CSCE 638 Natural Language Processing Foundation and Techniques

Lecture 11: Parameter-Efficient Fine-Tuning and Large Language Models

Kuan-Hao Huang

Spring 2025

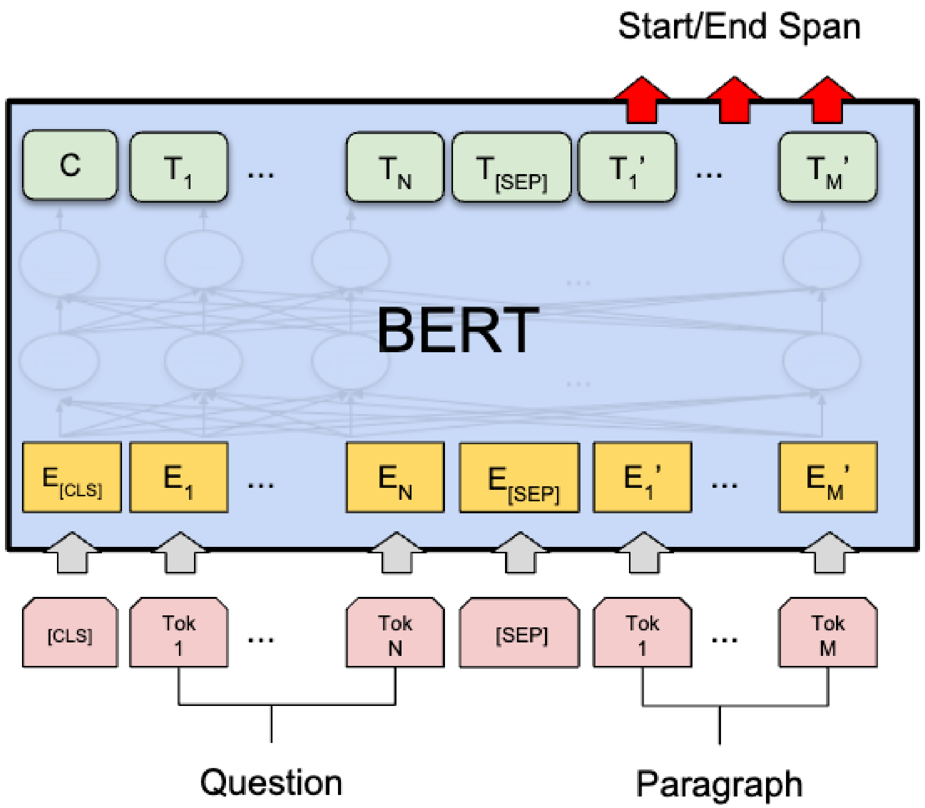
(Some slides adapted from Vivian Chen and Graham Neubig)

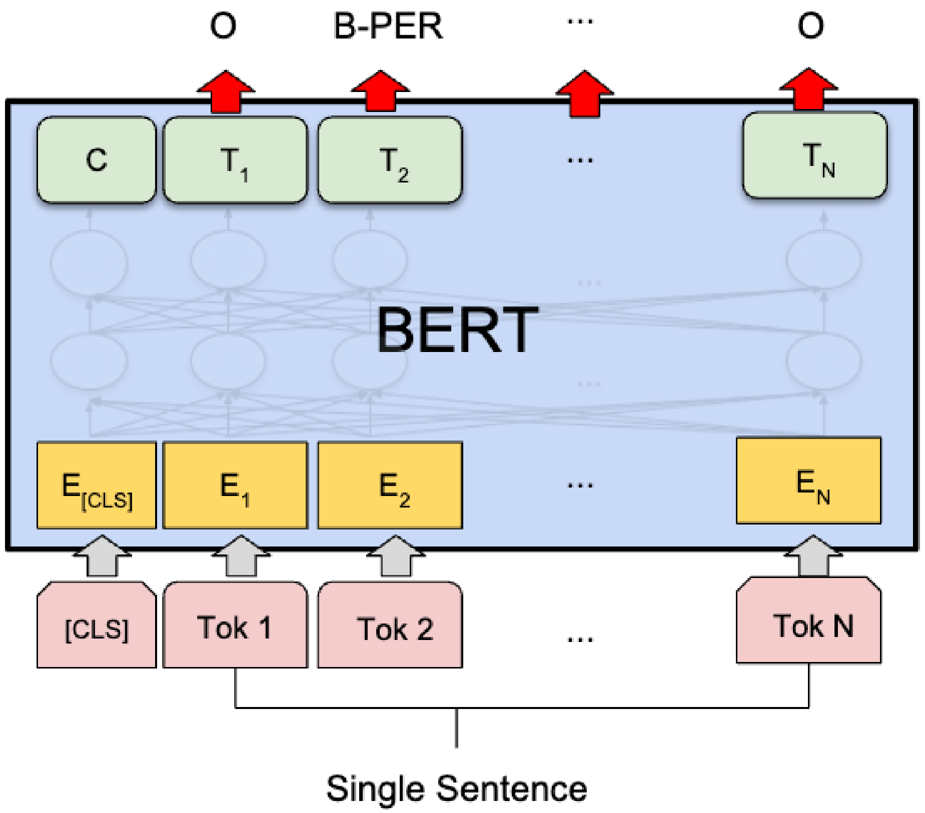
Lecture Plan

- Parameter-Efficient Fine-Tuning
  - Prompt Tuning
  - Prefix Tuning
  - Adapter
  - Mixture of Experts
  - LoRA
- Large Language Models

Look Back at Encoder: Fine-Tuning Token-Level Tasks

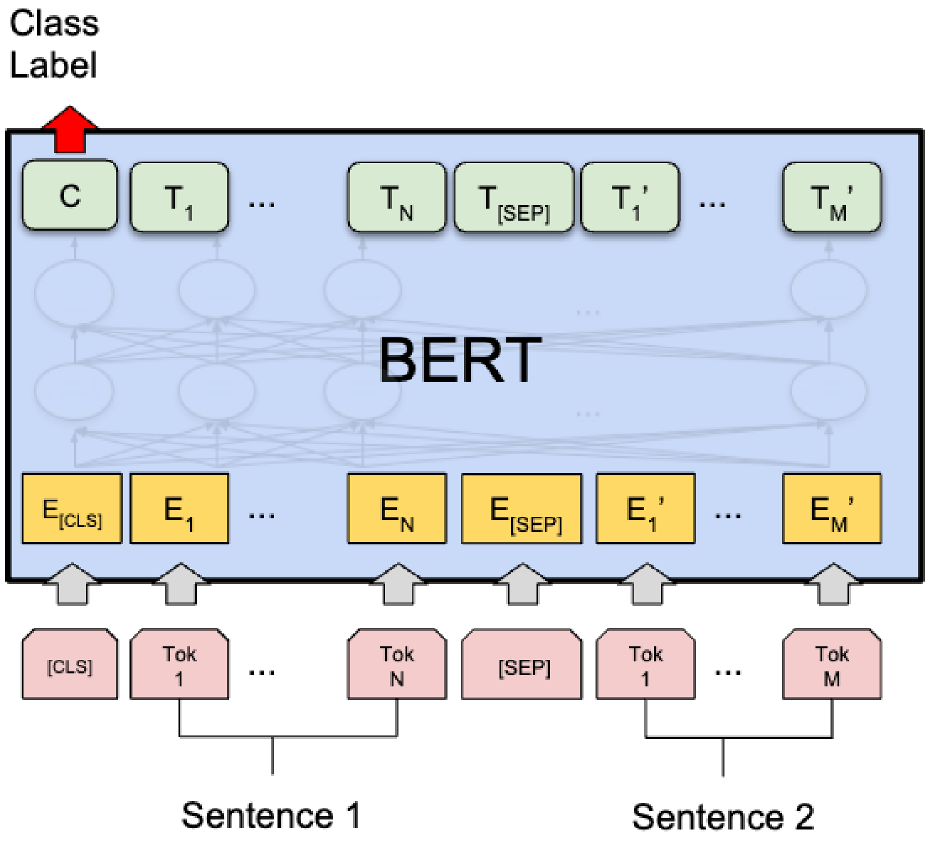
- Pre-training provides a good weight initialization

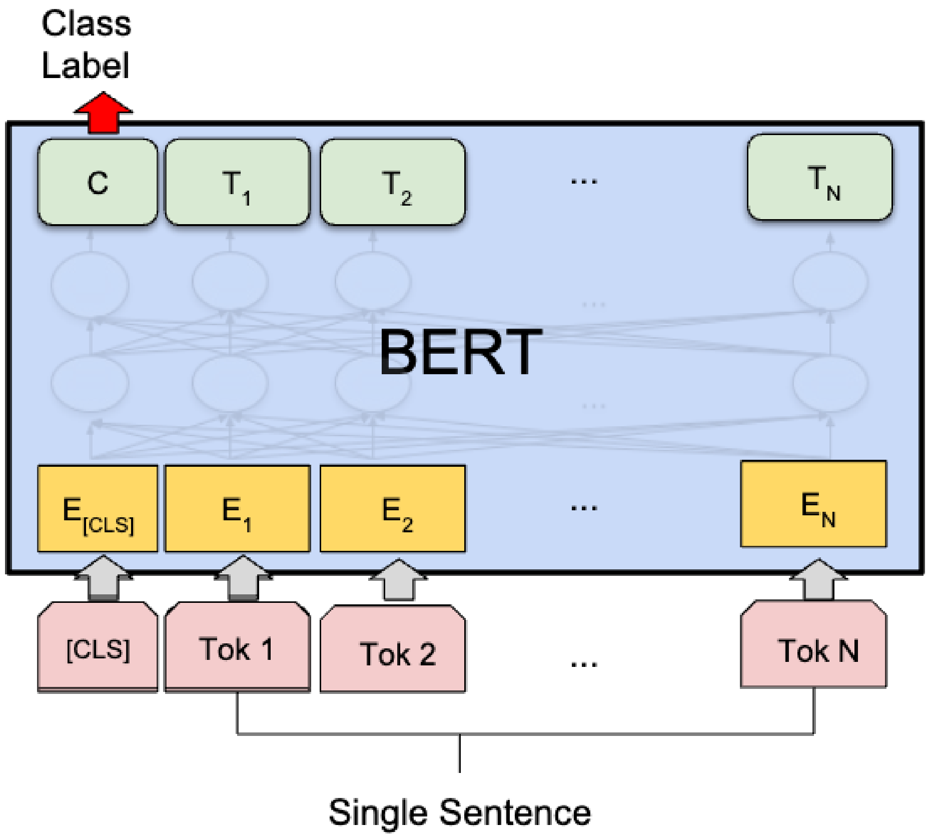




Look Back at Encoder: Fine-Tuning Sentence-Level Tasks

- Pre-training provides a good weight initialization

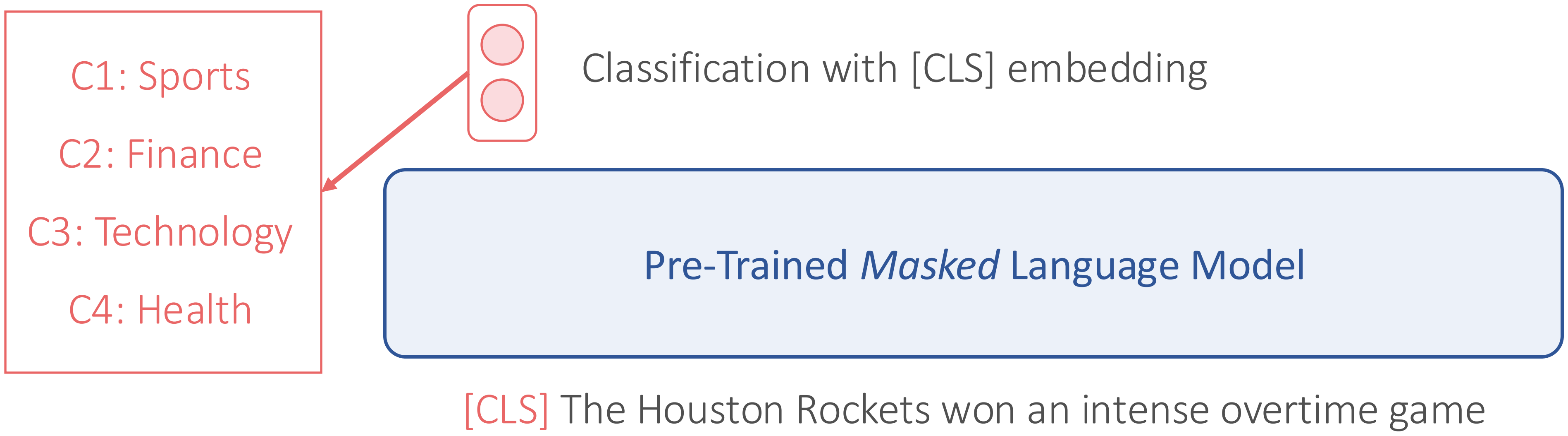




Classification with [CLS] Embedding

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



C1: Sports
C2: Finance
C3: Technology
C4: Health

Classification with [CLS] embedding

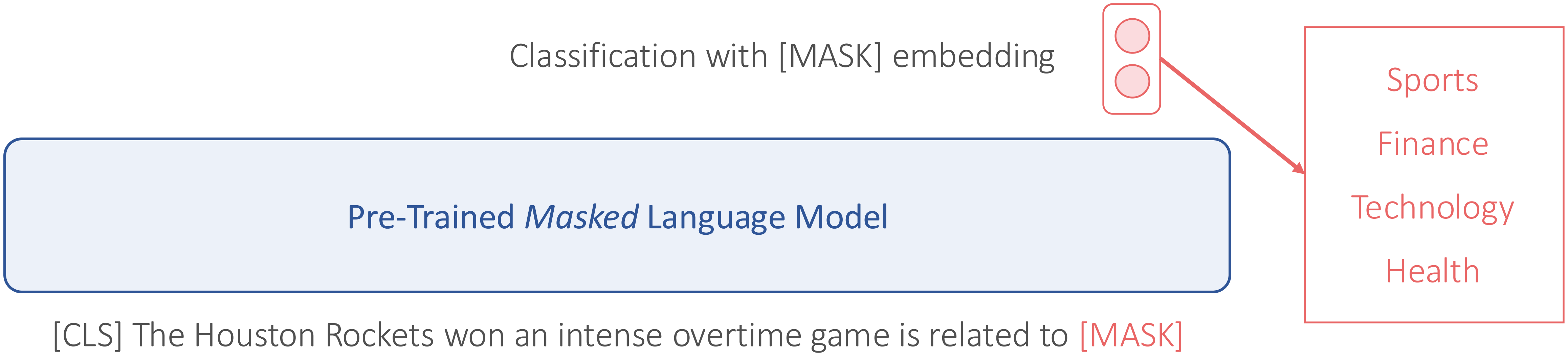
Pre-Trained *Masked* Language Model

[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]

Sports

Finance

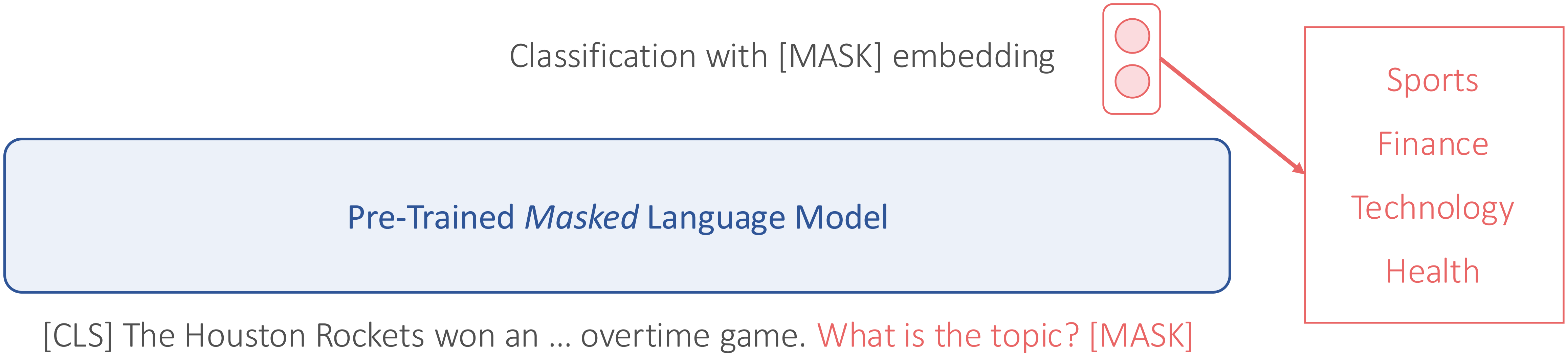
Technology

Health

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

Sports
Finance
Technology
Health

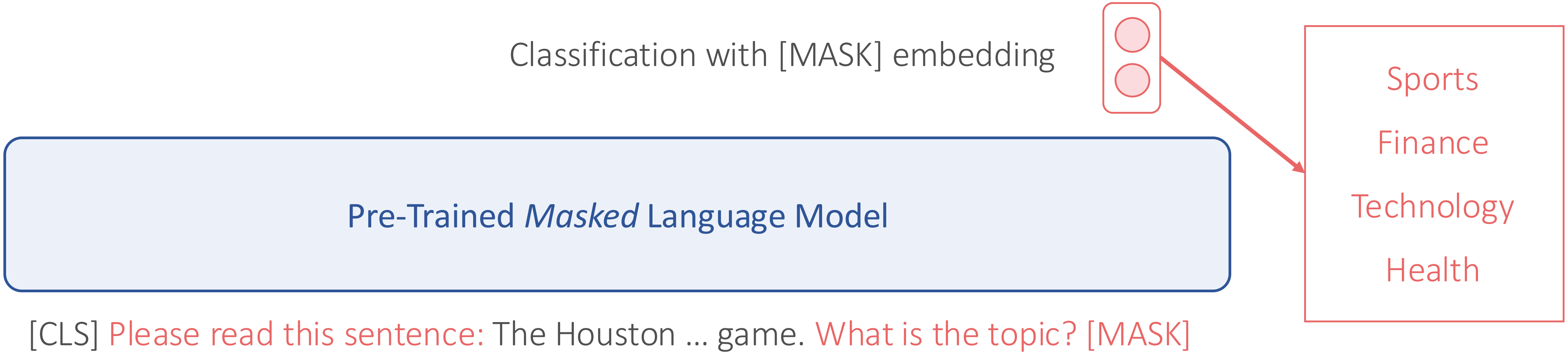
Pre-Trained *Masked* Language Model

[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

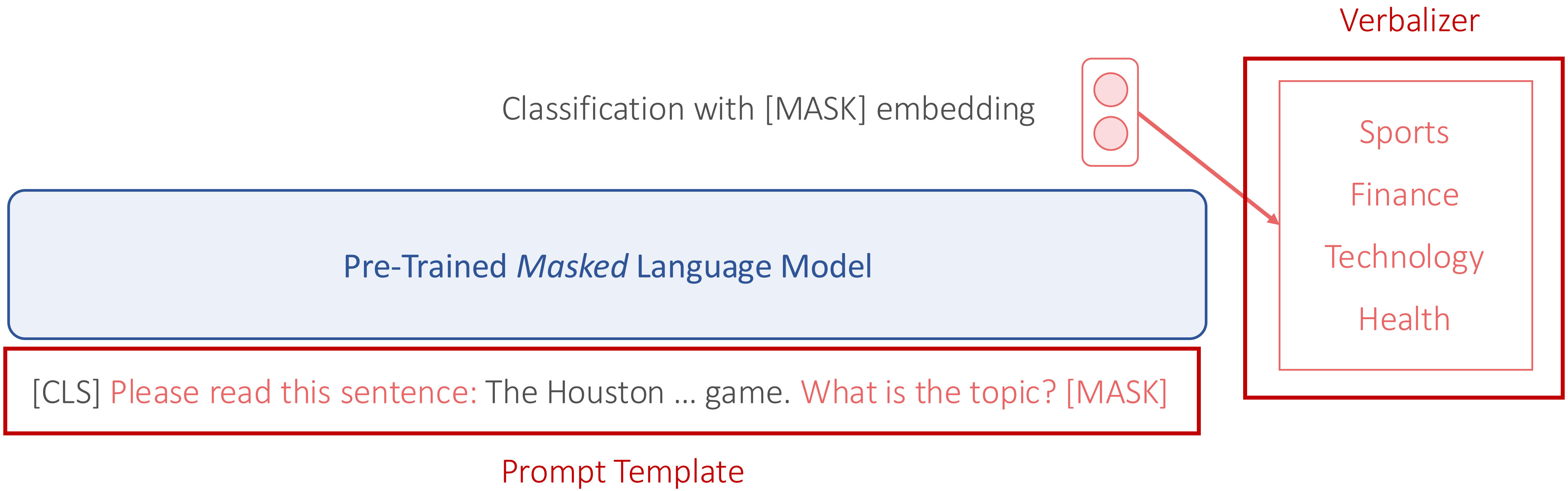


Classification with [MASK] embedding

Pre-Trained *Masked* Language Model

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

Prompt Tuning



Verbalizer

Classification with [MASK] embedding

Sports

Finance

Technology

Health

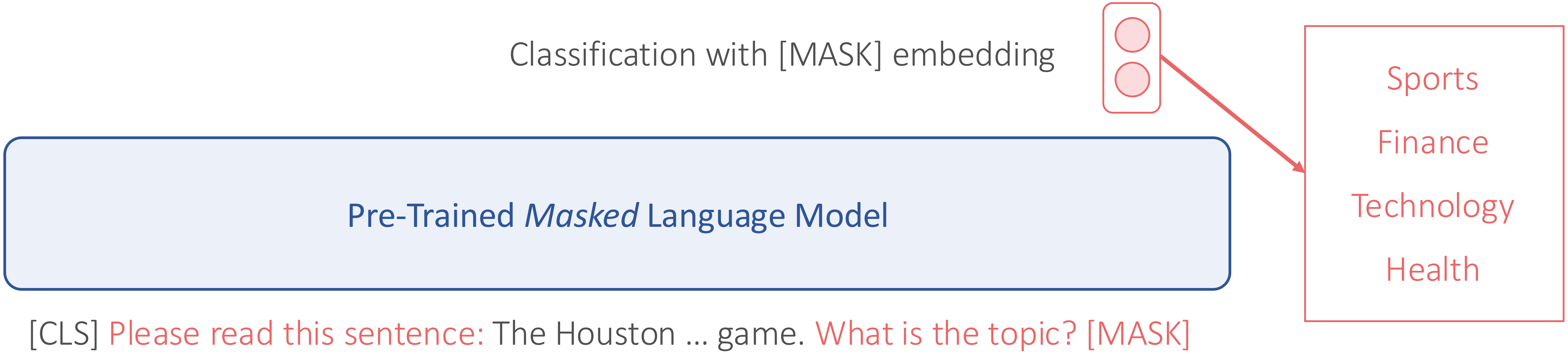
Pre-Trained *Masked* Language Model

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

Prompt Template

Prompt Tuning

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



Classification with [MASK] embedding

Pre-Trained *Masked* Language Model

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

Sports

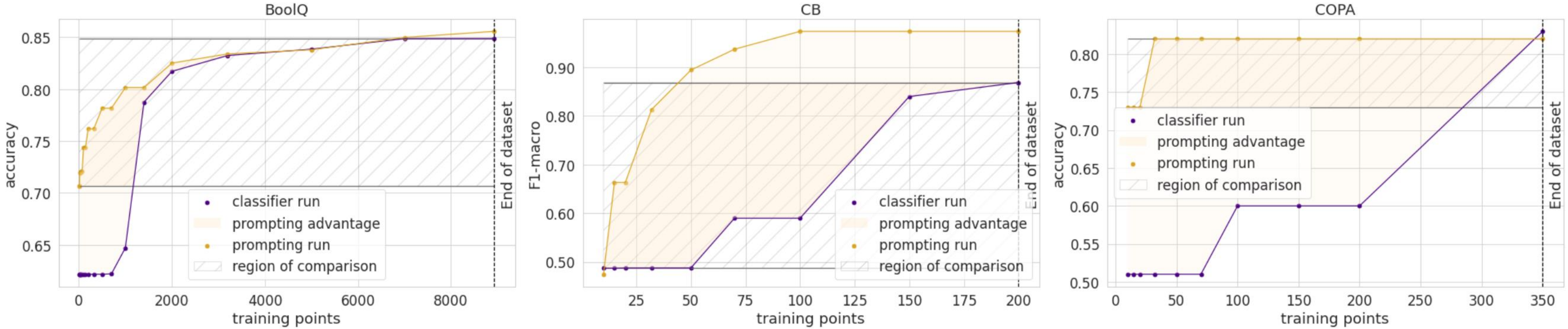
Finance

Technology

Health

Prompt Tuning

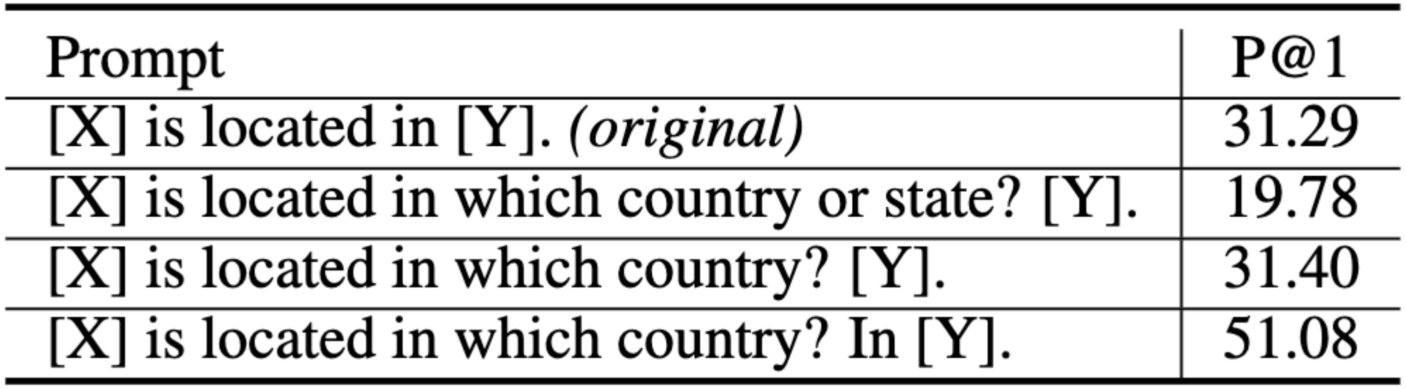
- Better utilize label semantics and pre-trained knowledge
  - Verbalizer
- Can make zero-shot predictions
- Require less training examples



