



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

COMP 4901B

Large Language Models

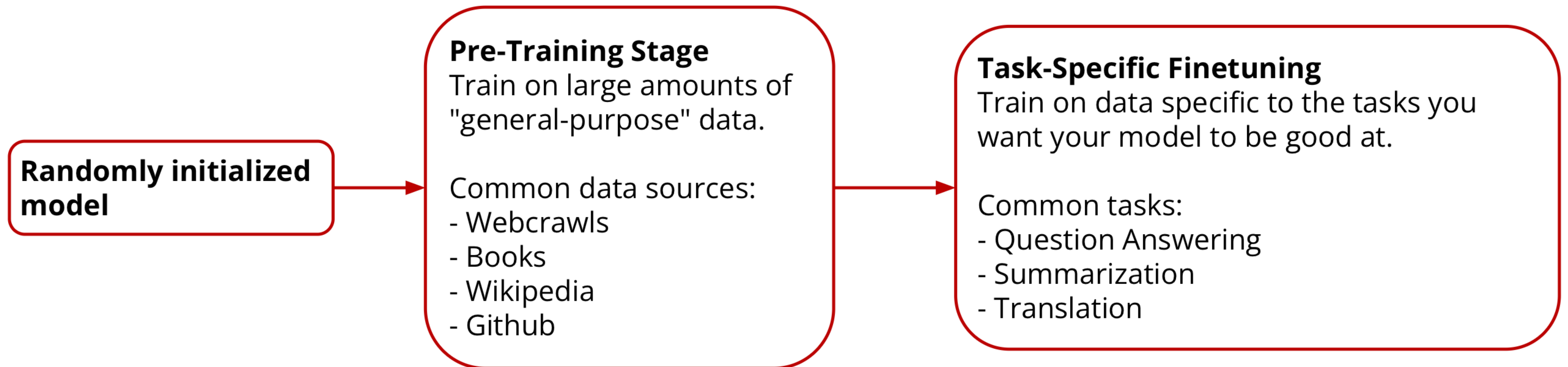
Parameter-Efficient Tuning and Evaluation

Junxian He

Oct 3, 2025

Review: Pretraining -> Fine-Tuning

Paradigm shift around 2018



Review: 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)	-	-	<u>89.3</u>	-	-	-
CAFE [58] (5x)	80.2	79.0	<u>89.3</u>	-	-	-
Stochastic Answer Network [35] (3x)	<u>80.6</u>	<u>80.1</u>	-	-	-	-
CAFE [58]	78.7	77.9	88.5	<u>83.3</u>		
GenSen [64]	71.4	71.3	-	-	<u>82.3</u>	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

Review: 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

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

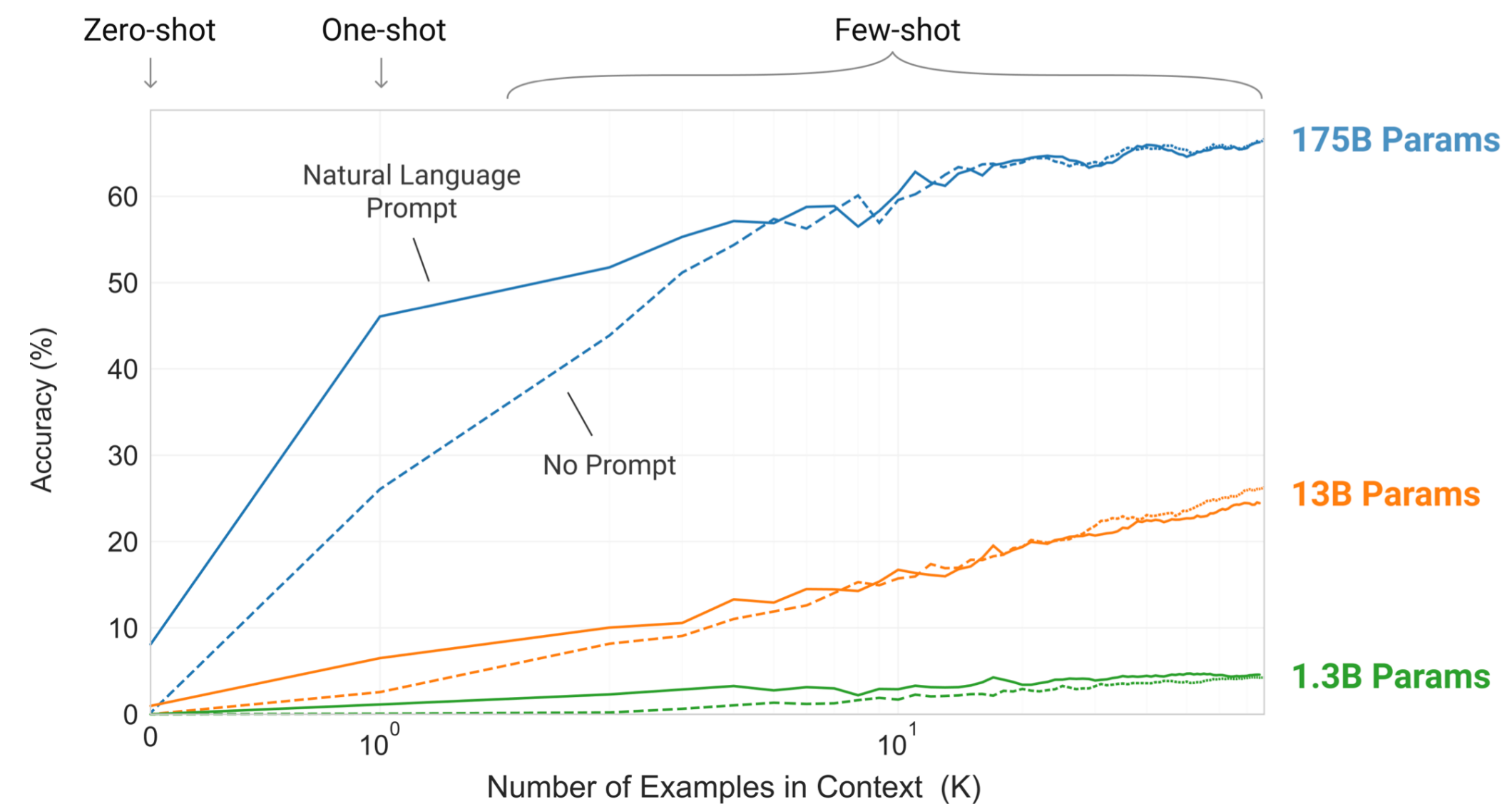
In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

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

Few-shot

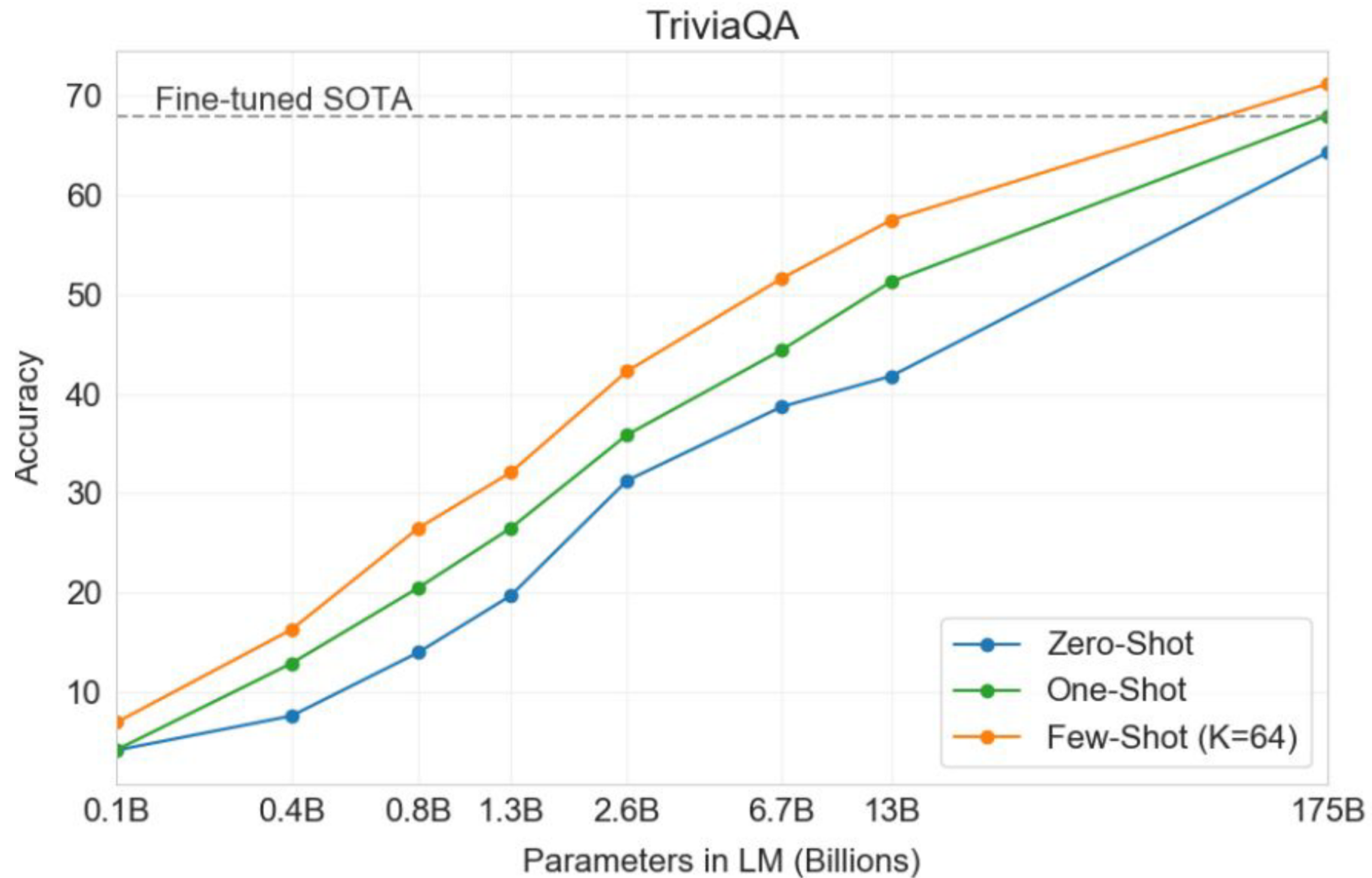
In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

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



In-Context Learning

Language Models Are Few-Shot Learners



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.

