Deep Learning for Vision & Language

Natural Language Processing I: Transformers III

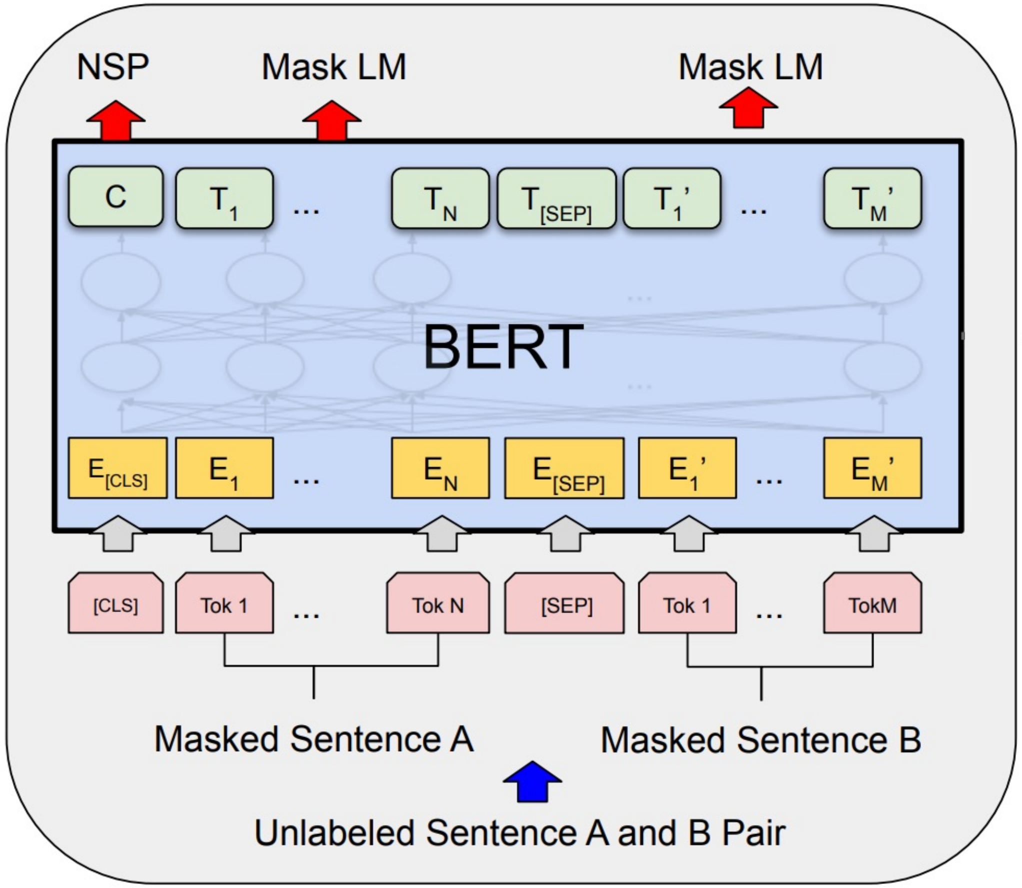
RICE UNIVERSITY

The BERT Encoder Model (October, 2018)

Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding . https://arxiv.org/abs/1810.04805

Important things to know

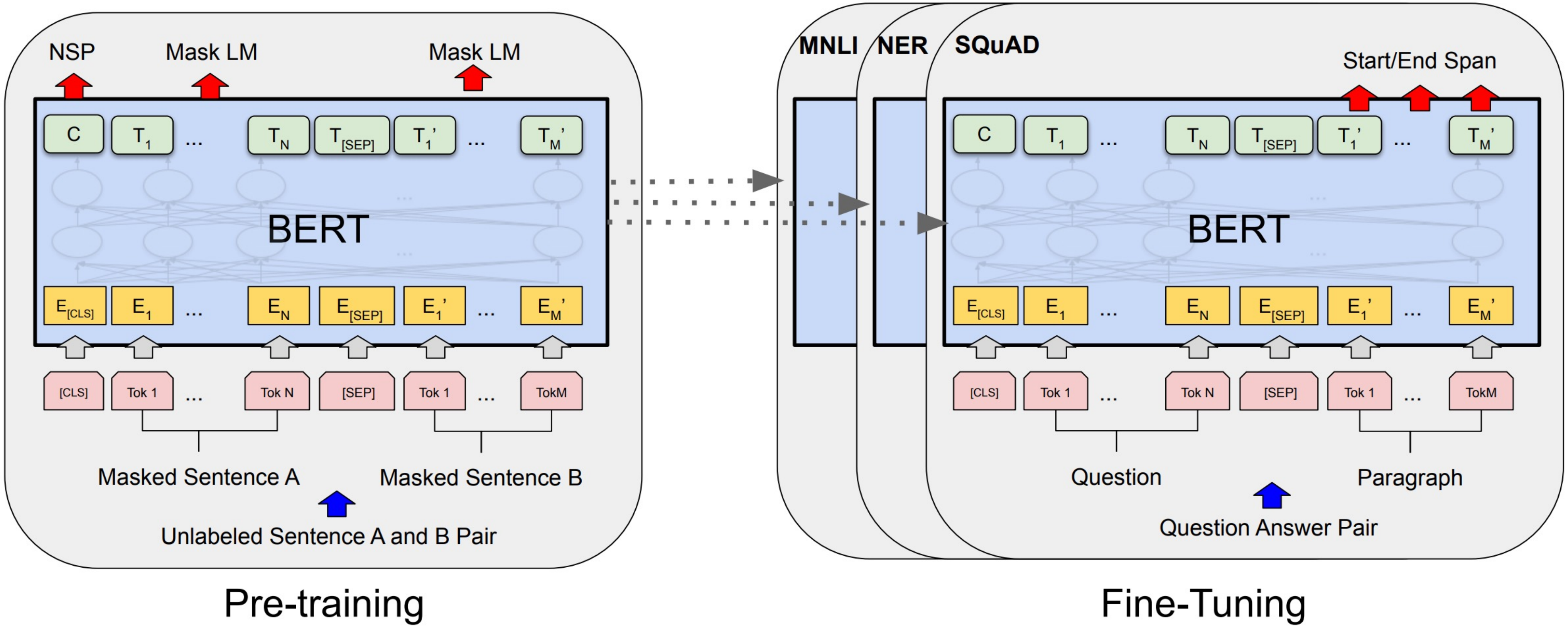
- No decoder
- Train the model to fill-in-the-blank by masking some of the input tokens and trying to recover the full sentence.
- The input is not one sentence but two sentences separated by a [SEP] token.
- Also try to predict whether these two input sentences are consecutive or not.



