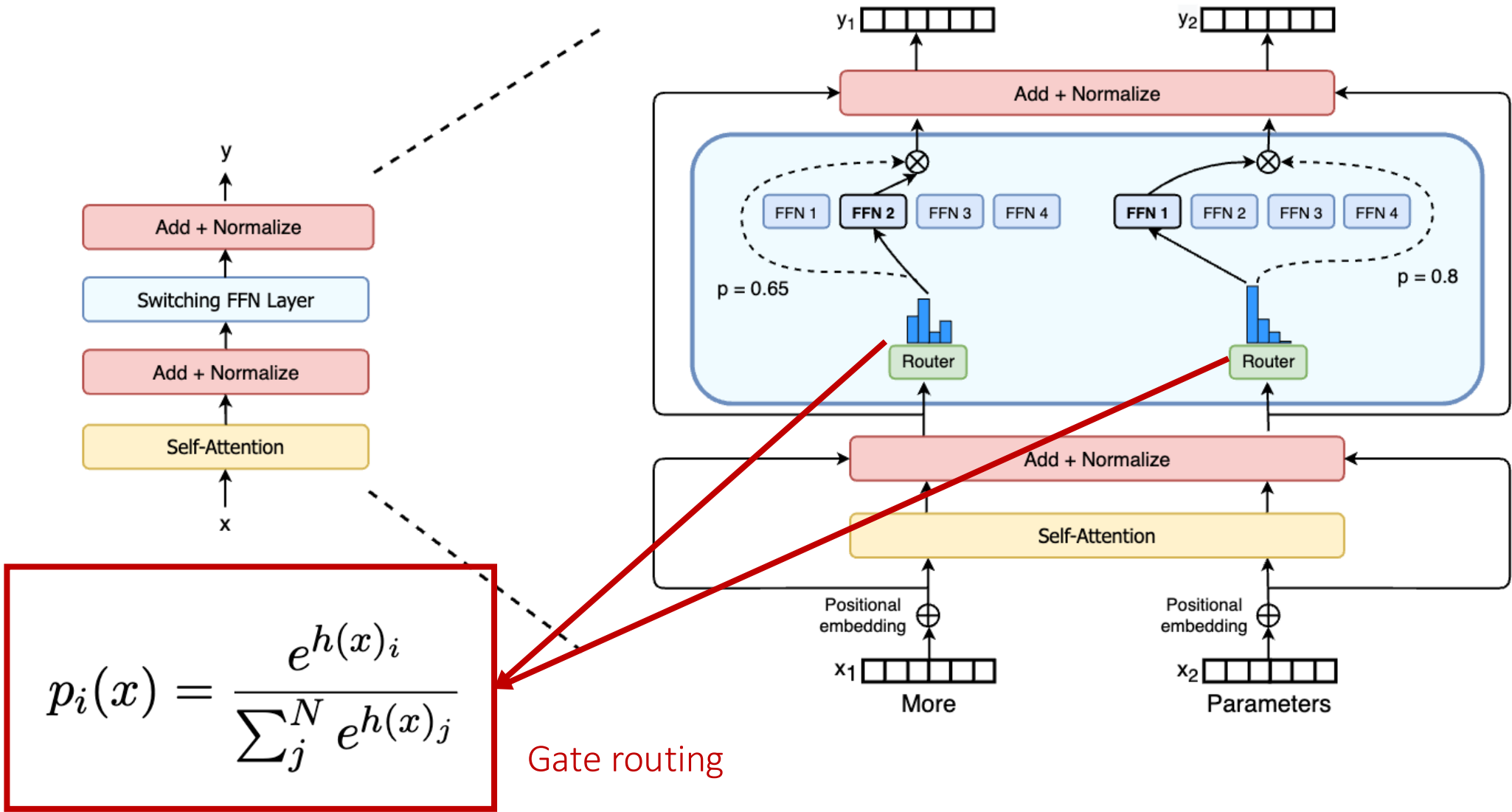
Mixture of Experts (MoE)

- Transformers with multiple specialized “expert” adapters
- A gating network selects experts per input
- Only only a subset of experts are activated per token/input

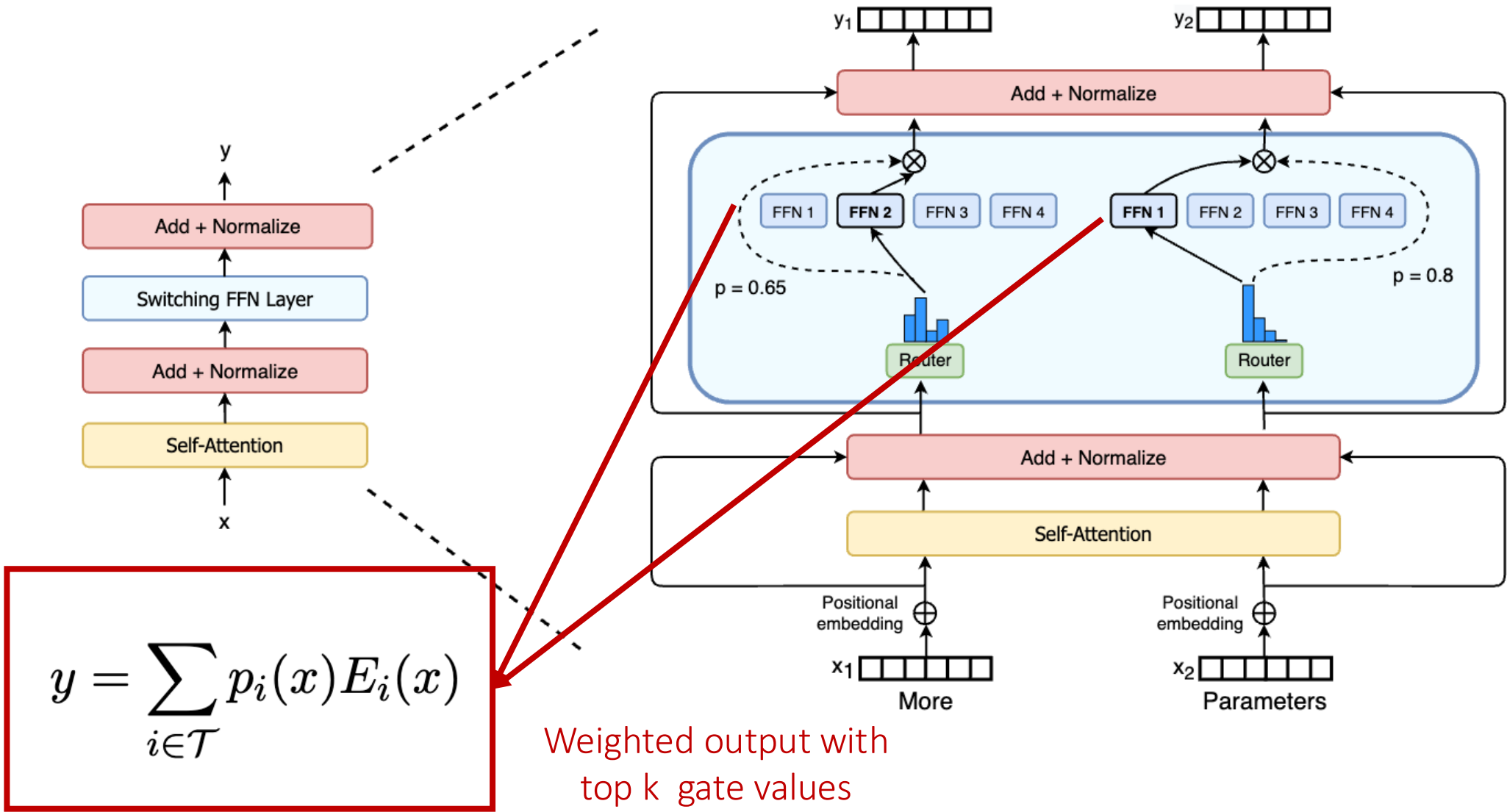
Mixture of Experts (MoE)



Mixture of Experts (MoE)



Mixture of Experts (MoE)

