香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

COMP 4901B

Large Language Models

Fine-tuning a Language Model: In-Context Learning and Parameter-Efficient Tuning

Junxian He

Sep 26, 2025

HW1 Out, Due Oct 8

- Post questions on Canvas
- You will try training a baby llama model, implement basic attention and sampling methods, and do inference with the trained model

Review: Pretraining Data

We want to start with clean text

- Wikipedia
- Books

History [edit]

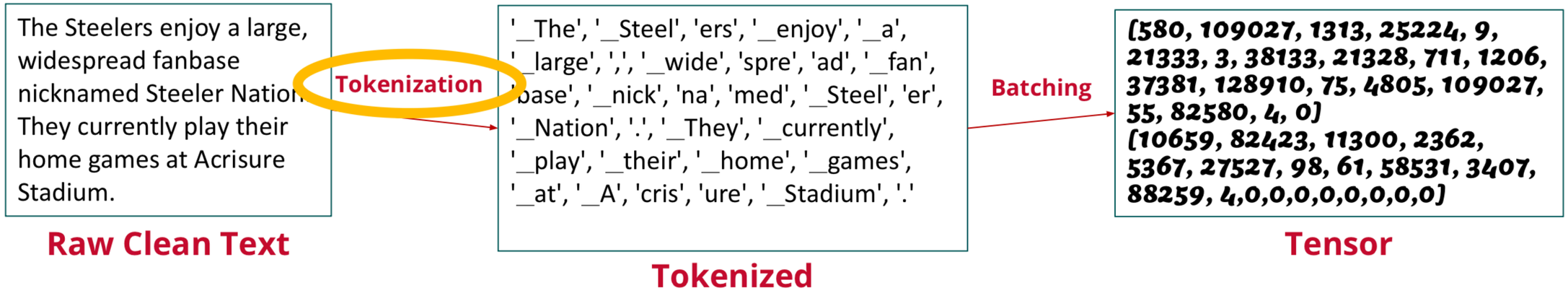
In the late 1980s, the Hong Kong Government anticipated a strong demand for university graduates to fuel an economy increasingly based on services. Sir Sze-Yuen Chung and the territory's governor, Sir Edward Youde, conceived the idea of establishing a third university, in addition to the pre-existing University of Hong Kong and Chinese University of Hong Kong.[7]

Planning for the "Third University", as the university was known provisionally, began in 1986. On 8 November 1989, Charles, Prince of Wales (now King Charles III) laid the foundation stone of the campus,[8] which was constructed at the Kohima Barracks site in Tai Po Tsai on the Clear Water Bay Peninsula. The site was earmarked for the construction of a new British Army garrison to house the 2nd King Edward VII's Own and 7th Duke of Edinburgh's Own Gurkha Rifles,[9] but plans for its construction were shelved after the 1984 signing of the Sino-British Joint Declaration resulted in the downsizing of army presence in Hong Kong.[10]

Originally scheduled to finish in 1994, the planning committee for the university decided in 1987 that the new institution should open its doors three years early, in keeping with the community's need and in fulfilment of the wishes of Youde, who died in 1986.[11][12] The university was officially opened by Youde's successor as governor, Sir David Wilson, on 10 October 1991.[13] Several leading scientists and researchers took up positions at the university in its early years, including physicist Leroy Chang who arrived in 1993 as Dean of Science and went on to become vice-president for academic affairs.[14] Thomas E. Stelson was also a founding member of the administration.[15]

Review: Preprocessing Clean Text

After the text is cleaned, now we need to convert it into a batch of training data



Tokenizing Text

A tokenizer takes text and turns it into a sequence of discrete tokens

A vocabulary is a list of all available tokens

Example: “A hippopotamus ate my homework”

Vocab Type	Example	Length
character-level	['A', ' ', 'h', 'i', 'p', 'p', 'o', 'p', 'o', 't', 'a', 'm', 'u', 's', ' ', 'a', 't', 'e', ' ', 'm', 'y', ' ', 'h', 'o', 'm', 'e', 'w', 'o', 'r', 'k', '.’]	31
subword-level	['A', 'hip', '##pop', '##ota', '##mus', 'ate', 'my', 'homework’, '.']	9
word-level	['A', 'hippopotamus', 'ate', 'my', 'homework', '.']	6

Word-Level Tokenization

rule-based (split text by spaces, punctuation, and other similar heuristics)

Challenges

- Open vocabulary problem
  - Many words may never appear in training data (becomes [UNK])
  - This is more severe in other low-resource languages
- Words with typos also get tokenized as [UNK]

Character-Level Tokenization

Vocab Type	Example	Length
character-level	['A', ' ', 'h', 'i', 'p', 'p', 'o', 'p', 'o', 't', 'a', 'm', 'u', 's', ' ', 'a', 't', 'e', ' ', 'm', 'y', ' ', 'h', 'o', 'm', 'e', 'w', 'o', 'r', 'k', '.’]	31
subword-level	['A', 'hip', '##pop', '##ota', '##mus', 'ate', 'my', 'homework’, '.']	9
word-level	['A', 'hippopotamus', 'ate', 'my', 'homework', '.']	6

Pro: No unseen tokens anymore

Con: Sequence is unnecessarily long, expensive to work with

Sub-word Tokenization

- Words get split into multiple tokens
- Vocabulary is build dynamically
  - Frequent words get assigned their own tokens
  - Rare words are split into subwords

Vocab Type	Example	Length
character-level	['A', ' ', 'h', 'i', 'p', 'p', 'o', 'p', 'o', 't', 'a', 'm', 'u', 's', ' ', 'a', 't', 'e', ' ', 'm', 'y', ' ', 'h', 'o', 'm', 'e', 'w', 'o', 'r', 'k', '.’]	31
subword-level	['A', 'hip', '##pop', '##ota', '##mus', 'ate', 'my', 'homework’, '.']	9
word-level	['A', 'hippopotamus', 'ate', 'my', 'homework', '.']	6

Byte Pair Encoding (BPE)

Main Idea

- Construct subword vocabulary by learning to merge characters
- Inspiration comes from compression algorithms