Pre-training

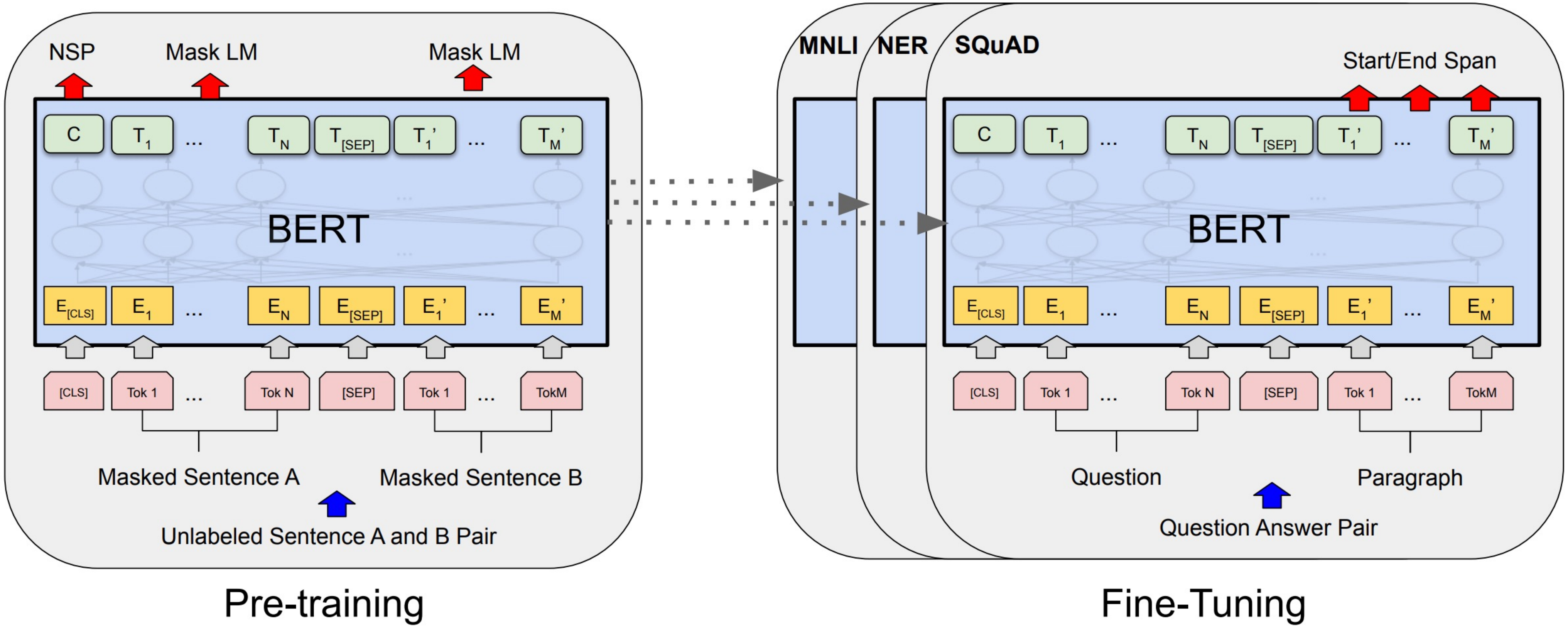
The BERT Encoder Model

Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding . https://arxiv.org/abs/1810.04805



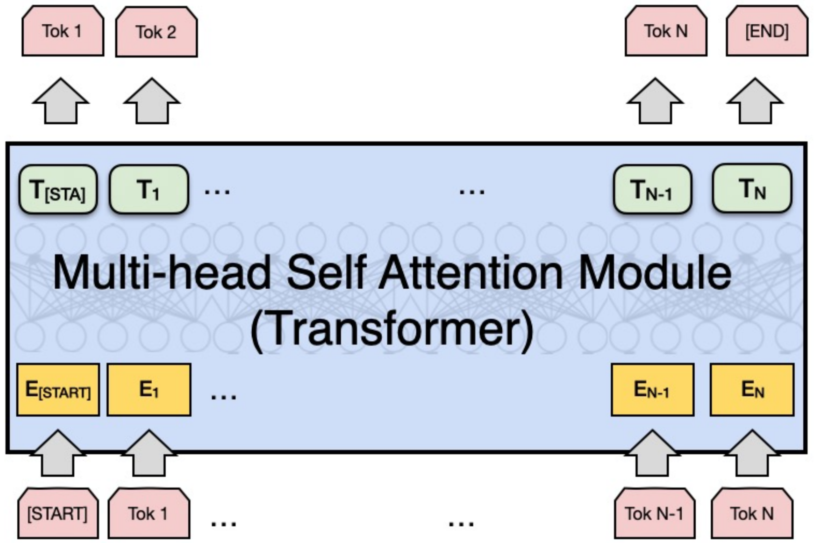
The BERT Encoder-only Model

Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding . https://arxiv.org/abs/1810.04805