Essentially, In-context Learning vs Fine-tuning?

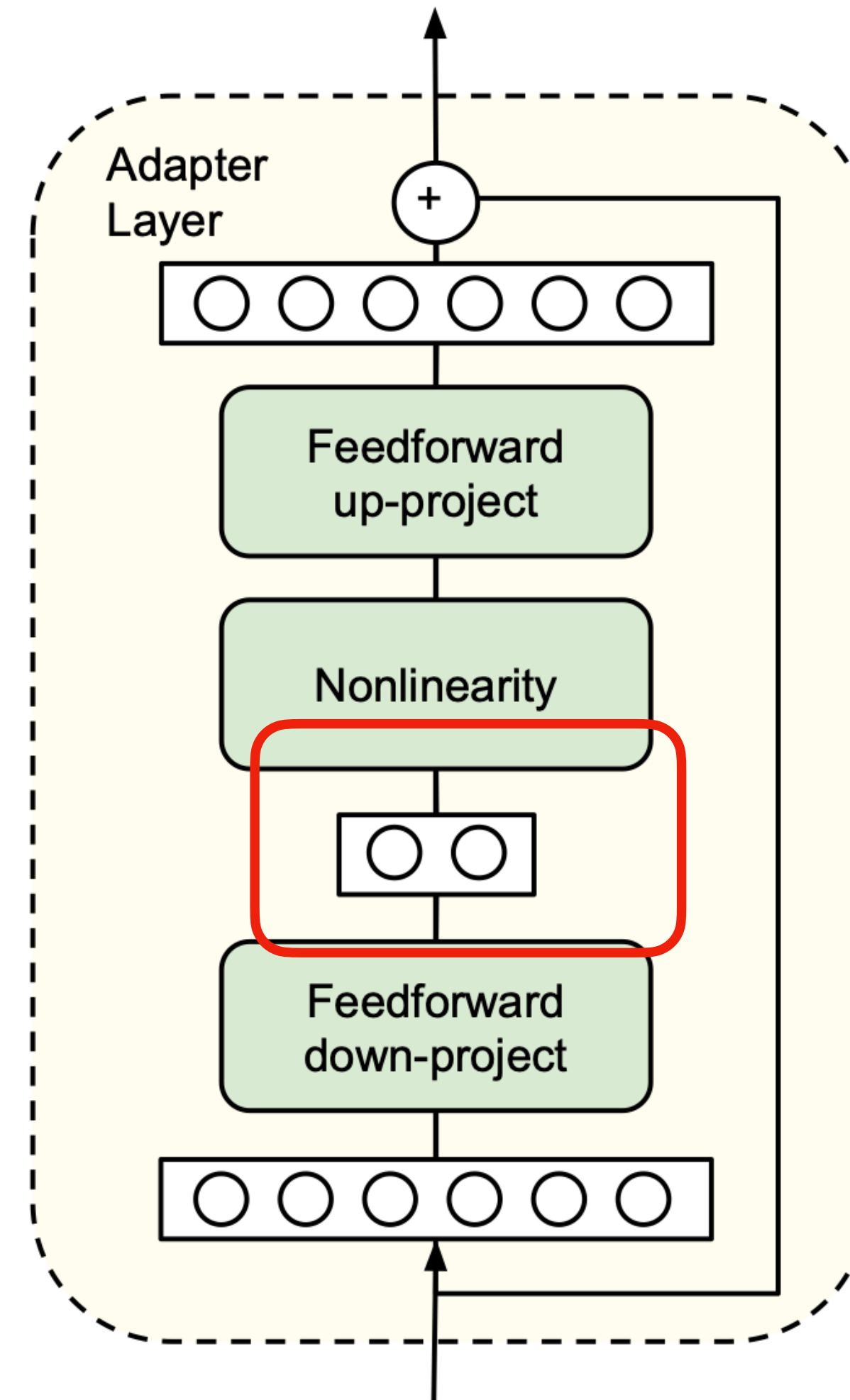
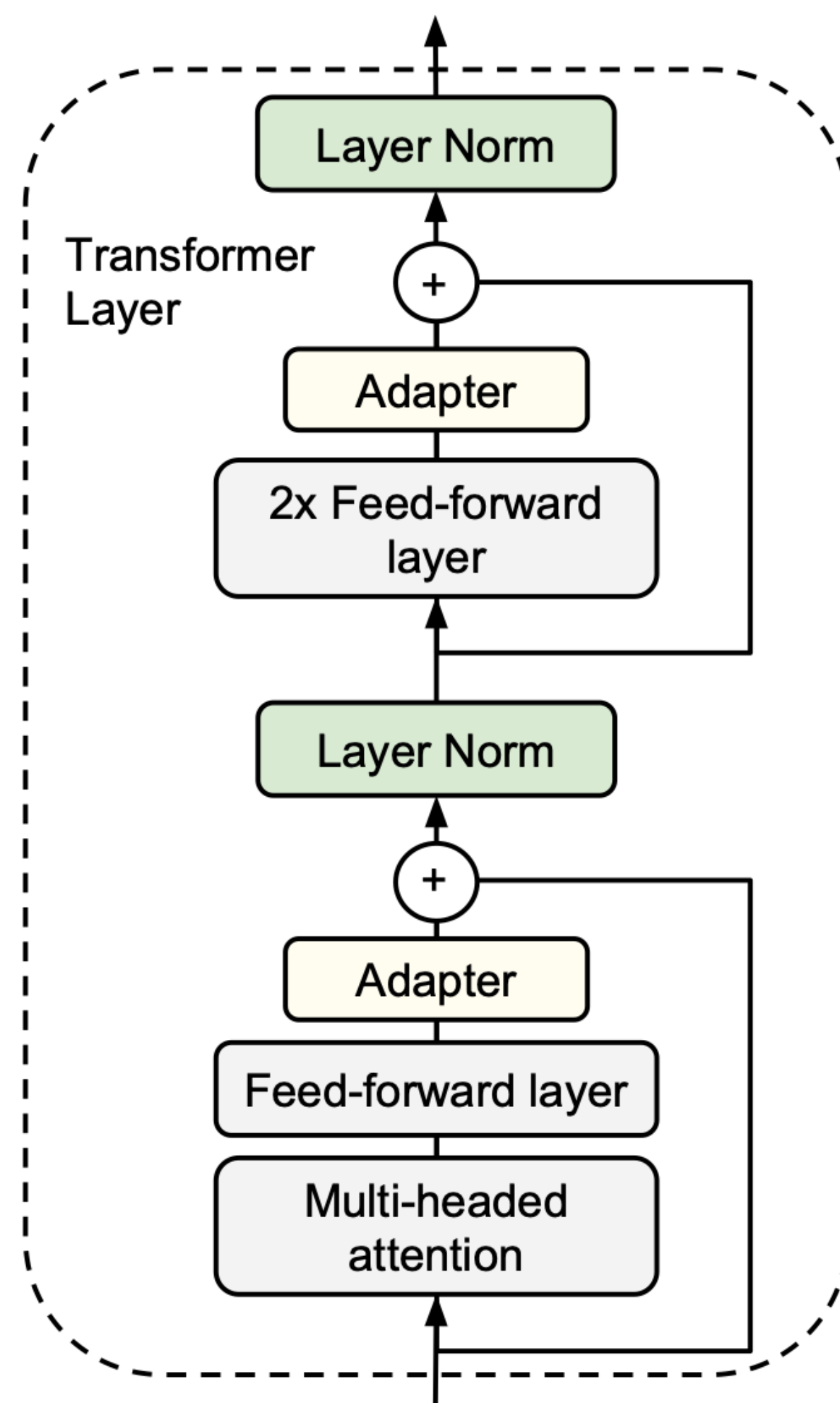
They are different ways of utilizing “annotated data”

Parameter-Efficient Fine-Tuning

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

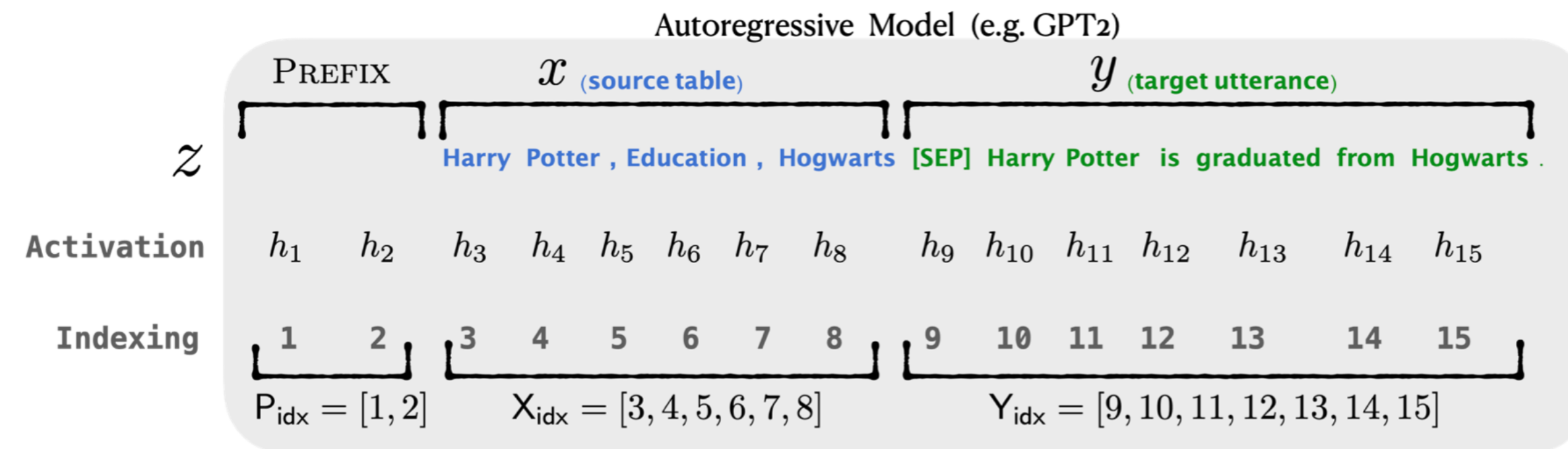
Storage savings

Adapter



Low-Rank

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.