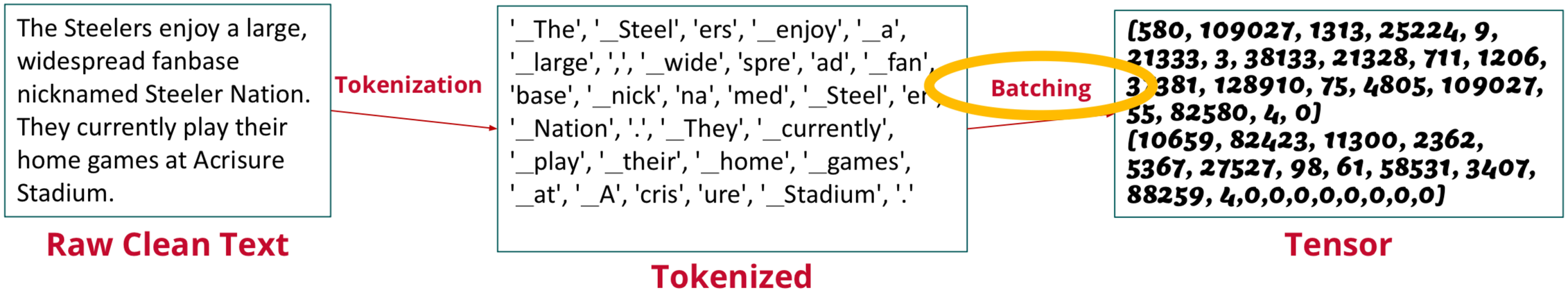
Training Steps

1. Initialize the vocabulary with characters as tokens (e.g., in English: alphabet, numbers, punctuation)
2. Merge the most frequent token pair in the corpus (vocabulary size +1)
3. Re-tokenize the corpus with the merged subword pair
4. Repeat steps 2 and 3 until the target vocabulary size is reached

Advantages of Subword Tokenization

- Controlled vocabulary size
- Strike a good balance between word-level and character-level
  - Frequent words kept whole
  - Tail words split to sub-words
    - More observations on sub-words
    - Utilization of morphology information