The GPT-2, GPT-3 Decoder-only Model

Decoder



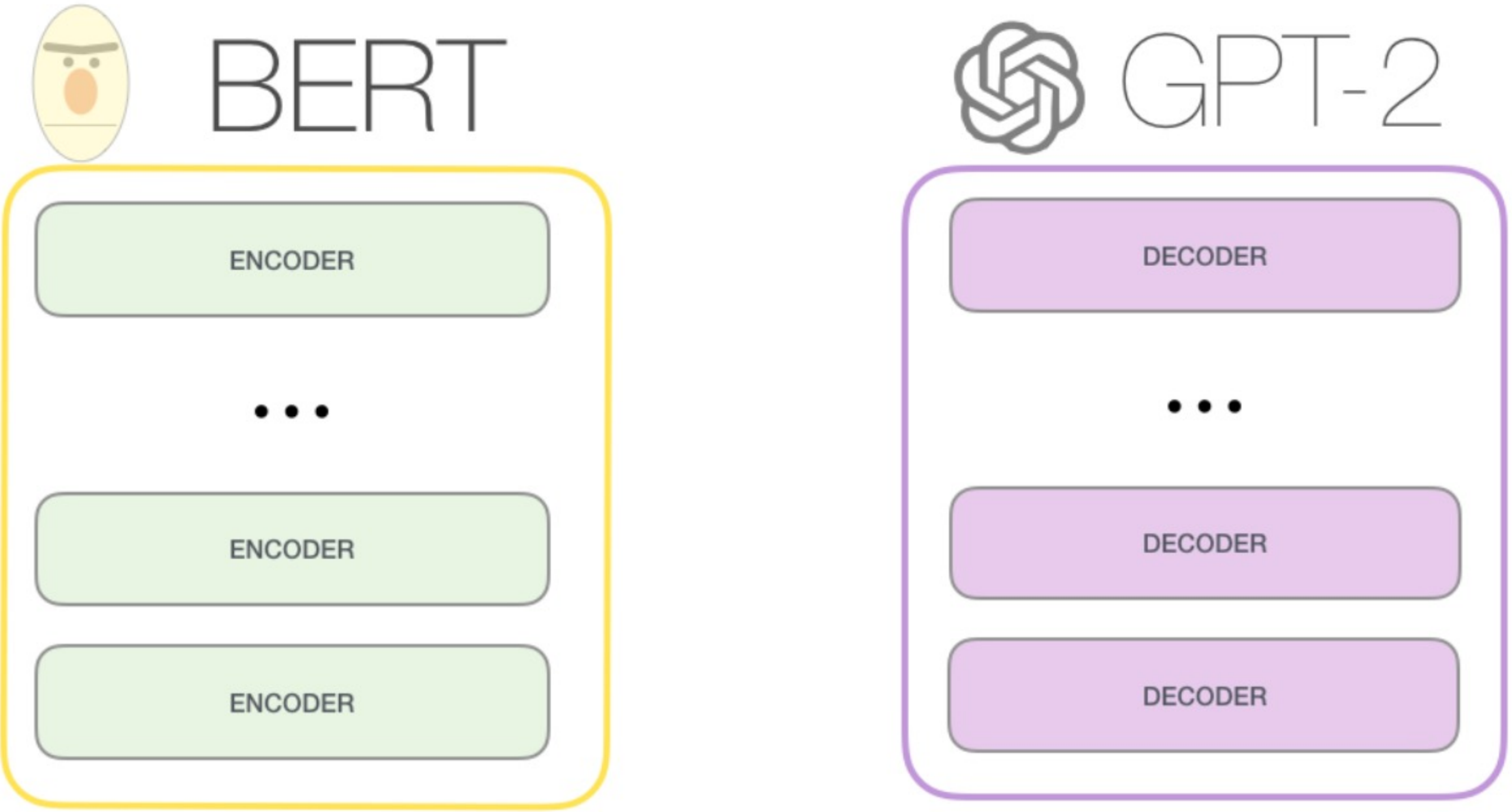
The GPT-2 Model (Feb, 2019)

Language Models are Unsupervised Multitask Learners

Alec Radford [*1] **Jeffrey Wu** [*1] **Rewon Child** [1] **David Luan** [1] **Dario Amodei** [**1] **Ilya Sutskever** [**1]