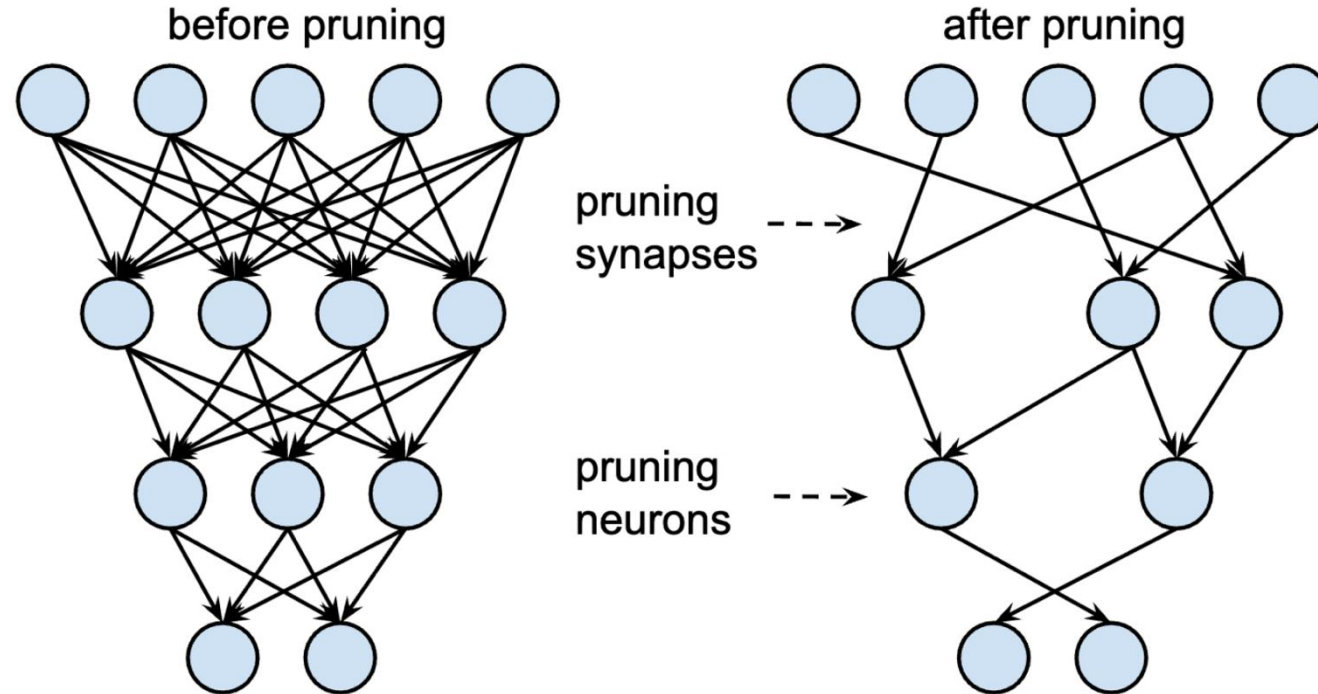


Mixture of Experts (MoE)

- Transformers with multiple specialized “expert” adapters
- A gating network selects experts per input
- Only only a subset of experts are activated per token/input

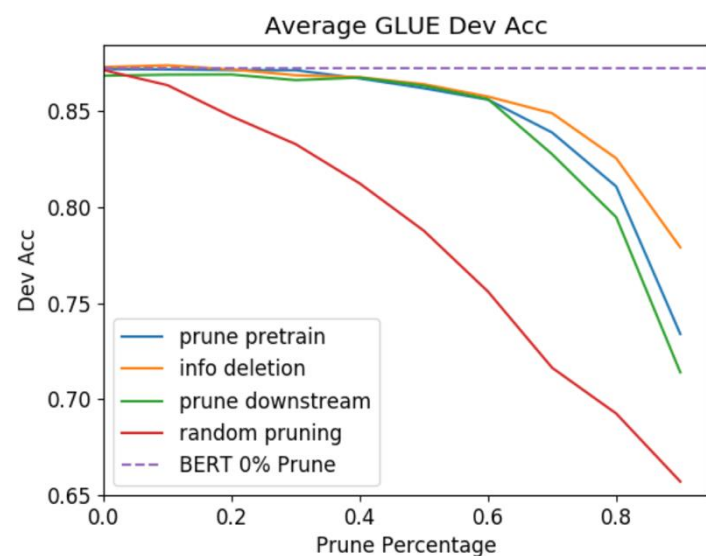
Pruning

- Reduces model size and computation
- Keeps performance with fewer weights



Transformer with Pruning

- Prune model weights
- Prune neurons
- Prune attention head
- Prune layers



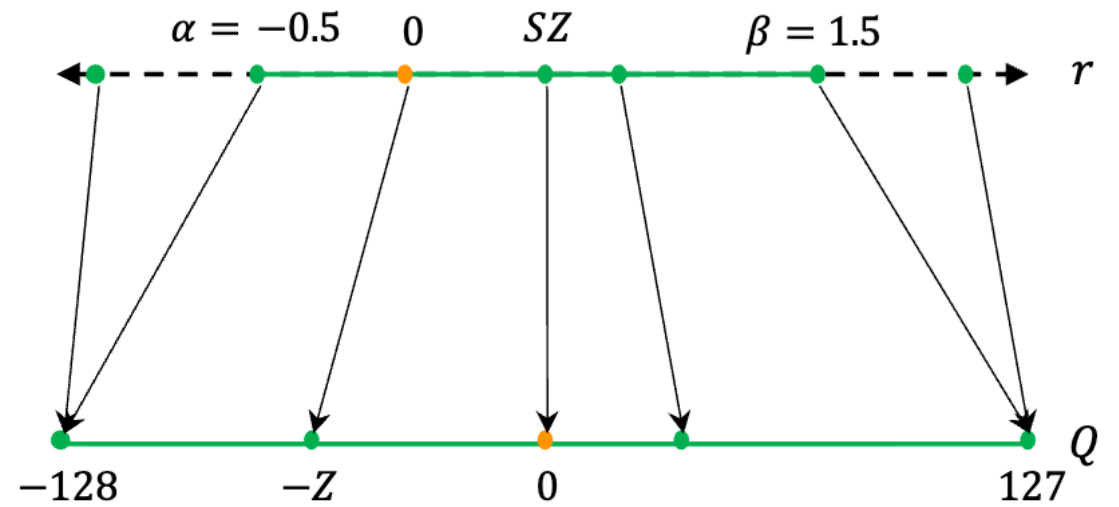
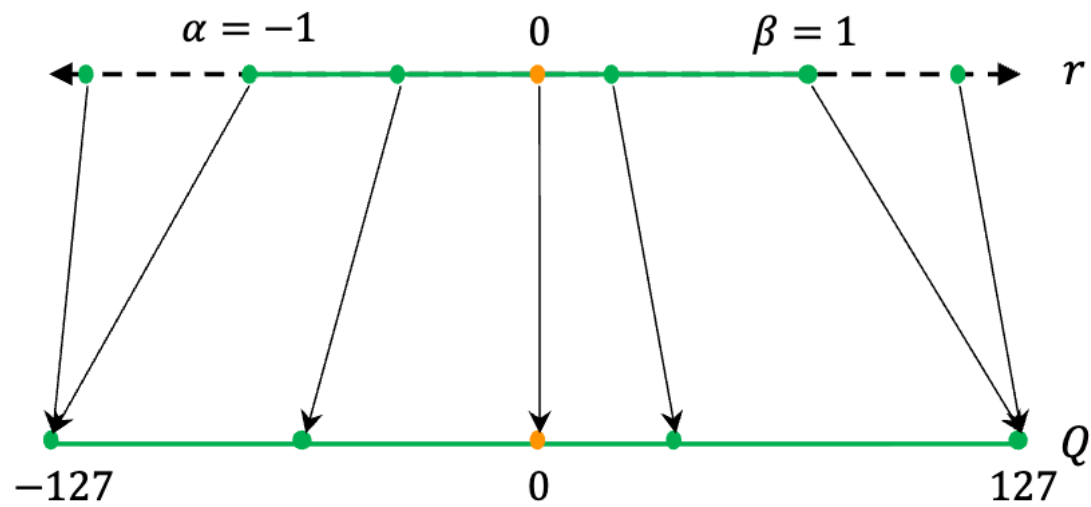
Drop.	SST-2	MNLI	QNLI	QQP	STS-B	RTE	MRPC
BERT							
0/12	92.43	84.04	91.12	91.07	88.79	67.87	87.99
2/12	92.20 (0.23↓)	83.26 (0.78↓)	89.84 (1.28↓)	90.92 (0.15↓)	88.70 (0.09↓)	62.82 (5.05↓)	86.27 (1.72↓)
4/12	90.60 (1.83↓)	82.51 (1.53↓)	89.68 (1.44↓)	90.63 (0.44↓)	88.64 (0.15↓)	67.87 (0.00)	79.41 (8.58↓)
6/12	90.25 (2.18↓)	81.13 (2.91↓)	87.63 (3.49↓)	90.35 (0.72↓)	88.45 (0.34↓)	64.98 (2.89↓)	80.15 (7.84↓)
RoBERTa							
0/12	92.20	86.44	91.73	90.48	89.87	68.95	88.48
2/12	93.46 (1.26↑)	86.53 (0.09↑)	91.23 (0.50↓)	91.02 (0.54↑)	90.21 (0.34↑)	71.84 (2.89↑)	89.71 (1.23↑)
4/12	93.00 (0.80↑)	86.20 (0.24↓)	90.57 (1.16↓)	91.12 (0.64↑)	89.77 (0.10↓)	70.40 (1.45↑)	87.50 (0.98↓)
6/12	91.97 (0.23↓)	84.44 (2.00↓)	90.00 (1.73↓)	90.91 (0.43↑)	88.92 (0.95↓)	64.62 (4.33↓)	85.78 (2.70↓)

Quantization

- Reduce numerical precision of model weights
- Converts high-precision values (fp32) → lower precision (int8, fp16, etc.)
- Reduces memory and computation requirements
- Model trained with fp32 → Quantization → Inference with int8