Batching Data

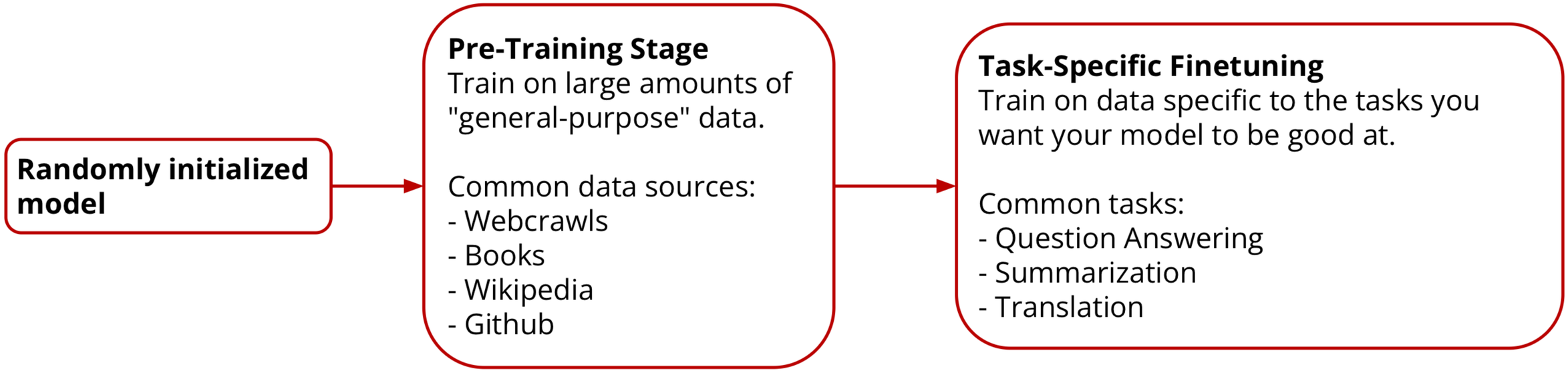


香港科技大學
THE HONG KONG
UNIVERSITY OF SCIENCE
AND TECHNOLOGY

In-Context Learning

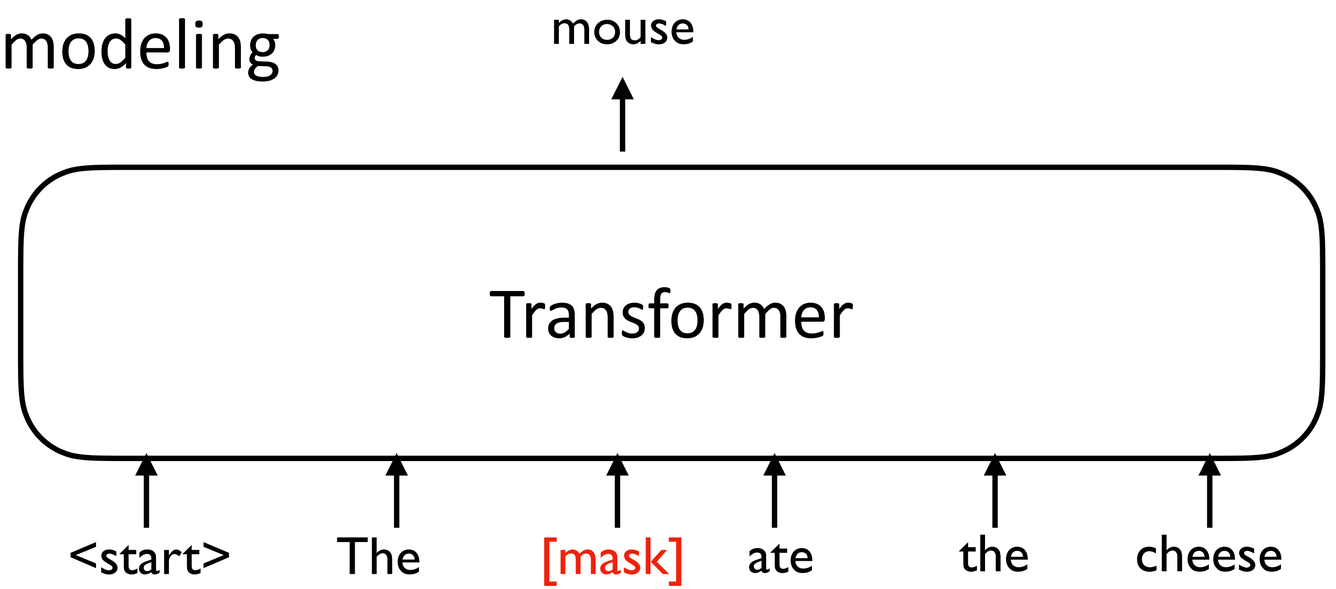
Pretraining -> Fine-Tuning

Paradigm shift around 2018



BERT

Mask language modeling



Self-supervised Learning

Construct a synthetic task from raw text only

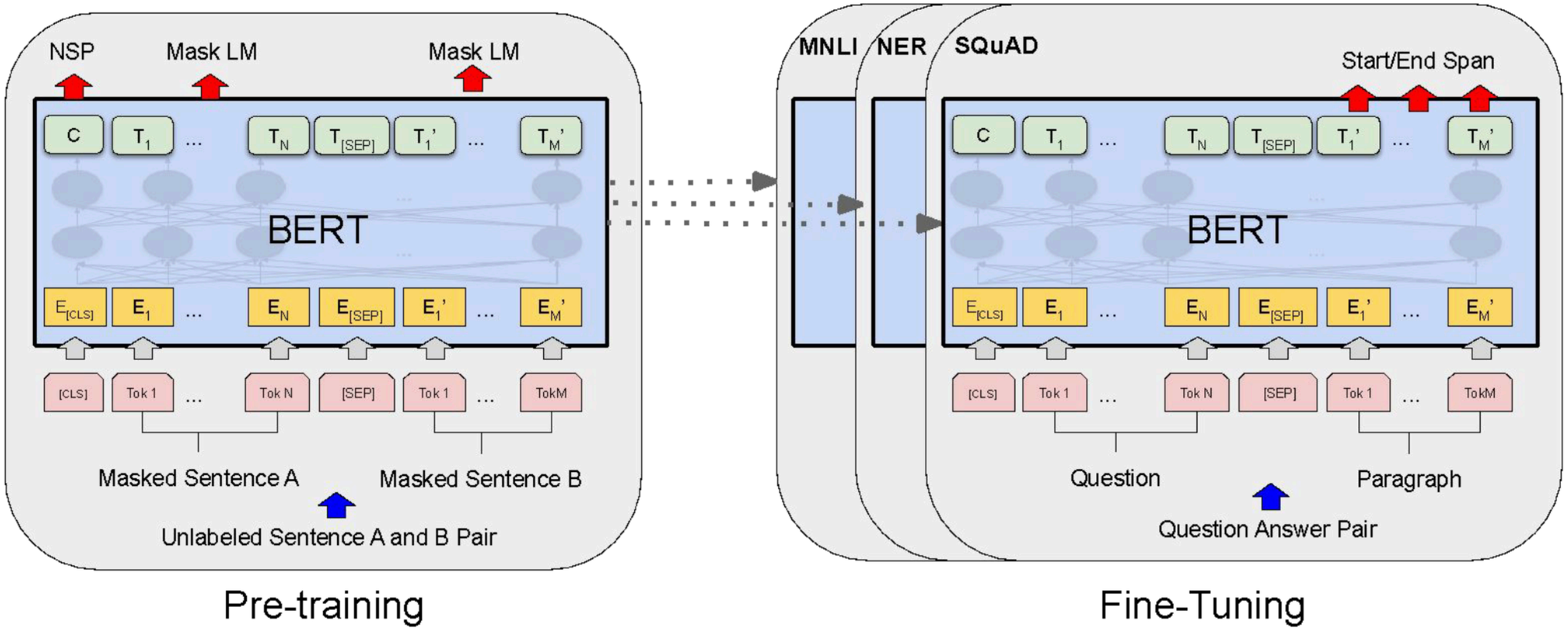
Can be made very large-scale

Is Bert a language model? Is it a generative model?

Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL 2019.

BERT

- Pre-train an encoder-only model with mask infilling and sentence ordering objectives.
- Finetune once per NLP task



GPT-1

- Pre-train a decoder-only LM with a language modelling objective.
- Finetune once per NLP task

Table 2: Experimental results on natural language inference tasks, comparing our model with current state-of-the-art methods. 5x indicates an ensemble of 5 models. All datasets use accuracy as the evaluation metric.

Method	MNLI-m	MNLI-mm	SNLI	SciTail	QNLI	RTE
ESIM + ELMo [44] (5x)	-	-	89.3	-	-	-
CAFE [58] (5x)	80.2	79.0	89.3	-	-	-
Stochastic Answer Network [35] (3x)	80.6	80.1	-	-	-	-
CAFE [58]	78.7	77.9	88.5	83.3		
GenSen [64]	71.4	71.3	-	-	82.3	59.2
Multi-task BiLSTM + Attn [64]	72.2	72.1	-	-	82.1	**61.7**
Finetuned Transformer LM (ours)	**82.1**	**81.4**	**89.9**	**88.3**	**88.1**	56.0

Natural Language Inference (NLI) Example

Met my first girlfriend that way.	FACE-TO-FACE **contradiction** C C N C	I didn't meet my first girlfriend until later.
8 million in relief in the form of emergency housing.	GOVERNMENT **neutral** N N N N	The 8 million dollars for emergency housing was still not enough to solve the problem.
Now, as children tend their gardens, they have a new appreciation of their relationship to the land, their cultural heritage, and their community.	LETTERS **neutral** N N N N	All of the children love working in their gardens.
At 8:34, the Boston Center controller received a third transmission from American 11	9/11 **entailment** E E E E	The Boston Center controller got a third transmission from American 11.
I am a lacto-vegetarian.	SLATE **neutral** N N E N	I enjoy eating cheese too much to abstain from dairy.
someone else noticed it and i said well i guess that's true and it was somewhat melodious in other words it wasn't just you know it was really funny	TELEPHONE **contradiction** C C C C	No one noticed and it wasn't funny at all.

Disadvantage of Fine-Tuning for Each Task

- One model per task is fine for small models, but not for today's big ones.
  - Training is expensive
  - Overfitting on small datasets
  - Storing one model for each task is expensive

Solutions

- Avoid fine-tuning entirely
  - In-context learning
- Parameter-efficient fine-tuning
- Multi-task fine-tuning -> instruction tuning

Is Next Token Prediction Useful?

Ok, language modeling can be used as pretraining, but is a language model itself useful for some tasks directly?

In the late 1980s the Hong Kong Government anticipated a strong demand for university graduates to fuel an economy increasingly based on services. Sir Sze-Yuen Chung and Sir Edward Youde, the then Governor of Hong Kong, conceived the idea of another university in addition to the pre-existing two universities, The University of Hong Kong and The Chinese University of Hong Kong.
Planning for the "Third University", named The Hong Kong University of Science and Technology later, began in 1986. Construction began at the Kohima Camp site in Tai Po Tsai on the Clear Water Bay Peninsula. The site was earmarked for the construction of a new [ ]

Completion

This task seems useless in practice

Language Models are Zero-Shot Learners

Context (human-written): In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

GPT-2: The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.

Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them – they were so close they could touch their horns.

While examining these bizarre creatures the scientists discovered that the creatures also spoke some fairly regular English. Pérez stated, "We can see, for example, that they have a common 'language,' something like a dialect or dialectic."

Dr. Pérez believes that the unicorns may have originated in Argentina, where the animals were believed to be descendants of a lost race of people who lived there before the arrival of humans in those parts of South America.

While their origins are still unclear, some believe that perhaps the creatures were created when a human and a unicorn met each other in a time before human civilization. According to Pérez, "In South America, such incidents seem to be quite common."

However, Pérez also pointed out that it is likely that the only way of knowing for sure if unicorns are indeed the descendants of a lost alien race is through DNA. "But they seem to be able to communicate in English quite well, which I believe is a sign of evolution, or at least a change in social organization," said the scientist.

Radford et al. Language Models are Unsupervised Multitask Learners. 2019 (GPT2)

GPT-2

Next token prediction can unify many tasks

Machine translation:

> Chinese: 今天是学期的最后一天。
> English:

Question answering:

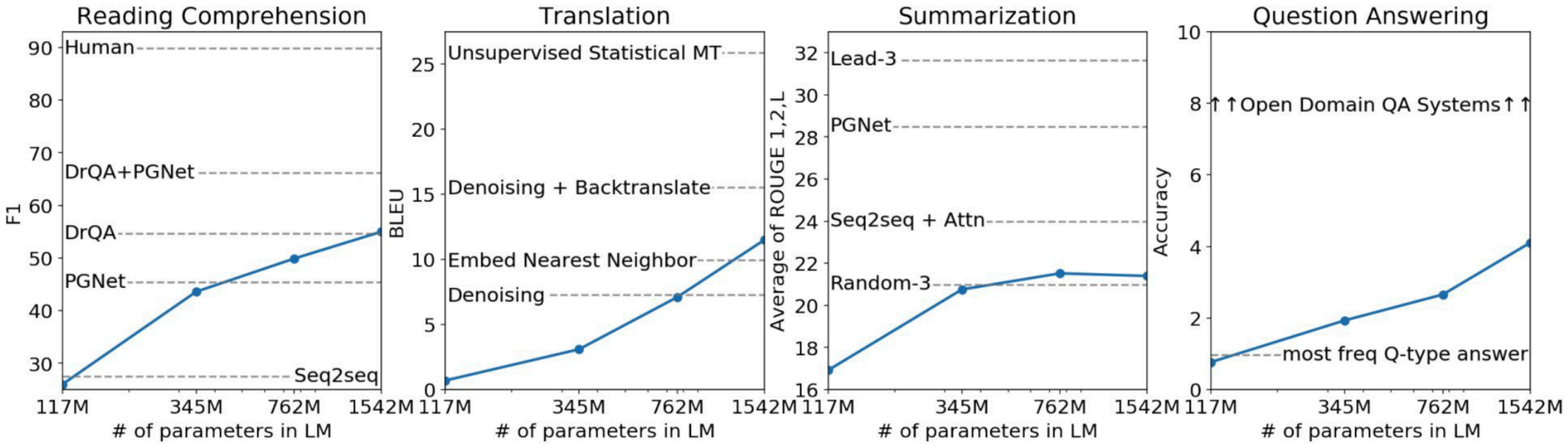
> Q: What is the capital of the United States?
> A:

Completion is very general

This was an early form of prompting, that is widely discussed today

Radford et al. Language Models are Unsupervised Multitask Learners. 2018.

Zero-Shot Performance of GPT-2



Radford et al. Language Models are Unsupervised Multitask Learners. 2018.

Language Models Are Few-Shot Learners

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  cheese =>                           ← prompt
```

One-shot

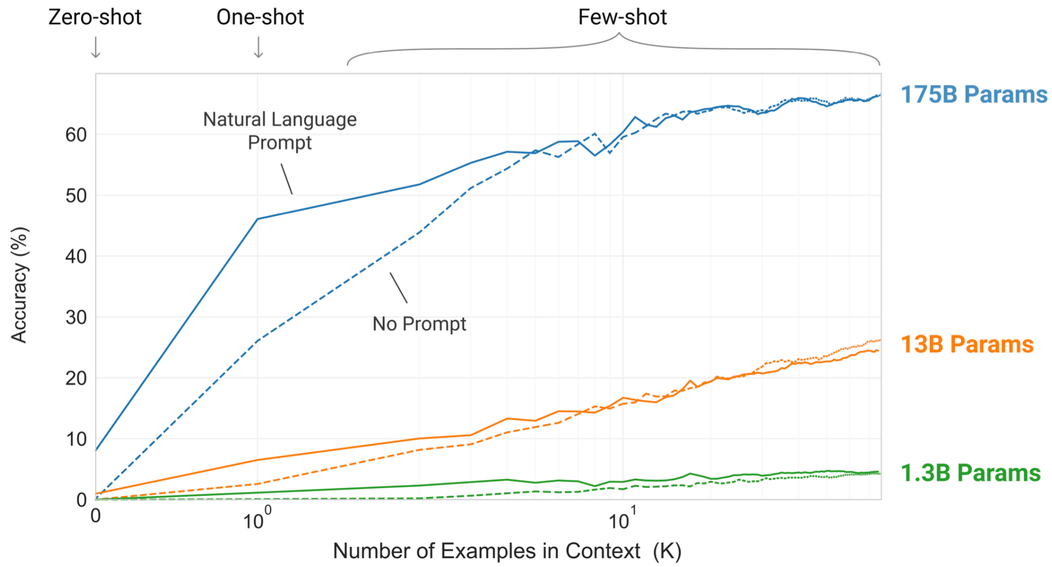
In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← example
3  cheese =>                           ← prompt
```