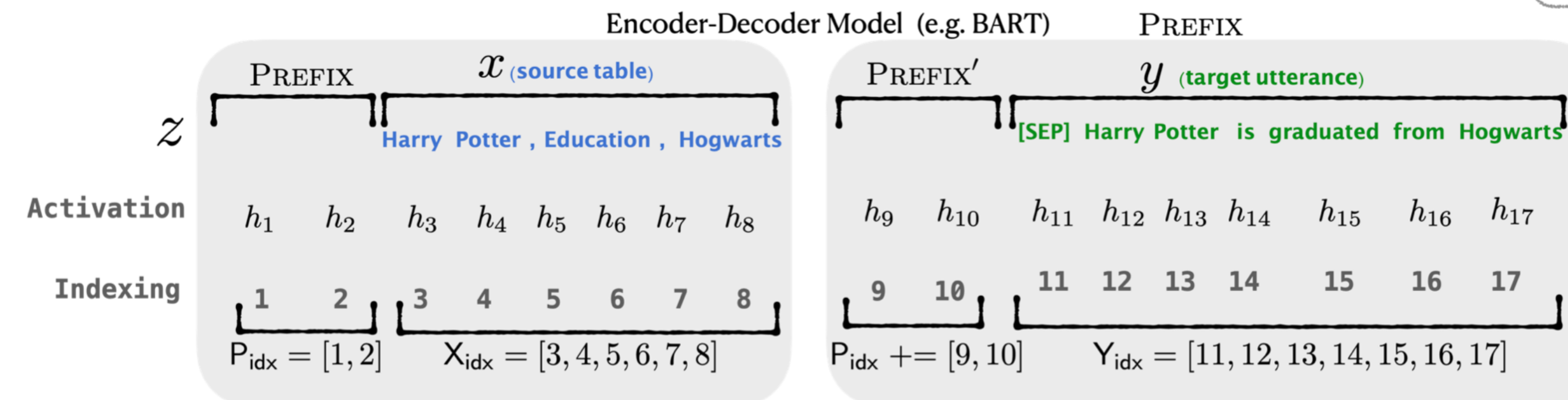


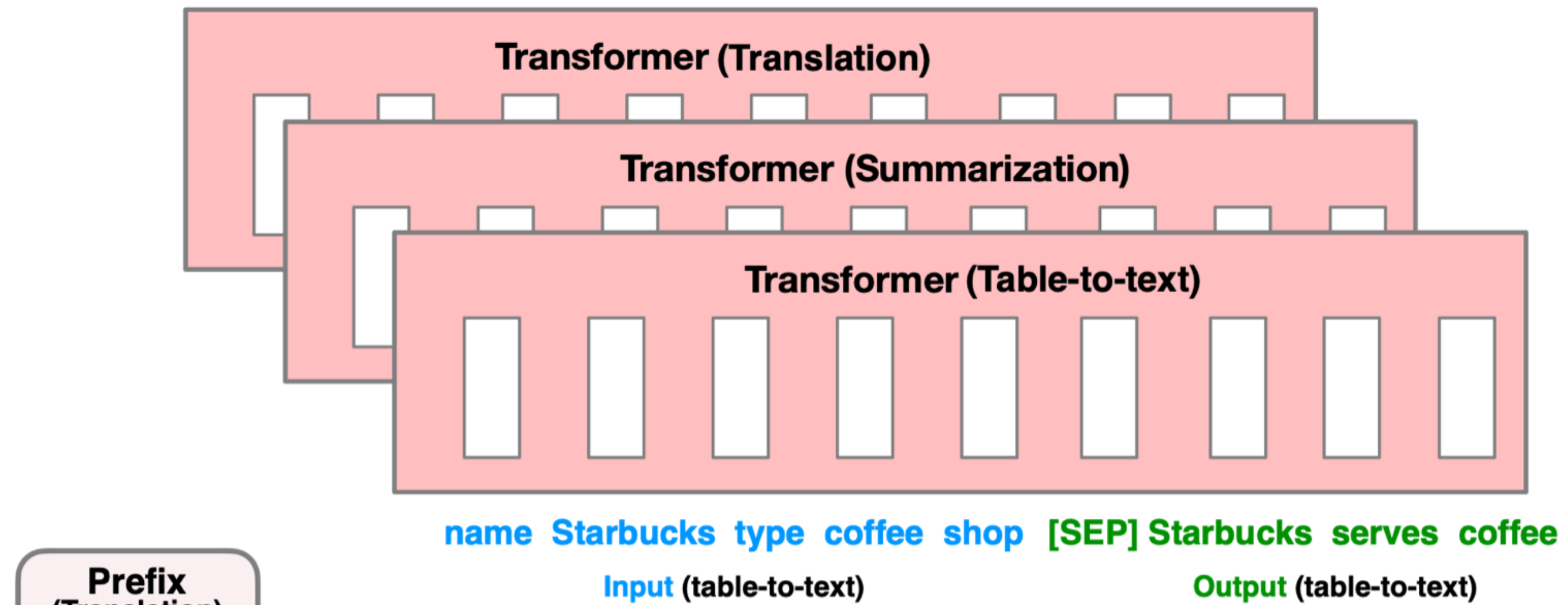
Table-to-text Example

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

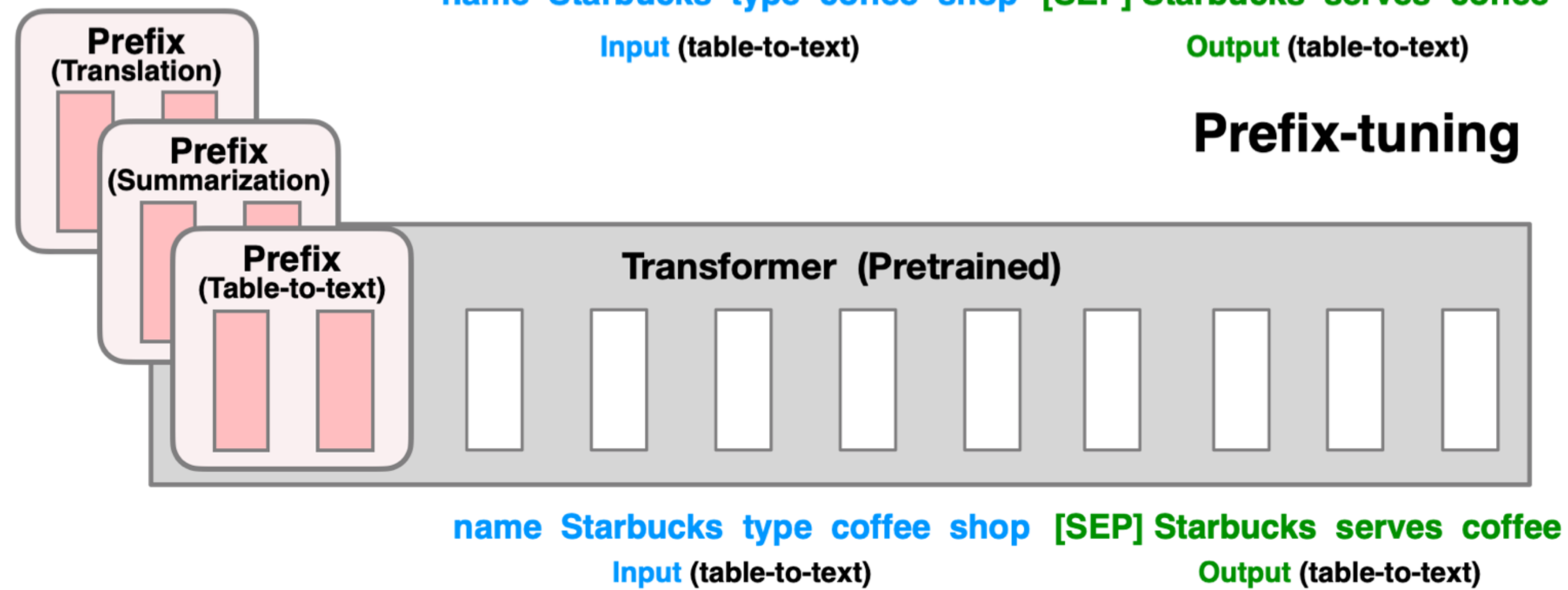
Textual Description: Clowns is a coffee shop in the riverside area near Clare Hall that has a rating 1 out of 5 . They serve Chinese food .