How Many Data Points is a Prompt Worth? 2021

Issues of Discrete/Hard Prompts

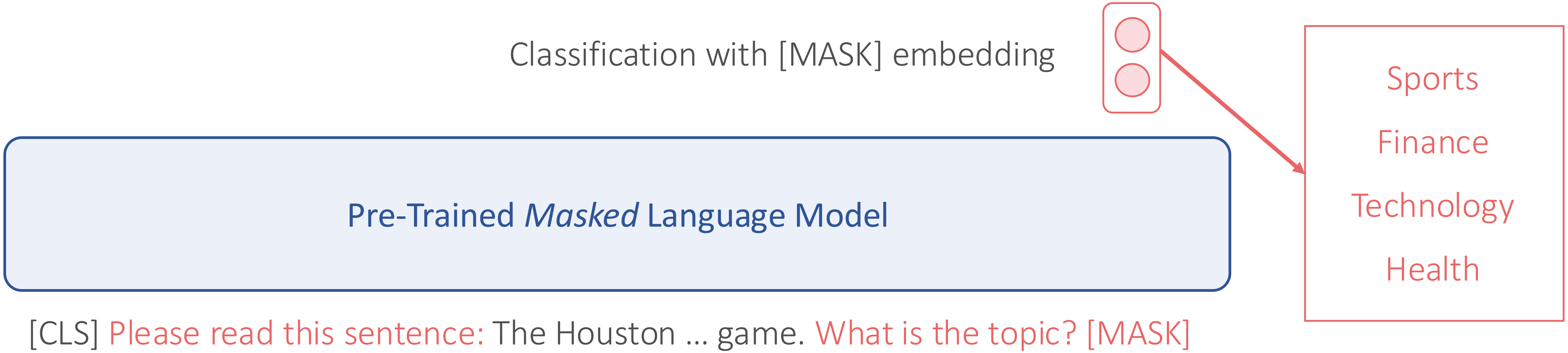
- Manually design prompts can be difficult
  - Which one is the best?
- Pre-trained models are sensitive to prompts



Prompt	P@1
[X] is located in [Y]. *(original)*	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

GPT Understands, Too, 2021

Hard Prompt Tuning



Classification with [MASK] embedding

Pre-Trained *Masked* Language Model

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

Sports

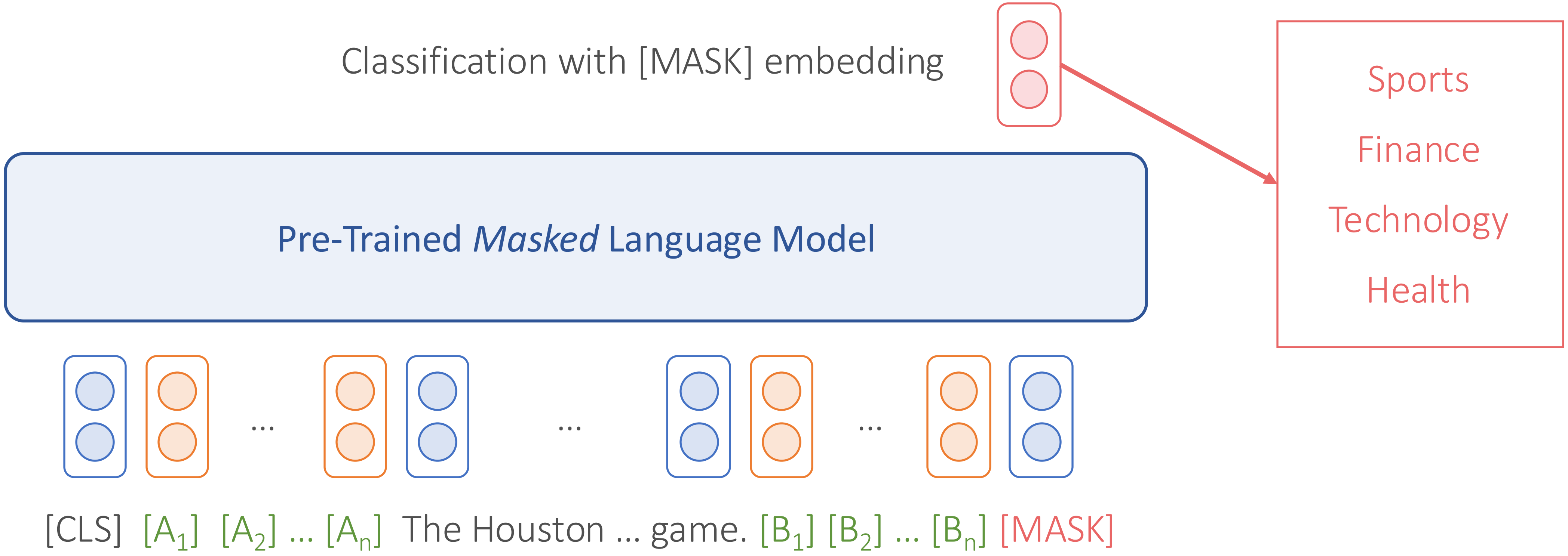
Finance

Technology

Health

Soft Prompt Tuning

- Let model learn good prompts by itself



Classification with [MASK] embedding

Pre-Trained *Masked* Language Model

Sports

Finance

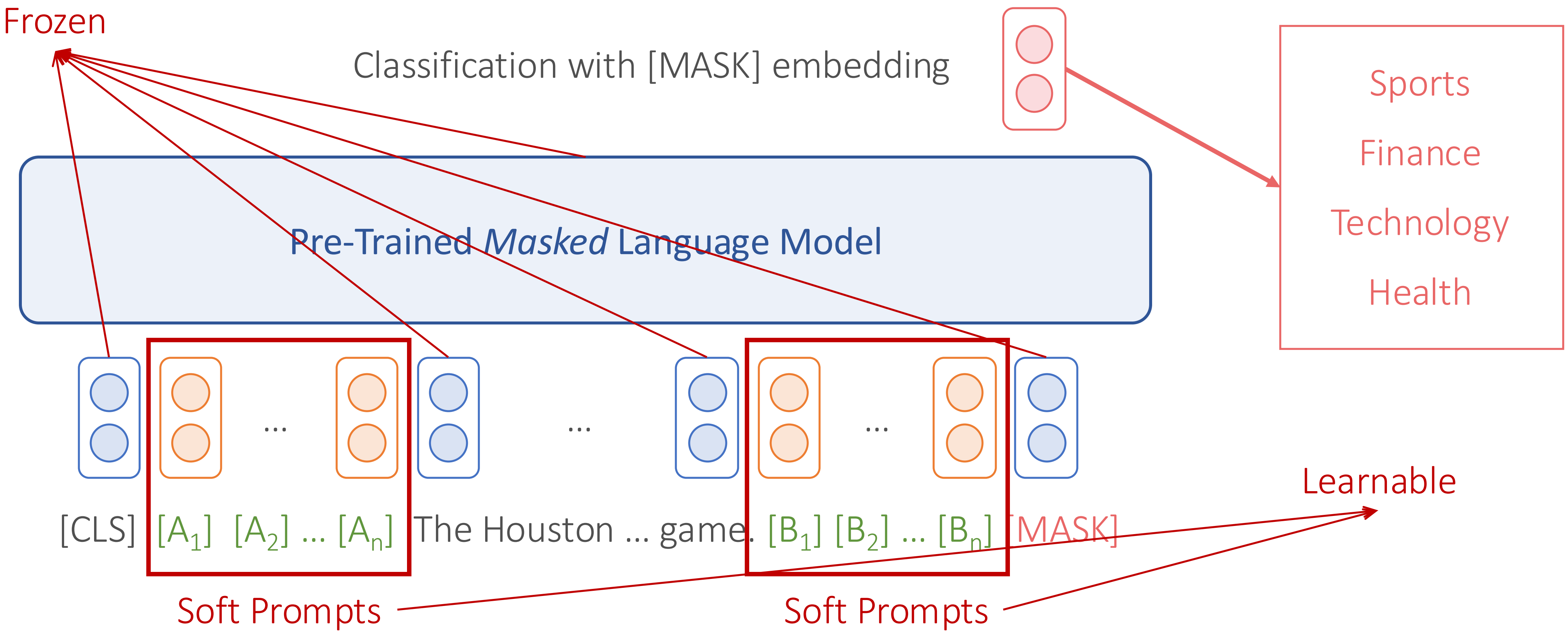
Technology

Health

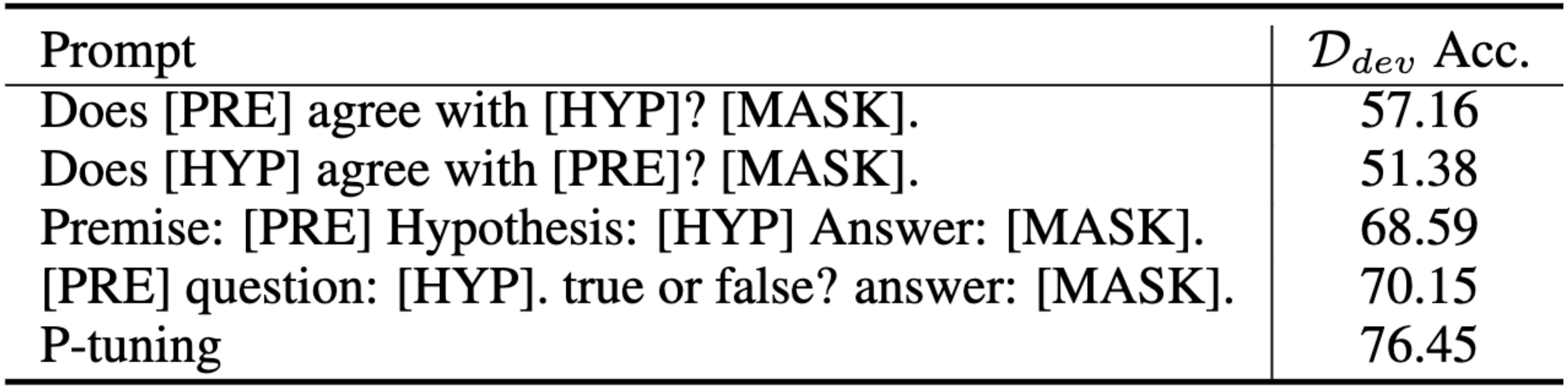
[CLS] [A_1] [A_2] … [A_n] The Houston … game. [B_1] [B_2] … [B_n] [MASK]

Soft Prompt Tuning

- Let model learn good prompts by itself


Frozen
Classification with [MASK] embedding
Pre-Trained *Masked* Language Model
Sports
Finance
Technology
Health
…
…
…
Learnable
[CLS] [A_1] [A_2] … [A_n] The Houston … game. [B_1] [B_2] … [B_n] [MASK]
Soft Prompts
Soft Prompts

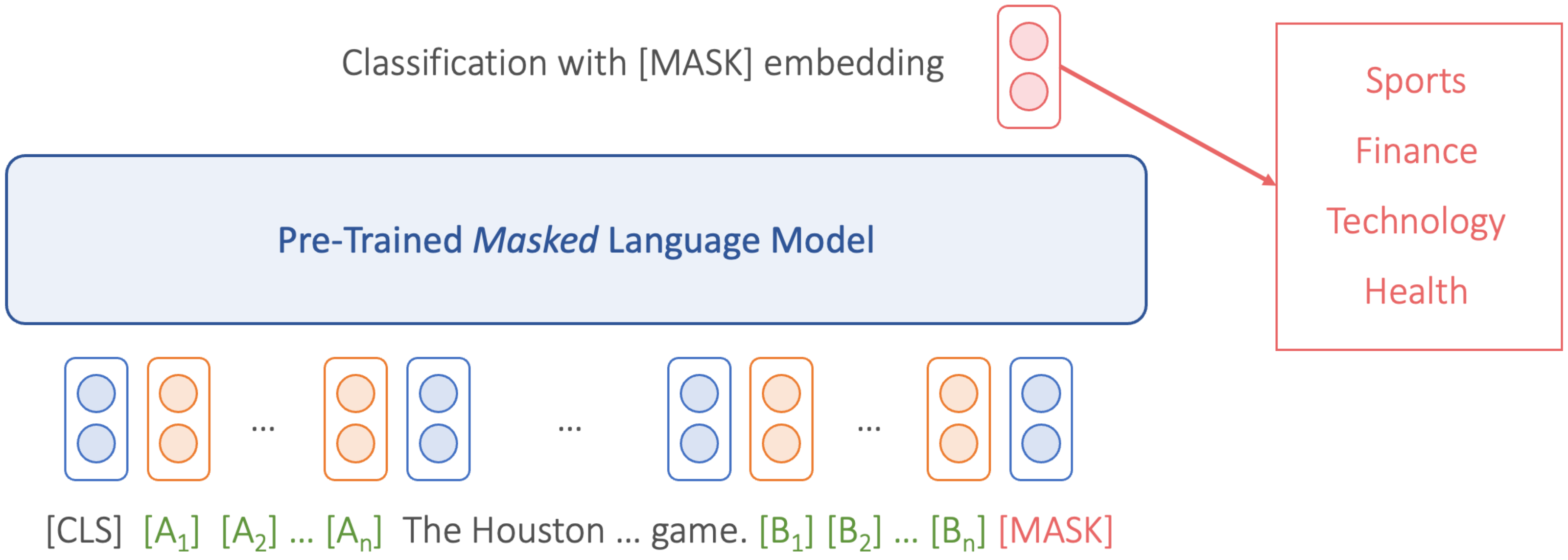
Soft Prompt Tuning



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

GPT Understands, Too, 2021

From Prompt Tuning to Prefix Tuning

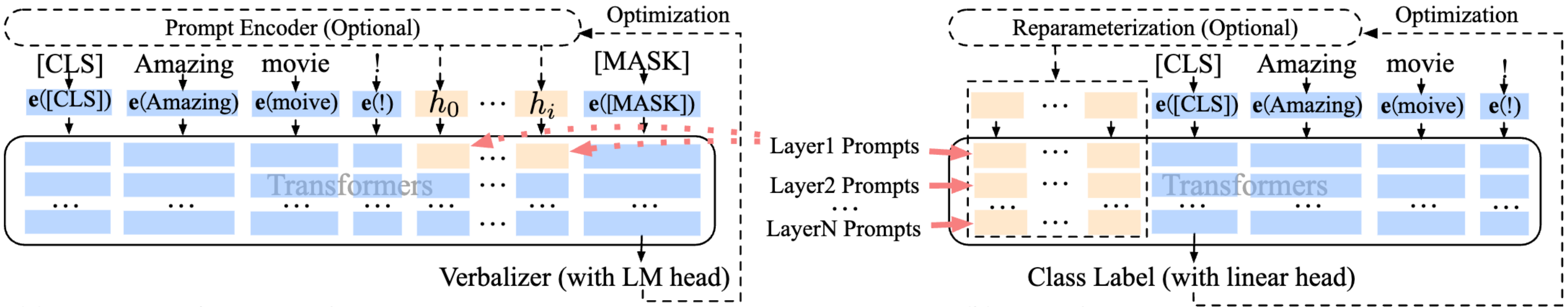


Classification with [MASK] embedding

Pre-Trained *Masked* Language Model

Sports
Finance
Technology
Health

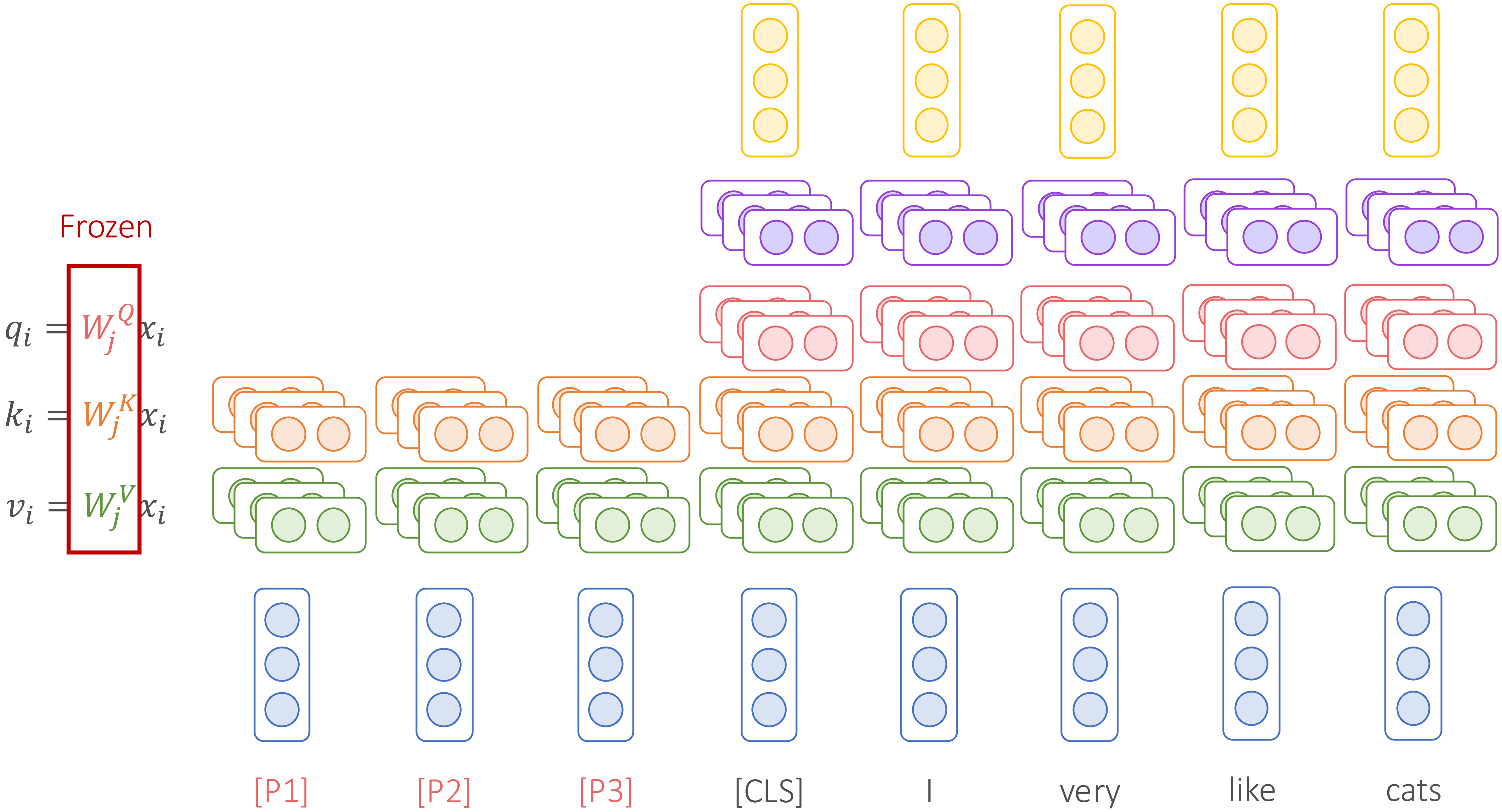
[CLS] [A_1] [A_2] … [A_n] The Houston … game. [B_1] [B_2] … [B_n] [MASK]



(a) Lester et al. & P-tuning (Frozen, 10-billion-scale, simple tasks) (b) P-tuning v2 (Frozen, most scales, most tasks)

P-Tuning v2: Prompt Tuning Can Be Comparable to Fine-tuning Universally Across Scales and Tasks, 2021