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

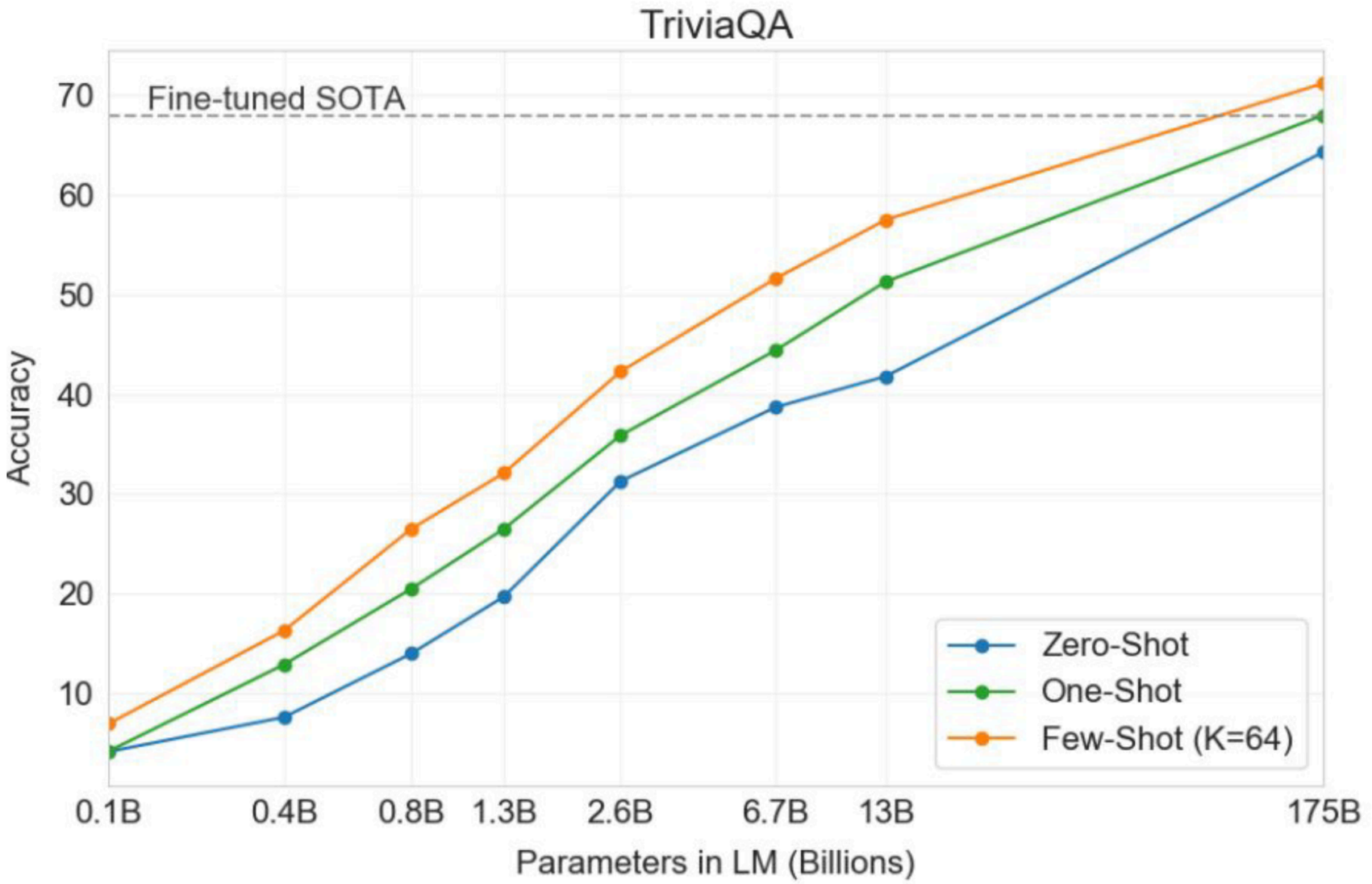
```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← examples
3  peppermint => menthe poivrée        ←
4  plush girafe => girafe peluche      ←
5  cheese =>                           ← prompt
```



In-Context Learning

Brown et al. Language models are few-shot learners. 2020

Language Models Are Few-Shot Learners



Brown et al. Language models are few-shot learners. 2020

Formally, In-Context Learning is…

- LLM zero-shot learning: a prompt that contains instructions for the task, but no actual examples of the task being performed.
- LLM few-shot learning: a prompt that contains both instructions as well as several examples of the task being performed.

A Prompt Example for Sentiment Classification

Prompt:

Review: Let there be no question: Alexions owns the best cheeseburger in the region and they have now for decades. Try a burger on Italian bread. The service is flawlessly friendly, the food is amazing, and the wings? Oh the wings... but it's still about the cheeseburger. The atmosphere is inviting, but you can't eat atmosphere... so go right now. Grab the car keys... you know you're hungry for an amazing cheeseburger, maybe some wings, and a cold beer! Easily, hands down, the best bar and grill in Pittsburgh.

On a 1 to 4 star scale, the reviewer would probably give this restaurant a

There are often different ways to verbalize a task

Prompt:

The dog chased a squirrel at the park. = 那只狗在公园里追一只松鼠。

I was late for class. = 我上课迟到了。

The hippopotamus ate my homework. =

Prompt with an Alternative Template:

Translate from English to Chinese.

The dog chased a squirrel at the park. = 那只狗在公园里追一只松鼠。

I was late for class. = 我上课迟到了。

The hippopotamus ate my homework. =

Prompt with an Alternative Template:

Translate from English to Chinese.

English: The dog chased a squirrel at the park.
Chinese: 那只狗在公园里追一只松鼠。

English: I was late for class.
Chinese: 我上课迟到了。

English: The hippopotamus ate my homework.
Chinese:

Different “Template”

Essentially, In-context Learning vs Fine-tuning?