Prefix-Tuning

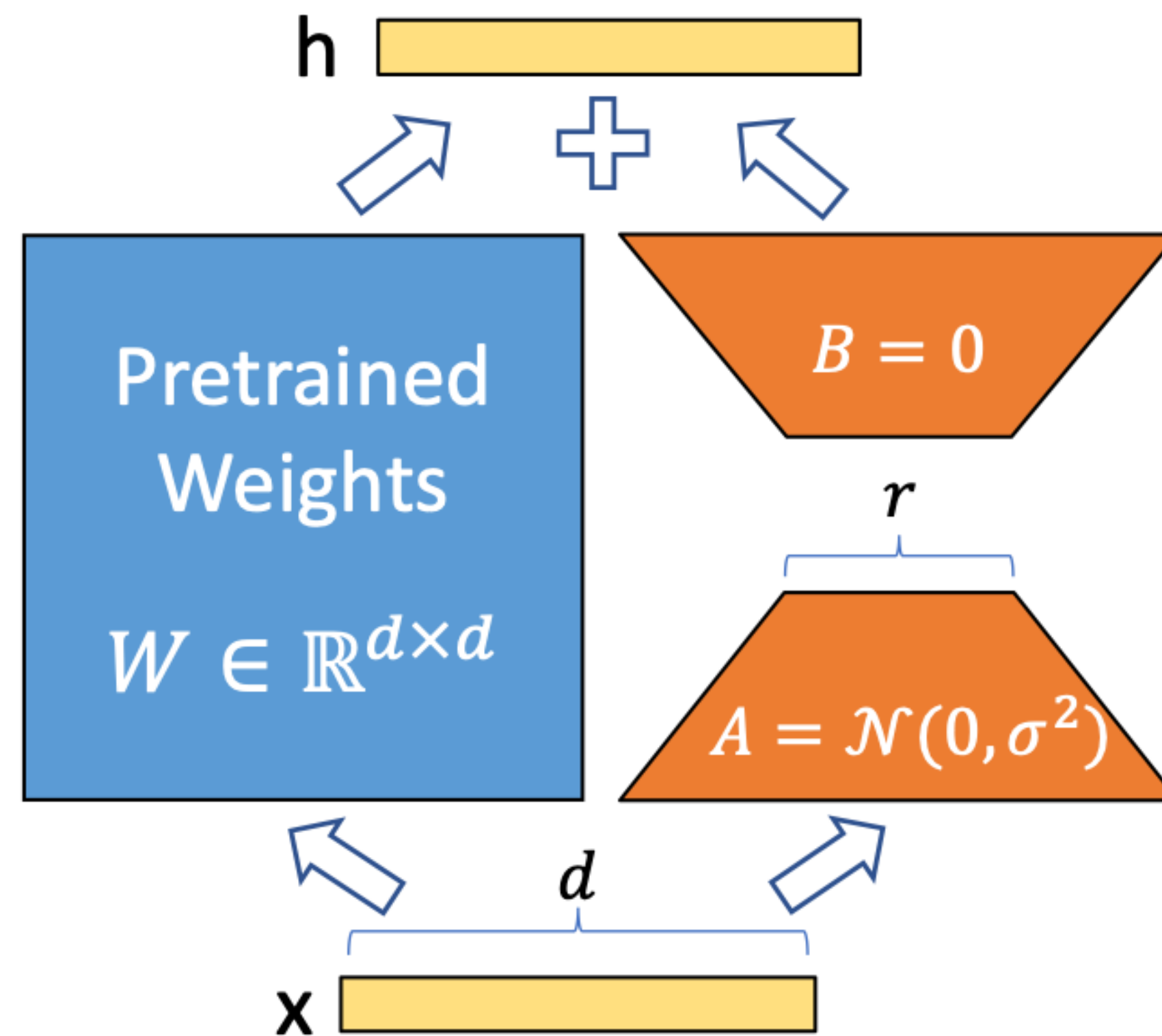
Fine-tuning



Prefix-tuning



LORA: LOW-RANK ADAPTATION



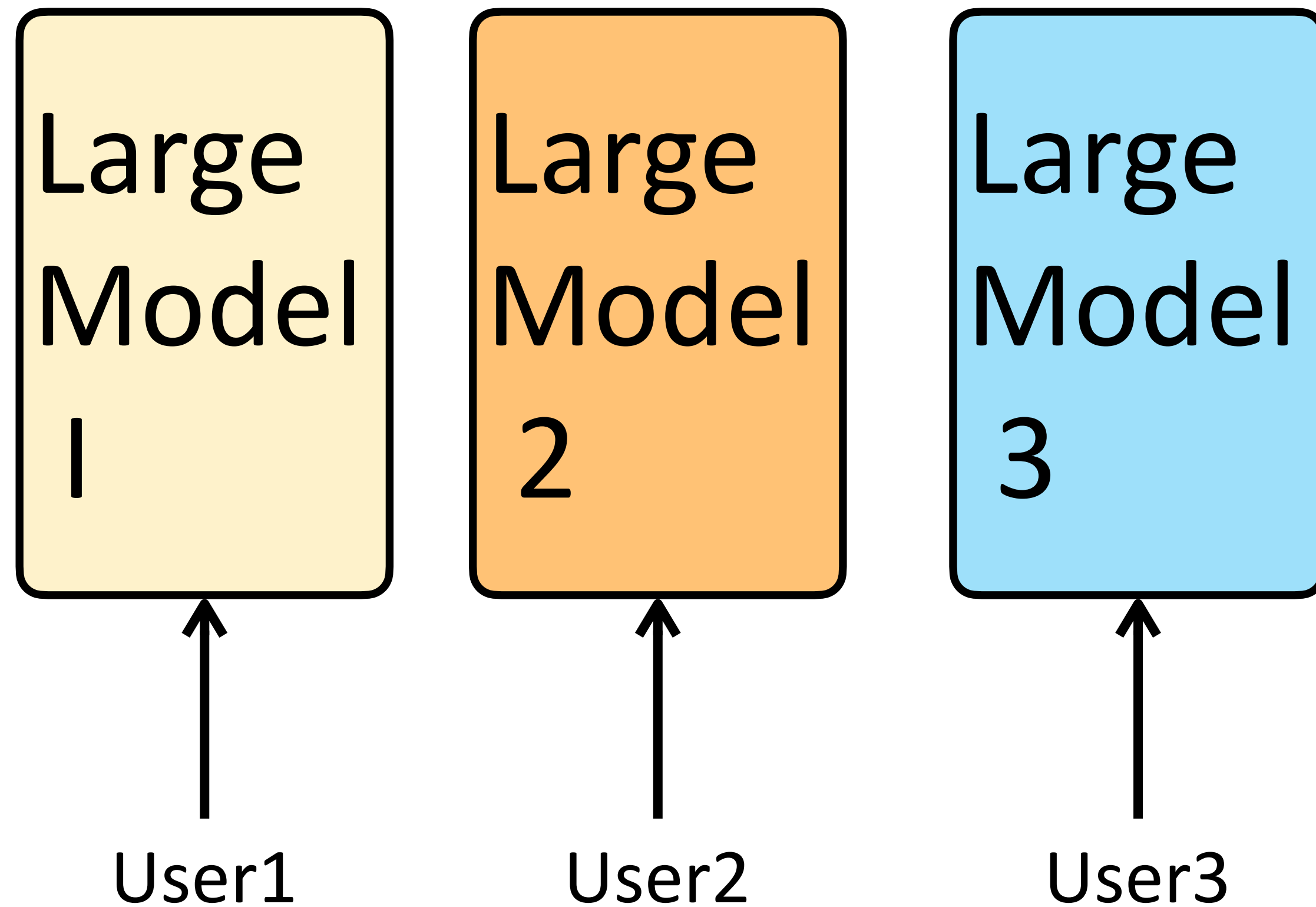
LORA: LOW-RANK ADAPTATION

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 (Adapter ^L)*	0.37M	66.3	8.41	45.0	69.8	2.40
GPT-2 M (Adapter ^L)*	11.09M	68.9	8.71	46.1	71.3	2.47
GPT-2 M (Adapter ^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	70.4$\pm$.1	8.85$\pm$.02	46.8$\pm$.2	71.8$\pm$.1	2.53$\pm$.02
GPT-2 L (FT)*	774.03M	68.5	8.78	46.0	69.9	2.45
GPT-2 L (Adapter ^L)	0.88M	69.1 $\pm$.1	8.68 $\pm$.03	46.3 $\pm$.0	71.4 $\pm$.2	2.49$\pm$.0
GPT-2 L (Adapter ^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	70.4$\pm$.1	8.89$\pm$.02	46.8$\pm$.2	72.0$\pm$.2	2.47 $\pm$.02