Prefix Tuning



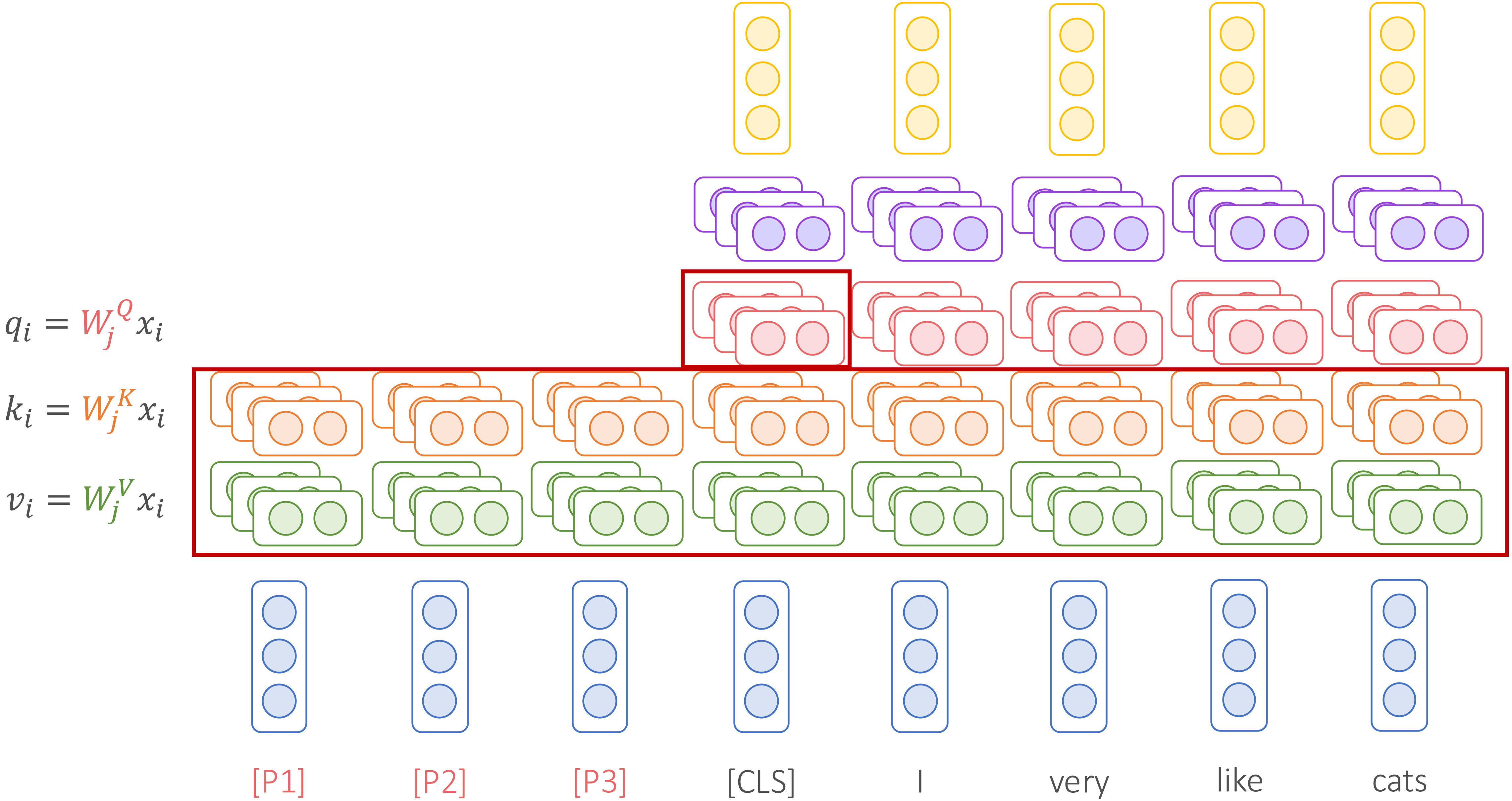
Frozen

$$q_i = W_j^Q x_i$$

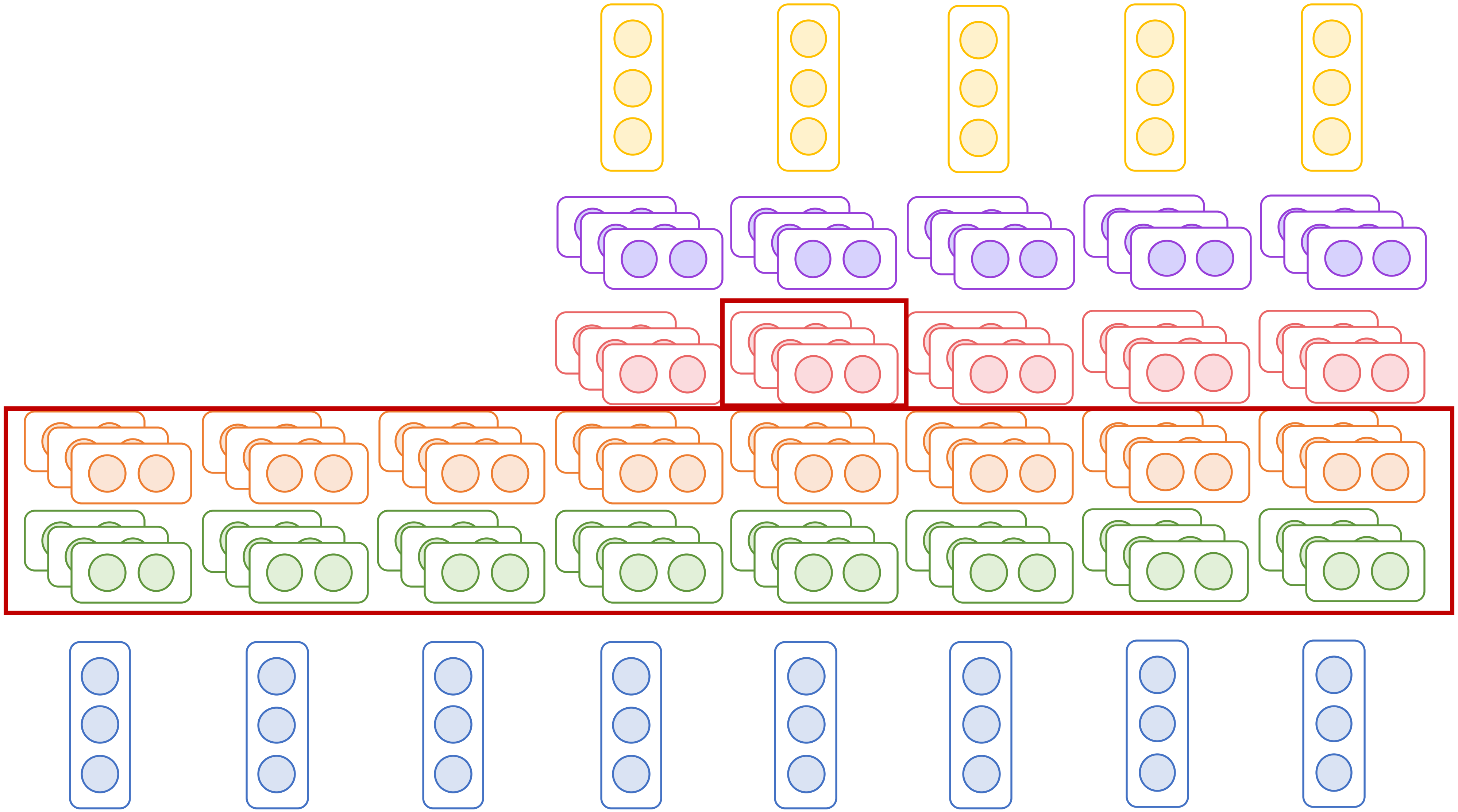
$$k_i = W_j^K x_i$$

$$v_i = W_j^V x_i$$

Prefix Tuning



Prefix Tuning

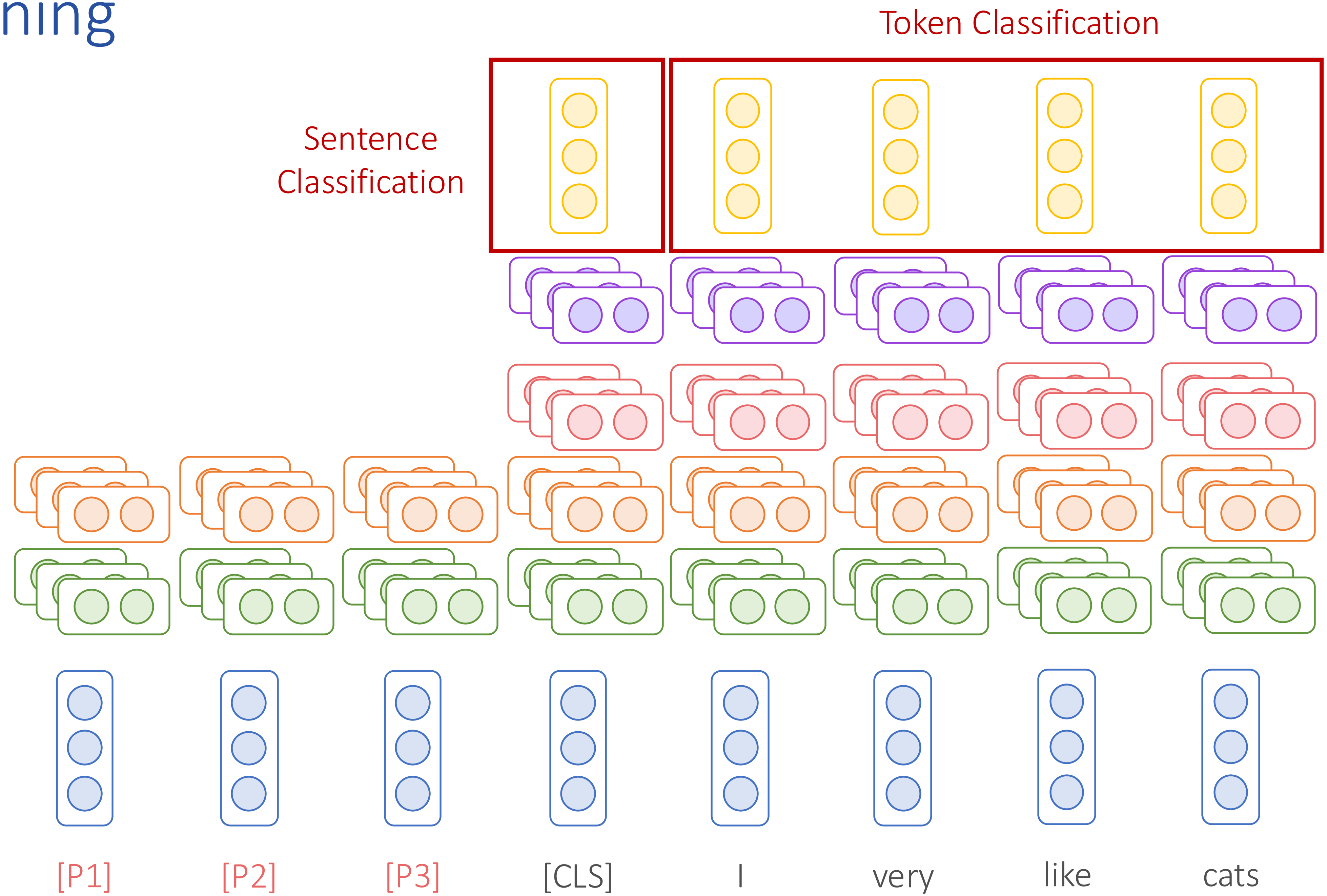


$q_i = W_j^Q x_i$

$k_i = W_j^K x_i$

$v_i = W_j^V x_i$

Prefix Tuning

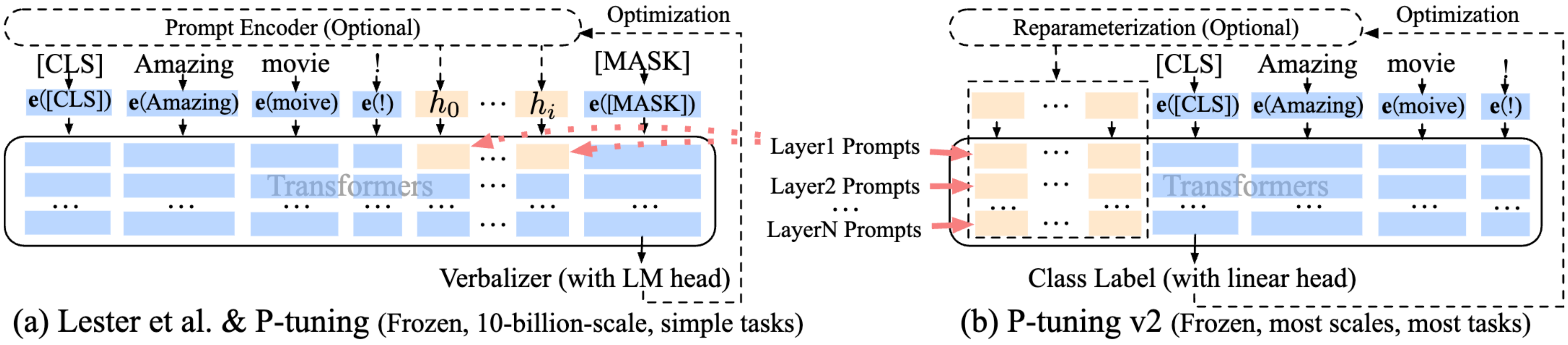


$q_i = W_j^Q x_i$

$k_i = W_j^K x_i$

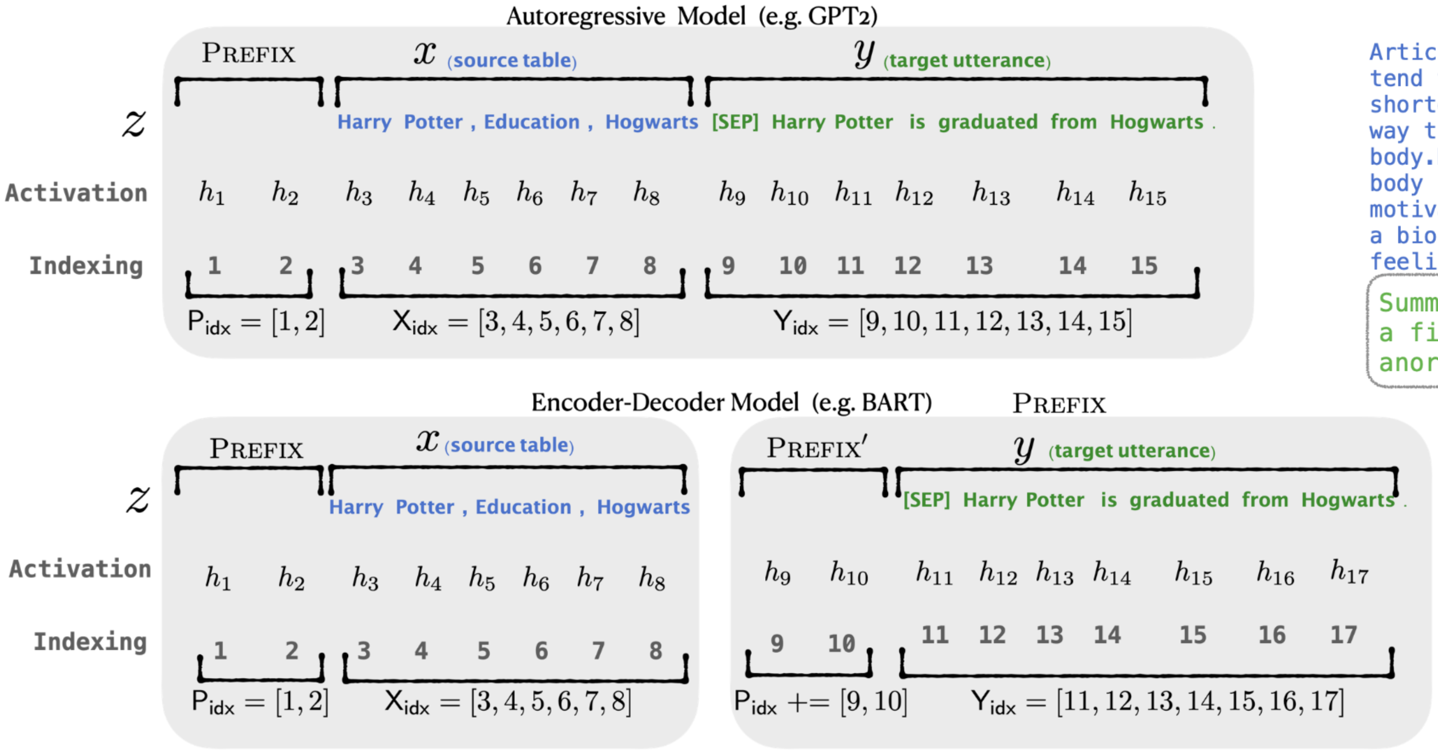
$v_i = W_j^V x_i$

Prefix Tuning


Prompt Encoder (Optional)
Optimization
[CLS]
Amazing
movie
!
[MASK]
e([CLS])
e(Amazing)
e(moive)
e(!)
h_0
···
h_i
e([MASK])
Transformers
Verbalizer (with LM head)
(a) Lester et al. & P-tuning (Frozen, 10-billion-scale, simple tasks)
Layer1 Prompts
Layer2 Prompts
···
LayerN Prompts
Reparameterization (Optional)
Optimization
[CLS]
Amazing
movie
!
e([CLS])
e(Amazing)
e(moive)
e(!)
Transformers
Class Label (with linear head)
(b) P-tuning v2 (Frozen, most scales, most tasks)

P-Tuning v2: Prompt Tuning Can Be Comparable to Fine-tuning Universally Across Scales and Tasks, 2021

Prefix Tuning for Generation


Autoregressive Model (e.g. GPT2)
PREFIX
x (source table)
y (target utterance)
z
Harry Potter , Education , Hogwarts [SEP] Harry Potter is graduated from Hogwarts .
Activation
h_1 h_2 h_3 h_4 h_5 h_6 h_7 h_8 h_9 h_{10} h_{11} h_{12} h_{13} h_{14} h_{15}
Indexing
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
$\mathrm{P_{idx}} = [1, 2]$
$\mathrm{X_{idx}} = [3, 4, 5, 6, 7, 8]$
$\mathrm{Y_{idx}} = [9, 10, 11, 12, 13, 14, 15]$
Encoder-Decoder Model (e.g. BART)
PREFIX
PREFIX
x (source table)
PREFIX′
y (target utterance)
z
Harry Potter , Education , Hogwarts
[SEP] Harry Potter is graduated from Hogwarts .
Activation
h_1 h_2 h_3 h_4 h_5 h_6 h_7 h_8
h_9 h_{10} h_{11} h_{12} h_{13} h_{14} h_{15} h_{16} h_{17}
Indexing
1 2 3 4 5 6 7 8
9 10 11 12 13 14 15 16 17
$\mathrm{P_{idx}} = [1, 2]$
$\mathrm{X_{idx}} = [3, 4, 5, 6, 7, 8]$
$\mathrm{P_{idx}} \mathrel{+}= [9, 10]$
$\mathrm{Y_{idx}} = [11, 12, 13, 14, 15, 16, 17]$

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.

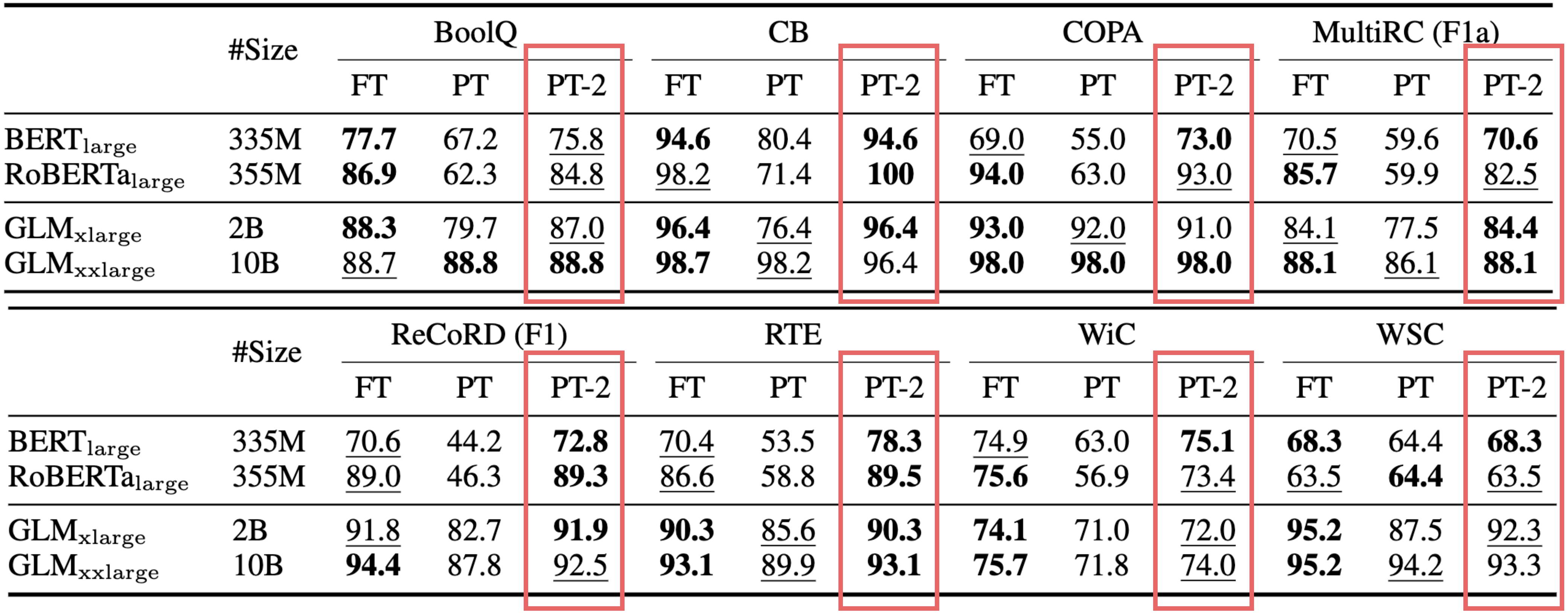
Table-to-text Example

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: Optimizing Continuous Prompts for Generation, 2021

Prefix Tuning

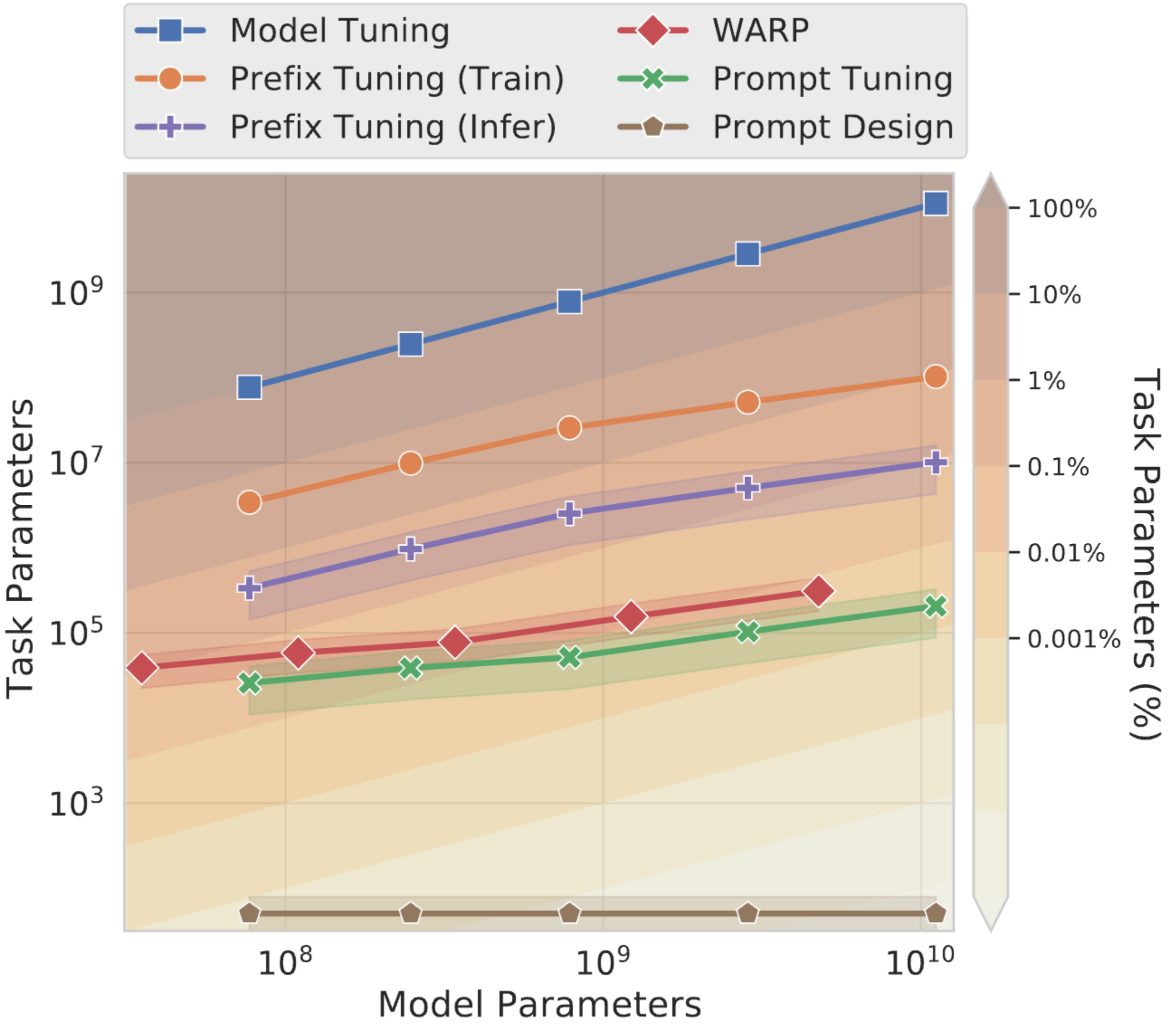


	#Size	BoolQ			CB			COPA			MultiRC (F1a)		
		FT	PT	PT-2	FT	PT	PT-2	FT	PT	PT-2	FT	PT	PT-2
BERT$_{\text{large}}$	335M	**77.7**	67.2	75.8	**94.6**	80.4	**94.6**	69.0	55.0	**73.0**	70.5	59.6	**70.6**
RoBERTa$_{\text{large}}$	355M	**86.9**	62.3	84.8	98.2	71.4	**100**	**94.0**	63.0	93.0	**85.7**	59.9	82.5
GLM$_{\text{xlarge}}$	2B	**88.3**	79.7	87.0	**96.4**	76.4	**96.4**	**93.0**	92.0	91.0	84.1	77.5	**84.4**
GLM$_{\text{xxlarge}}$	10B	88.7	**88.8**	**88.8**	**98.7**	98.2	96.4	**98.0**	**98.0**	**98.0**	**88.1**	86.1	**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$_{\text{large}}$	335M	70.6	44.2	**72.8**	70.4	53.5	**78.3**	74.9	63.0	**75.1**	**68.3**	64.4	**68.3**
RoBERTa$_{\text{large}}$	355M	89.0	46.3	**89.3**	86.6	58.8	**89.5**	**75.6**	56.9	73.4	63.5	**64.4**	63.5
GLM$_{\text{xlarge}}$	2B	91.8	82.7	**91.9**	**90.3**	85.6	**90.3**	**74.1**	71.0	72.0	**95.2**	87.5	92.3
GLM$_{\text{xxlarge}}$	10B	**94.4**	87.8	92.5	**93.1**	89.9	**93.1**	**75.7**	71.8	74.0	**95.2**	94.2	93.3

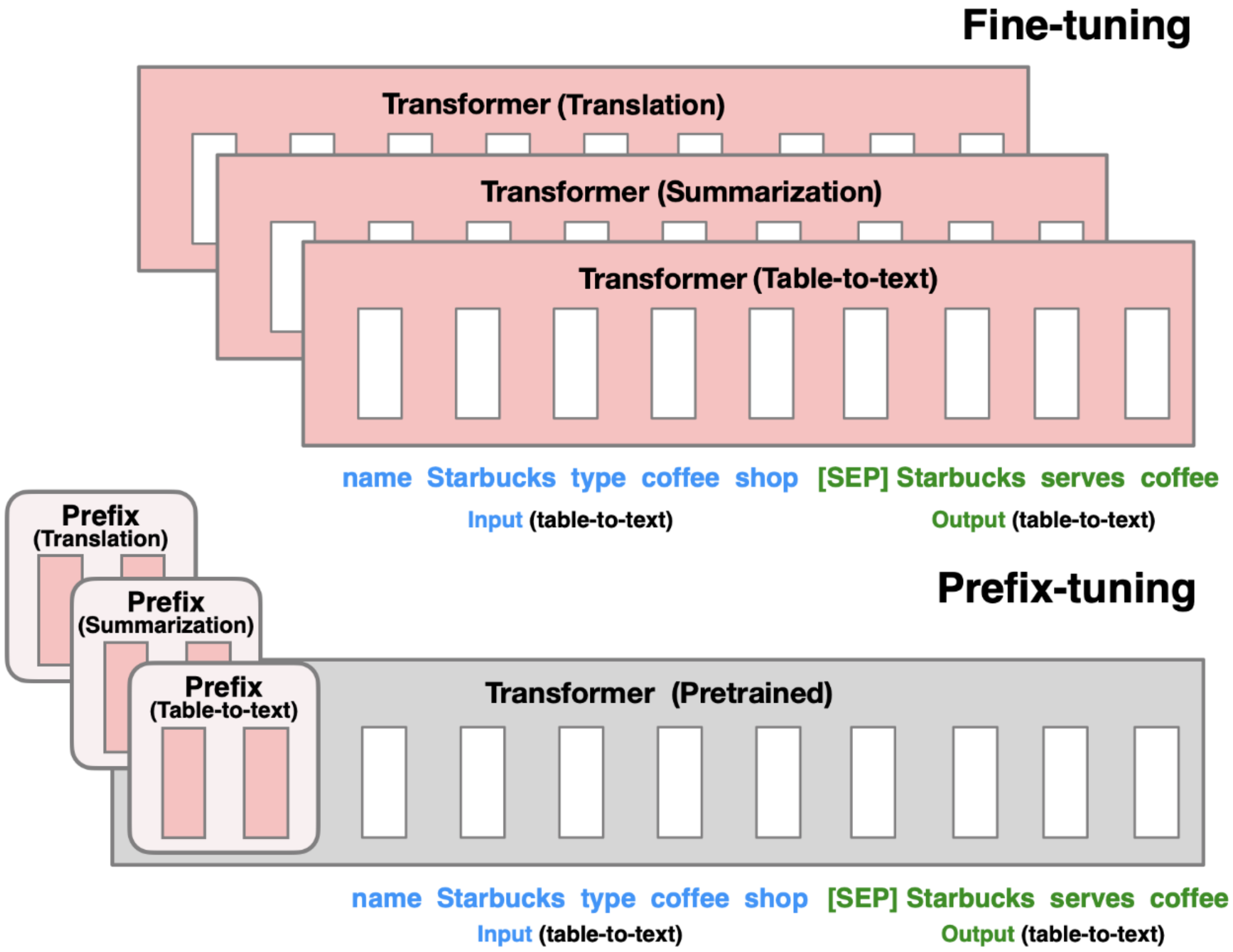
P-Tuning v2: Prompt Tuning Can Be Comparable to Fine-tuning Universally Across Scales and Tasks, 2021

Prefix Tuning – Parameter-Efficient



The Power of Scale for Parameter-Efficient Prompt Tuning, 2021

Prefix Tuning – Parameter-Efficient

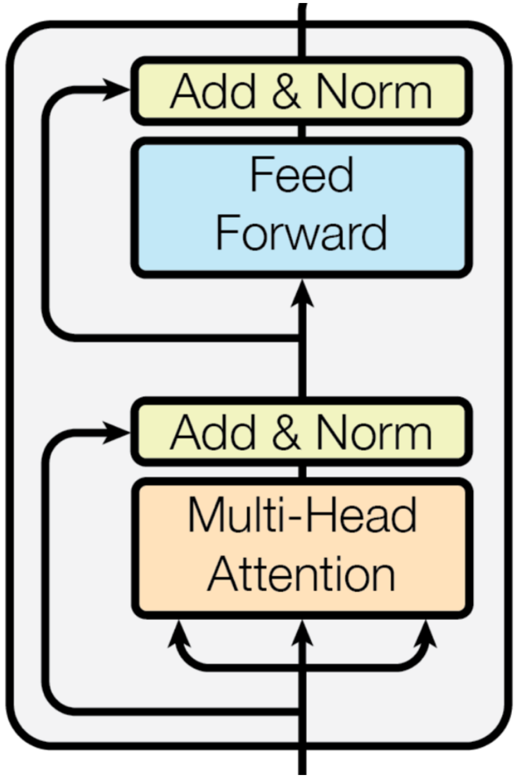

Fine-tuning
Transformer (Translation)
Transformer (Summarization)
Transformer (Table-to-text)
name Starbucks type coffee shop [SEP] Starbucks serves coffee
Input (table-to-text)
Output (table-to-text)
Prefix-tuning
Prefix (Translation)
Prefix (Summarization)
Prefix (Table-to-text)
Transformer (Pretrained)
name Starbucks type coffee shop [SEP] Starbucks serves coffee
Input (table-to-text)
Output (table-to-text)