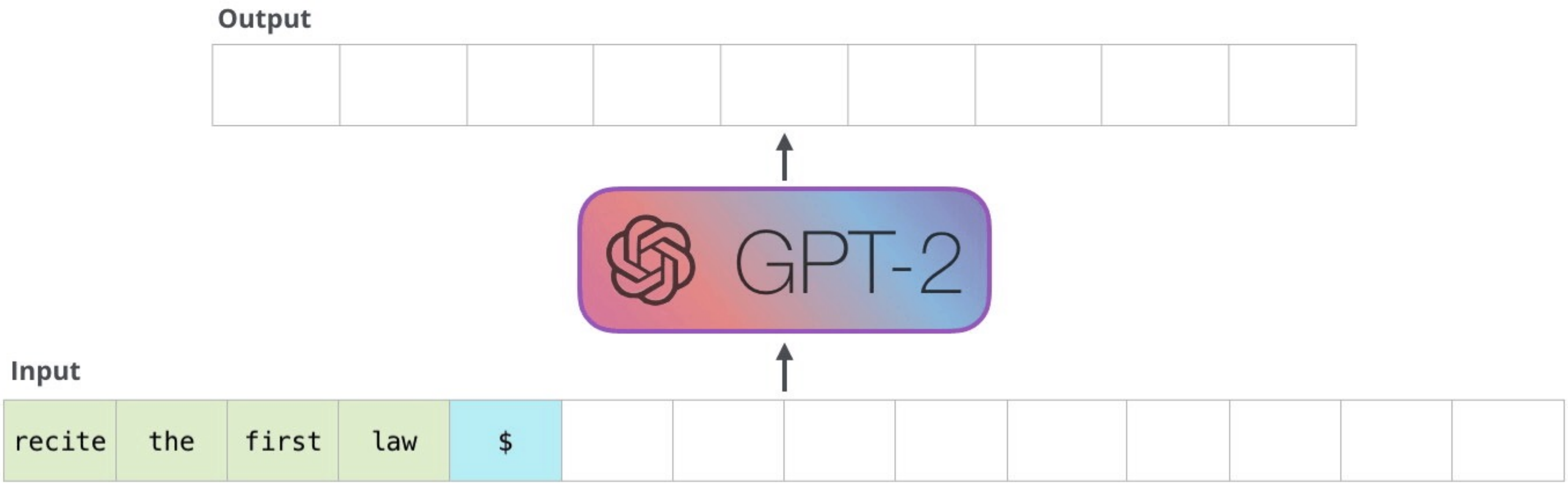
https://openai.com/blog/better-language-models/

The GPT-2 Model



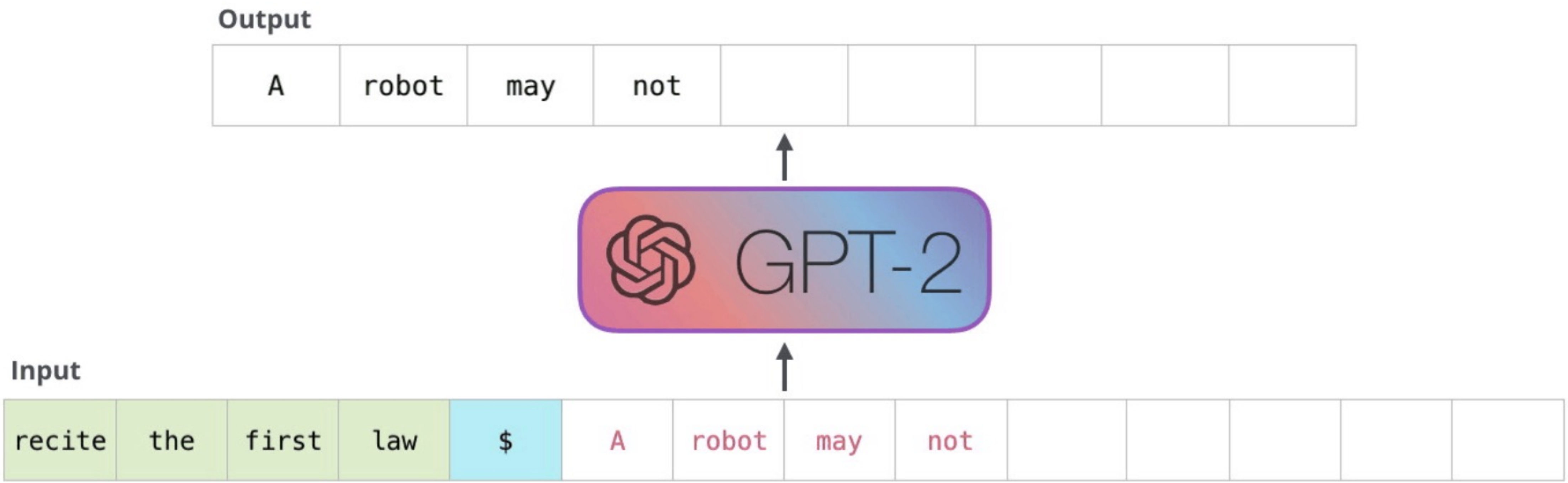
https://jalammar.github.io/illustrated-gpt2/