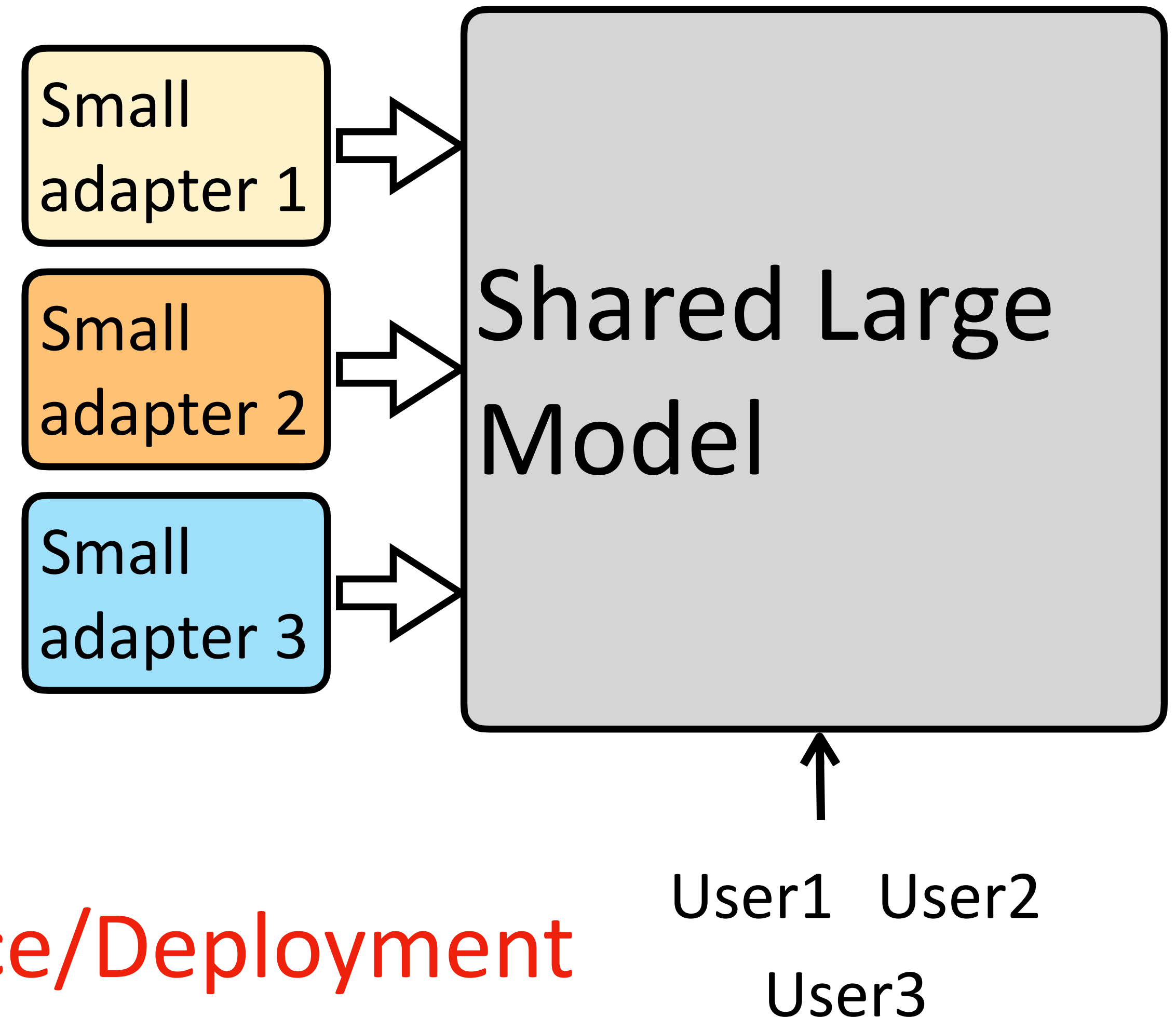
Table 3: GPT-2 medium (M) and large (L) with different adaptation methods on the E2E NLG Challenge. For all metrics, higher is better. LoRA outperforms several baselines with comparable or fewer trainable parameters. Confidence intervals are shown for experiments we ran. * indicates numbers published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

Why Parameter-Efficient Tuning

Before



After

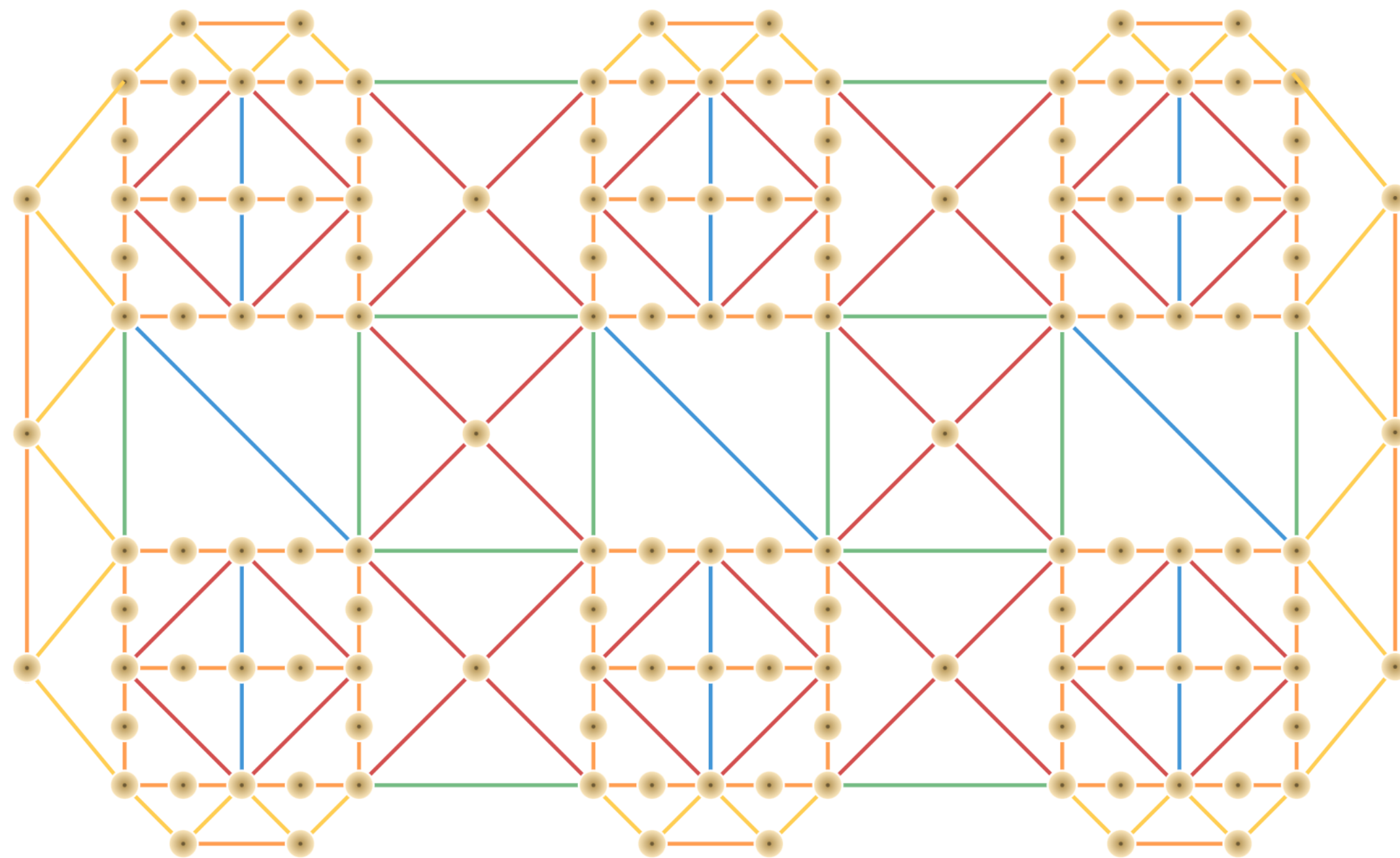


Hardware-Efficient Inference/Deployment

Announcing Tinker

Thinking Machines Lab

Oct 1, 2025



TinkerToy Computer invented by Daniel Hillis and Brian Silverman

Fine-Tuning as Service

Today, we are launching Tinker, a flexible API for fine-tuning language models. It empowers researchers and hackers to experiment with models by giving them control over the algorithms and data while we handle the complexity of distributed training. Tinker advances our mission of enabling more people to do research on cutting-edge models and customize them to their needs.



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

Evaluation of LLMs

**Evaluation needs to come first
before development starts**

How to Evaluate a Language Model

Loss on validation/test set

$$\log p(x)$$

Average over number of tokens

Perplexity

$$e^{\frac{-\log p(x)}{N}}$$

N: number of tokens

Evaluating for Downstream Tasks

Chat:

Which response do you prefer?
Your choice will help make ChatGPT better.

 Response 1

Thank you for providing the detailed error message and the latest code. The error appears to be related to how data is handled within the `collate_fn` function. Let's address this by correcting the `collate_fn` to properly handle the data format coming from the `GPT2Dataset`.

- Update to `collate_fn` Function:** The function should directly handle the list of tensors received from `GPT2Dataset` without trying to extract `input_ids` and `attention_mask` as if the data were in a dictionary format.
- Remove Diagnostics Prints:** The diagnostic print statements within `collate_fn` are not necessary anymore and can be removed to clean up the function.

Here's the updated `collate_fn`:

```
python Copy code
```

```
def collate_fn(data):  
    tokenizer = GPT2Tokenizer.from_pretrained('gpt2')  
    input_ids = pad_sequence(data, batch_first=True,  
                             attention_masks = torch.ones_like(input_ids)  
  
    return input_ids, attention_masks
```

Make sure to replace the existing `collate_fn` function in your script with this updated version. The `pad_sequence` function is used to ensure all sequences in a batch have the same length, and the `attention_masks` are created to be the same shape as `input_ids`, with all elements set to 1 (as all tokens are

 Response 2

Thank you for providing the detailed information and the current state of your code. It seems the issue lies in the `collate_fn` function. Specifically, the way `input_ids` and `attention_masks` are being extracted and processed.

In your `GPT2Dataset` class, each item in the `samples` list is a tensor representing tokenized input IDs. The `collate_fn` function should correctly handle these tensors. Here's the revised `collate_fn` and how it should be integrated:

- Update the `collate_fn` function:
 - Since each item in `data` is a tensor (not a dictionary), we directly work with these tensors.
 - We use `pad_sequence` to pad all tensors to the same length.
 - For attention masks, we create a mask that has the same shape as `input_ids` and has 1s where `input_ids` are not padding.

```
python Copy code
```

```
def collate_fn(data):  
    tokenizer = GPT2Tokenizer.from_pretrained('gpt2')  
    input_ids = pad_sequence(data, batch_first=True,  
                             attention_masks = (input_ids != tokenizer.pad_tok  
  
    return input_ids, attention_masks
```

- Make sure to use this `collate_fn` in your `DataLoader`:

Which one is better?