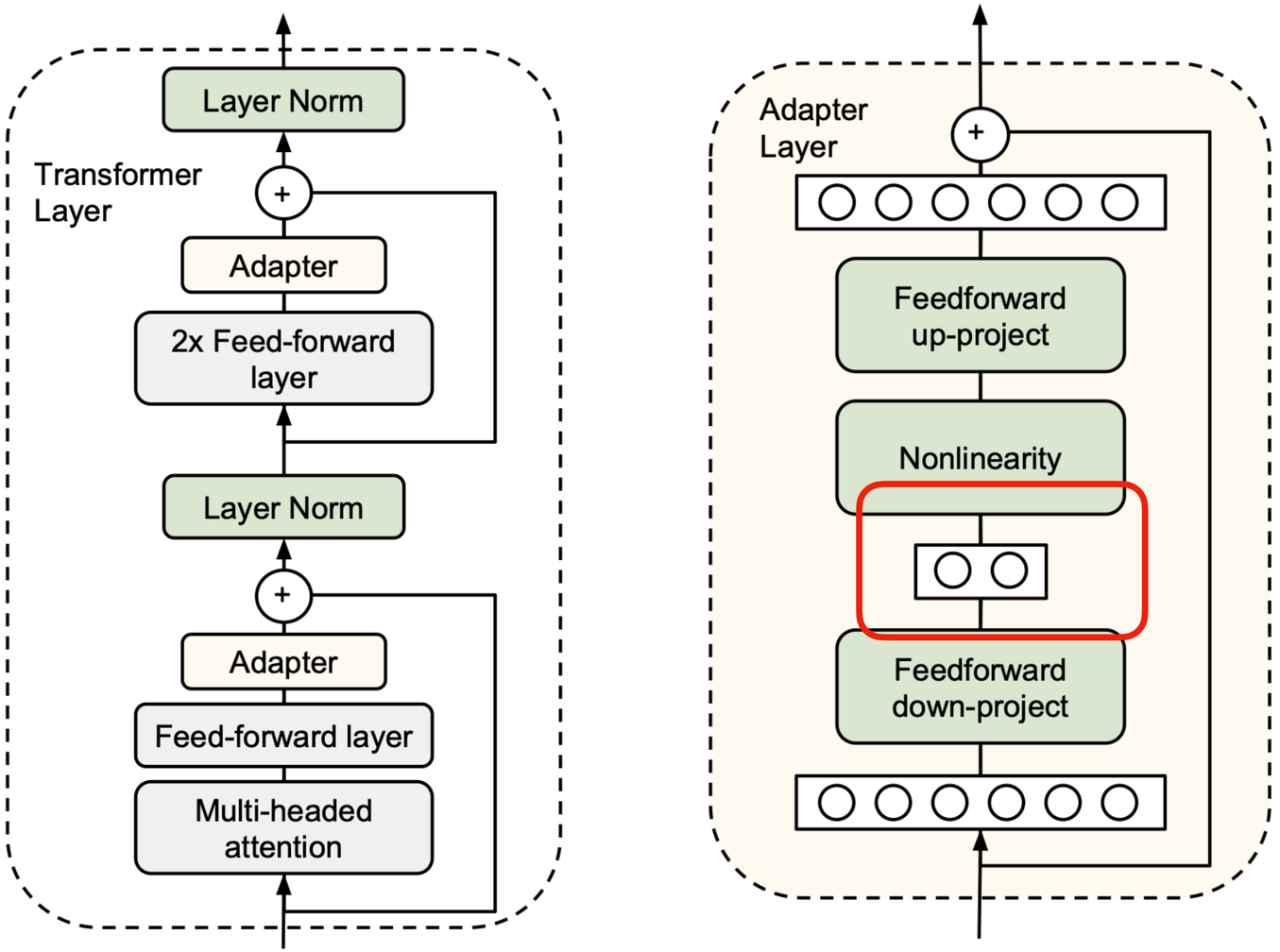
They are different ways of utilizing “annotated data”

Parameter-Efficient Fine-Tuning

Instead of fine-tuning the entire model, we just fine-tune a small amount of parameters

Storage savings

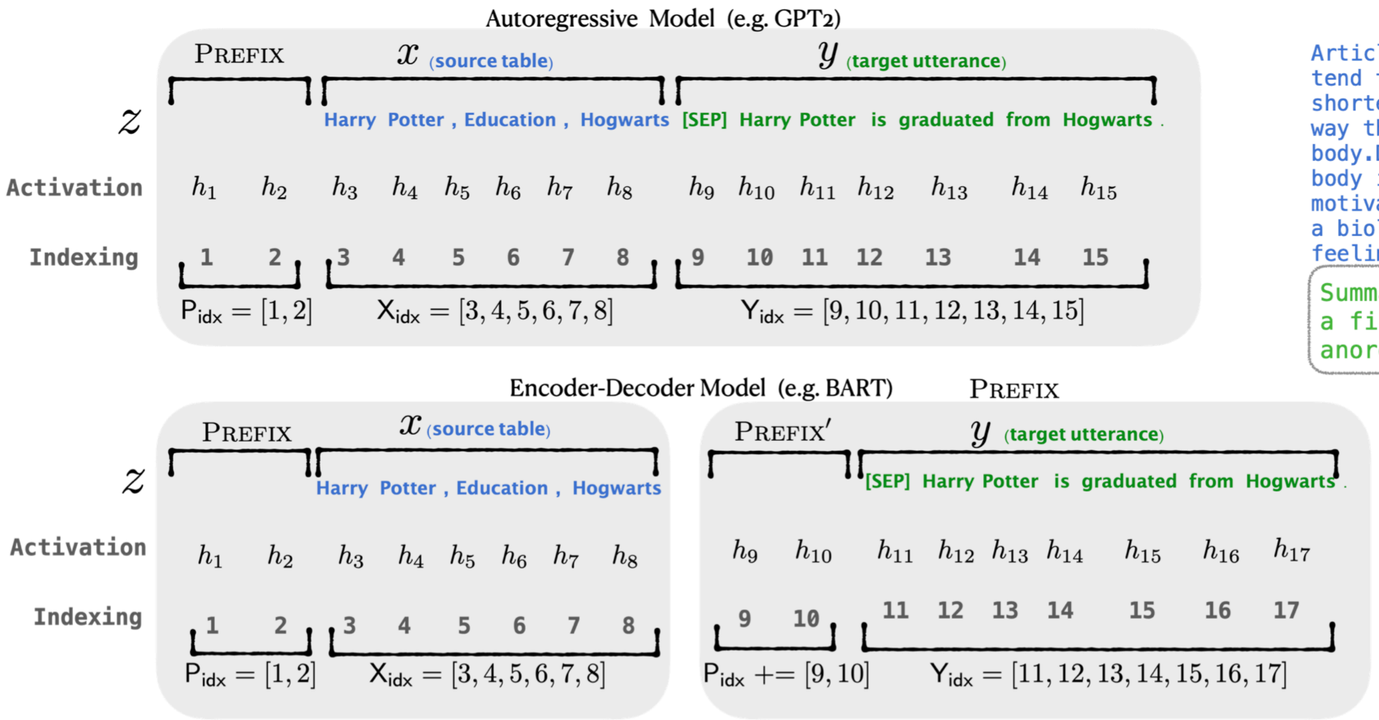
Adapter



Low-Rank

Houlsby et al. Parameter-Efficient Transfer Learning for NLP. 2019

Prefix-Tuning



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.