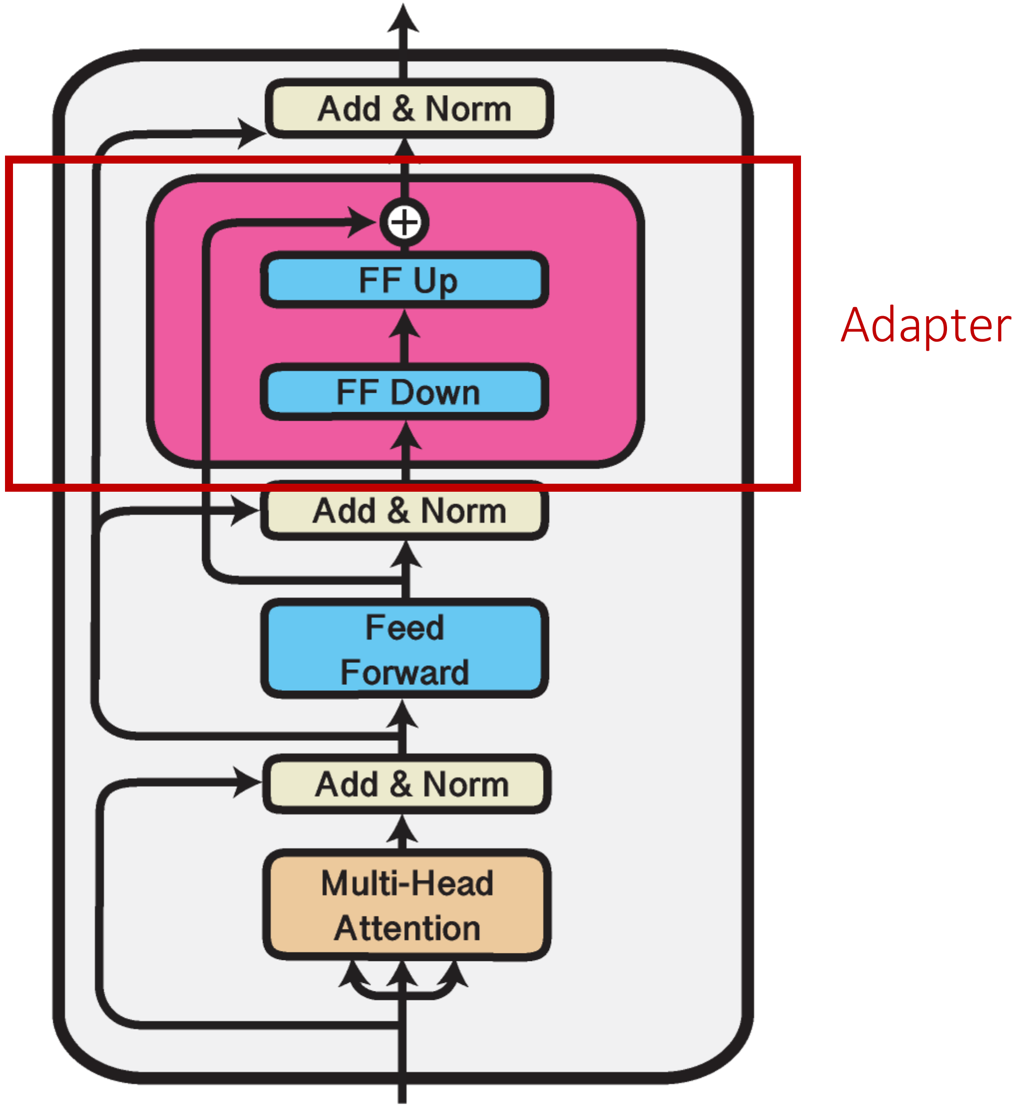
Prefix-Tuning: Optimizing Continuous Prompts for Generation, 2021

Parameter-Efficient Fine-Tuning

- Do not fine-tune the whole model
  - Most parameters are frozen
  - Fine-tune a small set of parameters
- Save GPU memory during training
- Save space for storing multiple models

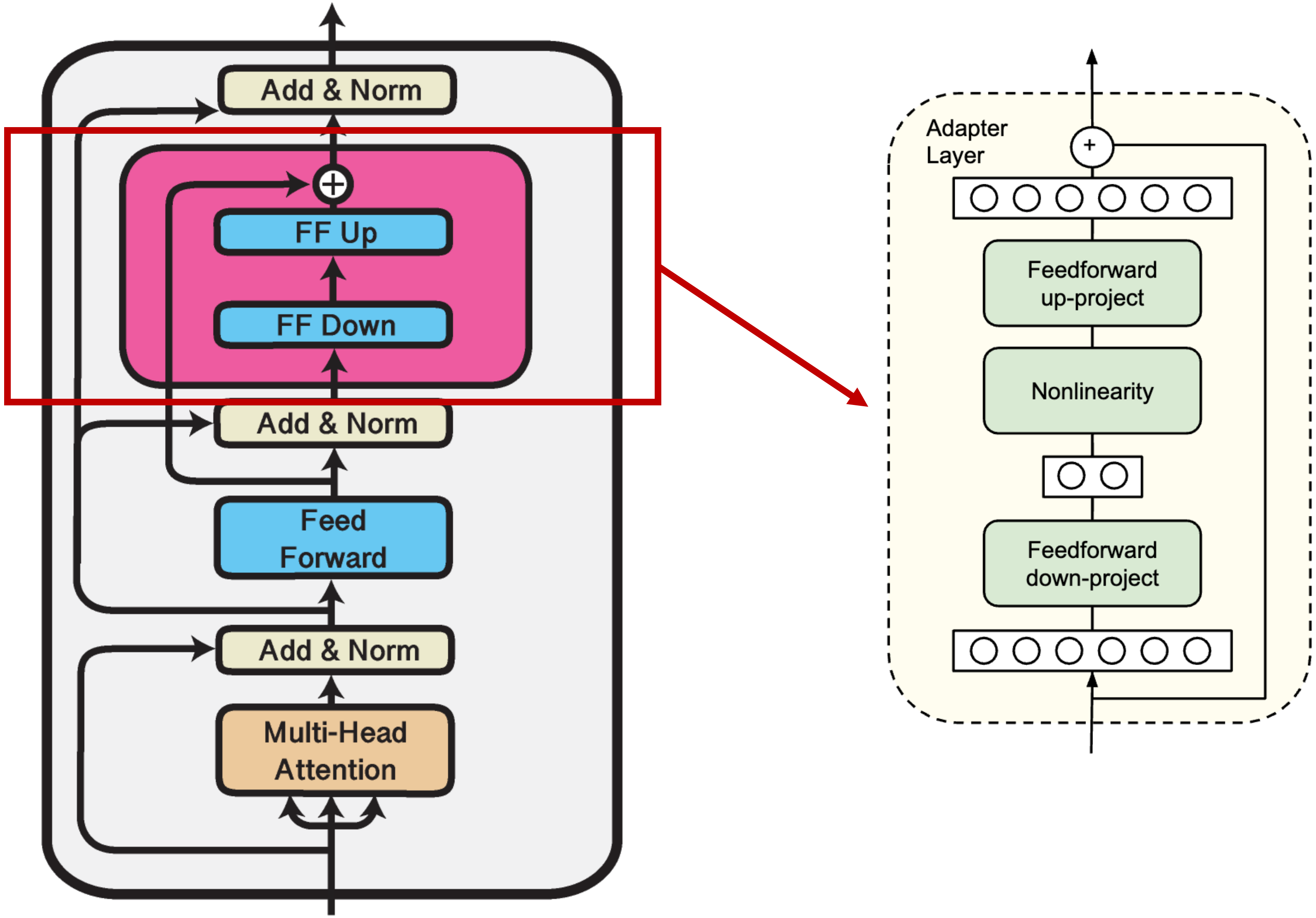
Adapter





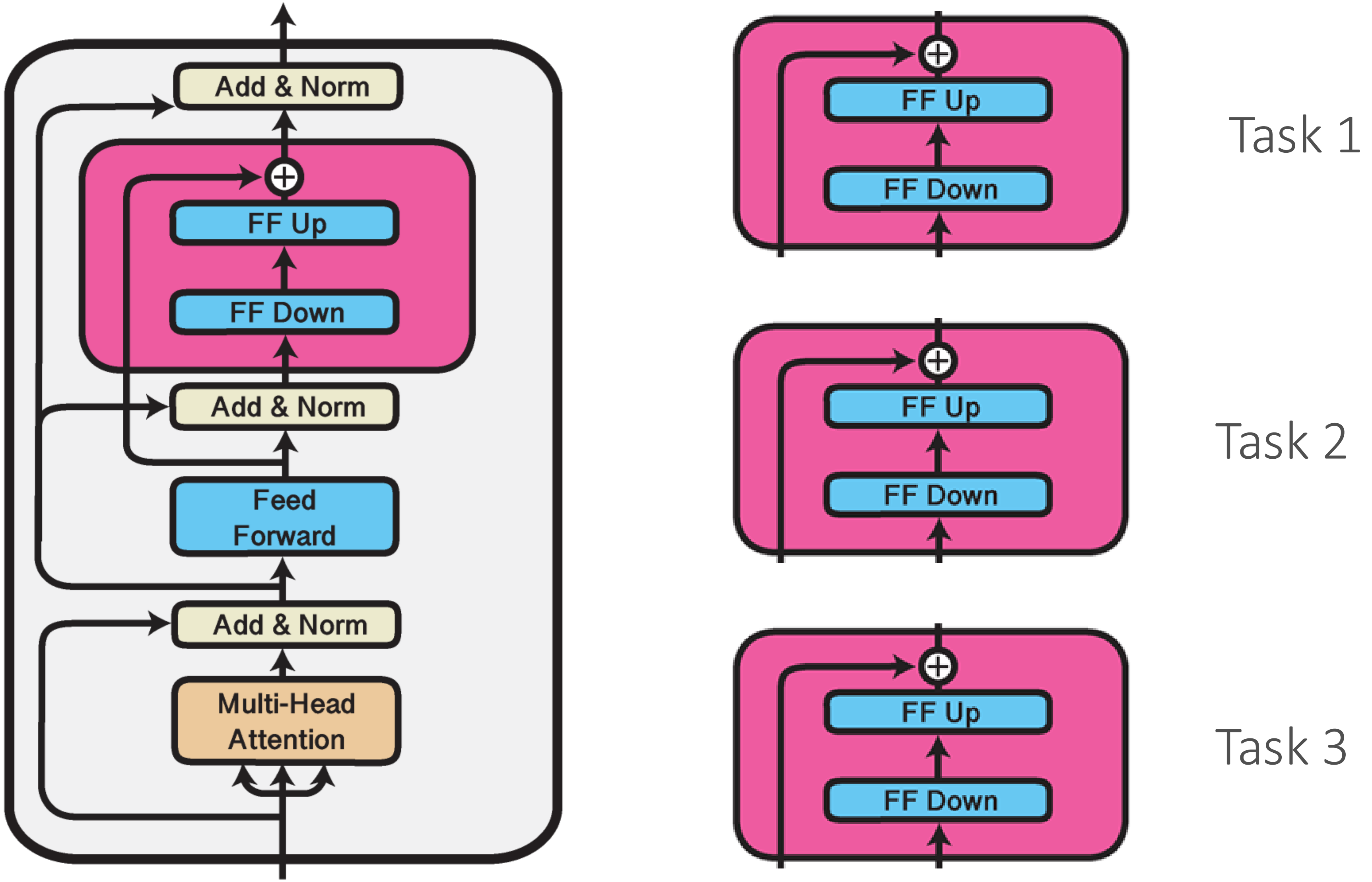
https://adapterhub.ml/blog/2022/03/adapter-transformers-v3-unifying-efficient-fine-tuning/

Adapter



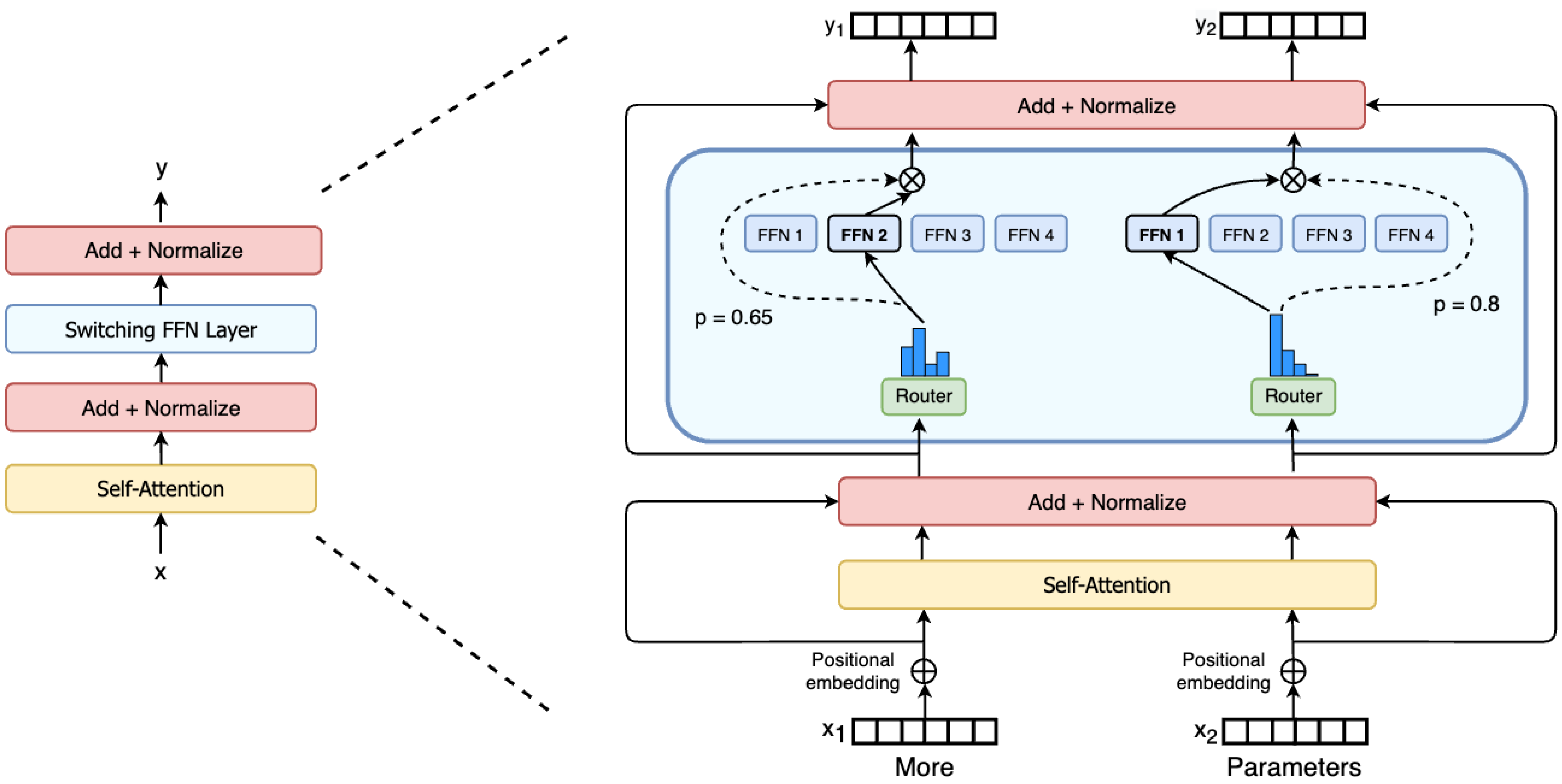
Parameter-Efficient Transfer Learning for NLP, 2019

Adapter



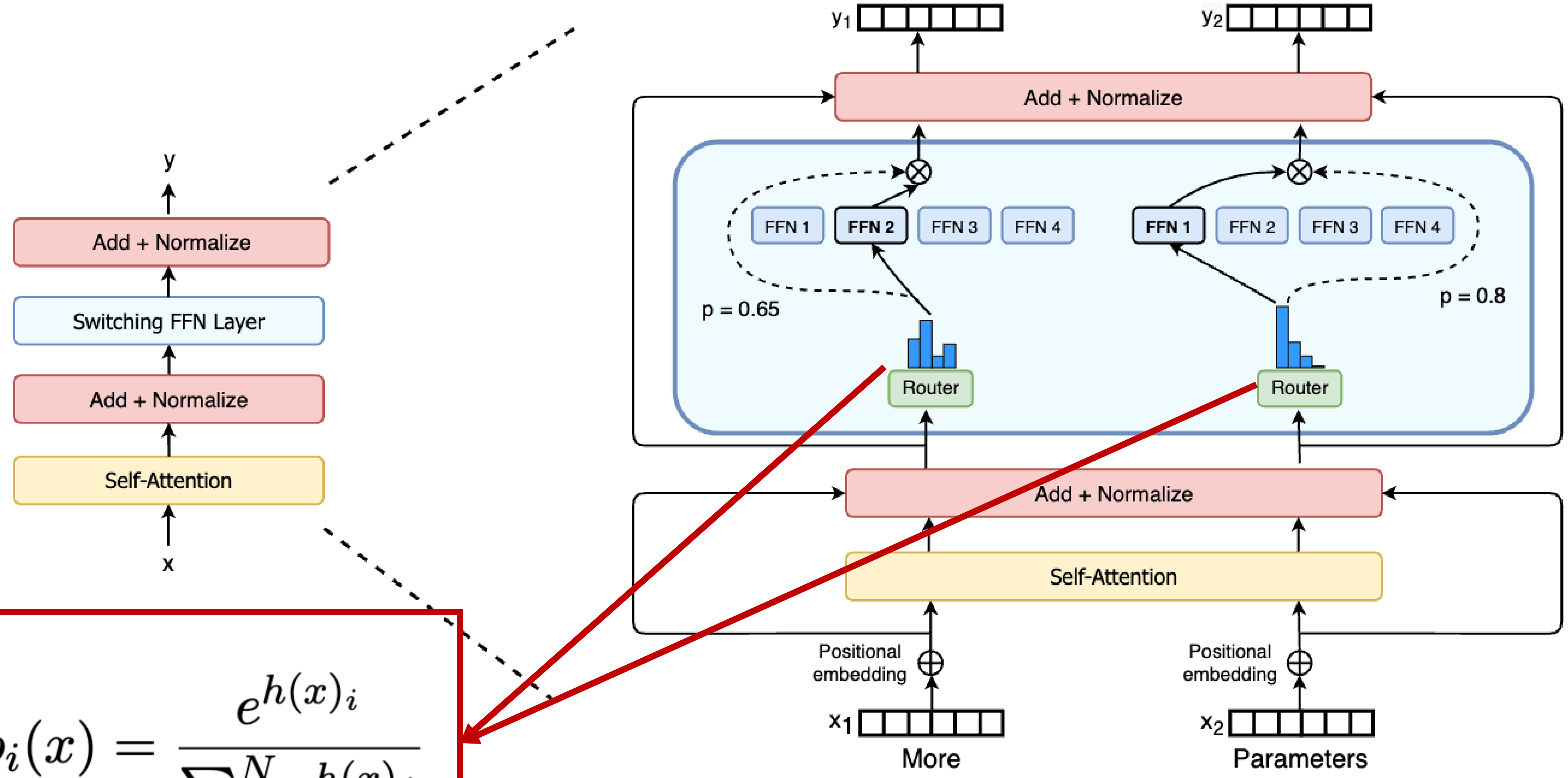
Parameter-Efficient Transfer Learning for NLP, 2019

Mixture of Experts (MoE)



Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity, 2022

Mixture of Experts (MoE)

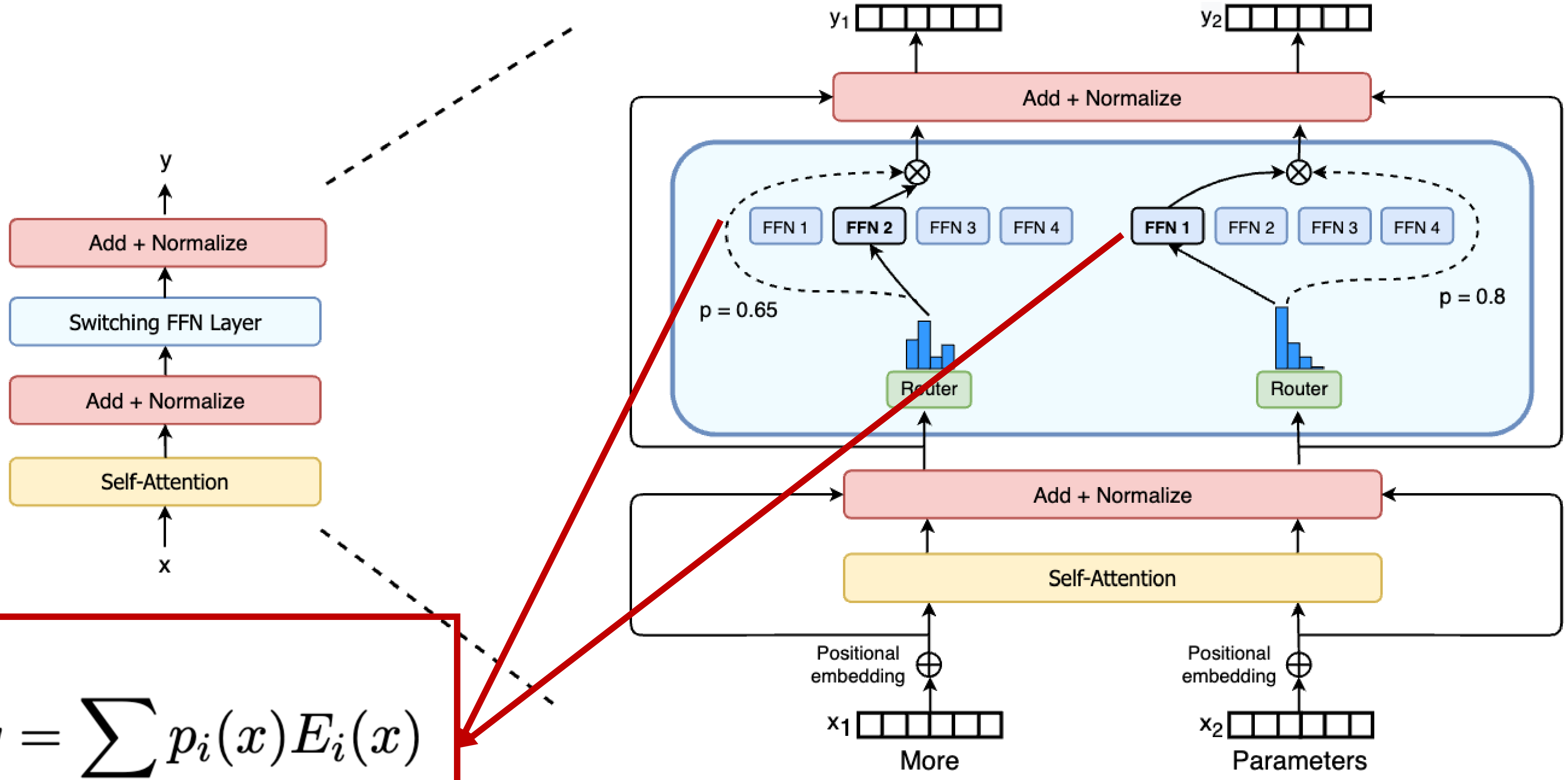


$$p_i(x) = \frac{e^{h(x)_i}}{\sum_j^N e^{h(x)_j}}$$

Gate routing

Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity, 2022

Mixture of Experts (MoE)