The GPT-2 Model



https://jalammar.github.io/illustrated-gpt2/

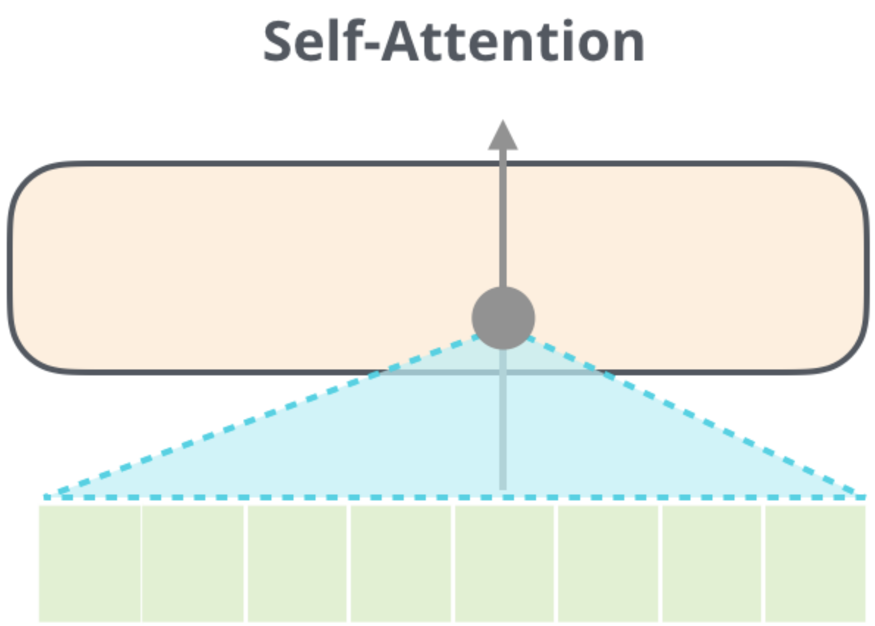
The GPT-2 Model



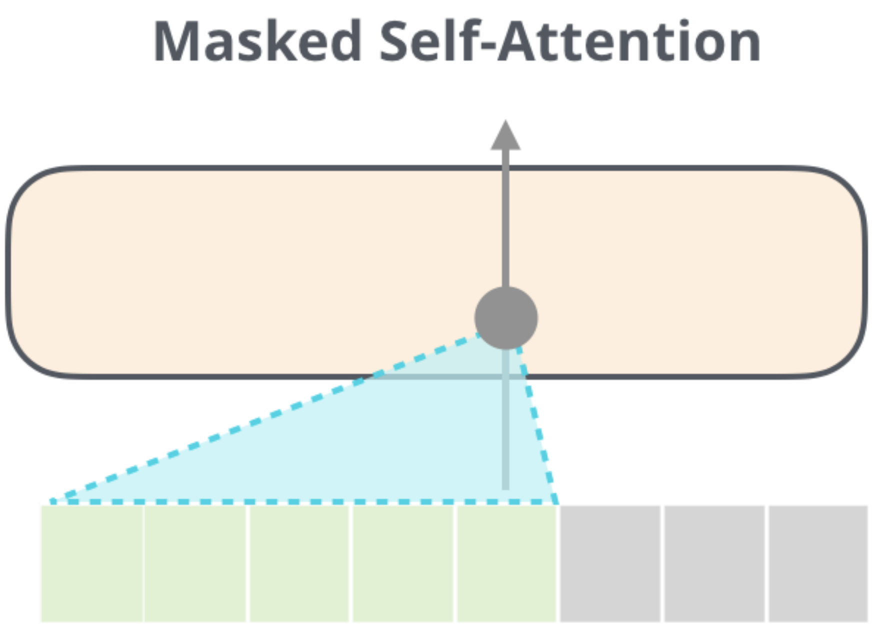
https://jalammar.github.io/illustrated-gpt2/