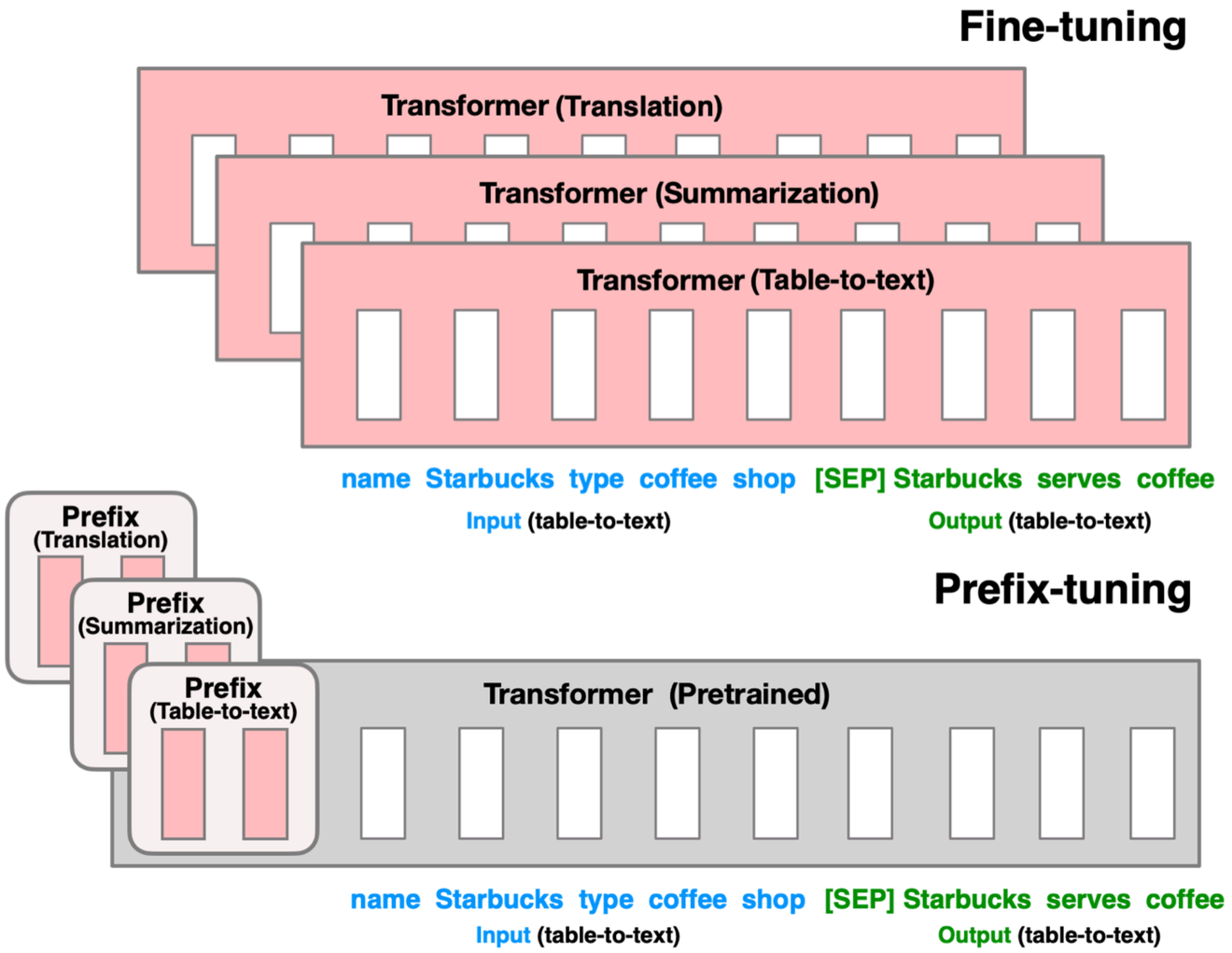
Table-to-text Example

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 .

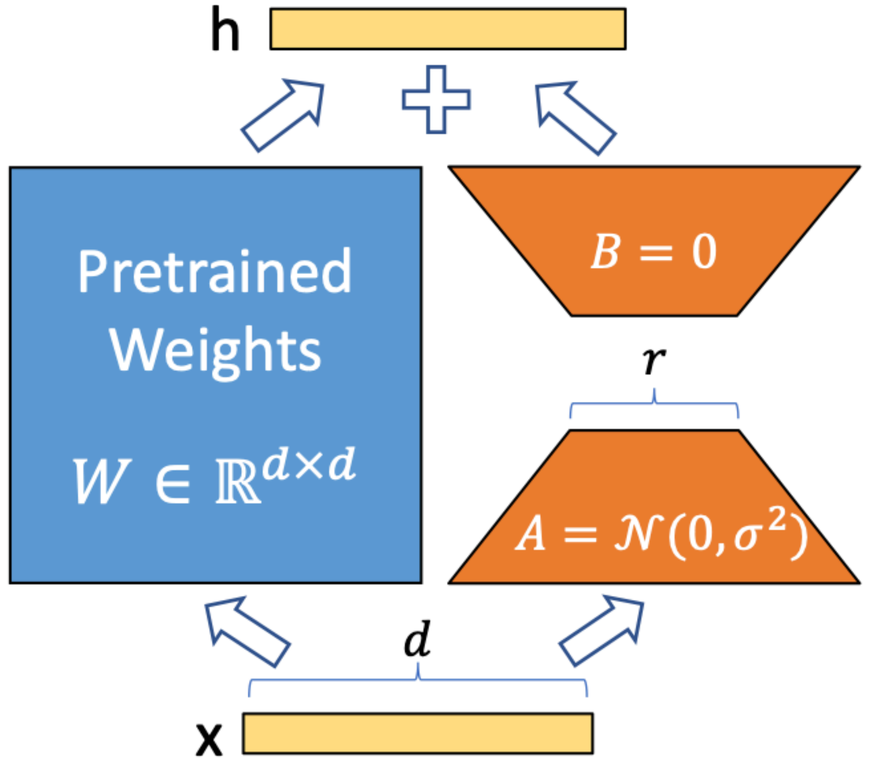
Li et al. Prefix-Tuning: Optimizing Continuous Prompts for Generation. 2021