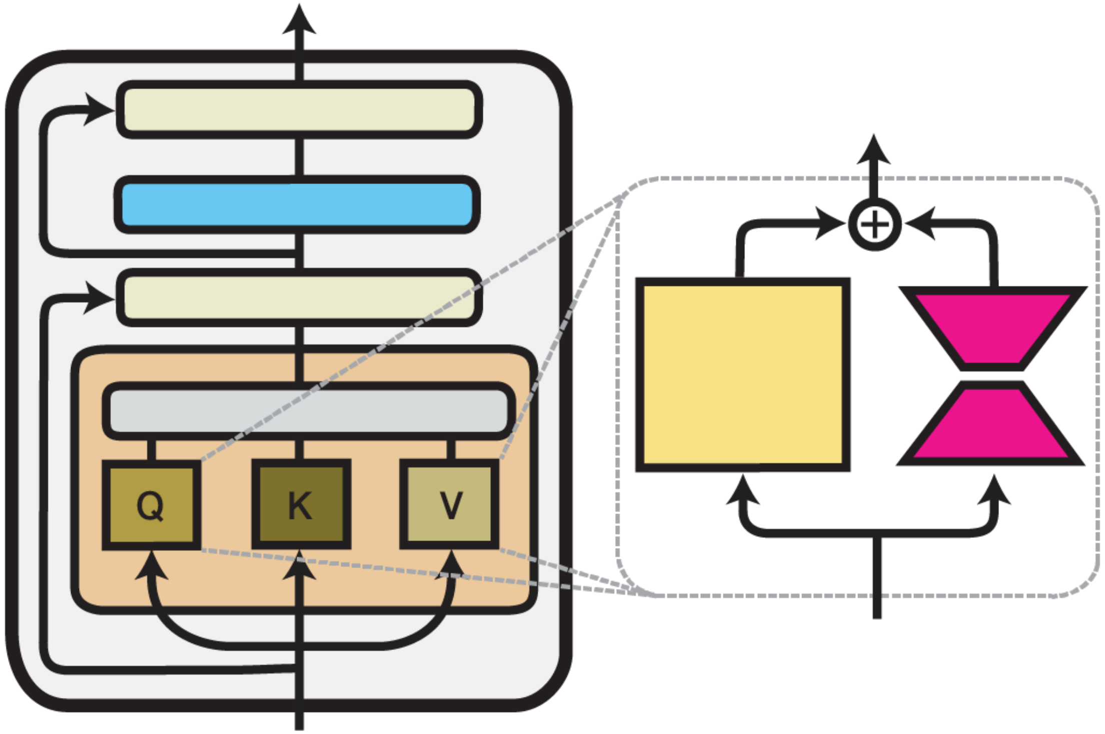


$$y = \sum_{i \in \mathcal{T}} p_i(x) E_i(x)$$

Weighted output with
top k gate values

Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity, 2022

LoRA: Low-Rank Adaptation



https://docs.adapterhub.ml/methods.html

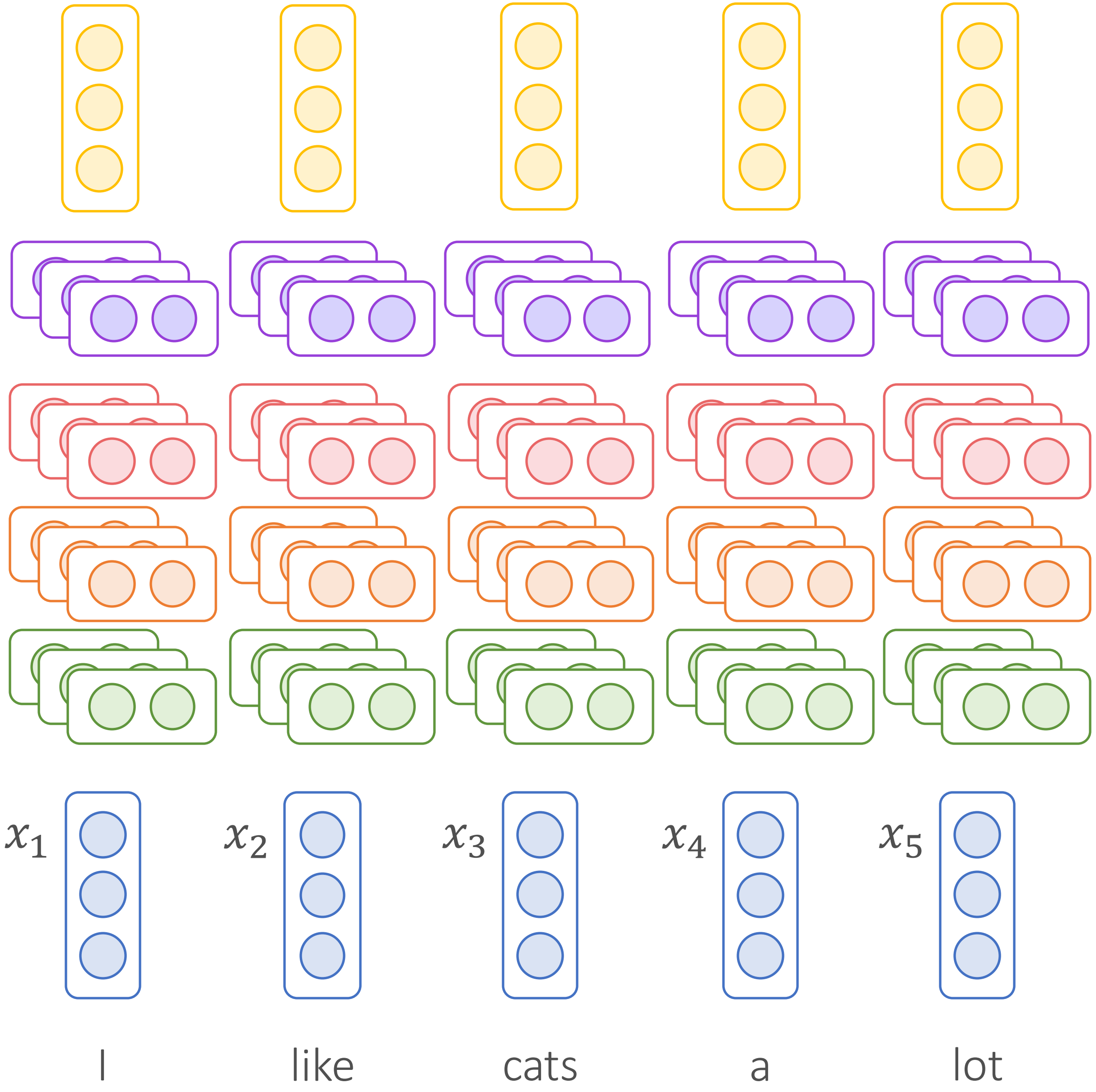
LoRA: Low-Rank Adaptation

Main Parameters

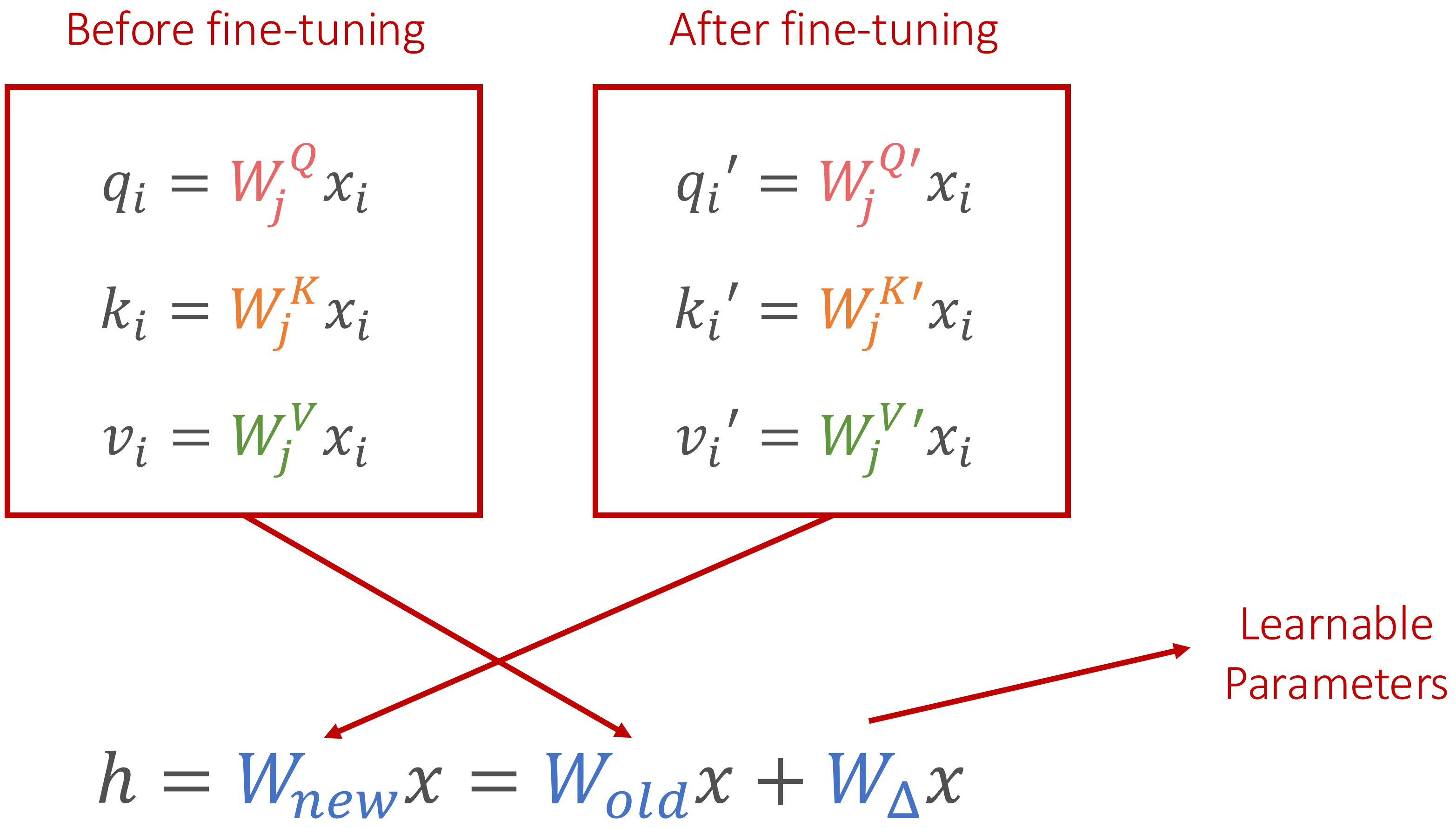
$$q_i = W_j^Q x_i$$

$$k_i = W_j^K x_i$$

$$v_i = W_j^V x_i$$



LoRA: Low-Rank Adaptation



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

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

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

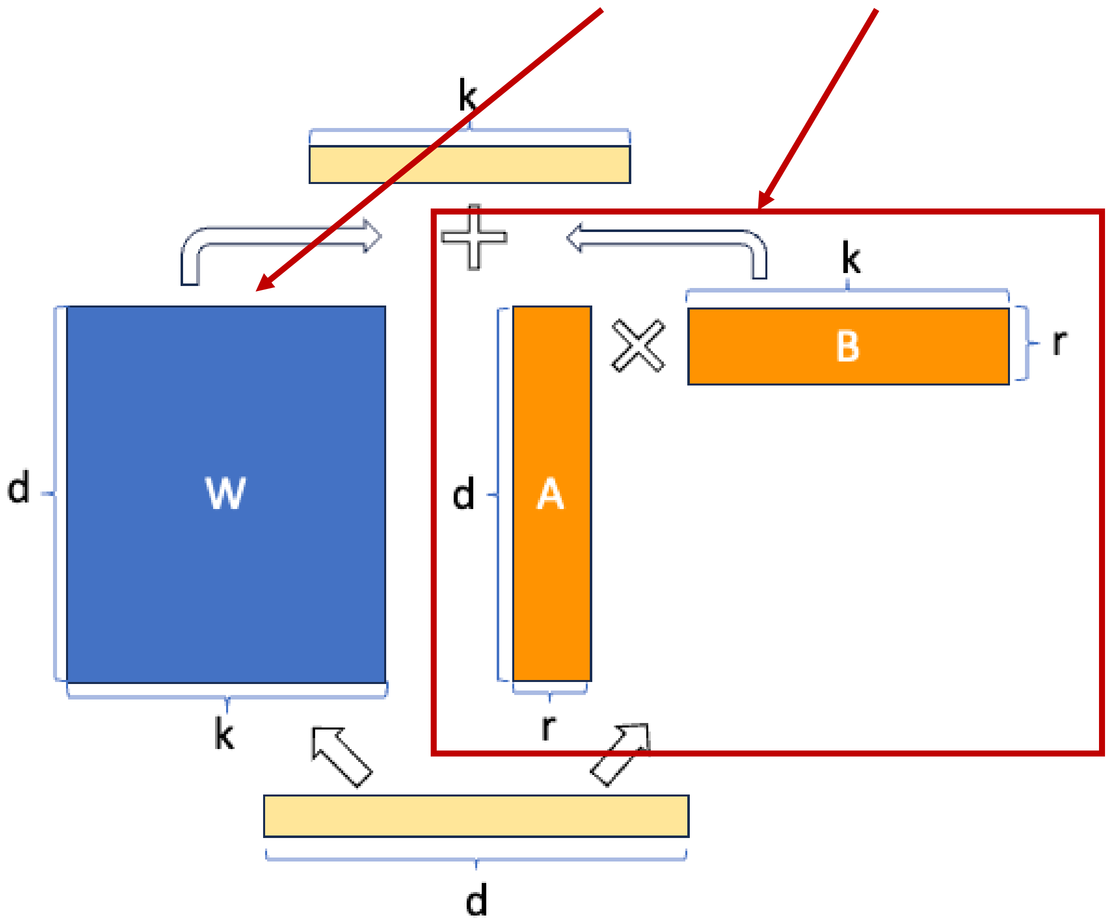
$$v_i' = W_j^{V'} x_i$$

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

Learnable Parameters

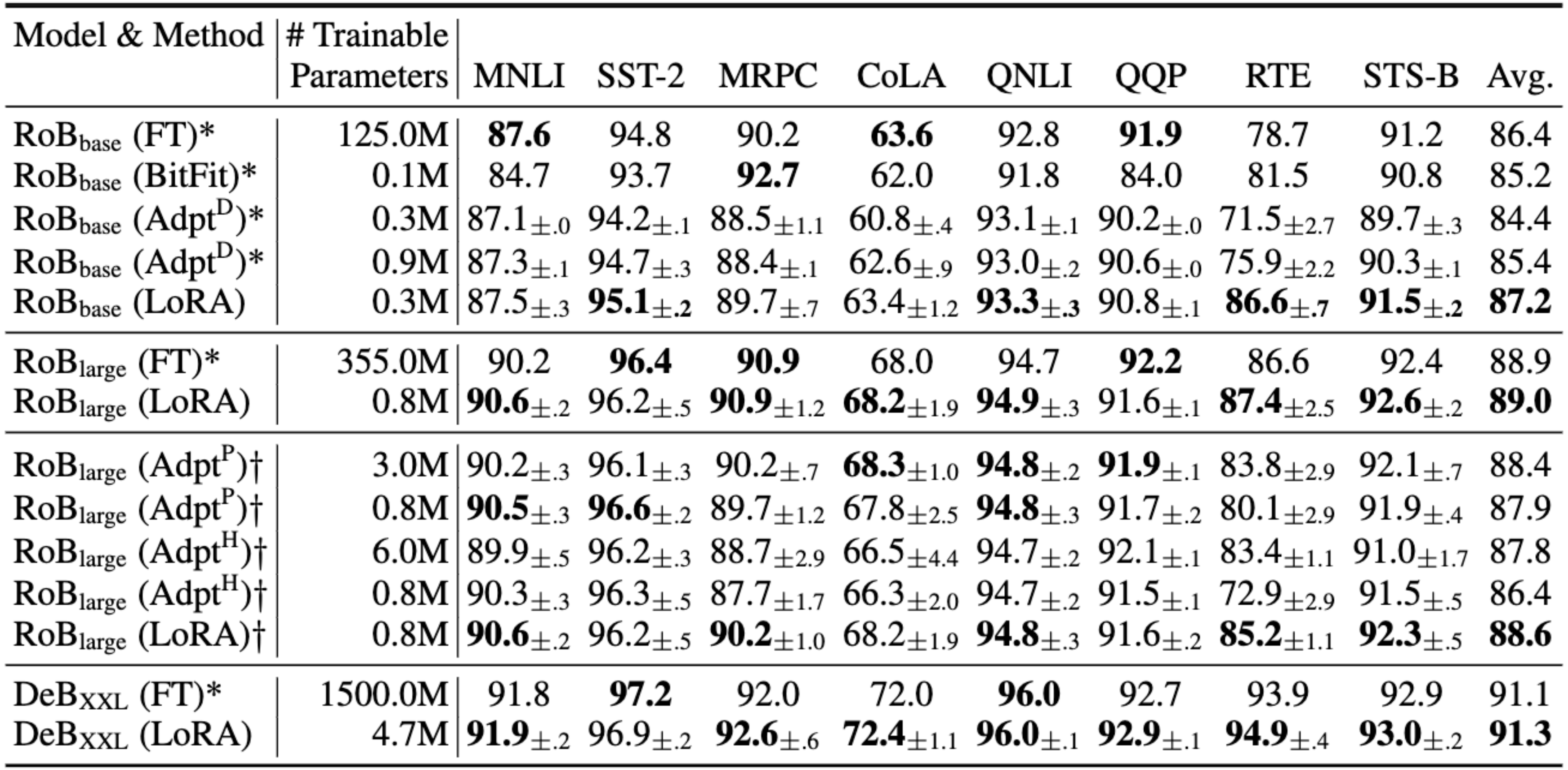
LoRA: Low-Rank Adaptation

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



https://dataman-ai.medium.com/fine-tune-a-gpt-lora-e9b72ad4ad3

LoRA: Low-Rank Adaptation



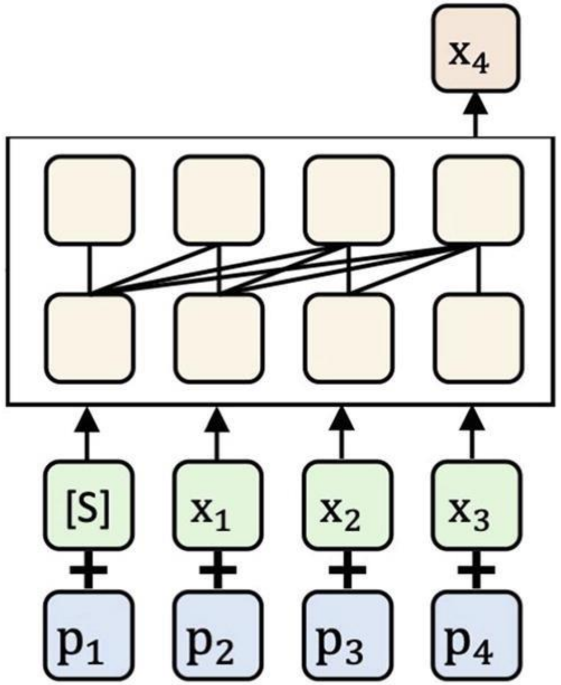
Model & Method	# Trainable Parameters	MNLI	SST-2	MRPC	CoLA	QNLI	QQP	RTE	STS-B	Avg.
RoB$_{\text{base}}$ (FT)*	125.0M	**87.6**	94.8	90.2	**63.6**	92.8	**91.9**	78.7	91.2	86.4
RoB$_{\text{base}}$ (BitFit)*	0.1M	84.7	93.7	**92.7**	62.0	91.8	84.0	81.5	90.8	85.2
RoB$_{\text{base}}$ (Adpt^{D})*	0.3M	87.1$_{\pm.0}$	94.2$_{\pm.1}$	88.5$_{\pm1.1}$	60.8$_{\pm.4}$	93.1$_{\pm.1}$	90.2$_{\pm.0}$	71.5$_{\pm2.7}$	89.7$_{\pm.3}$	84.4
RoB$_{\text{base}}$ (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$_{\pm2.2}$	90.3$_{\pm.1}$	85.4
RoB$_{\text{base}}$ (LoRA)	0.3M	87.5$_{\pm.3}$	**95.1**$_{\pm.2}$	89.7$_{\pm.7}$	63.4$_{\pm1.2}$	**93.3**$_{\pm.3}$	90.8$_{\pm.1}$	**86.6**$_{\pm.7}$	**91.5**$_{\pm.2}$	**87.2**
RoB$_{\text{large}}$ (FT)*	355.0M	90.2	**96.4**	**90.9**	68.0	94.7	**92.2**	86.6	92.4	88.9
RoB$_{\text{large}}$ (LoRA)	0.8M	**90.6**$_{\pm.2}$	96.2$_{\pm.5}$	**90.9**$_{\pm1.2}$	**68.2**$_{\pm1.9}$	**94.9**$_{\pm.3}$	91.6$_{\pm.1}$	**87.4**$_{\pm2.5}$	**92.6**$_{\pm.2}$	**89.0**
RoB$_{\text{large}}$ (Adpt^{P})†	3.0M	90.2$_{\pm.3}$	96.1$_{\pm.3}$	90.2$_{\pm.7}$	**68.3**$_{\pm1.0}$	**94.8**$_{\pm.2}$	**91.9**$_{\pm.1}$	83.8$_{\pm2.9}$	92.1$_{\pm.7}$	88.4
RoB$_{\text{large}}$ (Adpt^{P})†	0.8M	**90.5**$_{\pm.3}$	**96.6**$_{\pm.2}$	89.7$_{\pm1.2}$	67.8$_{\pm2.5}$	**94.8**$_{\pm.3}$	91.7$_{\pm.2}$	80.1$_{\pm2.9}$	91.9$_{\pm.4}$	87.9
RoB$_{\text{large}}$ (Adpt^{H})†	6.0M	89.9$_{\pm.5}$	96.2$_{\pm.3}$	88.7$_{\pm2.9}$	66.5$_{\pm4.4}$	94.7$_{\pm.2}$	92.1$_{\pm.1}$	83.4$_{\pm1.1}$	91.0$_{\pm1.7}$	87.8
RoB$_{\text{large}}$ (Adpt^{H})†	0.8M	90.3$_{\pm.3}$	96.3$_{\pm.5}$	87.7$_{\pm1.7}$	66.3$_{\pm2.0}$	94.7$_{\pm.2}$	91.5$_{\pm.1}$	72.9$_{\pm2.9}$	91.5$_{\pm.5}$	86.4
RoB$_{\text{large}}$ (LoRA)†	0.8M	**90.6**$_{\pm.2}$	96.2$_{\pm.5}$	**90.2**$_{\pm1.0}$	68.2$_{\pm1.9}$	**94.8**$_{\pm.3}$	91.6$_{\pm.2}$	**85.2**$_{\pm1.1}$	**92.3**$_{\pm.5}$	**88.6**
DeB$_{\text{XXL}}$ (FT)*	1500.0M	91.8	**97.2**	92.0	72.0	**96.0**	92.7	93.9	92.9	91.1
DeB$_{\text{XXL}}$ (LoRA)	4.7M	**91.9**$_{\pm.2}$	96.9$_{\pm.2}$	**92.6**$_{\pm.6}$	**72.4**$_{\pm1.1}$	**96.0**$_{\pm.1}$	**92.9**$_{\pm.1}$	**94.9**$_{\pm.4}$	**93.0**$_{\pm.2}$	**91.3**

LoRA: Low-Rank Adaptation of Large Language Models, 2021

Lecture Plan

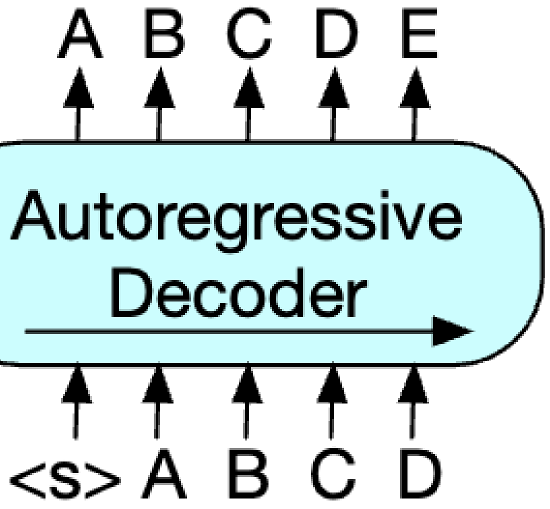
- Parameter-Efficient Fine-Tuning
  - Prompt Tuning
  - Prefix Tuning
  - Adapter
  - Mixture of Experts
  - LoRA
- Large Language Models

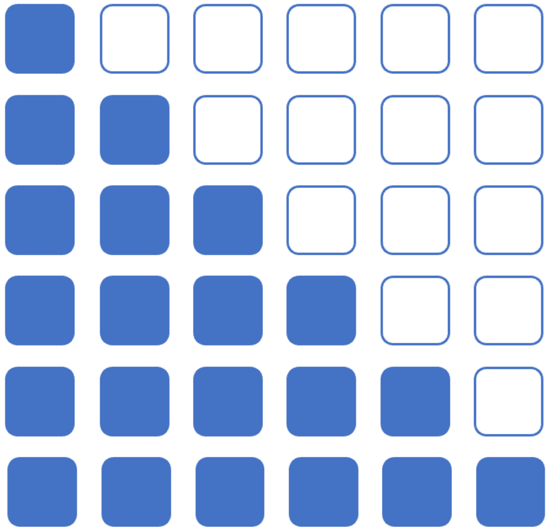
Pre-Trained Language Models with Decoder



$$\sum_{t=1}^{T} P(\mathrm{x}_t|\mathbf{x}_{<\mathrm{t}})$$

Decoder only

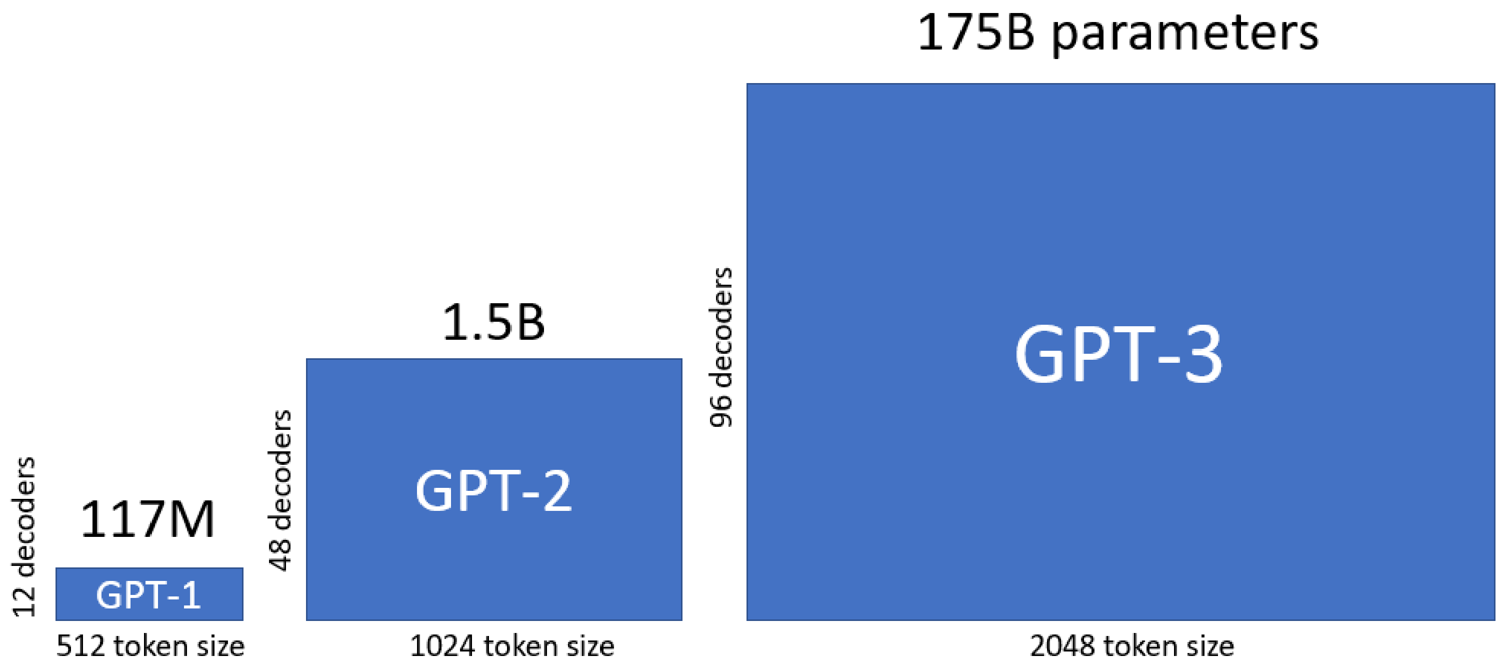




Causal Masking

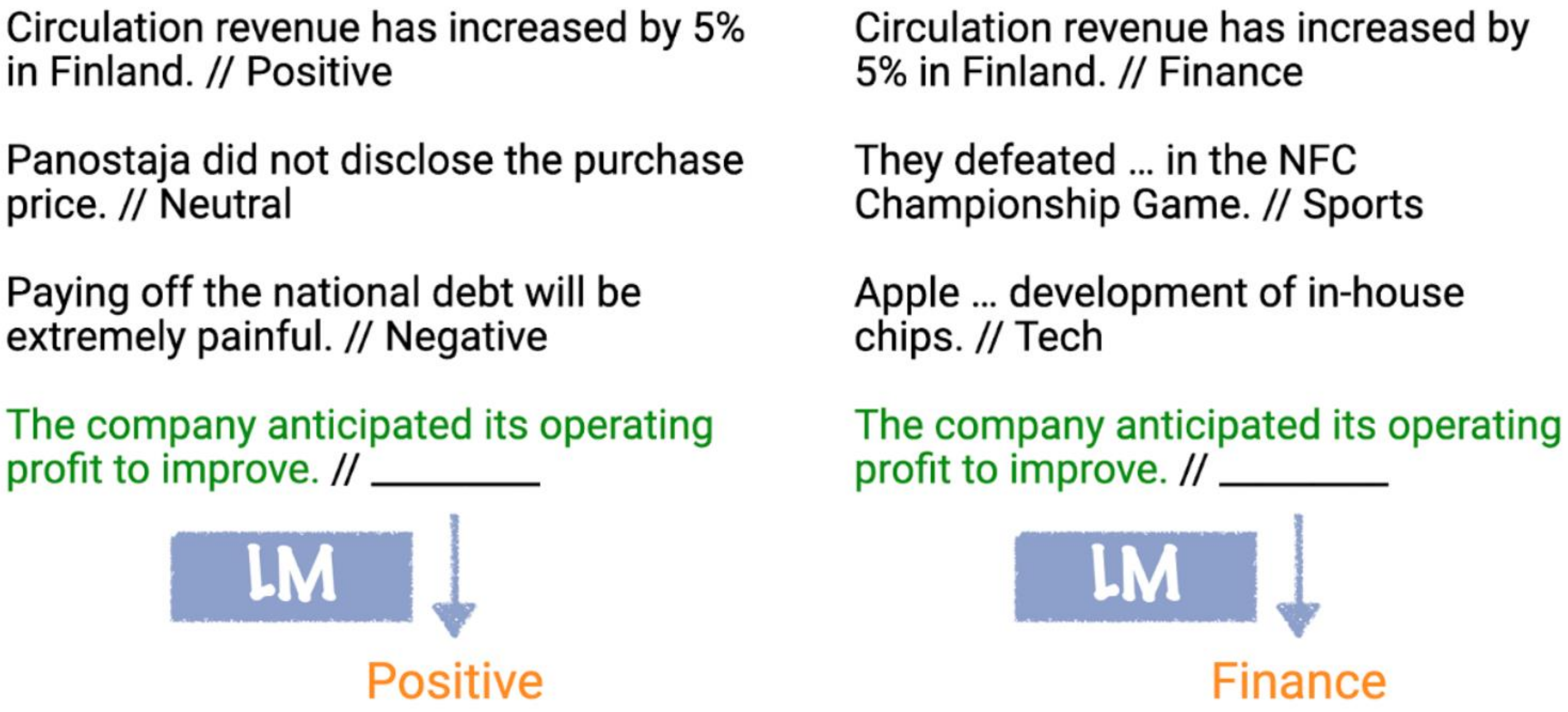
GPT-3: From Fine-Tuning to Few-Shot Learning