Quantization

- $123.45 \rightarrow 123$
- $10.789654 \rightarrow 10$



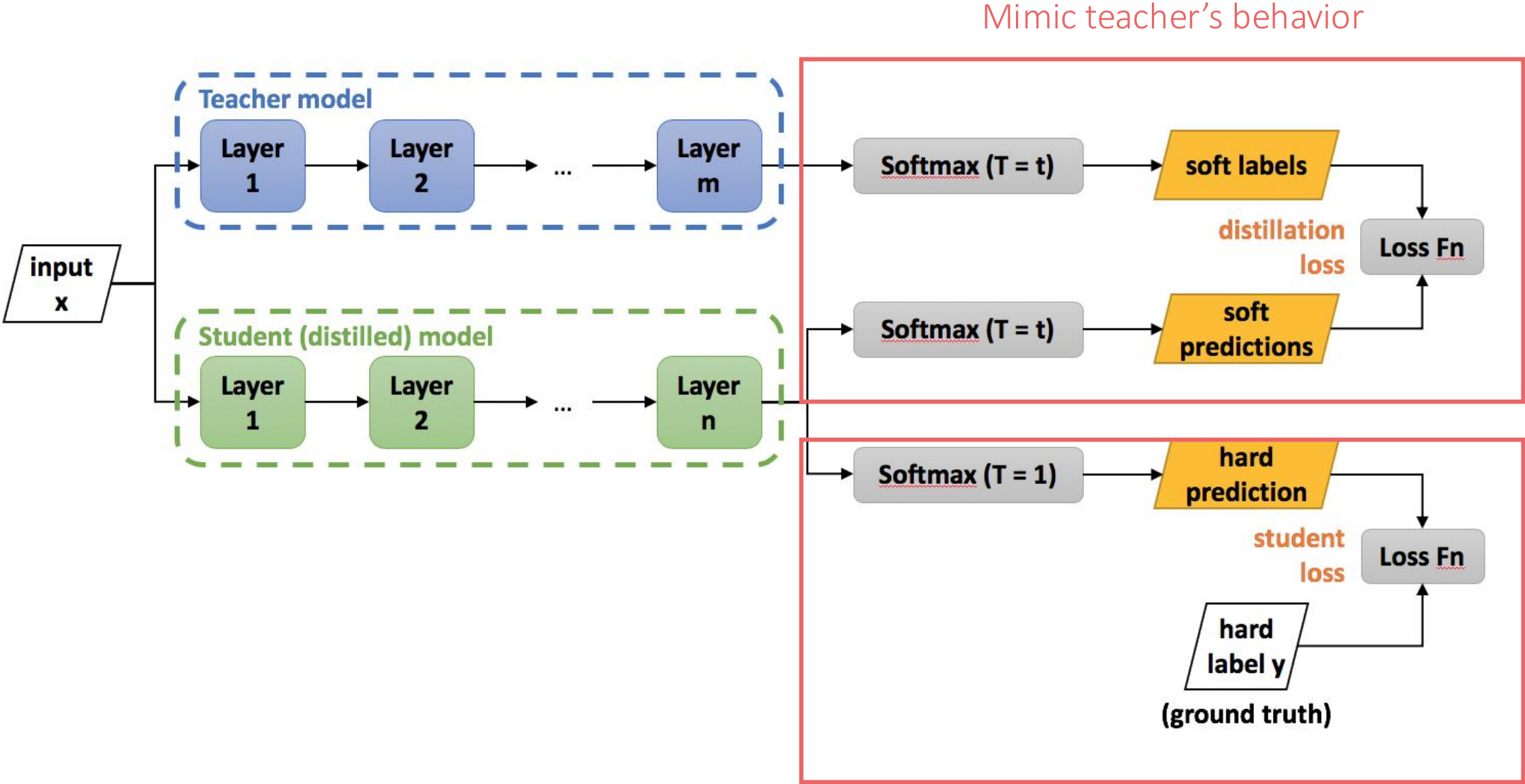
Quantization

Model	Method	W/A	Storage (GB)	OpenLLM Leaderboard-v1 ↑						
				ARC-c (25-shot) acc_norm	GSM8k (5-shot) acc	HellaSwag (10-shot) acc_norm	MMLU (5-shot) acc	TruthfulQA (0-shot) mc2	Winogrande (5-shot) acc	Avg.
Llama-3.1-8B-it	FP16	16 / 16	16	60.24	76.65	80.21	68.10	54.03	76.16	69.23
	FP8	8 / 8	8	61.52 (↑1.28)	74.75 (↓1.90)	80.12 (↓0.09)	68.52 (↑0.42)	53.81 (↓0.22)	77.43 (↑1.27)	69.36 (↑0.13)
	GPTQ*	4 / 16	4	61.43 (↑1.19)	72.33 (↓4.32)	78.36 (↓1.85)	66.85 (↓1.25)	53.60 (↓0.43)	75.22 (↓0.94)	67.97 (↓1.26)
	GPTQ**	4 / 16	4	59.81 (↓0.43)	69.98 (↓6.67)	78.53 (↓1.68)	66.07 (↓2.03)	50.45 (↓3.58)	76.64 (↑0.48)	66.91 (↓2.32)
	GPTQ**	8 / 16	8	61.01 (↑0.77)	75.81 (↓0.84)	80.27 (↓0.06)	68.21 (↑0.11)	54.03 (0.00)	77.19 (↑1.03)	69.42 (↑0.19)
	SmoothQuant	8 / 8	8	60.75 (↑0.51)	76.12 (↓0.53)	80.08 (↓0.13)	68.22 (↑0.12)	53.85 (↓0.18)	77.11 (↑0.95)	69.36 (↑0.13)
	AWQ	4 / 16	4	58.53 (↓1.71)	73.39 (↓3.26)	79.10 (↓1.11)	66.26 (↓1.84)	51.87 (↓2.16)	75.37 (↓0.79)	67.42 (↓1.81)
Llama-3.1-70B-it	FP16	16 / 16	140	69.54	88.70	86.74	82.30	59.85	85.40	78.76
	FP8	8 / 8	70	69.45 (↓0.09)	88.25 (↓0.45)	86.69 (↓0.05)	82.02 (↓0.28)	59.80 (↓0.05)	85.08 (↓0.32)	78.55 (↓0.21)
	GPTQ*	4 / 16	35	69.80 (↑0.26)	89.54 (↑0.84)	86.28 (↓0.46)	81.40 (↓0.90)	59.37 (↓0.48)	84.69 (↓0.71)	78.51 (↓0.25)
	GPTQ**	4 / 16	35	69.97 (↑0.43)	89.76 (↑1.06)	86.26 (↓0.48)	81.97 (↓0.33)	58.74 (↓1.11)	84.53 (↓0.87)	78.54 (↓0.22)
	GPTQ**	8 / 16	70	69.03 (↓0.51)	87.95 (↓0.75)	86.29 (↓0.45)	82.17 (↓0.13)	58.94 (↓0.91)	84.53 (↓0.87)	78.15 (↓0.61)
	SmoothQuant	8 / 8	70	70.05 (↑0.51)	88.55 (↓0.15)	86.56 (↓0.18)	82.10 (↓0.20)	60.39 (↑0.54)	85.24 (↓0.16)	78.82 (↑0.06)
	AWQ	4 / 16	35	69.80 (↑0.26)	90.83 (↑2.13)	86.18 (↓0.56)	81.33 (↓0.97)	59.68 (↓0.17)	84.37 (↓1.03)	78.70 (↓0.06)
Llama-3.1-405B-it	FP16	16 / 16	810	73.72	94.84	88.40	83.98	65.42	85.00	81.89
	FP8	8 / 8	405	73.12 (↓0.60)	95.38 (↑0.54)	88.32 (↓0.08)	85.91 (↑1.93)	64.79 (↓0.63)	85.63 (↑0.63)	82.19 (↑0.30)
	GPTQ**	4 / 16	202.5	72.10 (↓1.62)	94.24 (↓0.60)	88.17 (↓0.23)	85.79 (↑1.81)	64.80 (↓0.62)	85.48 (↑0.48)	81.76 (↓0.13)
	SmoothQuant	8 / 8	405	72.01 (↓1.71)	92.72 (↓2.12)	87.53 (↓0.87)	73.28 (↓10.70)	65.19 (↓0.23)	85.95 (↑0.95)	79.45 (↓2.44)
	AWQ	4 / 16	202.5	73.98 (↑0.26)	94.84 (0.00)	88.04 (↓0.36)	85.71 (↑1.73)	64.25 (↓1.17)	86.35 (↑1.35)	82.20 (↑0.31)