Evaluating for Downstream Tasks

Chat:

The screenshot shows the ChatGPT web interface. On the left is a sidebar with a list of chat topics: 'New chat', 'Digital Marketing Strate', 'Ingredients For Pancakes', 'New chat', 'Chocolate Pancake Recipe.', 'Popular New York City Attracti', 'General Greeting Help', 'Math book joke summary', '10 World Facts Summary', 'Clear conversations', 'Dark mode', 'OpenAI Discord', 'Updates & FAQ', and 'Log out'. The main area displays a prompt: 'Points to mention during a speech on digital marketing'. The response is a numbered list of 10 points, starting with 'The importance of having a strong online presence in today's digital age.' and ending with 'The importance of testing and experimenting with different digital marketing tactics to determine what works best for your business.' Below the response is a 'Regenerate response' button and a text input field. At the bottom, a footer note reads: 'ChatGPT Jan 9 Version. Free Research Preview. Our goal is to make AI systems more natural and safe to interact with. Your feedback will help us improve.'

+ New chat

Digital Marketing Strate

Ingredients For Pancakes

New chat

Chocolate Pancake Recipe.

Popular New York City Attracti

General Greeting Help

Math book joke summary

10 World Facts Summary

Clear conversations

Dark mode

OpenAI Discord

Updates & FAQ

Log out

Points to mention during a speech on digital marketing

1. The importance of having a strong online presence in today's digital age.
2. The various digital marketing channels available, such as social media, email, search engine optimization (SEO), pay-per-click (PPC) advertising, and content marketing.
3. The benefits of using data and analytics to track the performance of your digital marketing campaigns and make informed decisions.
4. The need to regularly update and adapt your digital marketing strategy to stay ahead of the competition and meet the changing needs of your target audience.
5. The role of customer engagement and building relationships through digital channels.
6. The importance of mobile optimization in digital marketing.
7. The potential of emerging technologies such as artificial intelligence and virtual reality in the field of digital marketing.
8. The importance of compliance with laws and regulations related to digital marketing, such as the General Data Protection Regulation (GDPR) and the Children's Online Privacy Protection Act (COPPA).
9. The importance of creating a consistent brand image across all digital channels.
10. The importance of testing and experimenting with different digital marketing tactics to determine what works best for your business.

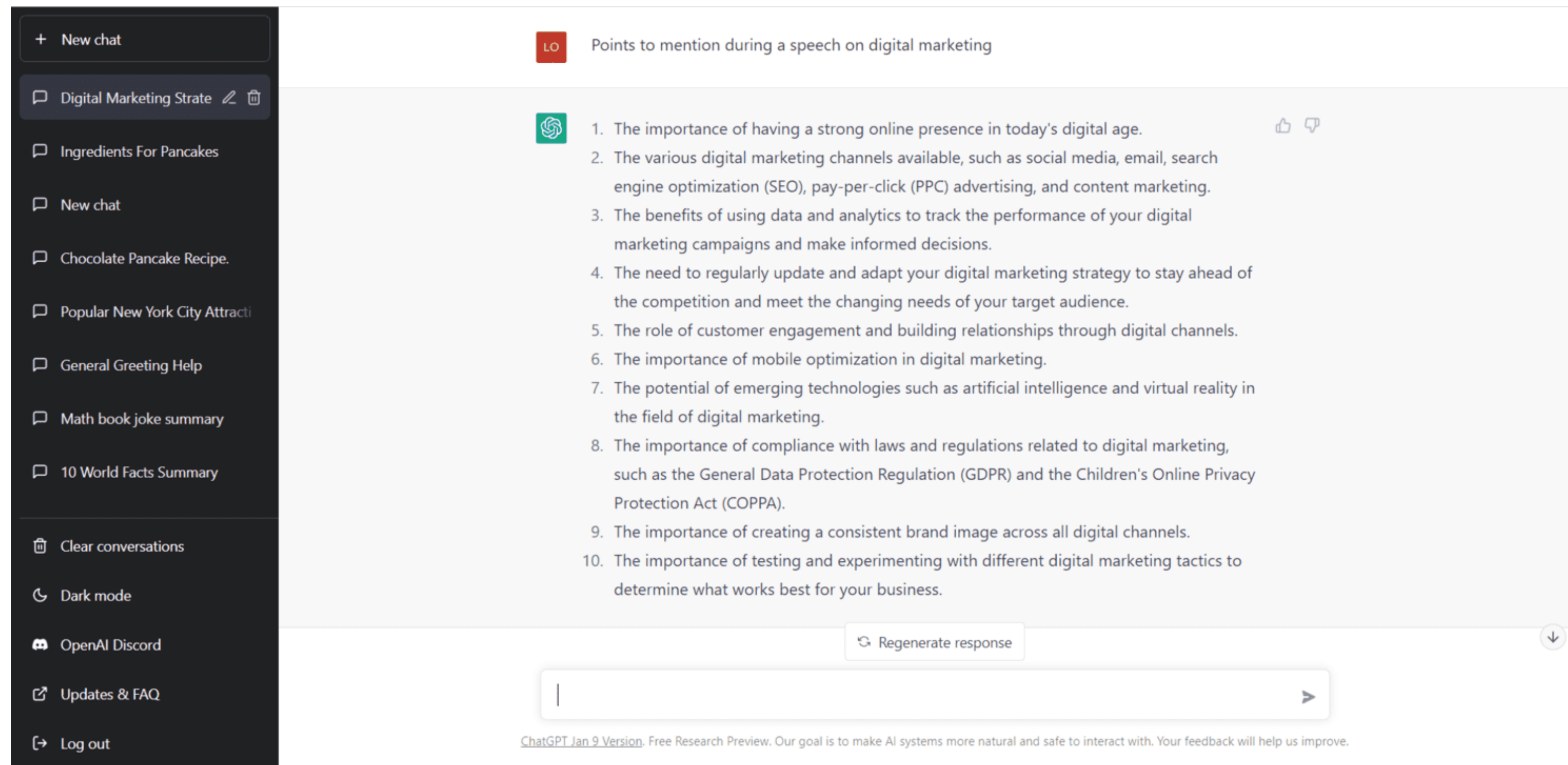
Regenerate response

ChatGPT Jan 9 Version. Free Research Preview. Our goal is to make AI systems more natural and safe to interact with. Your feedback will help us improve.

What is the quality of this response? It it helpful, factually correct, non-toxic?

Human Evaluation

Evaluating long-form generated text is hard and an old problem in NLP for a long time



The screenshot shows the ChatGPT web interface. On the left is a dark sidebar with a list of chat topics: 'New chat', 'Digital Marketing Strate', 'Ingredients For Pancakes', 'New chat', 'Chocolate Pancake Recipe.', 'Popular New York City Attracti', 'General Greeting Help', 'Math book joke summary', '10 World Facts Summary', 'Clear conversations', 'Dark mode', 'OpenAI Discord', 'Updates & FAQ', and 'Log out'. The main area displays a prompt 'Points to mention during a speech on digital marketing' and a response from GPT-4 listing 10 points. At the bottom, there is a 'Regenerate response' button and a text input field. A footer note reads: 'ChatGPT Jan 9 Version. Free Research Preview. Our goal is to make AI systems more natural and safe to interact with. Your feedback will help us improve.'

Points to mention during a speech on digital marketing

1. The importance of having a strong online presence in today's digital age.
2. The various digital marketing channels available, such as social media, email, search engine optimization (SEO), pay-per-click (PPC) advertising, and content marketing.
3. The benefits of using data and analytics to track the performance of your digital marketing campaigns and make informed decisions.
4. The need to regularly update and adapt your digital marketing strategy to stay ahead of the competition and meet the changing needs of your target audience.
5. The role of customer engagement and building relationships through digital channels.
6. The importance of mobile optimization in digital marketing.
7. The potential of emerging technologies such as artificial intelligence and virtual reality in the field of digital marketing.
8. The importance of compliance with laws and regulations related to digital marketing, such as the General Data Protection Regulation (GDPR) and the Children's Online Privacy Protection Act (COPPA).
9. The importance of creating a consistent brand image across all digital channels.
10. The importance of testing and experimenting with different digital marketing tactics to determine what works best for your business.

Regenerate response

ChatGPT Jan 9 Version. Free Research Preview. Our goal is to make AI systems more natural and safe to interact with. Your feedback will help us improve.

Q: Is this response helpful?

A. Very helpful; B. Helpful. C. Not useful at all

Why multi-choice questions?

Human Evaluation is Subjective

Comparative human evaluation is more reliable

Which response do you prefer?
Your choice will help make ChatGPT better.

Response 1

Thank you for providing the detailed error message and the latest code. The error appears to be related to how data is handled within the `collate_fn` function. Let's address this by correcting the `collate_fn` to properly handle the data format coming from the `GPT2Dataset`.

- Update to `collate_fn` Function:** The function should directly handle the list of tensors received from `GPT2Dataset` without trying to extract `input_ids` and `attention_mask` as if the data were in a dictionary format.
- Remove Diagnostics Prints:** The diagnostic print statements within `collate_fn` are not necessary anymore and can be removed to clean up the function.

Here's the updated `collate_fn`:

```
python
def collate_fn(data):
    tokenizer = GPT2Tokenizer.from_pretrained('gpt2')
    input_ids = pad_sequence(data, batch_first=True,
                             attention_masks = torch.ones_like(input_ids)

    return input_ids, attention_masks
```

Make sure to replace the existing `collate_fn` function in your script with this updated version. The `pad_sequence` function is used to ensure all sequences in a batch have the same length, and the `attention_masks` are created to be the same shape as `input_ids`, with all elements set to 1 (as all tokens are

Response 2

Thank you for providing the detailed information and the current state of your code. It seems the issue lies in the `collate_fn` function. Specifically, the way `input_ids` and `attention_masks` are being extracted and processed.