- Even larger training data, even larger model size



GPT-3: From Fine-Tuning to Few-Shot Learning

- Solve entirely new tasks by few-shot learning (in-context learning)



Circulation revenue has increased by 5% in Finland. // Positive

Panostaja did not disclose the purchase price. // Neutral

Paying off the national debt will be extremely painful. // Negative

The company anticipated its operating profit to improve. // ________

Positive

Circulation revenue has increased by 5% in Finland. // Finance

They defeated … in the NFC Championship Game. // Sports

Apple … development of in-house chips. // Tech

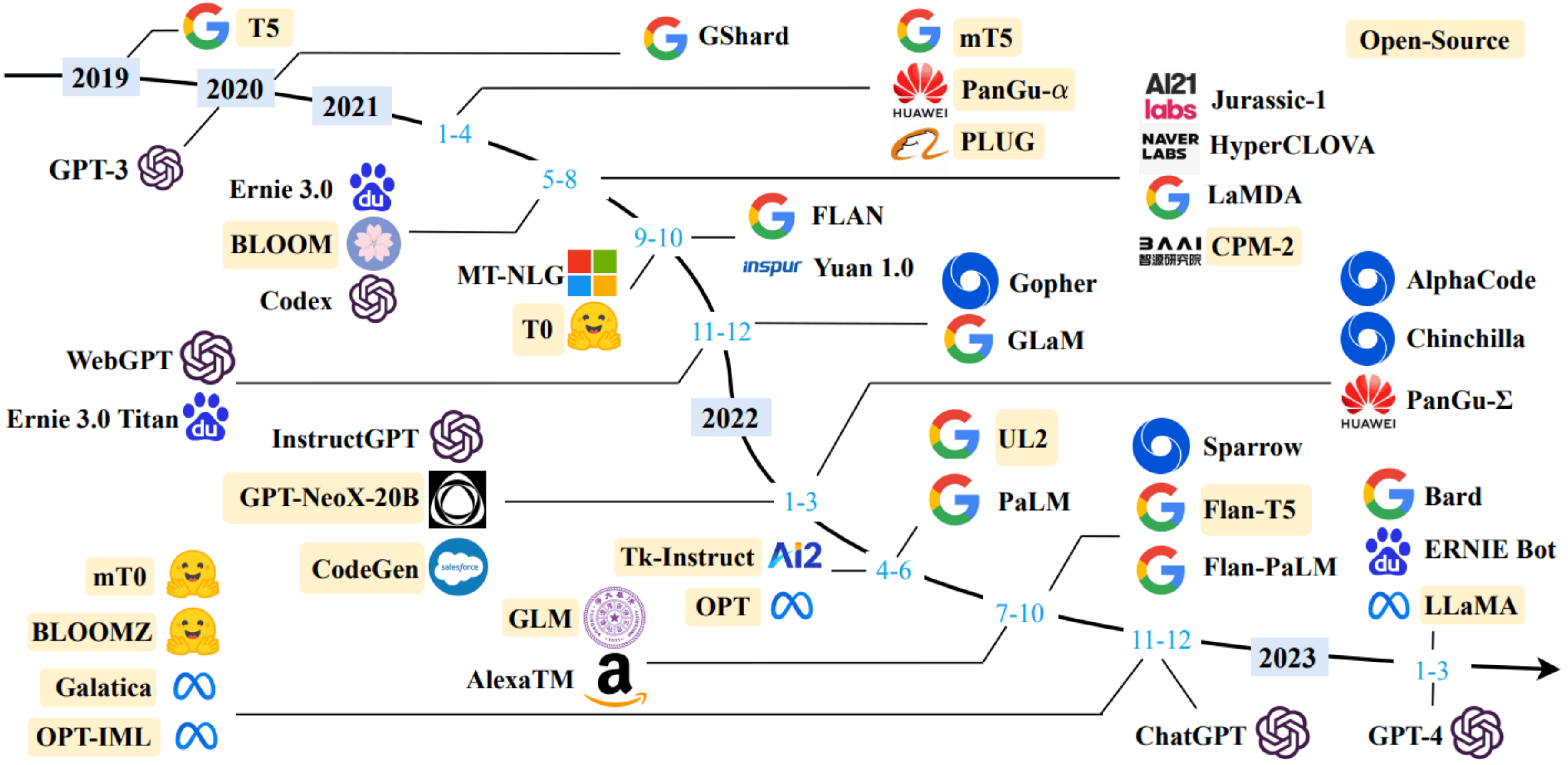
The company anticipated its operating profit to improve. // ________

Finance

https://ai.stanford.edu/blog/understanding-incontext/

Large Language Models (LLMs)

LLMs = (Large Scale) Transformers + Language Models + Pre-Training



A Survey of Large Language Models, 2023

What Makes an LLM?

- Architecture decisions
- Data decisions
- Training decisions

Open Access vs. Closed Access

- Model Weights
  - Open / Described / Closed
- Data
  - Open / Described / Closed
- Training Code
  - Open / Described / Closed

Open Access vs. Closed Access

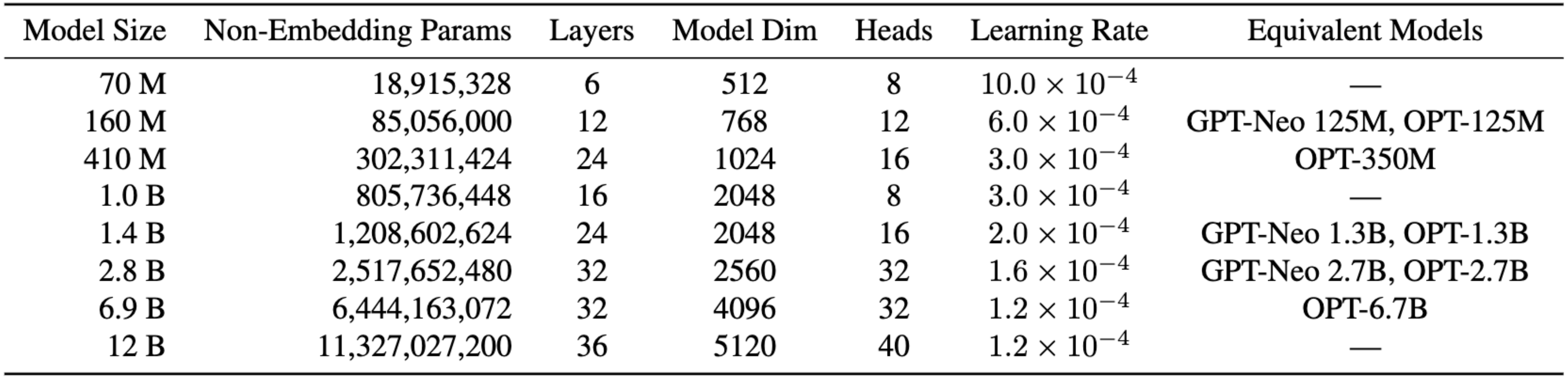
- Open-source LLMs
- Open-weight LLMs
- Closed LLMs

Open-Source / Reproducible LLMs

- Pythia
  - Fully open, many sizes/checkpoints
- OLMo
  - Possibly strongest reproducible model

Pythia

- **Creator:** https://github.com/EleutherAI/pythia
- **Goal:** Joint understanding of model training dynamics and scaling
- **Unique features:** 8 model sizes 70M-12B, 154 checkpoints for each

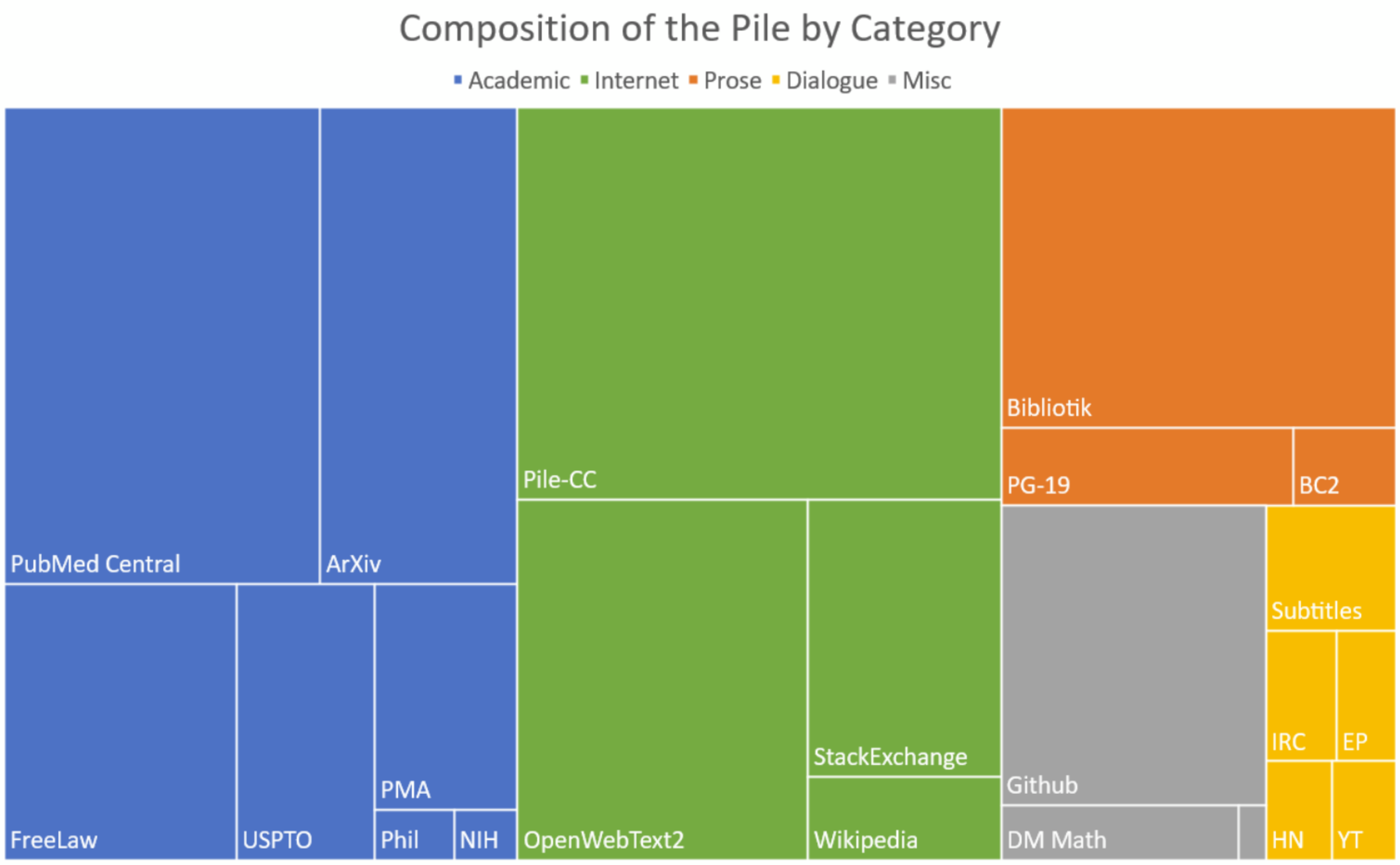


Model Size	Non-Embedding Params	Layers	Model Dim	Heads	Learning Rate	Equivalent Models
70 M	18,915,328	6	512	8	10.0×10^{-4}	—
160 M	85,056,000	12	768	12	6.0×10^{-4}	GPT-Neo 125M, OPT-125M
410 M	302,311,424	24	1024	16	3.0×10^{-4}	OPT-350M
1.0 B	805,736,448	16	2048	8	3.0×10^{-4}	—
1.4 B	1,208,602,624	24	2048	16	2.0×10^{-4}	GPT-Neo 1.3B, OPT-1.3B
2.8 B	2,517,652,480	32	2560	32	1.6×10^{-4}	GPT-Neo 2.7B, OPT-2.7B
6.9 B	6,444,163,072	32	4096	32	1.2×10^{-4}	OPT-6.7B
12 B	11,327,027,200	36	5120	40	1.2×10^{-4}	—

Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling, 2023

Pythia: The Pile

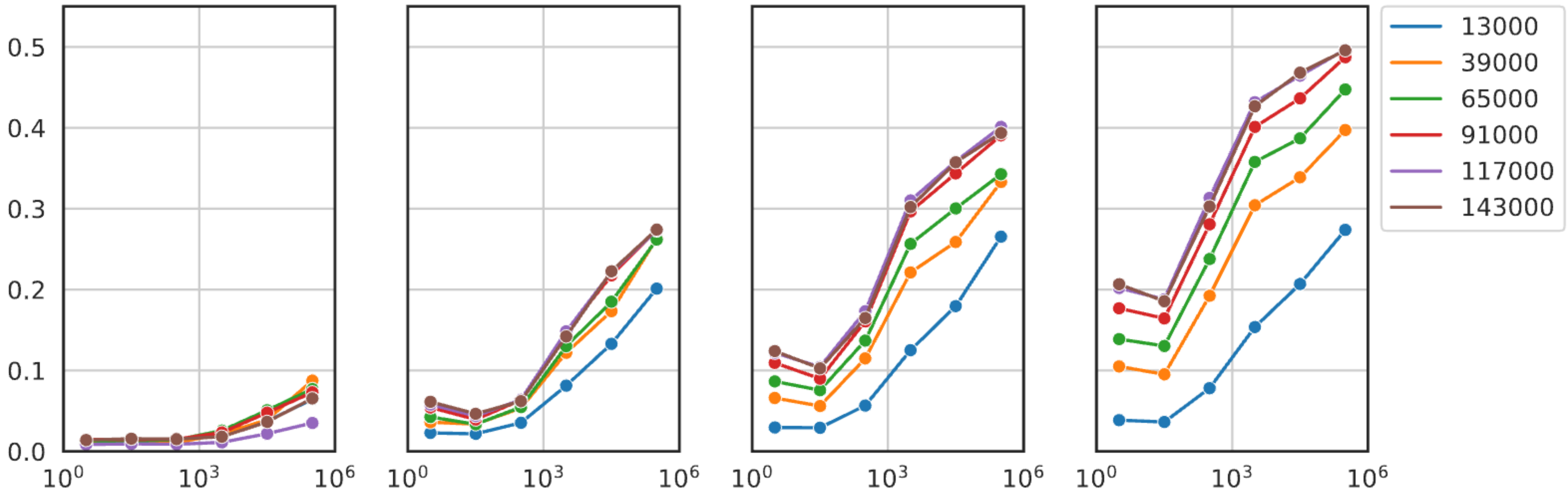
- An 800GB Dataset of Diverse Text for Language Modeling



The Pile: An 800GB Dataset of Diverse Text for Language Modeling, 2020

Pythia: Findings

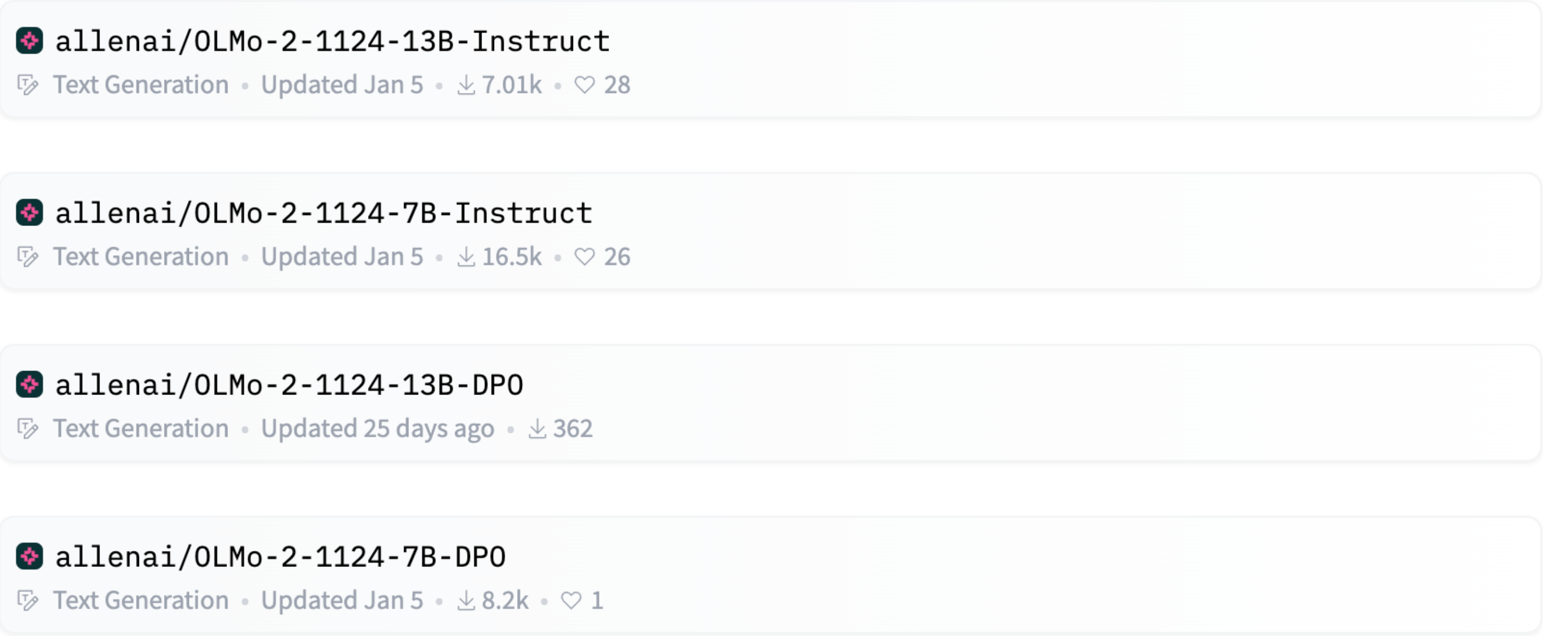
- Some insights into training dynamics, e.g. larger models memorize facts more quickly



(a) 160 M (b) 1.0 B (c) 2.8 B (d) 12 B

OLMo

- **Creator:** Ai2 https://allenai.org/olmo
- **Goal:** Better science of state-of-the-art LMs
- **Unique features:** Top performance of fully documented model, instruction tuned etc.



allenai/OLMo-2-1124-13B-Instruct
Text Generation • Updated Jan 5 • 7.01k • ♡ 28

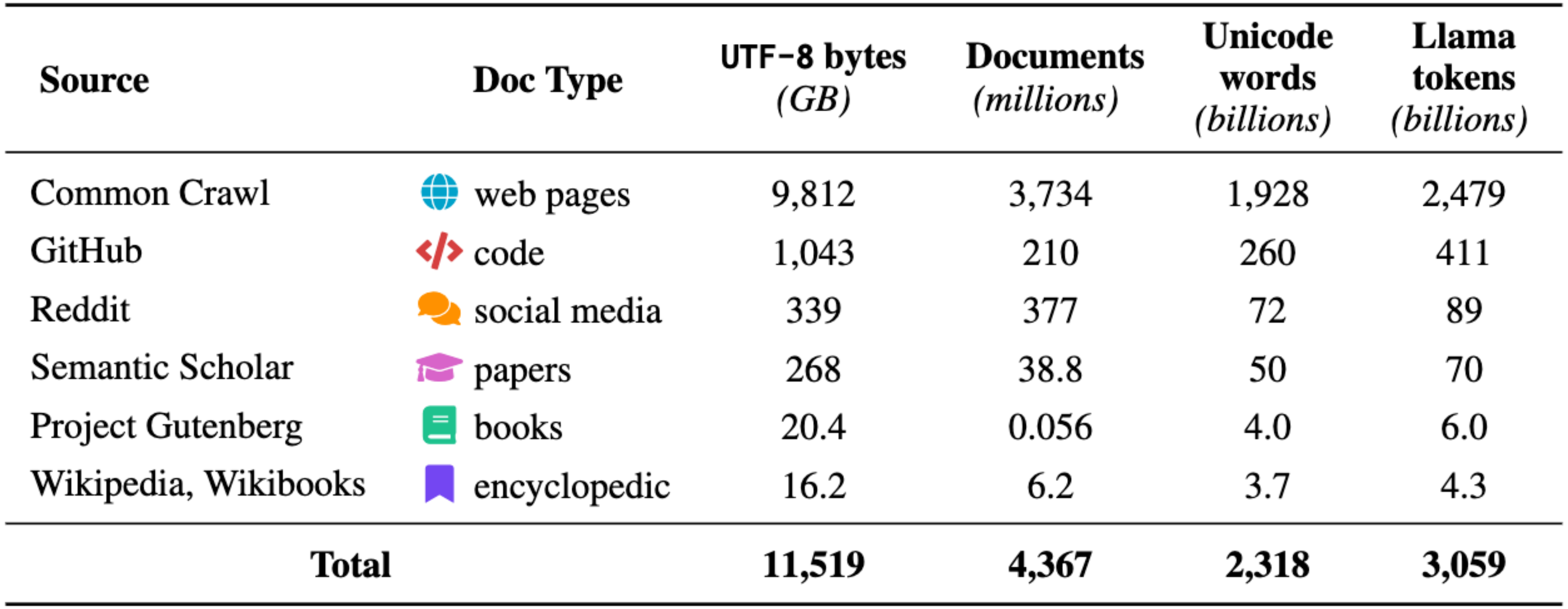
allenai/OLMo-2-1124-7B-Instruct
Text Generation • Updated Jan 5 • 16.5k • ♡ 26

allenai/OLMo-2-1124-13B-DPO
Text Generation • Updated 25 days ago • 362

allenai/OLMo-2-1124-7B-DPO
Text Generation • Updated Jan 5 • 8.2k • ♡ 1

OLMo: Dolma

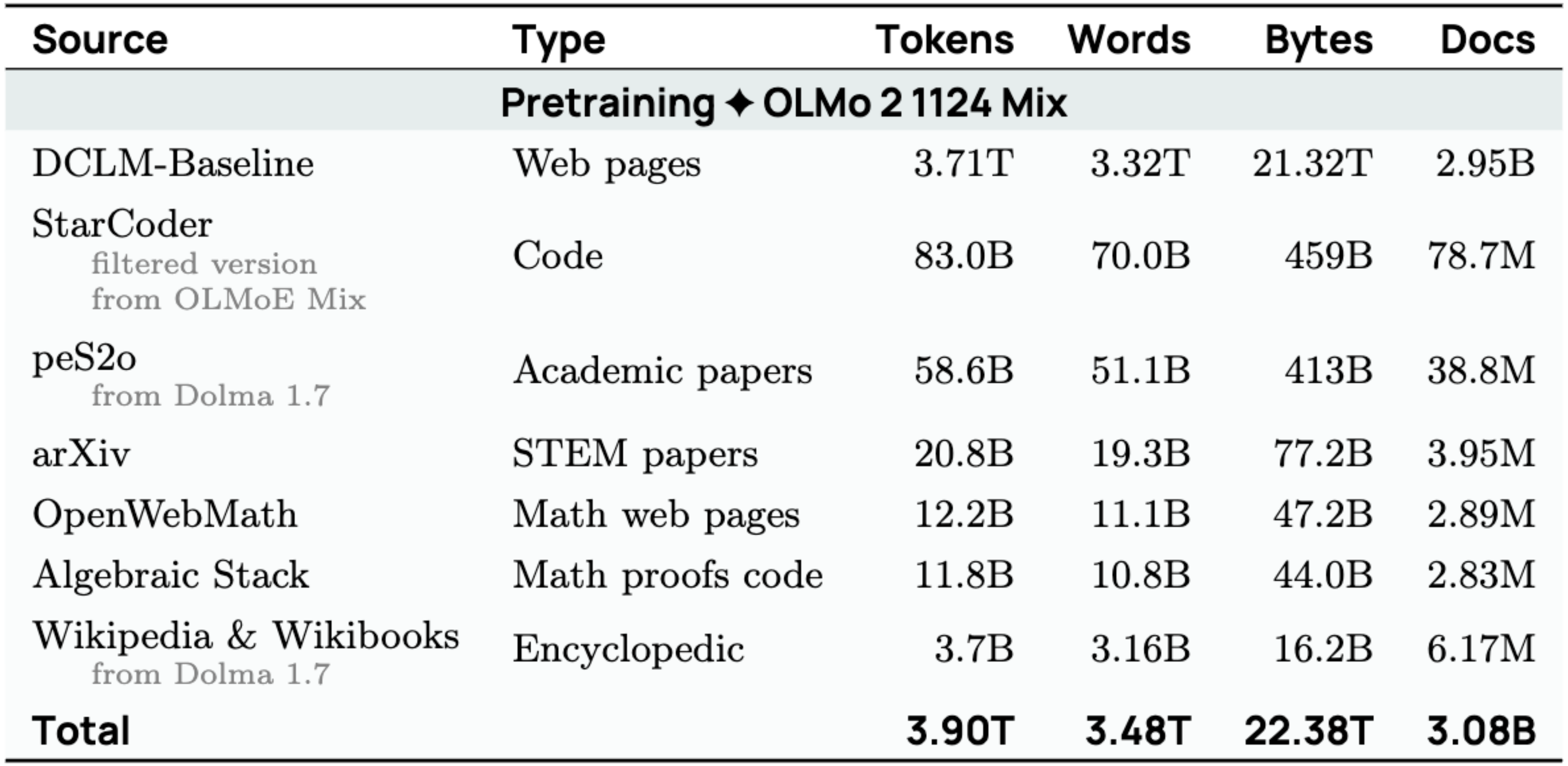
- 3T token corpus created and released by AI2 for LM training
- A pipeline of (1) language filtering, (2) quality filtering, (3) content filtering, (4) deduplication, (5) multi-source mixing, and (6) tokenization