The GPT-2 Model

BERT



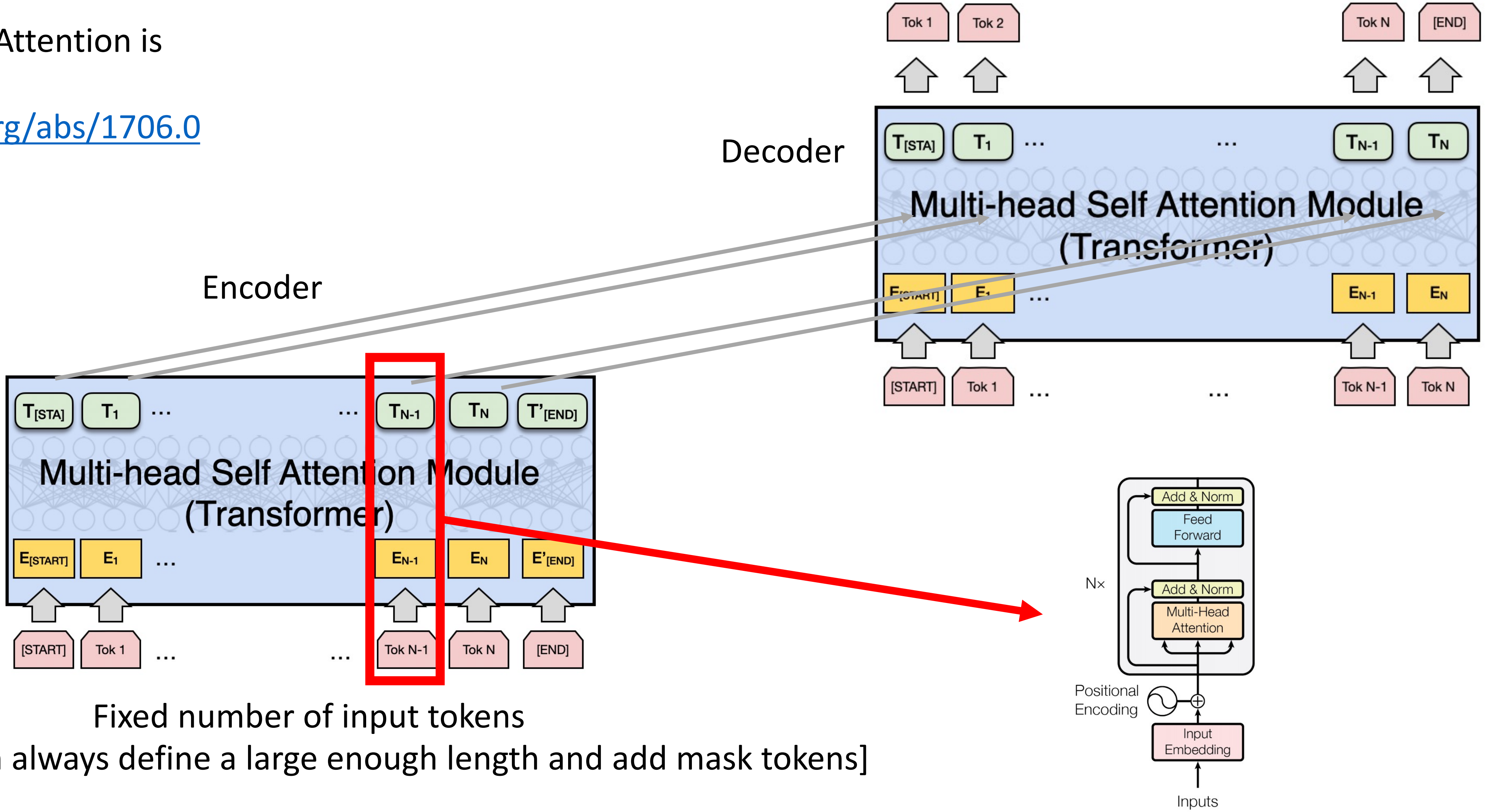
GPT



https://jalammar.github.io/illustrated-gpt2/

Attention is All you Need

Vaswani et al. Attention is all you need https://arxiv.org/abs/1706.03762

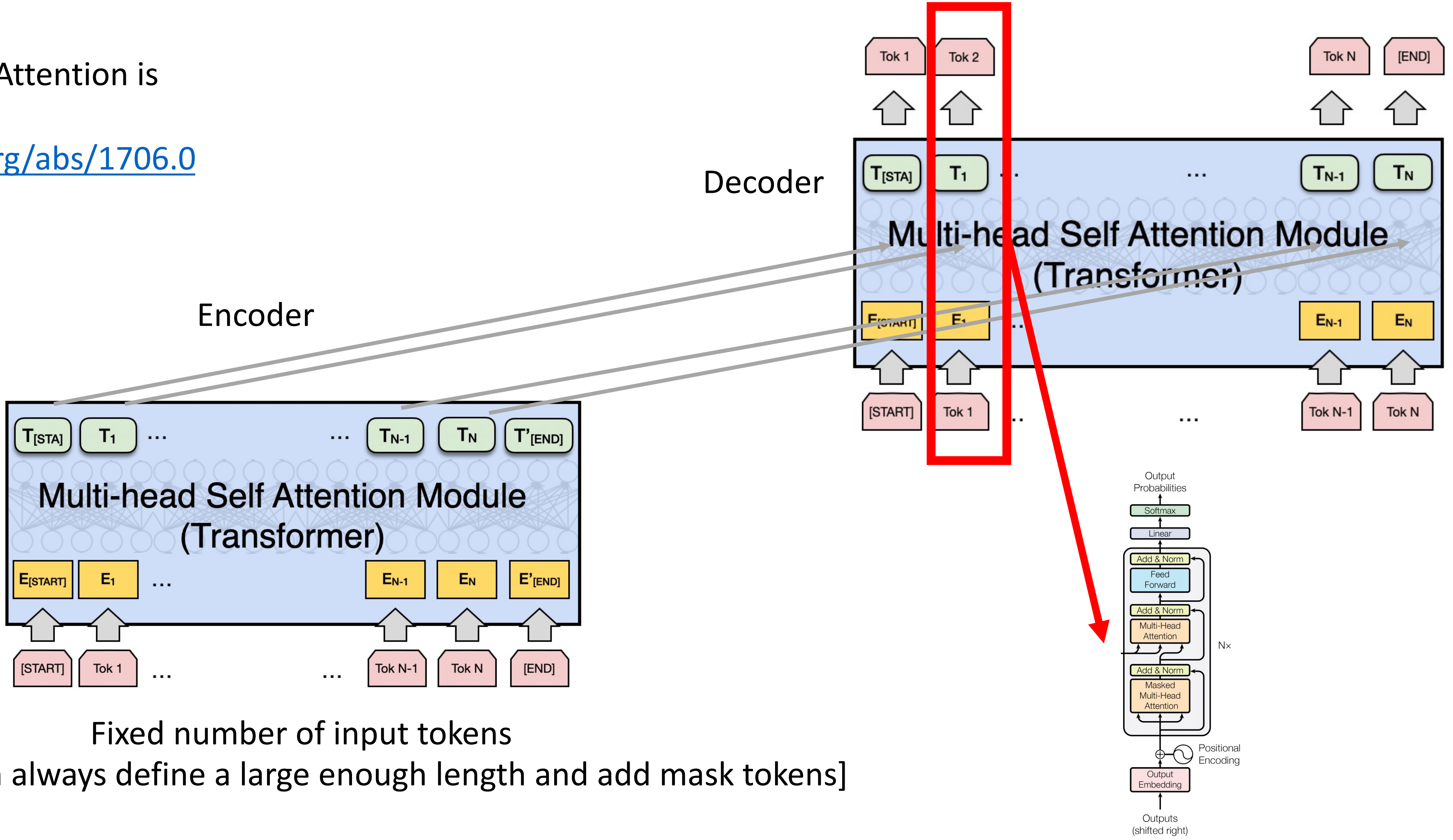


Fixed number of input tokens

[but hey! we can always define a large enough length and add mask tokens]