In your `GPT2Dataset` class, each item in the `samples` list is a tensor representing tokenized input IDs. The `collate_fn` function should correctly handle these tensors. Here's the revised `collate_fn` and how it should be integrated:

- Update the `collate_fn` function:
 - Since each item in `data` is a tensor (not a dictionary), we directly work with these tensors.
 - We use `pad_sequence` to pad all tensors to the same length.
 - For attention masks, we create a mask that has the same shape as `input_ids` and has 1s where `input_ids` are not padding.

```
python
def collate_fn(data):
    tokenizer = GPT2Tokenizer.from_pretrained('gpt2')
    input_ids = pad_sequence(data, batch_first=True,
                             attention_masks = (input_ids != tokenizer.pad_t

    return input_ids, attention_masks
```

- Make sure to use this `collate_fn` in your `DataLoader`:

Which one is better?



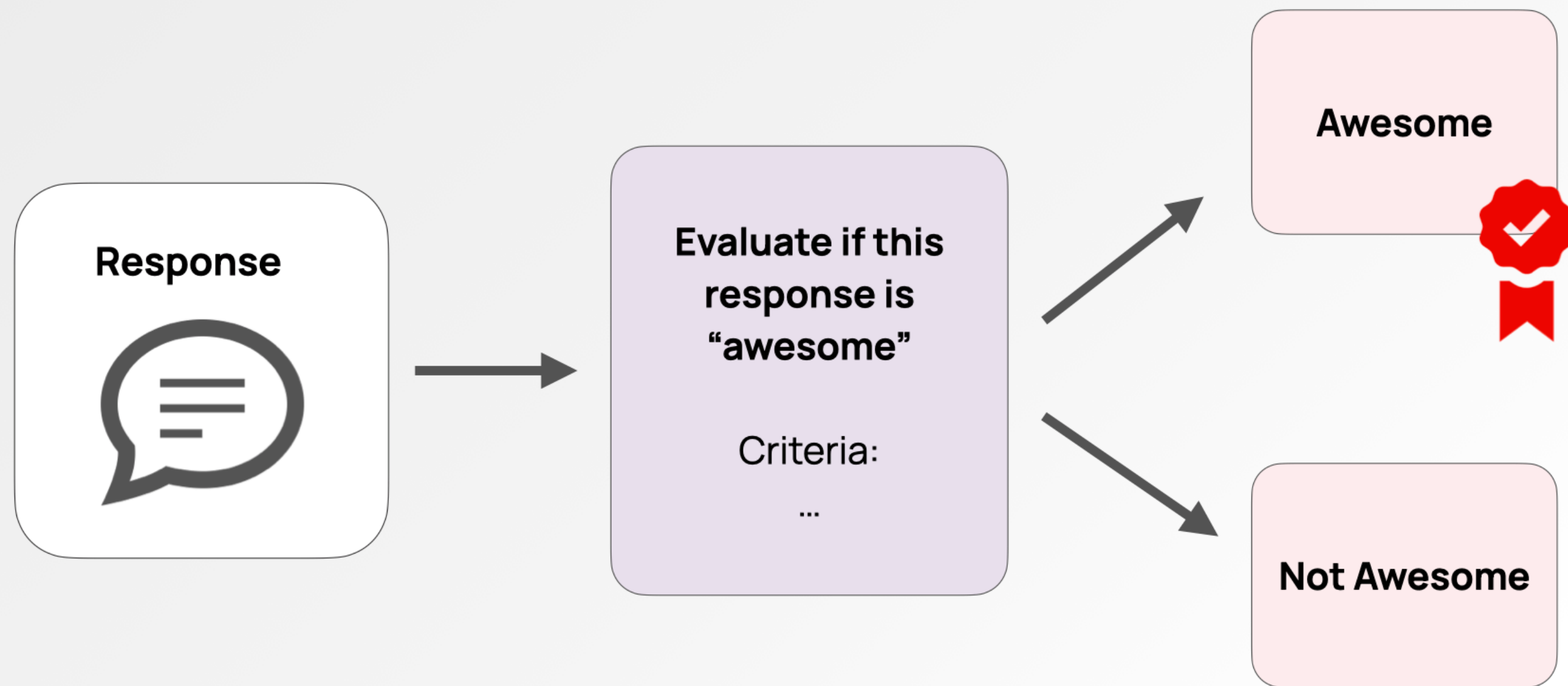
Sometimes Evaluation is Easier than Generation

Example: Humans can judge which essay is better from two model-written essays, but the humans may not generate such essays with similar quality

Humans can still judge AI in some tasks even though AI much stronger than humans

LLM as Judge

Human evaluation is slow, expensive, and not scalable



1. Sometimes people just prompt models for judgement
2. Sometimes people train specific judgement model



LLM as judge may be inaccurate and uninterpretable.

Crowd-Sourcing Human Evaluation

https://lmarena.ai/

Imagine any two models can battle a game, whose response is better

Assistant A

How are you today

I'm doing great, thanks—always ready to chat or help out! How about you?

Assistant B

How are you today

As an AI, I don't experience emotions or have personal feelings, so I don't have a 'how are you' in the human sense.

However, I am fully operational and ready to assist you! How can I help you today?

🏆..Overall

Search by model name...

Default

Rank (UB) ↑	Model ↑↓	Score ↑↓	95% CI (±) ↑↓	Votes ↑↓	Organization ↑↓	Licens
1	 gemini-2.5-pro	1452	±5	50,309	Google	Propr
1	 claude-opus-4-1-20250805-thinking-16k	1449	±6	18,383	Anthropic	Propr
1	 claude-sonnet-4-5-20250929	1440	±12	2,504	Anthropic	Propr
2	 chatgpt-4o-latest-20250326	1441	±4	36,600	OpenAI	Propr
2	 gpt-4.5-preview-2025-02-27	1440	±6	14,644	OpenAI	Propr
2	 o3-2025-04-16	1440	±4	48,558	OpenAI	Propr
2	 gpt-5-high	1439	±6	20,426	OpenAI	Propr
2	 claude-opus-4-1-20250805	1439	±5	29,104	Anthropic	Propr
3	 qwen3-max-preview	1434	±6	15,187	Alibaba	Propr

← Left is Better

It's a tie 🤝

Both are bad ☹️

Right is Better →

Crowd-Sourcing Human Evaluation

🏆 ..Overall

▼

🔍

Search by model name...

/

Default

▼

Rank (UB) ↑	Model ↑↓	Score ↑↓	95% CI (±) ↑↓	Votes ↑↓	Organization ↑↓	License ↑↓
1	 gemini-2.5-pro	1452	±5	50,309	Google	Proprietary
1	 claude-opus-4-1-20250805-thinking-16k	1449	±6	18,383	Anthropic	Proprietary
1	 claude-sonnet-4-5-20250929	1440	±12	2,504	Anthropic	Proprietary
2	 chatgpt-4o-latest-20250326	1441	±4	36,600	OpenAI	Proprietary
2	 gpt-4.5-preview-2025-02-27	1440	±6	14,644	OpenAI	Proprietary
2	 o3-2025-04-16	1440	±4	48,558	OpenAI	Proprietary
2	 gpt-5-high	1439	±6	20,426	OpenAI	Proprietary
2	 claude-opus-4-1-20250805	1439	±5	29,104	Anthropic	Proprietary
3	 qwen3-max-preview	1434	±6	15,187	Alibaba	Proprietary

Elo scores as in sports

Evaluate Language Model Knowledge

Automatic evaluations typically seek for objective metrics

Example from MMLU:

Microeconomics

One of the reasons that the government discourages and regulates monopolies is that

(A) producer surplus is lost and consumer surplus is gained.

(B) monopoly prices ensure productive efficiency but cost society allocative efficiency.

(C) monopoly firms do not engage in significant research and development.

(D) consumer surplus is lost with higher prices and lower levels of output.



Multi-choice QA

Conceptual Physics

When you drop a ball from rest it accelerates downward at 9.8 m/s^2 . If you instead throw it downward assuming no air resistance its acceleration immediately after leaving your hand is

(A) 9.8 m/s^2

(B) more than 9.8 m/s^2

(C) less than 9.8 m/s^2

(D) Cannot say unless the speed of throw is given.



Ranging from middle school to college level

Mathematical Reasoning

Example from GSM8K:

Problem: Beth bakes 4, 2 dozen batches of cookies in a week. If these cookies are shared amongst 16 people equally, how many cookies does each person consume? Solution: Beth bakes 4 2 dozen batches of cookies for a total of $4 \times 2 = 8$ dozen cookies There are 12 cookies in a dozen and she makes 8 dozen cookies for a total of $12 \times 8 = 96$ cookies She splits the 96 cookies equally amongst 16 people so they each eat $96 / 16 = 6$ cookies Final Answer: 6	
Problem: Mrs. Lim milks her cows twice a day. Yesterday morning, she got 68 gallons of milk and in the evening, she got 82 gallons. This morning, she got 18 gallons fewer than she had yesterday morning. After selling some gallons of milk in the afternoon, Mrs. Lim has only 24 gallons left. How much was her revenue for the milk if each gallon costs \$3.50? Mrs. Lim got 68 gallons - 18 gallons = 50 gallons this morning. So she was able to get a total of 68 gallons + 82 gallons + 50 gallons = 200 gallons. She was able to sell 200 gallons - 24 gallons = 176 gallons. Thus, her total revenue for the milk is \$3.50/gallon x 176 gallons = \$616. Final Answer: 616	
Problem: Tina buys 3 12-packs of soda for a party. Including Tina, 6 people are at the party. Half of the people at the party have 3 sodas each, 2 of the people have 4, and 1 person has 5. How many sodas are left over when the party is over? Solution: Tina buys 3 12-packs of soda, for $3 \times 12 = 36$ sodas 6 people attend the party, so half of them is $6 / 2 = 3$ people Each of those people drinks 3 sodas, so they drink $3 \times 3 = 9$ sodas Two people drink 4 sodas, which means they drink $2 \times 4 = 8$ sodas With one person drinking 5, that brings the total drank to $5 + 9 + 8 + 3 = 25$ sodas As Tina started off with 36 sodas, that means there are $36 - 25 = 11$ sodas left Final Answer: 11	

Short-answer

Mathematical Reasoning

Example from AIME (American Invitational Mathematics Examination):

Problem

Let the sequence of rationals $x_1, x_2, \dots$ be defined such that $x_1 = \frac{25}{11}$ and

$$x_{k+1} = \frac{1}{3} \left(x_k + \frac{1}{x_k} - 1 \right).$$

x_{2025} can be expressed as $\frac{m}{n}$ for relatively prime positive integers m and n . Find the remainder when $m + n$ is divided by 1000.