Source	**Doc Type**	**UTF-8 bytes** *(GB)*	**Documents** *(millions)*	**Unicode words** *(billions)*	**Llama tokens** *(billions)*
Common Crawl	web pages	9,812	3,734	1,928	2,479
GitHub	code	1,043	210	260	411
Reddit	social media	339	377	72	89
Semantic Scholar	papers	268	38.8	50	70
Project Gutenberg	books	20.4	0.056	4.0	6.0
Wikipedia, Wikibooks	encyclopedic	16.2	6.2	3.7	4.3
Total		**11,519**	**4,367**	**2,318**	**3,059**

Dolma: an Open Corpus of Three Trillion Tokens for Language Model Pretraining Research, 2024

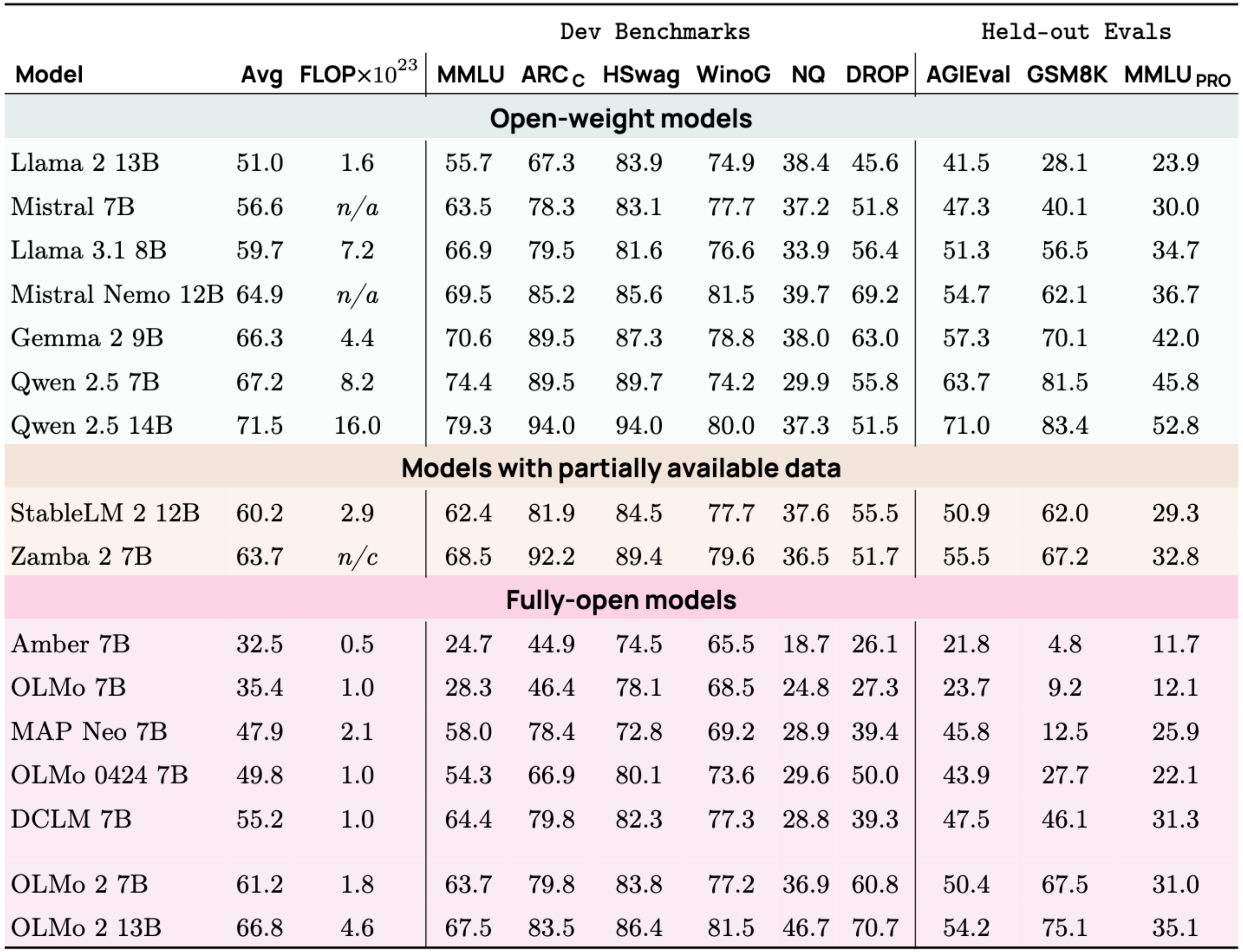
OLMo 2



Source	**Type**	**Tokens**	**Words**	**Bytes**	**Docs**
	Pretraining ✦ OLMo 2 1124 Mix				
DCLM-Baseline	Web pages	3.71T	3.32T	21.32T	2.95B
StarCoder filtered version from OLMoE Mix	Code	83.0B	70.0B	459B	78.7M
peS2o from Dolma 1.7	Academic papers	58.6B	51.1B	413B	38.8M
arXiv	STEM papers	20.8B	19.3B	77.2B	3.95M
OpenWebMath	Math web pages	12.2B	11.1B	47.2B	2.89M
Algebraic Stack	Math proofs code	11.8B	10.8B	44.0B	2.83M
Wikipedia & Wikibooks from Dolma 1.7	Encyclopedic	3.7B	3.16B	16.2B	6.17M
Total		**3.90T**	**3.48T**	**22.38T**	**3.08B**

2 OLMo 2 Furious, 2024

OLMo 2



Model	Avg	FLOP×10^{23}	Dev Benchmarks						Held-out Evals		
			MMLU	ARC$_C$	HSwag	WinoG	NQ	DROP	AGIEval	GSM8K	MMLU$_{PRO}$
Open-weight models											
Llama 2 13B	51.0	1.6	55.7	67.3	83.9	74.9	38.4	45.6	41.5	28.1	23.9
Mistral 7B	56.6	*n/a*	63.5	78.3	83.1	77.7	37.2	51.8	47.3	40.1	30.0
Llama 3.1 8B	59.7	7.2	66.9	79.5	81.6	76.6	33.9	56.4	51.3	56.5	34.7
Mistral Nemo 12B	64.9	*n/a*	69.5	85.2	85.6	81.5	39.7	69.2	54.7	62.1	36.7
Gemma 2 9B	66.3	4.4	70.6	89.5	87.3	78.8	38.0	63.0	57.3	70.1	42.0
Qwen 2.5 7B	67.2	8.2	74.4	89.5	89.7	74.2	29.9	55.8	63.7	81.5	45.8
Qwen 2.5 14B	71.5	16.0	79.3	94.0	94.0	80.0	37.3	51.5	71.0	83.4	52.8
Models with partially available data											
StableLM 2 12B	60.2	2.9	62.4	81.9	84.5	77.7	37.6	55.5	50.9	62.0	29.3
Zamba 2 7B	63.7	*n/c*	68.5	92.2	89.4	79.6	36.5	51.7	55.5	67.2	32.8
Fully-open models											
Amber 7B	32.5	0.5	24.7	44.9	74.5	65.5	18.7	26.1	21.8	4.8	11.7
OLMo 7B	35.4	1.0	28.3	46.4	78.1	68.5	24.8	27.3	23.7	9.2	12.1
MAP Neo 7B	47.9	2.1	58.0	78.4	72.8	69.2	28.9	39.4	45.8	12.5	25.9
OLMo 0424 7B	49.8	1.0	54.3	66.9	80.1	73.6	29.6	50.0	43.9	27.7	22.1
DCLM 7B	55.2	1.0	64.4	79.8	82.3	77.3	28.8	39.3	47.5	46.1	31.3
OLMo 2 7B	61.2	1.8	63.7	79.8	83.8	77.2	36.9	60.8	50.4	67.5	31.0
OLMo 2 13B	66.8	4.6	67.5	83.5	86.4	81.5	46.7	70.7	54.2	75.1	35.1

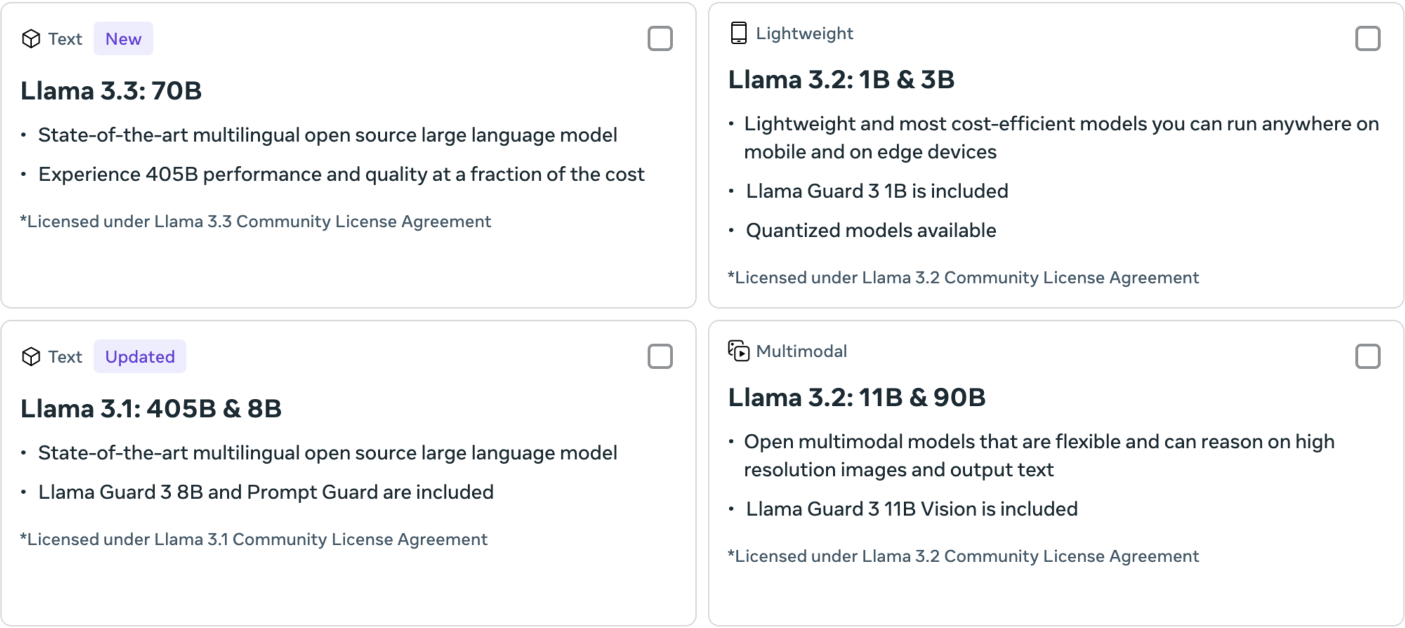
2 OLMo 2 Furious, 2024

Open-Weight LLMs

- LLaMa Series
- Mistral/Mixtral
- Qwen Series
- DeepSeek Series

LLaMa Series

- **Creator:** Meta https://ai.meta.com/blog/meta-llama-3/
- **Goal:** Strong and safe open language model
- **Unique features:** Open models with strong safeguards and chat tuning, good performance



Text New ☐

Llama 3.3: 70B

- State-of-the-art multilingual open source large language model
- Experience 405B performance and quality at a fraction of the cost

*Licensed under Llama 3.3 Community License Agreement

Lightweight ☐

Llama 3.2: 1B & 3B

- Lightweight and most cost-efficient models you can run anywhere on mobile and on edge devices
- Llama Guard 3 1B is included
- Quantized models available

*Licensed under Llama 3.2 Community License Agreement

Text Updated ☐

Llama 3.1: 405B & 8B

- State-of-the-art multilingual open source large language model
- Llama Guard 3 8B and Prompt Guard are included

*Licensed under Llama 3.1 Community License Agreement

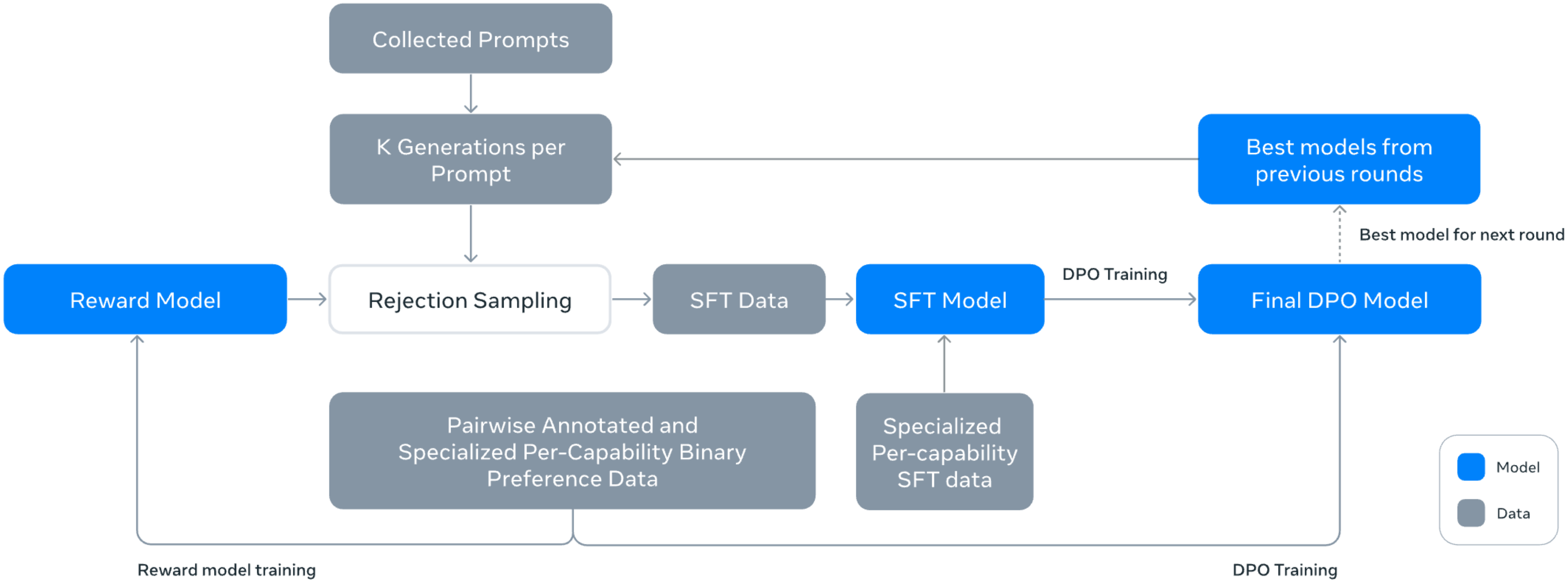
Multimodal ☐

Llama 3.2: 11B & 90B

- Open multimodal models that are flexible and can reason on high resolution images and output text
- Llama Guard 3 11B Vision is included

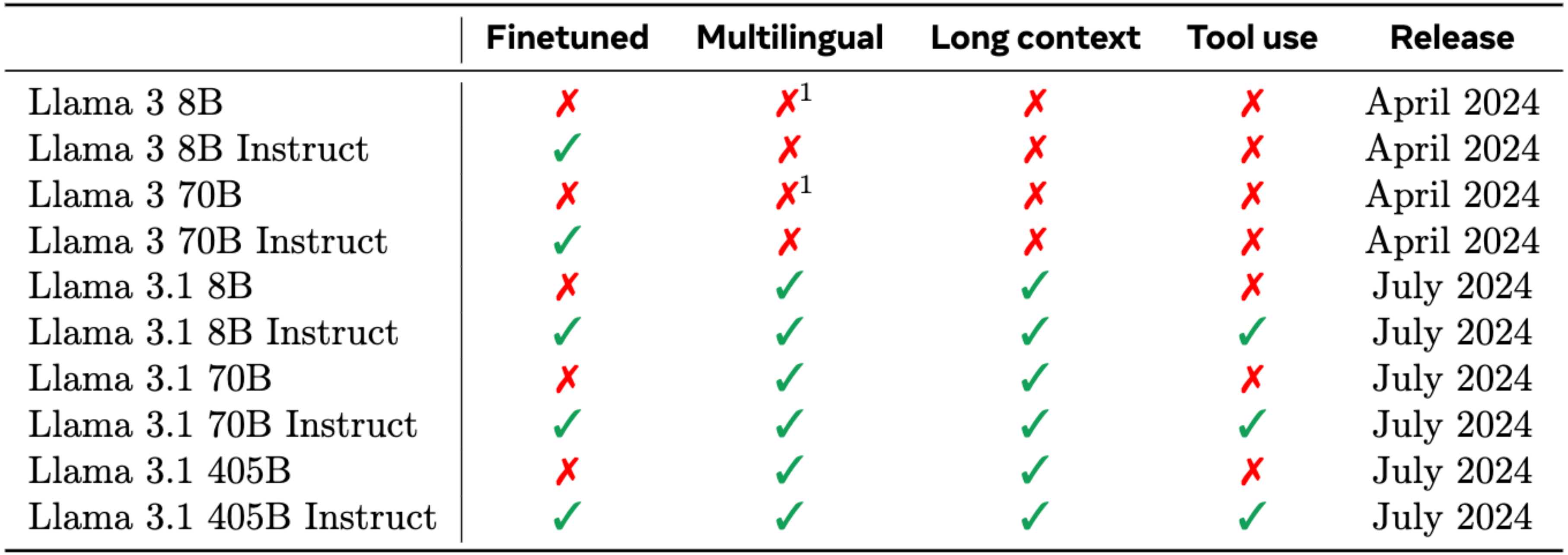
*Licensed under Llama 3.2 Community License Agreement

LLaMa 3: Post-Training Alignment


Collected Prompts
K Generations per Prompt
Best models from previous rounds
Best model for next round
Reward Model
Rejection Sampling
SFT Data
SFT Model
DPO Training
Final DPO Model
Pairwise Annotated and Specialized Per-Capability Binary Preference Data
Specialized Per-capability SFT data
Model
Data
Reward model training
DPO Training

The Llama 3 Herd of Models, 2024

LLaMa 3: Post-Training Alignment

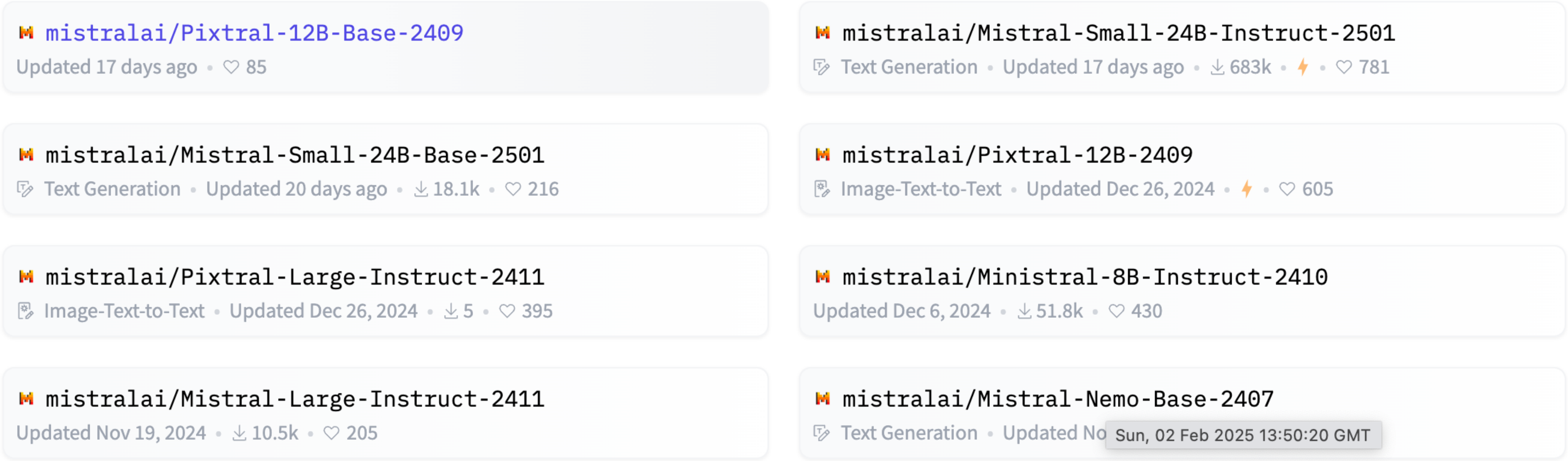


	Finetuned	**Multilingual**	**Long context**	**Tool use**	**Release**
Llama 3 8B	✗	✗[1]	✗	✗	April 2024
Llama 3 8B Instruct	✓	✗	✗	✗	April 2024
Llama 3 70B	✗	✗[1]	✗	✗	April 2024
Llama 3 70B Instruct	✓	✗	✗	✗	April 2024
Llama 3.1 8B	✗	✓	✓	✗	July 2024
Llama 3.1 8B Instruct	✓	✓	✓	✓	July 2024
Llama 3.1 70B	✗	✓	✓	✗	July 2024
Llama 3.1 70B Instruct	✓	✓	✓	✓	July 2024
Llama 3.1 405B	✗	✓	✓	✗	July 2024
Llama 3.1 405B Instruct	✓	✓	✓	✓	July 2024

The Llama 3 Herd of Models, 2024

Mistral/Mixtral

- **Creator:** MISTRAL AI_ https://mistral.ai/en/news/mixtral-of-experts
- **Goal:** Strong and somewhat multilingual open language model
- **Unique features:** Speed optimizations, including GQA and Mixture of Experts



mistralai/Pixtral-12B-Base-2409
Updated 17 days ago • ♡ 85

mistralai/Mistral-Small-24B-Instruct-2501
Text Generation • Updated 17 days ago • ⤓ 683k • ⚡ • ♡ 781

mistralai/Mistral-Small-24B-Base-2501
Text Generation • Updated 20 days ago • ⤓ 18.1k • ♡ 216

mistralai/Pixtral-12B-2409
Image-Text-to-Text • Updated Dec 26, 2024 • ⚡ • ♡ 605

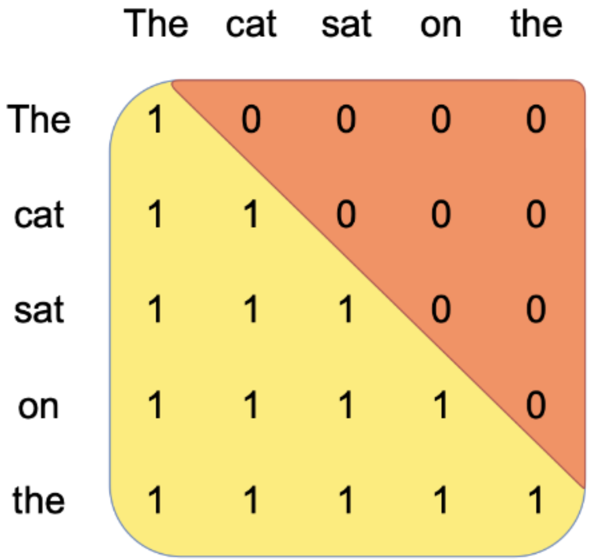
mistralai/Pixtral-Large-Instruct-2411
Image-Text-to-Text • Updated Dec 26, 2024 • ⤓ 5 • ♡ 395

mistralai/Ministral-8B-Instruct-2410
Updated Dec 6, 2024 • ⤓ 51.8k • ♡ 430

mistralai/Mistral-Large-Instruct-2411
Updated Nov 19, 2024 • ⤓ 10.5k • ♡ 205

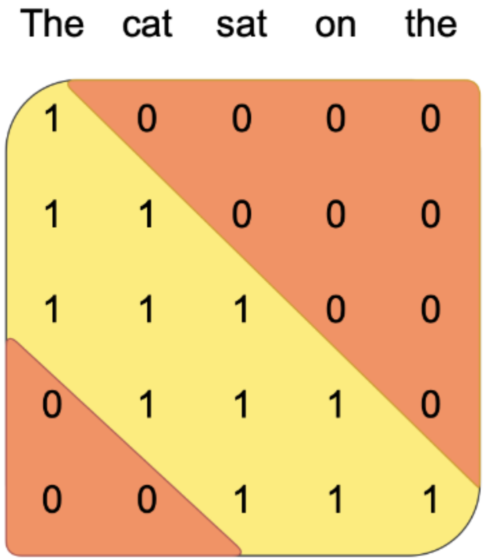
mistralai/Mistral-Nemo-Base-2407
Text Generation • Updated No Sun, 02 Feb 2025 13:50:20 GMT