Model Distillation

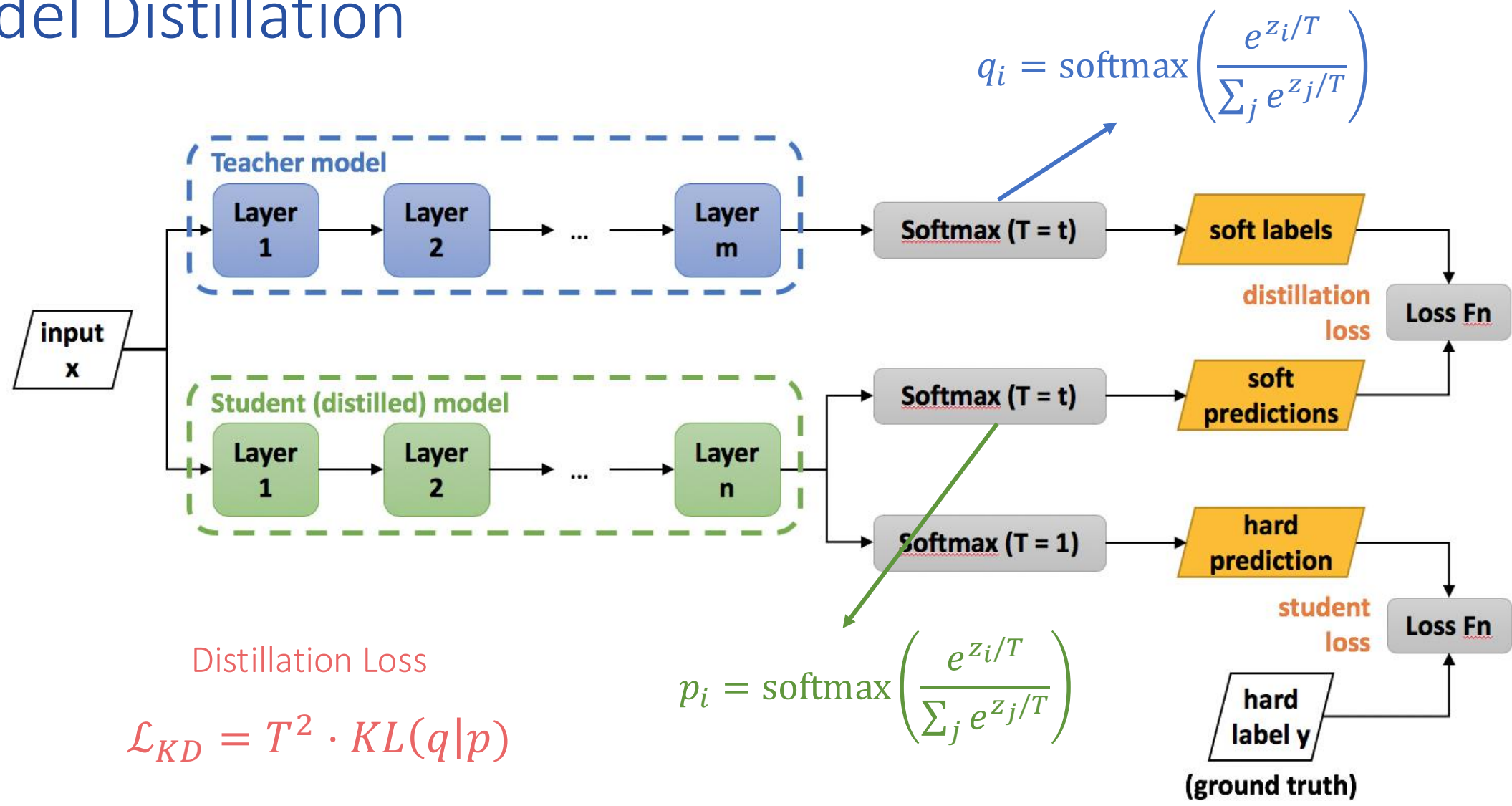
- Distill knowledge from a large model to a small model while maintaining similar capability
 - **Large model:** teacher model
 - **Small model:** student model
- Train a student model to mimic the behavior of the teacher model
- Reduce the number of parameters