Attention is All you Need

Vaswani et al. Attention is all you need https://arxiv.org/abs/1706.03762



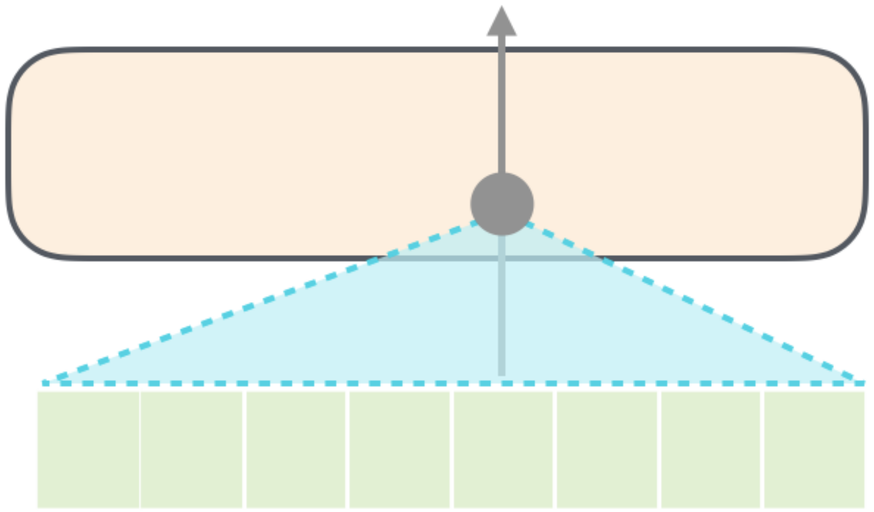
Fixed number of input tokens

[but hey! we can always define a large enough length and add mask tokens]

The GPT-2 Model

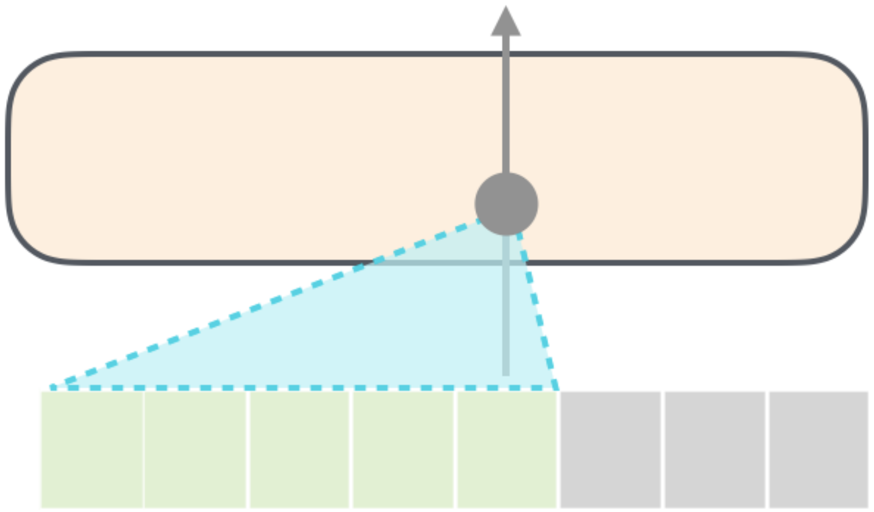
BERT

Self-Attention



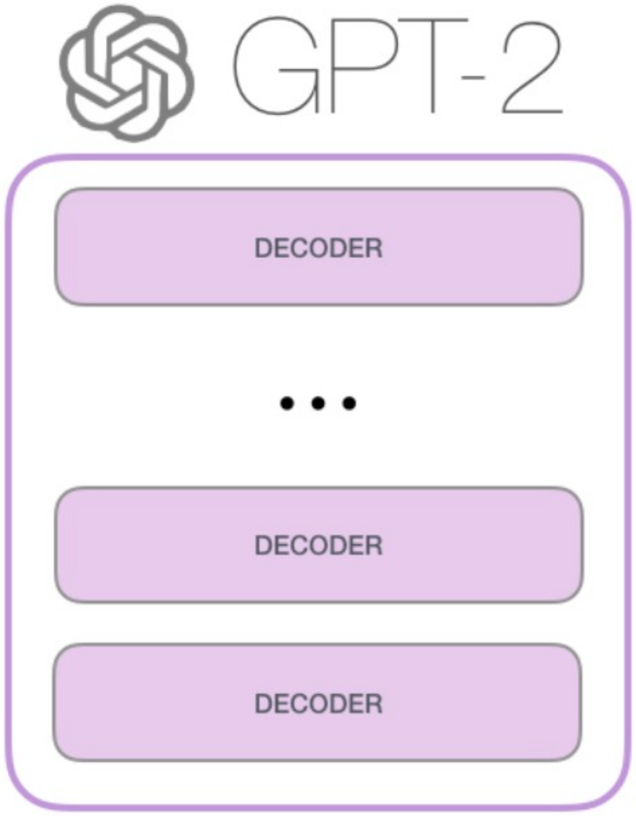
GPT

Masked Self-Attention



https://jalammar.github.io/illustrated-gpt2/

The GPT-2 Model



https://jalammar.github.io/illustrated-gpt2/

GPT-1 vs GPT-2 vs GPT-3

	GPT-1	GPT-2	GPT-3
Parameters	117 Million	1.5 Billion	175 Billion
Decoder Layers	12	48	96
Context Token Size	512	1024	2048
Hidden Layer	768	1600	12288
Batch Size	64	512	3.2M

https://360digitmg.com/blog/types-of-gpt-in-artificial-intelligence

GPT-3 (July, 2020)