Mistral/Mixtral: Sliding Window Attention



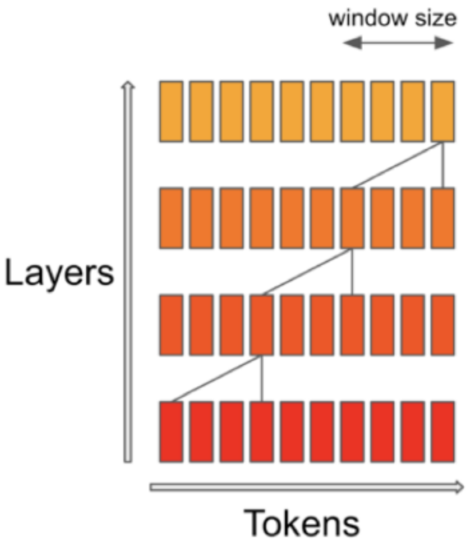
	The	cat	sat	on	the
The	1	0	0	0	0
cat	1	1	0	0	0
sat	1	1	1	0	0
on	1	1	1	1	0
the	1	1	1	1	1

Vanilla Attention



	The	cat	sat	on	the
The	1	0	0	0	0
cat	1	1	0	0	0
sat	1	1	1	0	0
on	0	1	1	1	0
the	0	0	1	1	1

Sliding Window Attention

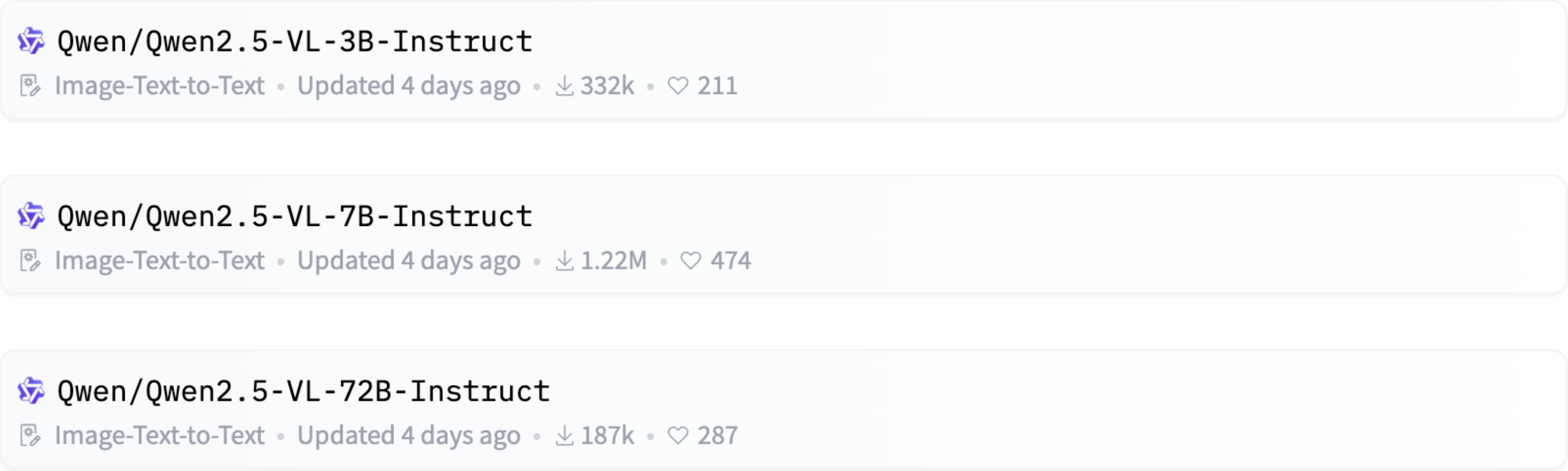


Effective Context Length

Mistral 7B, 2023

Qwen Series

- **Creator:** Alibaba https://github.com/QwenLM/Qwen2.5
- **Goal:** Strong multilingual (esp. English and Chinese) language model
- **Unique features:** Large vocabulary for multilingual support, strong performance

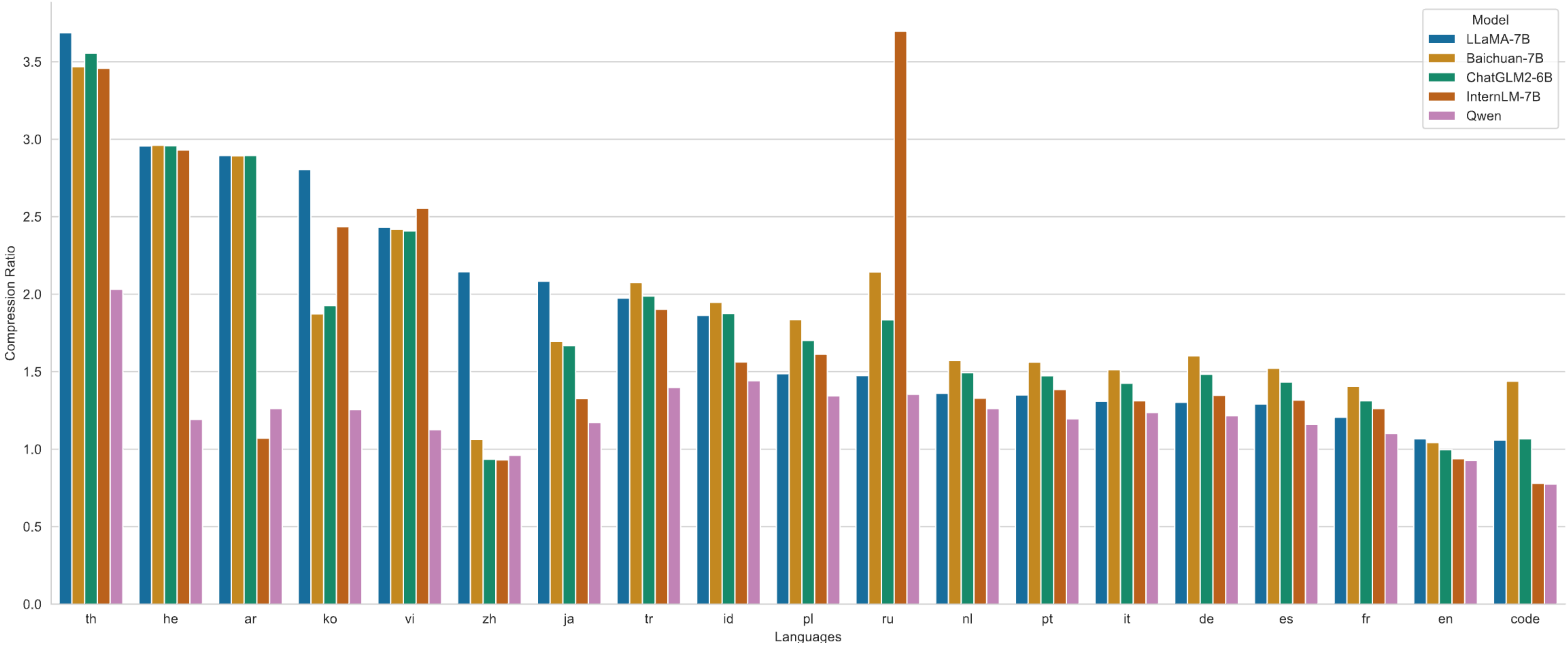


Qwen/Qwen2.5-VL-3B-Instruct
Image-Text-to-Text • Updated 4 days ago • ⤓ 332k • ♡ 211

Qwen/Qwen2.5-VL-7B-Instruct
Image-Text-to-Text • Updated 4 days ago • ⤓ 1.22M • ♡ 474

Qwen/Qwen2.5-VL-72B-Instruct
Image-Text-to-Text • Updated 4 days ago • ⤓ 187k • ♡ 287

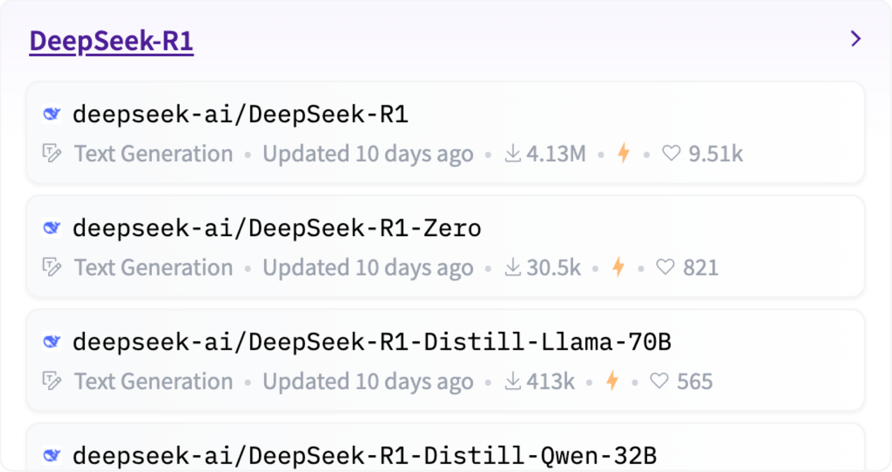
Qwen: Multilinguality



Qwen Technical Report, 2023

DeepSeek Series

- **Creator:** deepseek https://www.deepseek.com/
- **Goal:** Strongest open-weight language model so far
- **Unique features:** Relatively low-cost reinforcement-learning-based alignment for reasoning



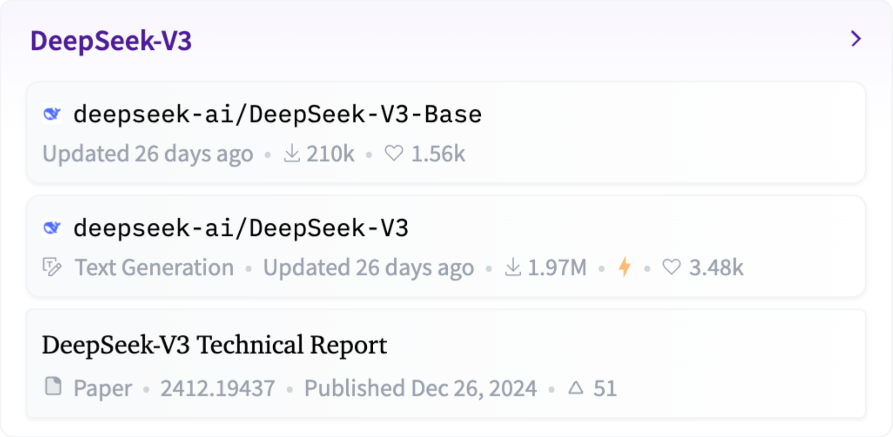
DeepSeek-R1

deepseek-ai/DeepSeek-R1
Text Generation • Updated 10 days ago • 4.13M • • 9.51k

deepseek-ai/DeepSeek-R1-Zero
Text Generation • Updated 10 days ago • 30.5k • • 821

deepseek-ai/DeepSeek-R1-Distill-Llama-70B
Text Generation • Updated 10 days ago • 413k • • 565

deepseek-ai/DeepSeek-R1-Distill-Qwen-32B



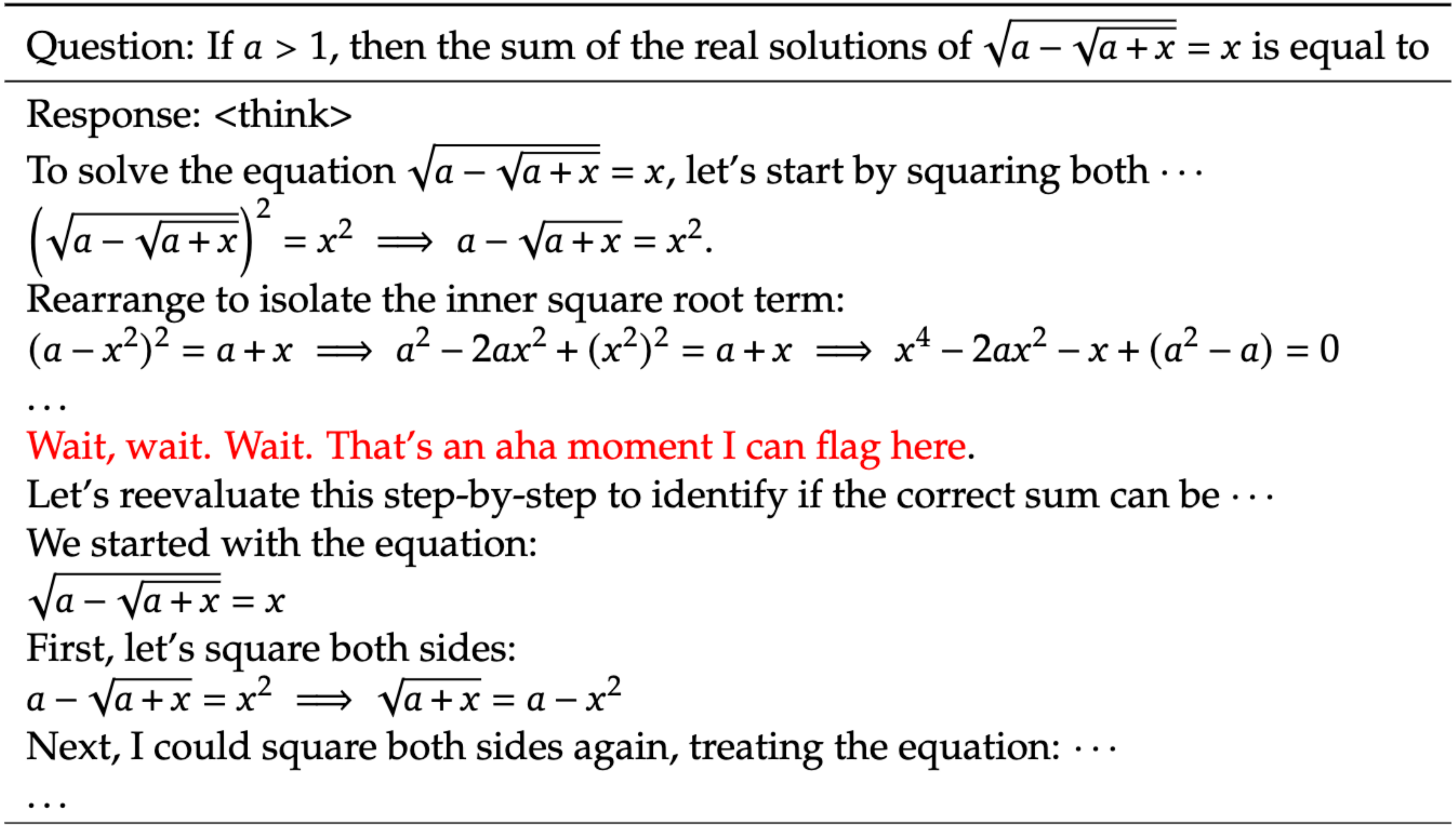
DeepSeek-V3

deepseek-ai/DeepSeek-V3-Base
Updated 26 days ago • 210k • 1.56k

deepseek-ai/DeepSeek-V3
Text Generation • Updated 26 days ago • 1.97M • • 3.48k

DeepSeek-V3 Technical Report
Paper • 2412.19437 • Published Dec 26, 2024 • 51

DeepSeek-R1: Aha Moment



Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a + x}} = x$ is equal to

Response: <think>

To solve the equation $\sqrt{a - \sqrt{a + x}} = x$, let's start by squaring both $\cdots$

$$\left(\sqrt{a - \sqrt{a + x}}\right)^2 = x^2 \implies a - \sqrt{a + x} = x^2.$$

Rearrange to isolate the inner square root term:

$$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$$

...

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step-by-step to identify if the correct sum can be $\cdots$

We started with the equation:

$$\sqrt{a - \sqrt{a + x}} = x$$

First, let's square both sides:

$$a - \sqrt{a + x} = x^2 \implies \sqrt{a + x} = a - x^2$$

Next, I could square both sides again, treating the equation: $\cdots$

...

DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, 2025

Other Open-Weight Models

- Code
  - StarCoder 2 BigCode
  - Code Llama Meta
  - DeepSeek-Coder-V2 deepseek
- Math
  - LLeMa EleutherAI
  - DeepSeek Math deepseek
- Science
  - Galactica Meta

Closed Models

- GPT Series
- Gemini Series
- Claude Series
- Grok Series

GPT Series

- **Creator:** OpenAI https://openai.com/
- ChatGPT-4o
  - Most popular LLM so far
- o1
  - Optimize for deep reasoning

Gemini Series

- **Creator:** DeepMind https://gemini.google.com/
- Gemini 2.0

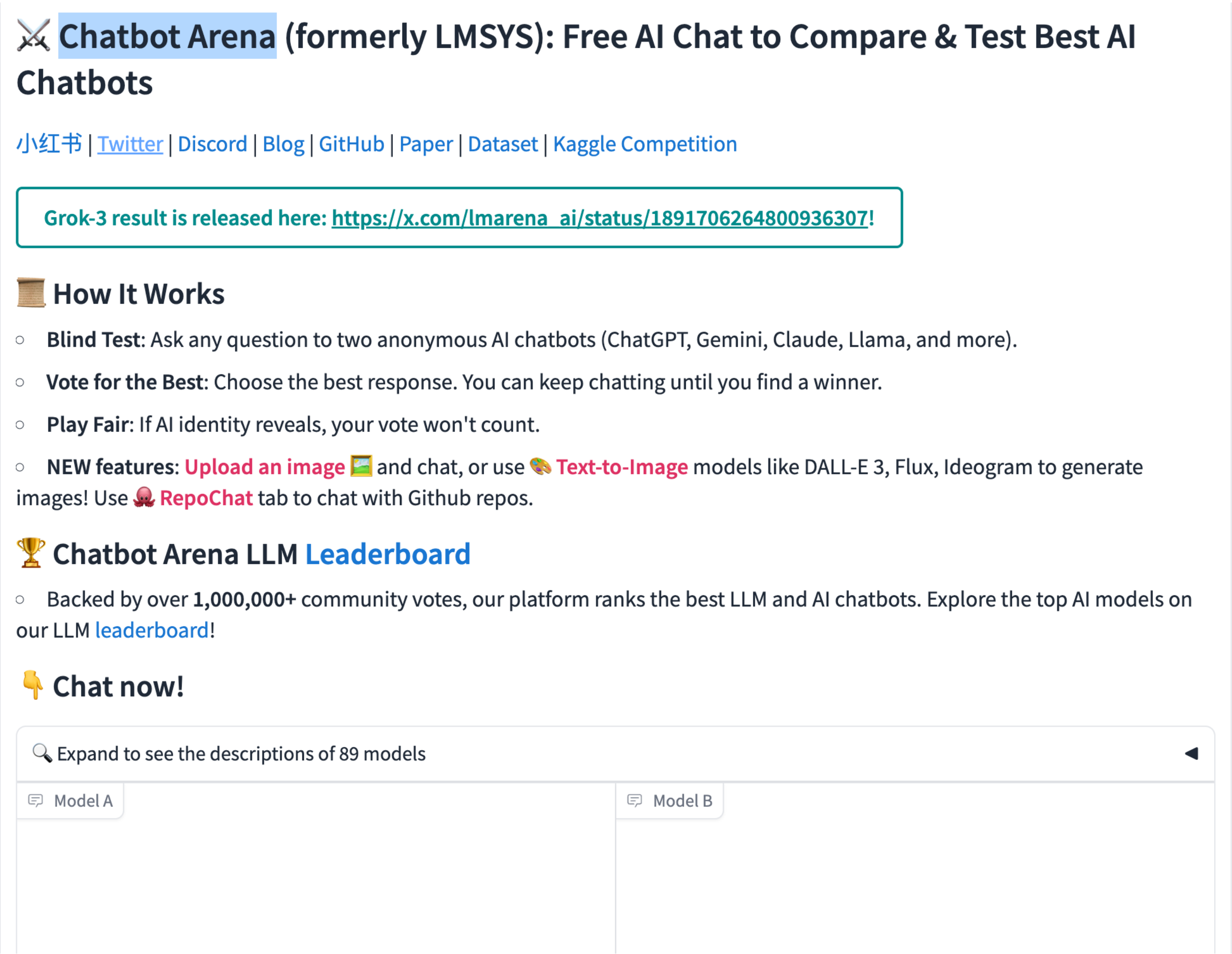
Claude Series

- **Creator:** ANTHROP\C https://claude.ai/
- Claude 3.5 Sonnet
    - Optimize for safety

Grok Series

- **Creator:** https://x.ai/
- Grok 3
  - Recently released, rank 1 in many benchmarks

Chatbot Arena



⚔️ Chatbot Arena (formerly LMSYS): Free AI Chat to Compare & Test Best AI Chatbots

小红书 | Twitter | Discord | Blog | GitHub | Paper | Dataset | Kaggle Competition

Grok-3 result is released here: https://x.com/lmarena_ai/status/1891706264800936307!

📜 How It Works

- **Blind Test**: Ask any question to two anonymous AI chatbots (ChatGPT, Gemini, Claude, Llama, and more).
- **Vote for the Best**: Choose the best response. You can keep chatting until you find a winner.
- **Play Fair**: If AI identity reveals, your vote won't count.
- **NEW features**: **Upload an image** 🖼️ and chat, or use 🎨 **Text-to-Image** models like DALL-E 3, Flux, Ideogram to generate images! Use 🐙 **RepoChat** tab to chat with Github repos.

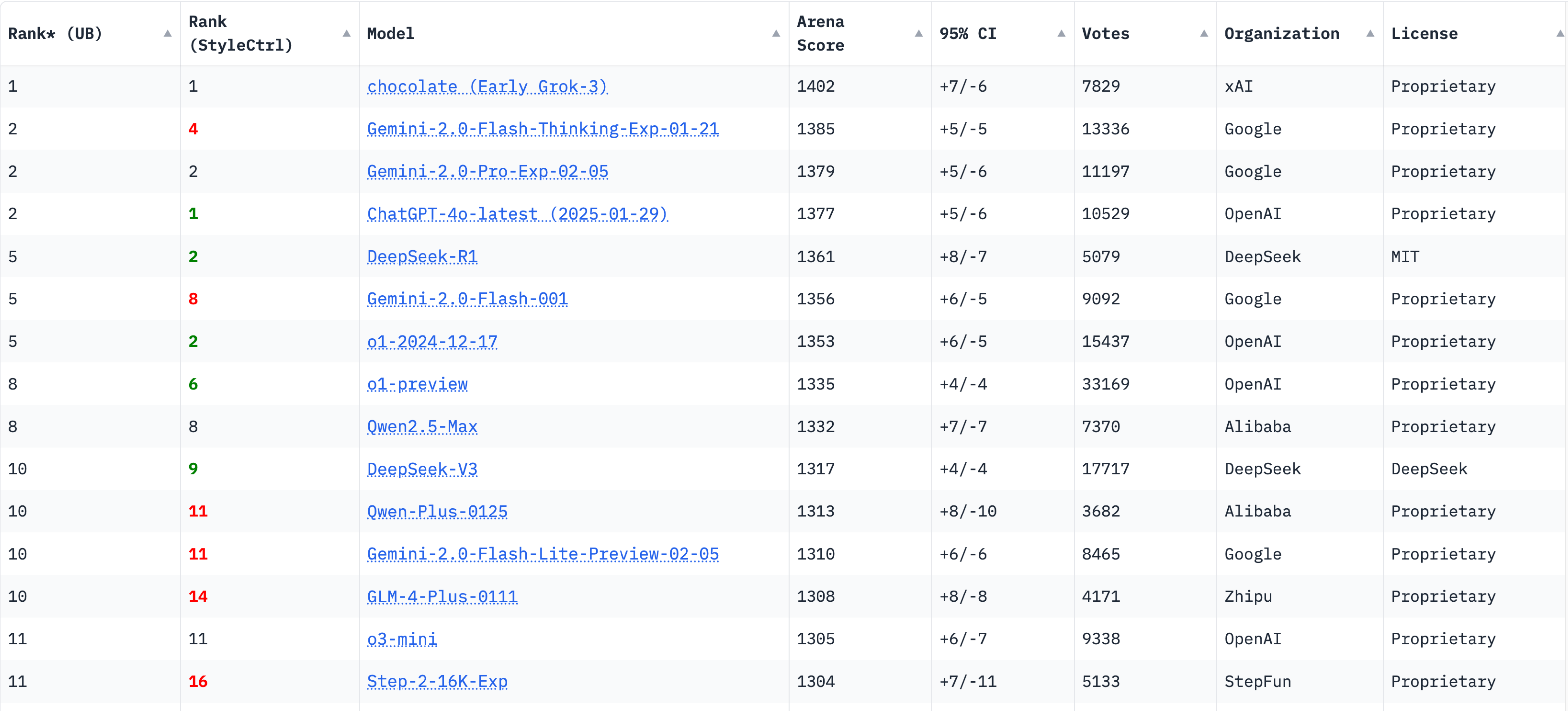
🏆 Chatbot Arena LLM Leaderboard

- Backed by over **1,000,000+** community votes, our platform ranks the best LLM and AI chatbots. Explore the top AI models on our LLM leaderboard!

👇 Chat now!

🔍 Expand to see the descriptions of 89 models ◀

Chatbot Arena Leaderboard



Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score	95% CI	Votes	Organization	License
1	1	chocolate (Early Grok-3)	1402	+7/-6	7829	xAI	Proprietary
2	4	Gemini-2.0-Flash-Thinking-Exp-01-21	1385	+5/-5	13336	Google	Proprietary
2	2	Gemini-2.0-Pro-Exp-02-05	1379	+5/-6	11197	Google	Proprietary
2	1	ChatGPT-4o-latest (2025-01-29)	1377	+5/-6	10529	OpenAI	Proprietary
5	2	DeepSeek-R1	1361	+8/-7	5079	DeepSeek	MIT
5	8	Gemini-2.0-Flash-001	1356	+6/-5	9092	Google	Proprietary
5	2	o1-2024-12-17	1353	+6/-5	15437	OpenAI	Proprietary
8	6	o1-preview	1335	+4/-4	33169	OpenAI	Proprietary
8	8	Qwen2.5-Max	1332	+7/-7	7370	Alibaba	Proprietary
10	9	DeepSeek-V3	1317	+4/-4	17717	DeepSeek	DeepSeek
10	11	Qwen-Plus-0125	1313	+8/-10	3682	Alibaba	Proprietary
10	11	Gemini-2.0-Flash-Lite-Preview-02-05	1310	+6/-6	8465	Google	Proprietary
10	14	GLM-4-Plus-0111	1308	+8/-8	4171	Zhipu	Proprietary
11	11	o3-mini	1305	+6/-7	9338	OpenAI	Proprietary
11	16	Step-2-16K-Exp	1304	+7/-11	5133	StepFun	Proprietary

Lecture Plan

- Parameter-Efficient Fine-Tuning
  - Prompt Tuning
  - Prefix Tuning
  - Adapter
  - Mixture of Experts
  - LoRA
- Large Language Models