Model Distillation

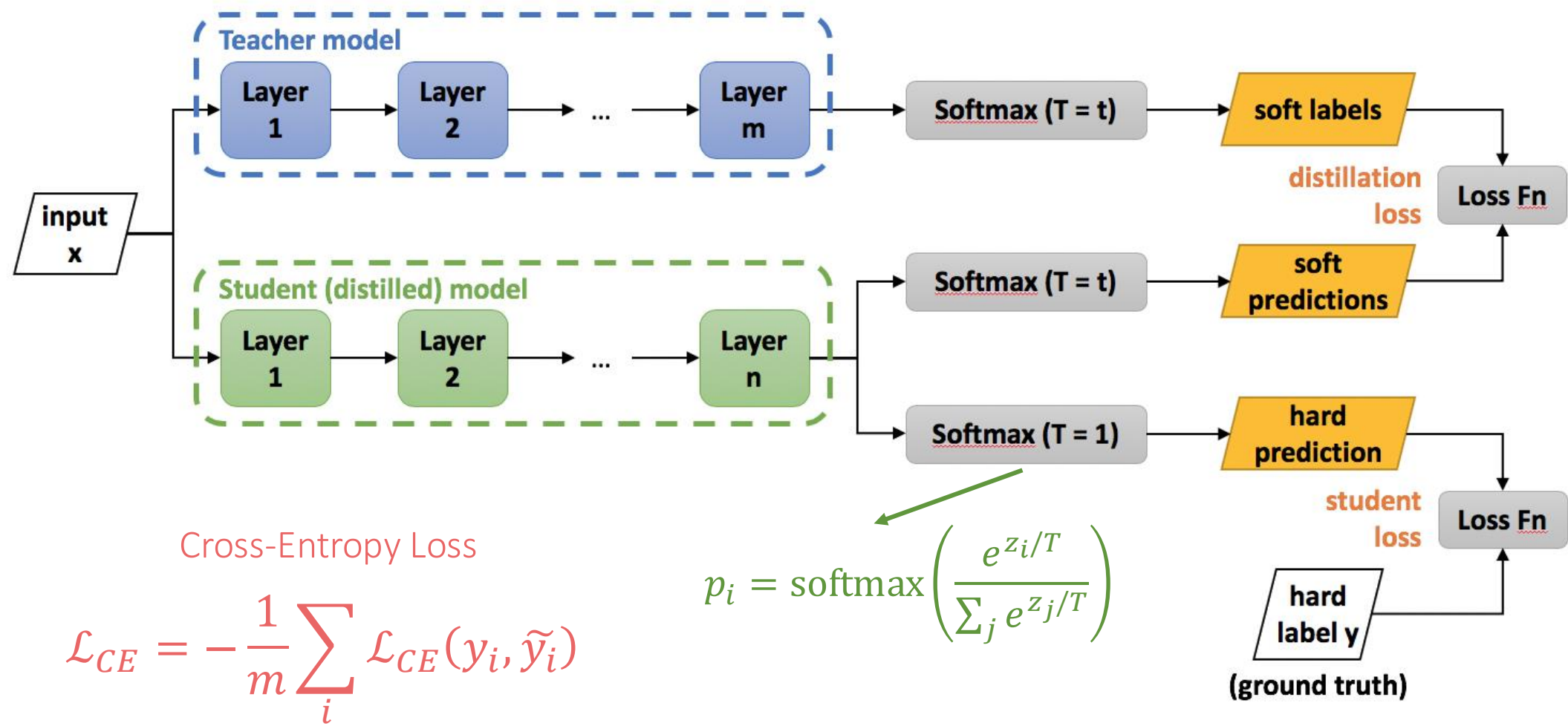


Learn from data

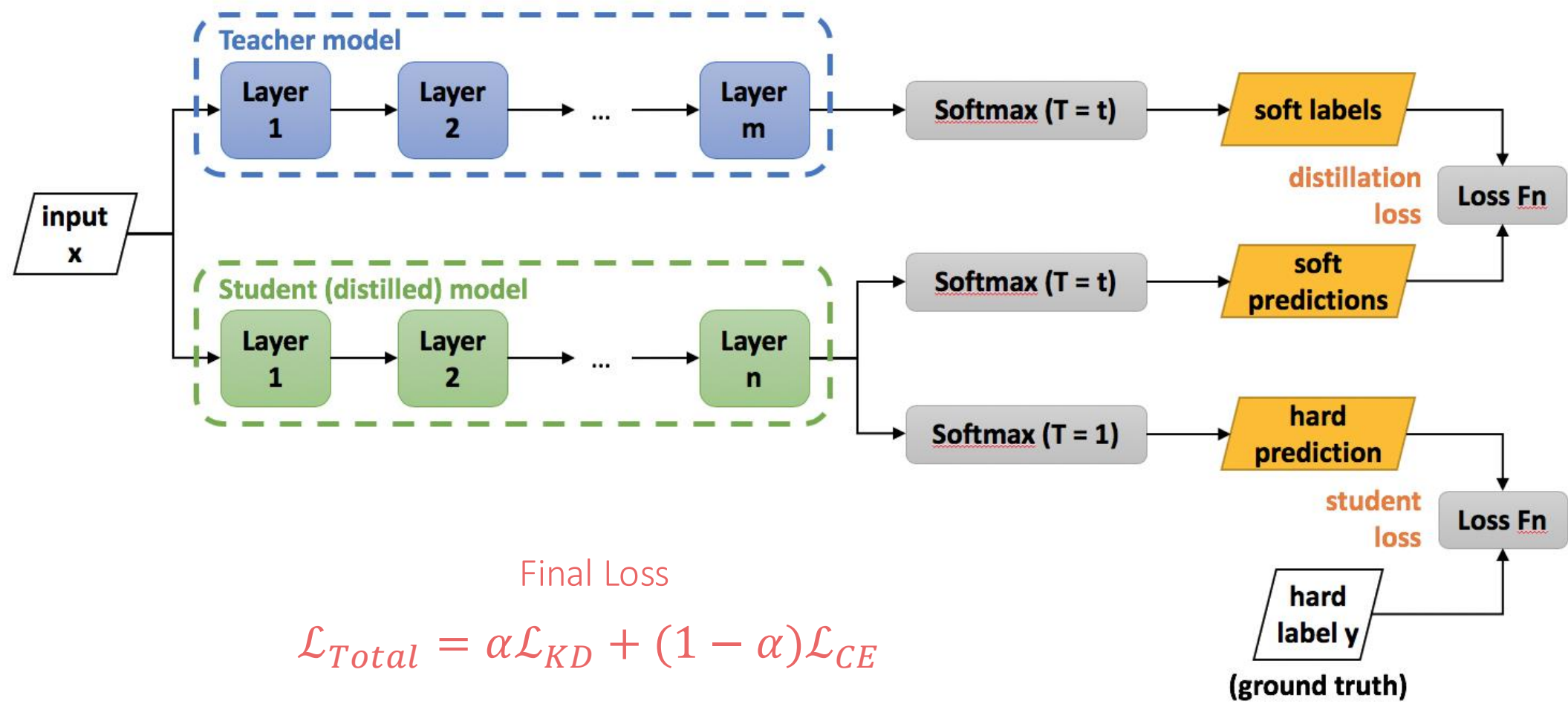
Model Distillation



Model Distillation



Model Distillation



DistilBERT

Smaller Size

Model	# param. (Millions)	Inf. time (seconds)
ELMo	180	895
BERT-base	110	668
DistilBERT	66	410

- BERT-base
 - 12 layers, hidden size = 768, 12 attention heads
- DistilBERT
 - 6 layers, hidden size = 768, 12 attention heads

Almost similar performance

Model	Score	CoLA	MNLI	MRPC	QNLI	QQP	RTE	SST-2	STS-B	WNLI
ELMo	68.7	44.1	68.6	76.6	71.1	86.2	53.4	91.5	70.4	56.3
BERT-base	79.5	56.3	86.7	88.6	91.8	89.6	69.3	92.7	89.0	53.5
DistilBERT	77.0	51.3	82.2	87.5	89.2	88.5	59.9	91.3	86.9	56.3

LLM Distillation

- Smaller and faster
 - Sacrifice a little bit performance
- Distillation → Copy training data

 **deepseek-ai/DeepSeek-R1-Distill-Llama-70B**

like 742

Follow DeepSeek 119k

Text Generation

Transformers

Safetensors

llama

conversational

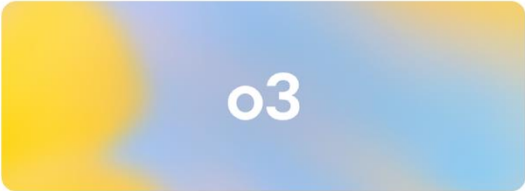
text-generation-inference

arxiv:2501.12948

License: mit

o3

o3-mini





Reasoning model for complex tasks, succeeded by GPT-5

A small model alternative to o3

Learn more

Learn more

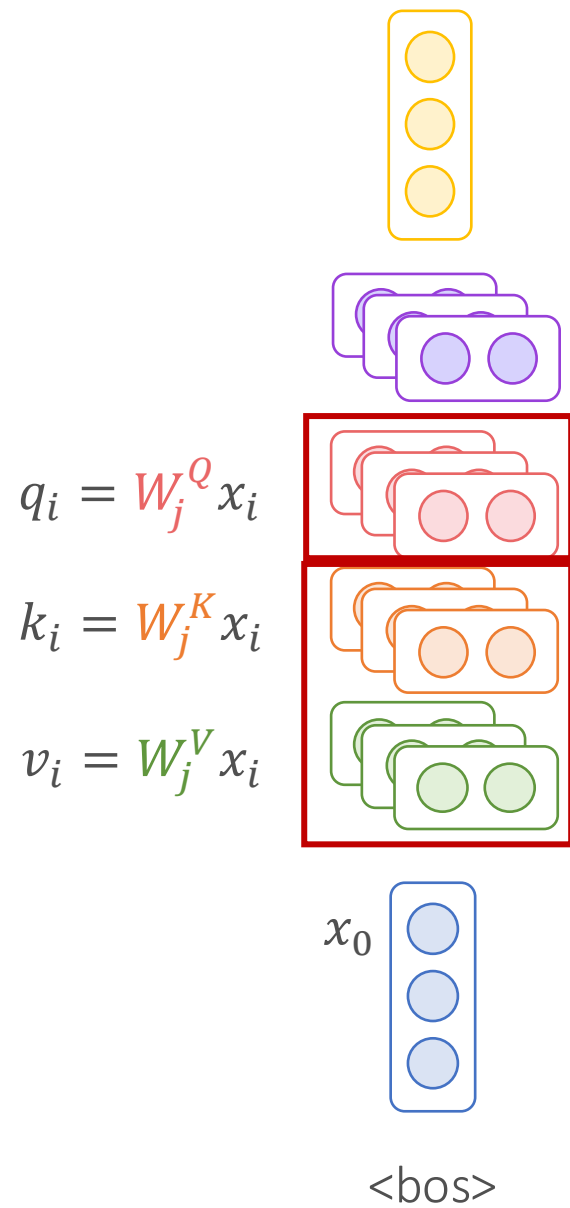
Playground

Playground

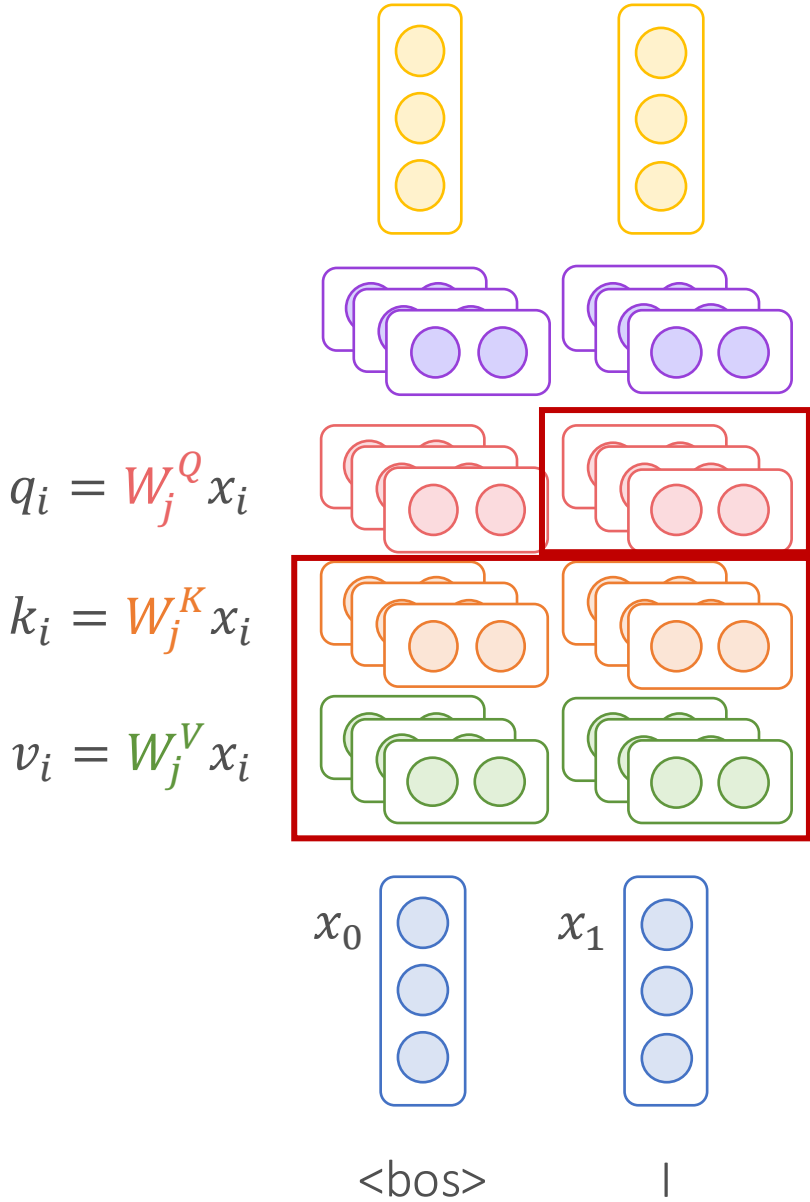
Reasoning	● ● ● ● ●
Speed	⚡
Input	T 📄 🗑️ 🗑️
Output	T 🗑️ 🗑️ 🗑️
Reasoning tokens	✓