Prefix-Tuning



Li et al. Prefix-Tuning: Optimizing Continuous Prompts for Generation. 2021

LORA: LOW-RANK ADAPTATION



Hu et al. LoRA: Low-Rank Adaptation of Large Language Models 2021

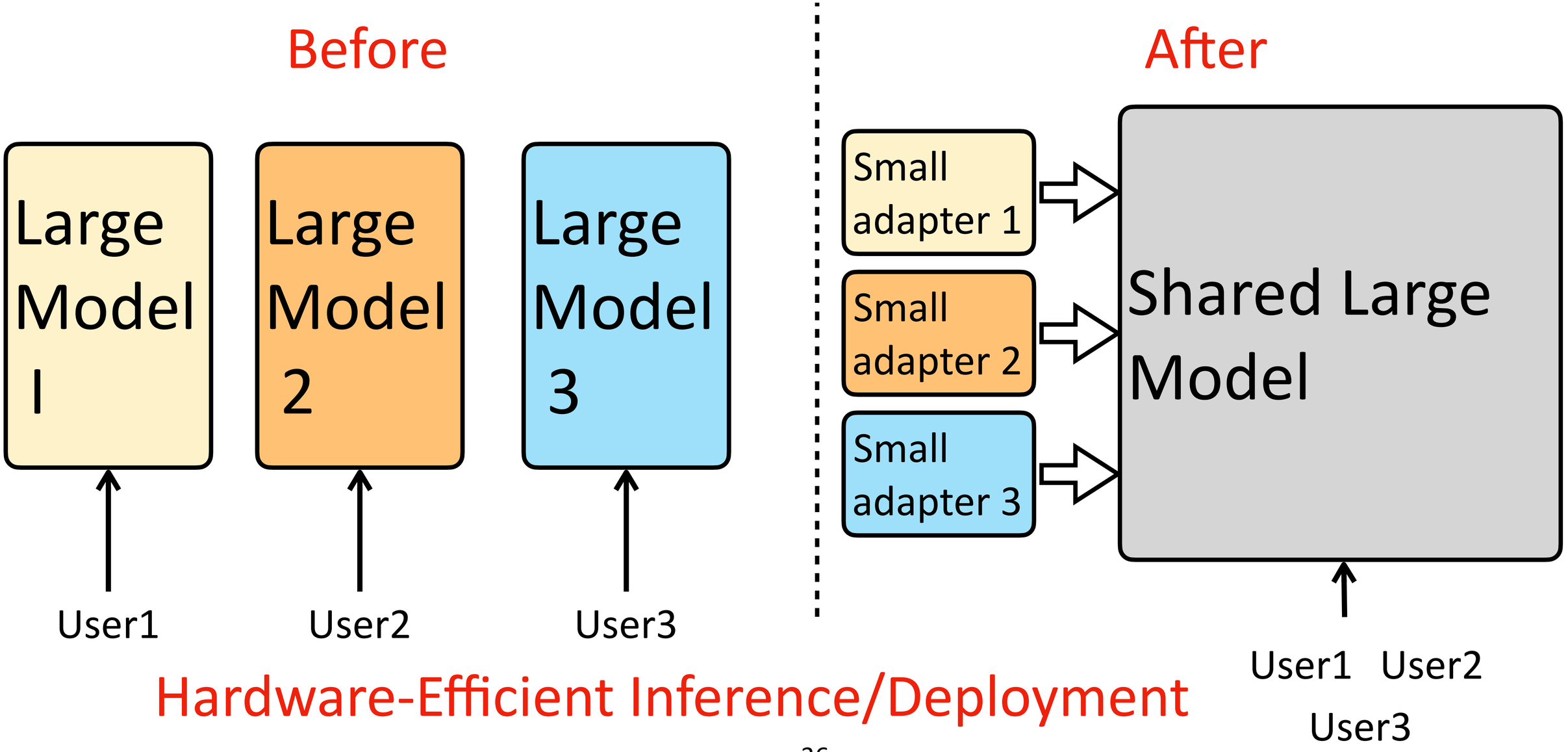
LORA: LOW-RANK ADAPTATION

Model & Method	# Trainable Parameters	E2E NLG Challenge BLEU	NIST	MET	ROUGE-L	CIDEr
GPT-2 M (FT)*	354.92M	68.2	8.62	46.2	71.0	2.47
GPT-2 M ($\text{Adapter}^{\text{L}}$)*	0.37M	66.3	8.41	45.0	69.8	2.40
GPT-2 M ($\text{Adapter}^{\text{L}}$)*	11.09M	68.9	8.71	46.1	71.3	2.47
GPT-2 M ($\text{Adapter}^{\text{H}}$)	11.09M	$67.3_{\pm.6}$	$8.50_{\pm.07}$	$46.0_{\pm.2}$	$70.7_{\pm.2}$	$2.44_{\pm.01}$
GPT-2 M (FT^{Top2})*	25.19M	68.1	8.59	46.0	70.8	2.41
GPT-2 M (PreLayer)*	0.35M	69.7	8.81	46.1	71.4	2.49
GPT-2 M (LoRA)	0.35M	$\mathbf{70.4}_{\pm.1}$	$\mathbf{8.85}_{\pm.02}$	$\mathbf{46.8}_{\pm.2}$	$\mathbf{71.8}_{\pm.1}$	$\mathbf{2.53}_{\pm.02}$
GPT-2 L (FT)*	774.03M	68.5	8.78	46.0	69.9	2.45
GPT-2 L ($\text{Adapter}^{\text{L}}$)	0.88M	$69.1_{\pm.1}$	$8.68_{\pm.03}$	$46.3_{\pm.0}$	$71.4_{\pm.2}$	$\mathbf{2.49}_{\pm.0}$
GPT-2 L ($\text{Adapter}^{\text{L}}$)	23.00M	$68.9_{\pm.3}$	$8.70_{\pm.04}$	$46.1_{\pm.1}$	$71.3_{\pm.2}$	$2.45_{\pm.02}$
GPT-2 L (PreLayer)*	0.77M	70.3	8.85	46.2	71.7	2.47
GPT-2 L (LoRA)	0.77M	$\mathbf{70.4}_{\pm.1}$	$\mathbf{8.89}_{\pm.02}$	$\mathbf{46.8}_{\pm.2}$	$\mathbf{72.0}_{\pm.2}$	$2.47_{\pm.02}$

Table 3: GPT-2 medium (M) and large (L) with different adaptation methods on the E2E NLG Challenge. For all metrics, higher is better. LoRA outperforms several baselines with comparable or fewer trainable parameters. Confidence intervals are shown for experiments we ran. * indicates numbers published in prior works.

Hu et al. LoRA: Low-Rank Adaptation of Large Language Models 2021

Why Parameter-Efficient Tuning



Hardware-Efficient Inference/Deployment

Thank You!