Solution 1 (complete)

This problem can be split into three parts, listed below:

Part 1: Analyzing Fractions

Let $x_k = \frac{a_k}{b_k}$, where a_k, b_k are relatively prime positive integers. First, we analyze the moduli of the problem. Plugging in for x_2 yields

$x_2 = \frac{157}{275}$. Notice that in both x_1 and x_2 , the numerator is equivalent to 1 and the denominator is equivalent to 2 modulus 3. We see that $x_2 = \frac{1}{3} \cdot \frac{(a_1 - b_1)^2 + a_1 b_1}{a_1 b_1}$. Specifically, we know that

$$(a_1 - b_1)^2 + a_1 b_1 \equiv (1 - 2)^2 + 1 \cdot 2 \equiv 0 \pmod{3}$$

Then this is always divisible by 3 for all x_k (it can be shown that for all x_k , we have $a_k \equiv 1 \pmod{3}$ and $b_k \equiv 2 \pmod{3}$ by using mod 9).

Thus, $x_2 = \frac{\frac{1}{3}((a_1 - b_1)^2 + a_1 b_1)}{a_1 b_1}$, and the numerator and denominator of the right-hand side (RHS) correspond to the numerator and denominator of x_2 in simplest form. (To further prove that the top and bottom are relatively prime, consider that a_k and b_k are by definition relatively prime, so $(a_k - b_k)^2$ and $a_k b_k$ share no factors.)

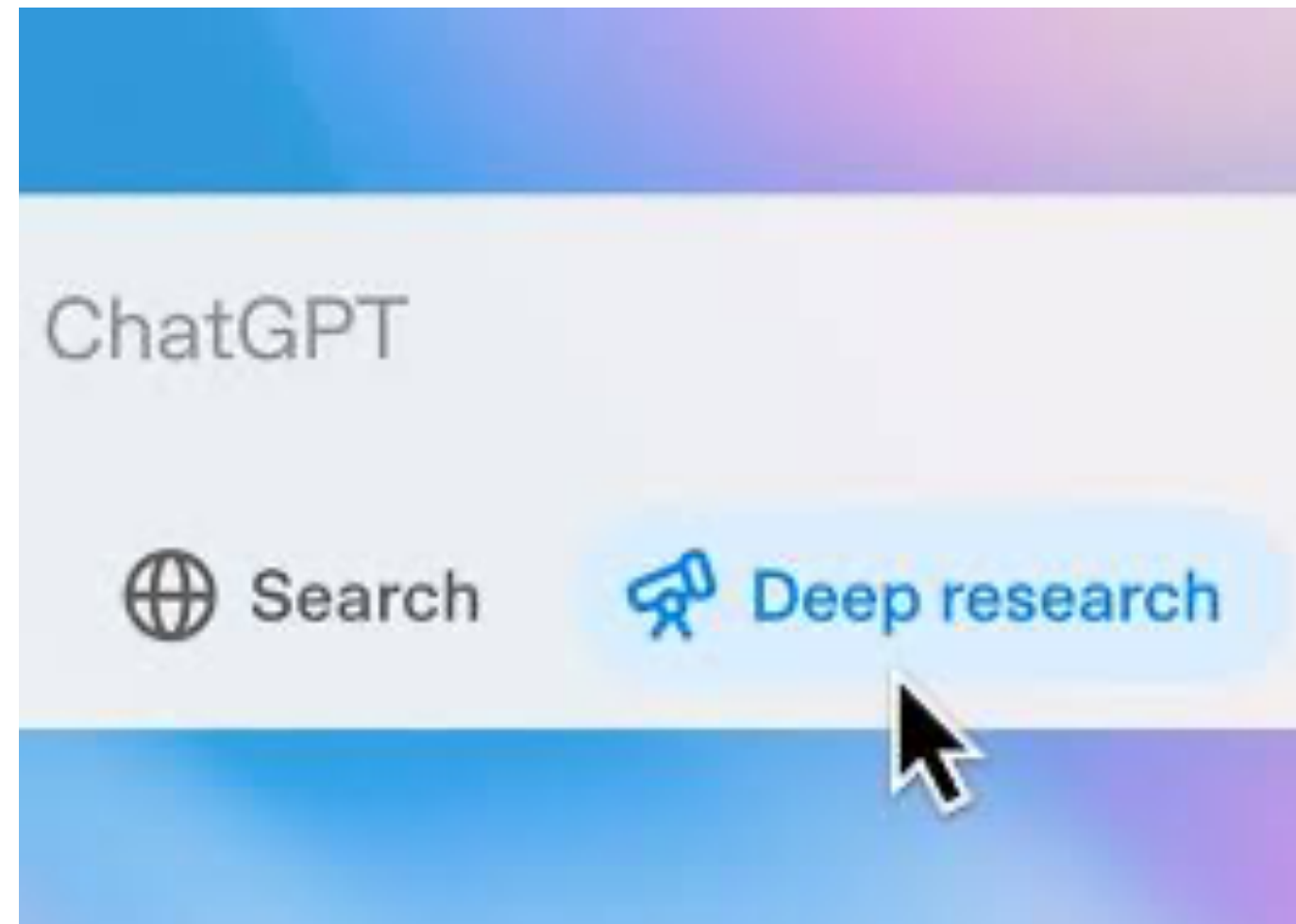
Notice that the above do not just apply to x_1 ; we did not use any specific properties of x_1 . Then we may generalize the above, finding that:

$$a_k = \frac{1}{3}((a_{k-1} - b_{k-1})^2 + a_{k-1} b_{k-1})$$

$$b_k = a_{k-1} b_{k-1}$$

Deep Research

<https://openai.com/index/introducing-deep-research/>



Deep research

Help me find iOS and android adoption rates, % who want to learn another language, and change in mobile penetration, over the past 10 years, for top 10 developed and top 10 developing countries by GDP. Lay this info out in a table and separate stats into columns, and include recommendations on markets to target for a new iOS translation app from ChatGPT, focusing on markets ChatGPT is currently active in.

 **Top 10 Developed Economies (by GDP) – Smartphone OS Share, Language-Learning Interest, and Mobile Penetration Growth**

Country	iOS Market	Android Market	Interest in New Language	Mobile Penetration
---------	------------	----------------	--------------------------	--------------------

Hard to Evaluate

Evaluating Deep Research

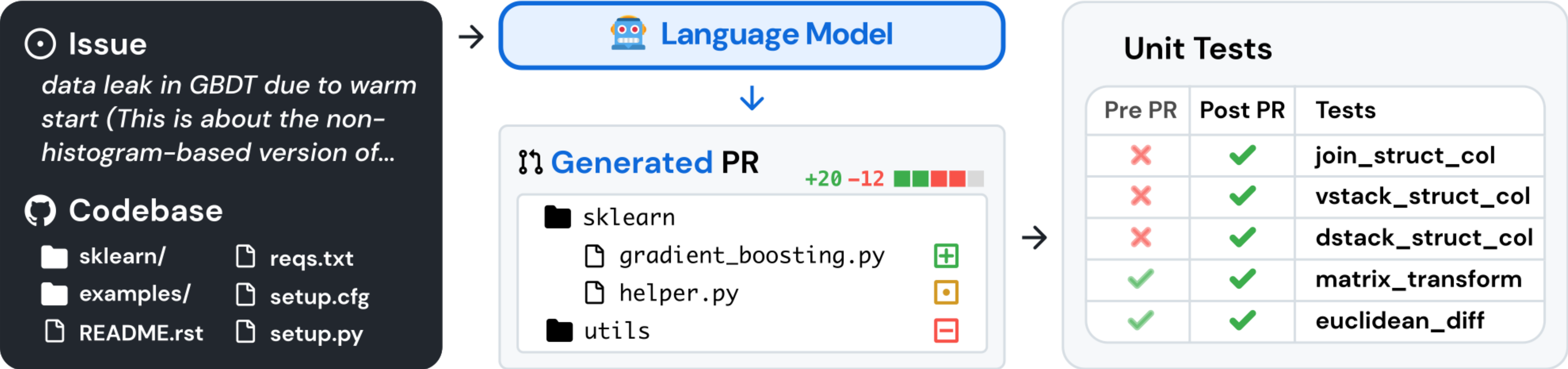
<https://openai.com/index/browsecomp/>

Please identify the fictional character who occasionally breaks the fourth wall with the audience, has a backstory involving help from selfless ascetics, is known for his humor, and had a TV show that aired between the 1960s and 1980s with fewer than 50 episodes.



Answer: Plastic Man

Software Engineering



Benchmarking code generation by running tests

Thank You!