Reasoning	● ● ● ● ●
Speed	⚡ ⚡ ⚡
Input	T 📄 🗑️ 🗑️
Output	T 🗑️ 🗑️ 🗑️
Reasoning tokens	✓

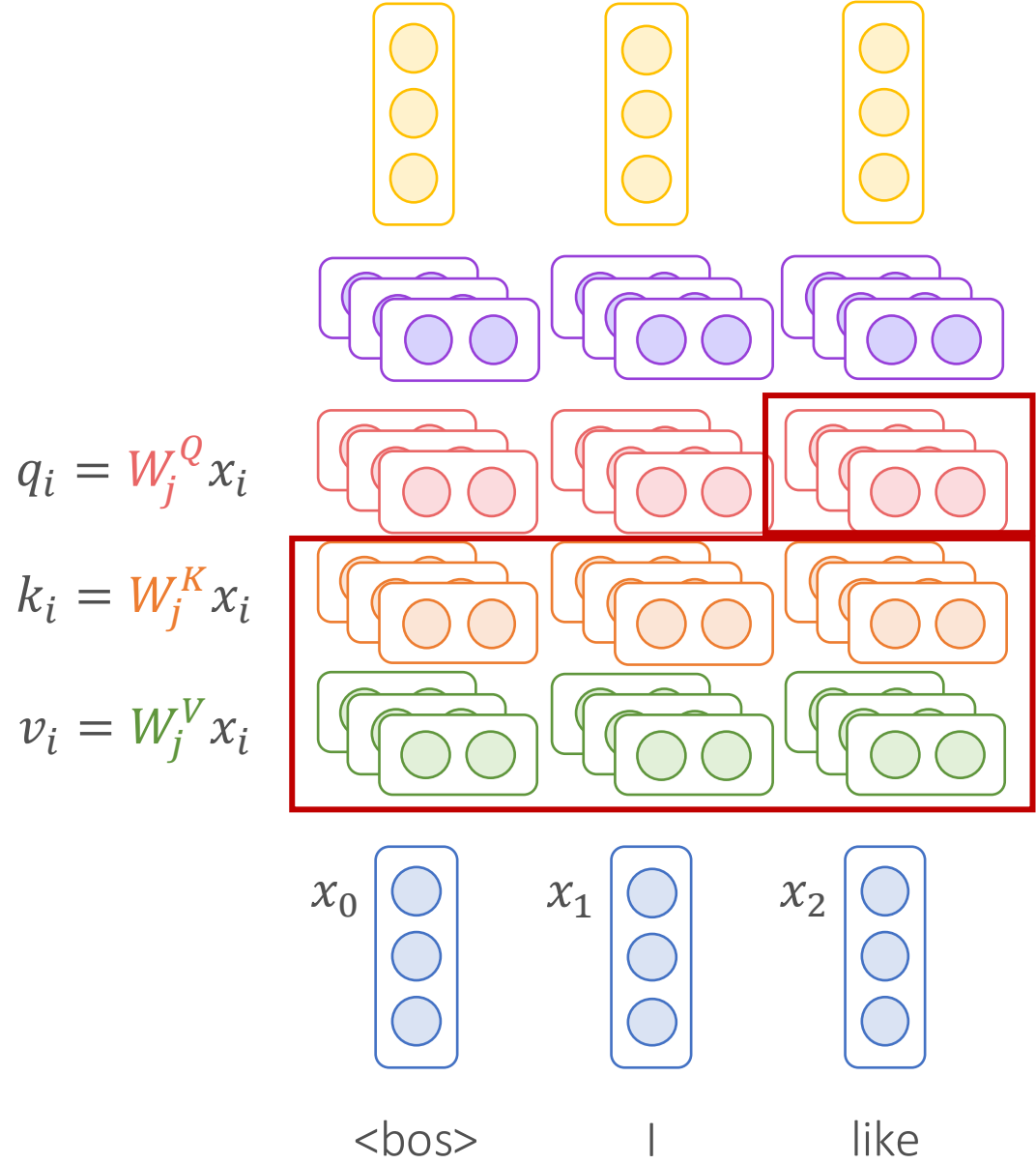
LLM Inference



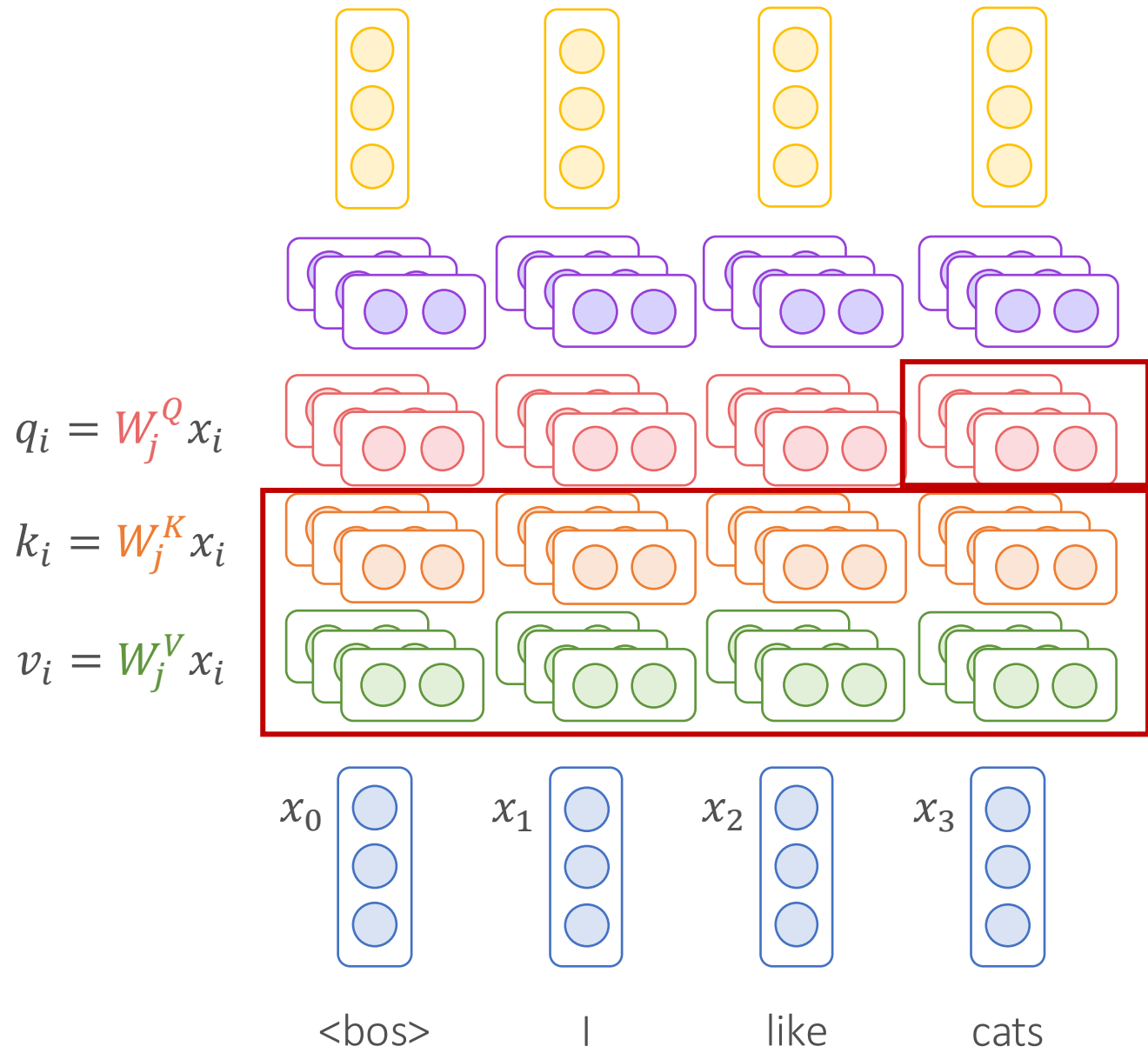
LLM Inference



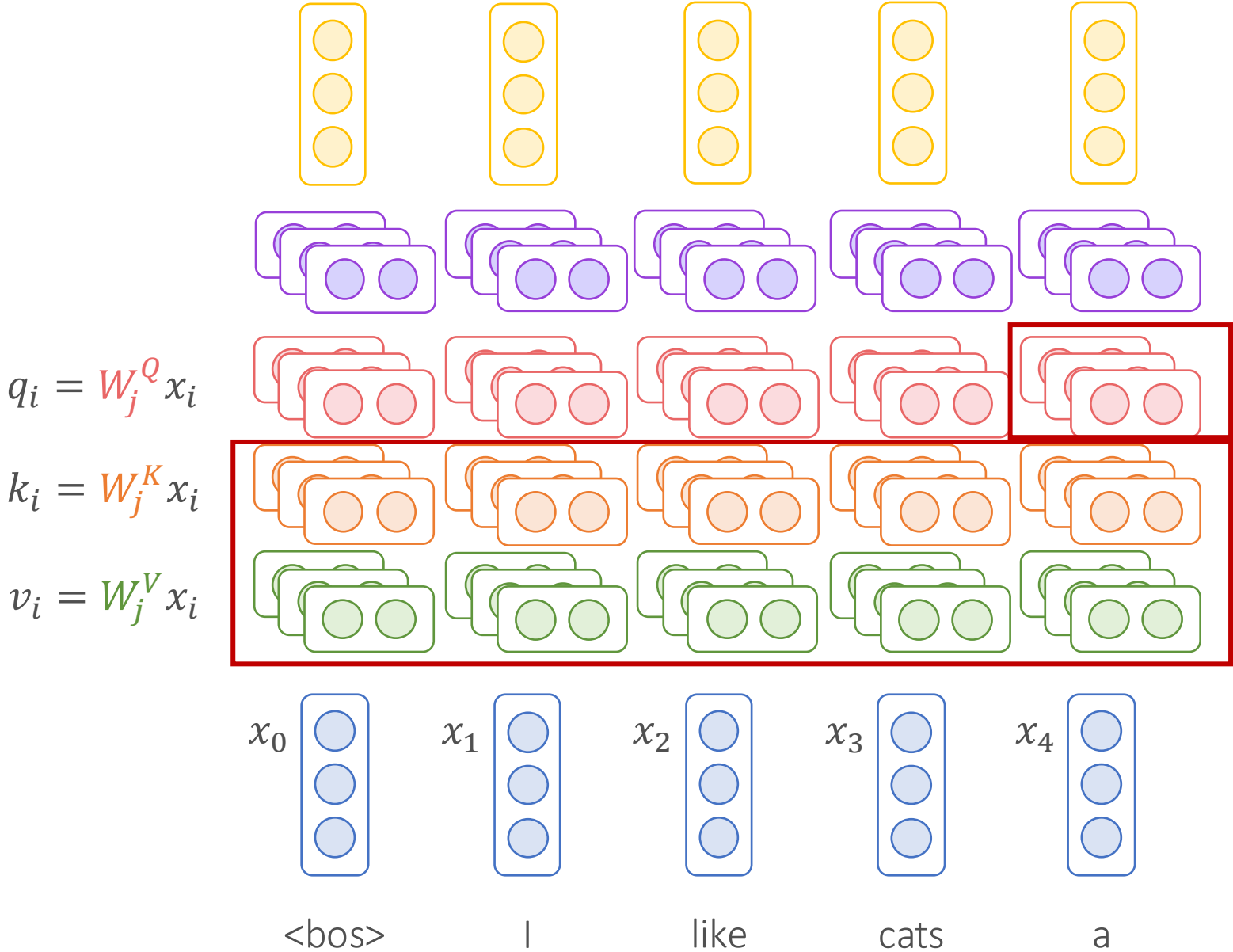
LLM Inference



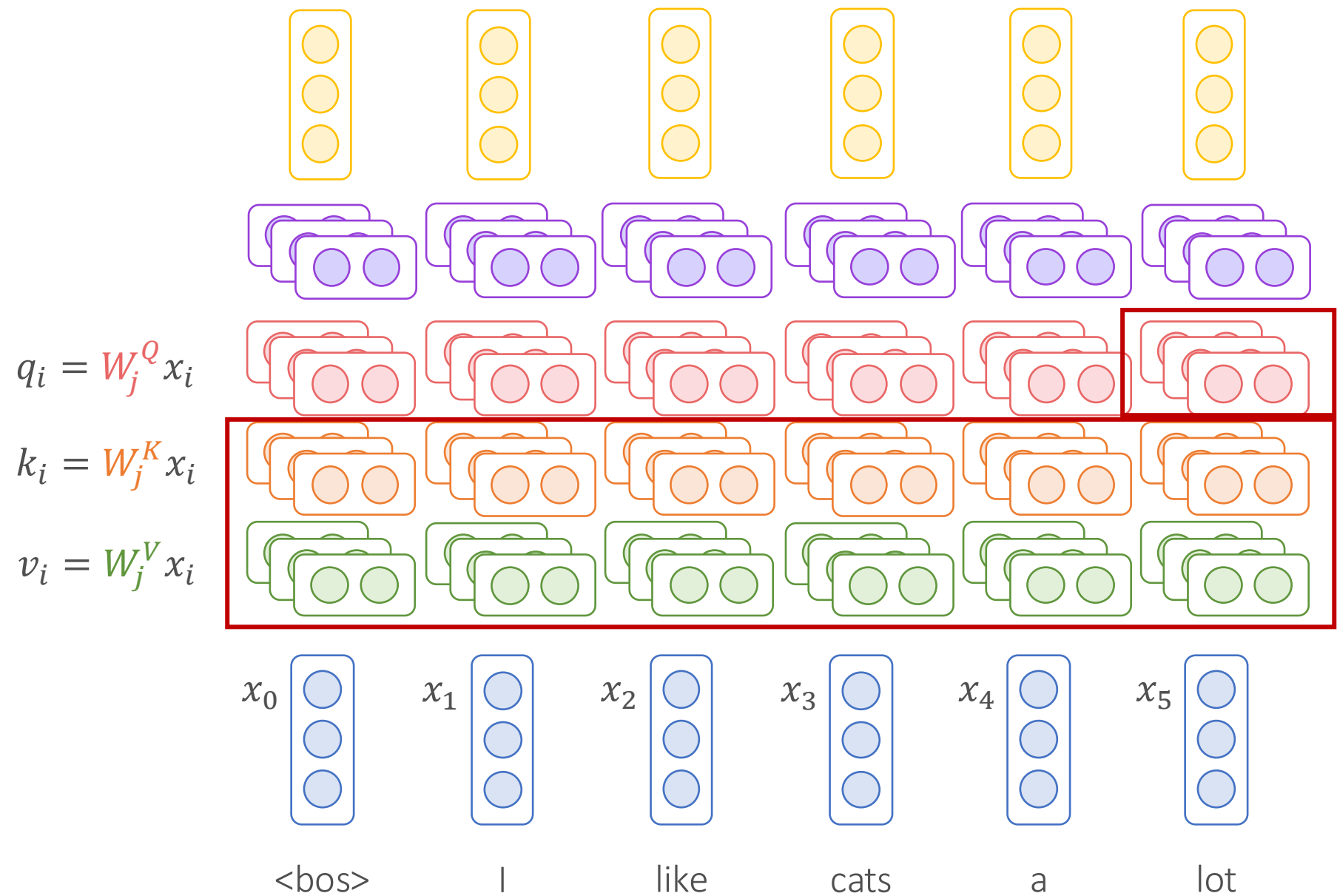
LLM Inference



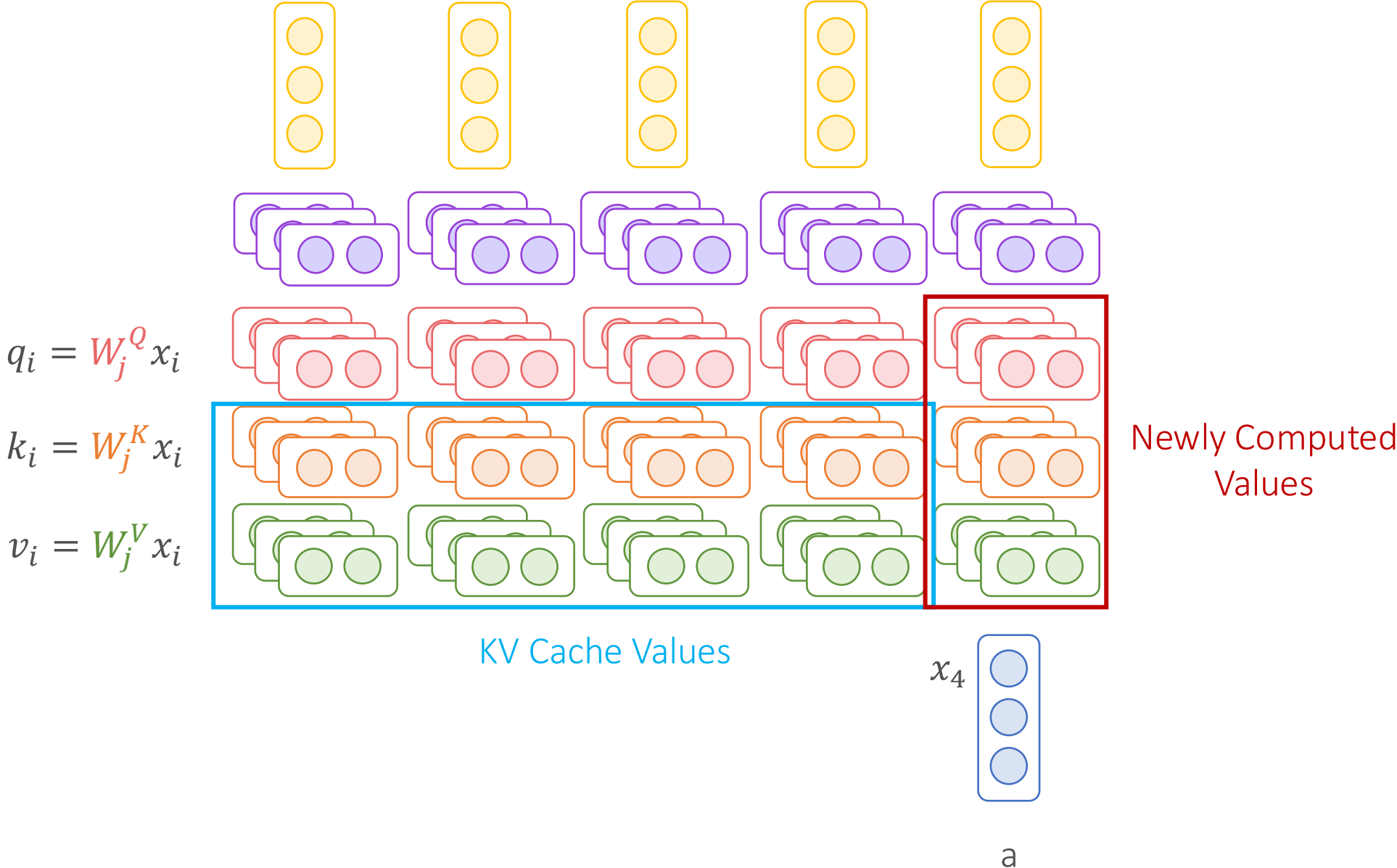
LLM Inference



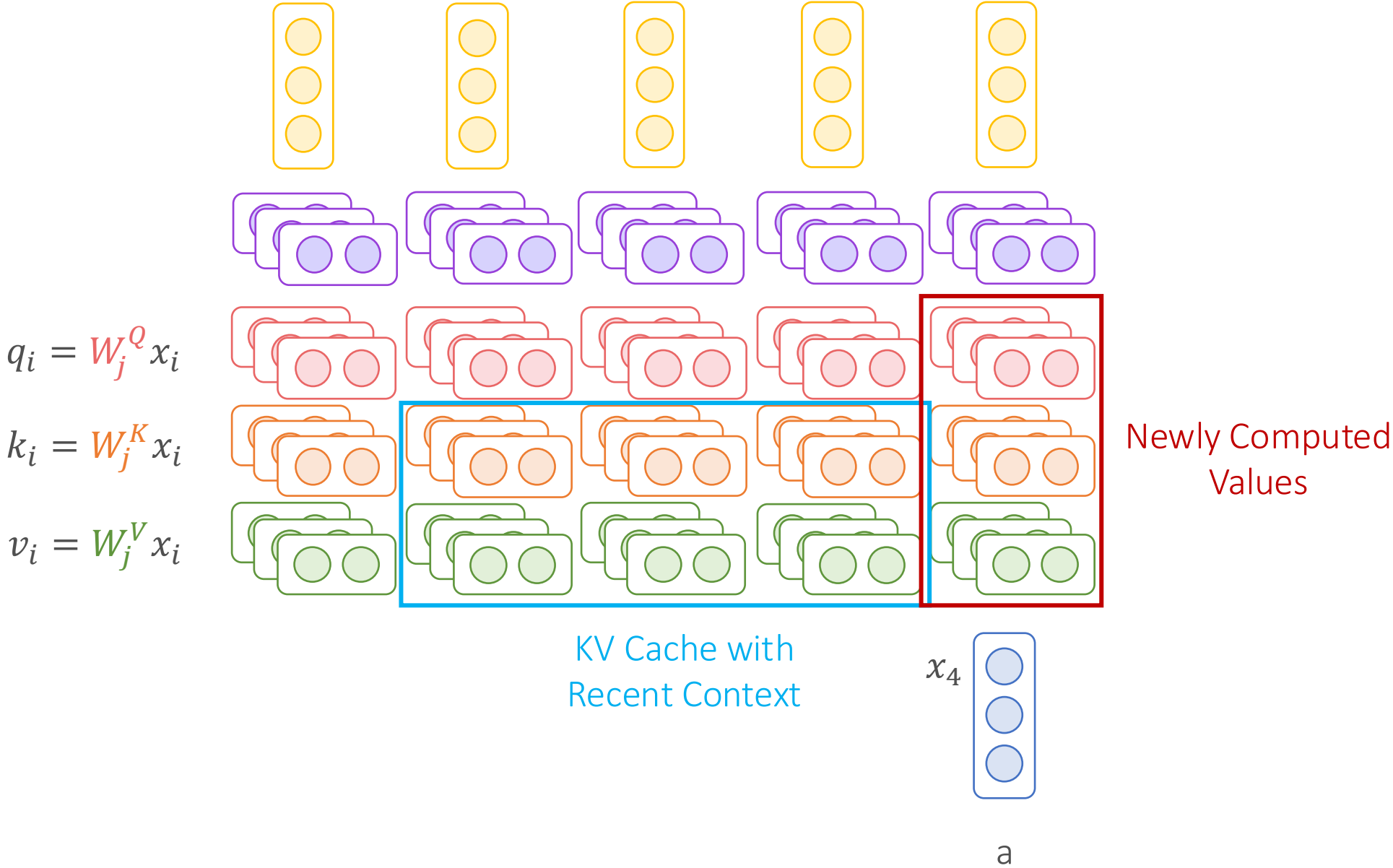
LLM Inference



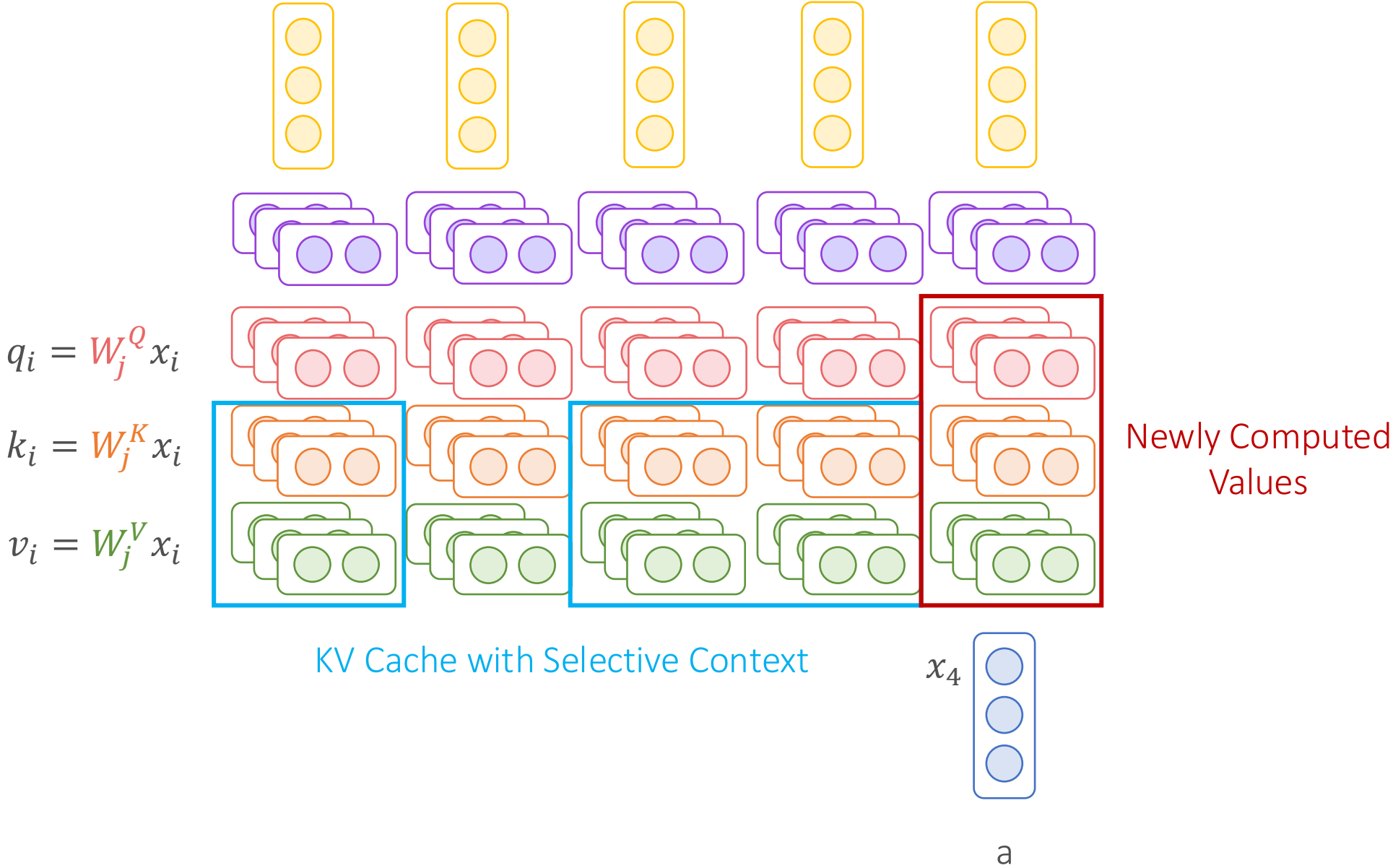
KV Cache



Selective KV Cache



Selective KV Cache



Questions?