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

https://arxiv.org/abs/2005.14165

GPT family keeps growing

- GPT-3.5
- GPT-3.5-turbo
- GPT-4
- GPT-4-turbo

Competitors

- PaLM (Google)
- Gemini family (Gemini Pro) (Google)
- Mistral 7xMoE (Open Source by Mistral.ai)
- Llama-2 70B (Open Source by Meta AI)

Rank	Model	Arena Elo	95% CI	Votes	Organization	License
1	GPT-4-0125-preview	1253	+10/-11	3922	OpenAI	Proprietary
2	GPT-4-1106-preview	1252	+5/-6	35385	OpenAI	Proprietary
3	Bard (Gemini Pro)	1224	+9/-9	9081	Google	Proprietary
4	GPT-4-0314	1190	+5/-6	18945	OpenAI	Proprietary
5	GPT-4-0613	1162	+4/-5	29950	OpenAI	Proprietary
6	Mistral Medium	1150	+6/-7	15447	Mistral	Proprietary
7	Claude-1	1149	+6/-6	18189	Anthropic	Proprietary
8	Claude-2.0	1132	+6/-7	12131	Anthropic	Proprietary
9	Gemini Pro (Dev API)	1120	+7/-7	7616	Google	Proprietary
10	Claude-2.1	1119	+5/-6	25494	Anthropic	Proprietary
11	GPT-3.5-Turbo-0613	1118	+5/-5	33617	OpenAI	Proprietary
12	Mixtral-8x7b-Instruct-v0.1	1118	+5/-7	15705	Mistral	Apache 2.0
13	Yi-34B-Chat	1115	+7/-8	6710	01 AI	Yi License
14	Gemini Pro	1114	+7/-8	6969	Google	Proprietary
15	Claude-Instant-1	1109	+4/-7	18689	Anthropic	Proprietary
16	WizardLM-70B-v1.0	1105	+6/-7	8483	Microsoft	Llama 2 Community
17	GPT-3.5-Turbo-0314	1105	+10/-9	5960	OpenAI	Proprietary

LLM Leader Board https://chat.lmsys.org/

Rank	Model	Arena Elo	95% CI	Votes	Organization	License
1	GPT-4-0125-preview	1253	+10/-11	3922	OpenAI	Proprietary
3	Bard (Gemini Pro)	1224	+9/-9	9081	Google	Proprietary
5	GPT-4-0613	1162	+4/-5	29950	OpenAI	Proprietary
12	Mixtral-8x7b-Instruct-v0.1	1118	+5/-7	15705	Mistral	Apache 2.0
27	GPT-3.5-Turbo-1106	1071	+4/-5	15711	OpenAI	Proprietary
41	Llama-2-7b-chat	1024	+8/-9	7722	Meta	Llama 2 Community
43	Mistral-7B-Instruct-v0.1	1006	+8/-7	7919	Mistral	Apache 2.0

LLM Leader Board https://chat.lmsys.org/

Prompt Engineering

```
Poor English input:  I eated the purple berries.
Good English output:  I ate the purple berries.
Poor English input:  Thank you for picking me as your designer.  I'd appreciate it.
Good English output:  Thank you for choosing me as your designer.  I appreciate it.
Poor English input:  The mentioned changes have done.  or I did the alteration that you
requested.  or I changed things you wanted and did the modifications.
Good English output:  The requested changes have been made.  or I made the alteration that you
requested.  or I changed things you wanted and made the modifications.
Poor English input:  I'd be more than happy to work with you in another project.
```
Good English output: I'd be more than happy to work with you on another project.

Language Models are Few-Shot Learners

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, Dario Amodei

https://arxiv.org/abs/2005.14165

Prompt Engineering

```
Translate English to French:

sea otter => loutre de mer

peppermint => menthe poivrée

plush girafe => girafe peluche

cheese => ...........................
```

Language Models are Few-Shot Learners

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, Dario Amodei

https://arxiv.org/abs/2005.14165

Prompt Engineer

Prompt engineering

文A 12 languages

Article Talk Read Edit View history Tools

From Wikipedia, the free encyclopedia

Prompt engineering is a concept in artificial intelligence (AI), particularly natural language processing (NLP). In prompt engineering, the description of the task that the AI is supposed to accomplish is embedded in the input, e.g., as a question, instead of it being implicitly given. Prompt engineering typically works by converting one or more tasks to a prompt-based dataset and training a language model with what has been called "prompt-based learning" or just "prompt learning".[1][2]

History [edit]

The GPT-2 and GPT-3 language models[3] were important steps in prompt engineering. In 2021, multitask[*jargon*] prompt engineering using multiple NLP datasets showed good performance on new tasks.[4] In a method called chain-of-thought (CoT) prompting, few-shot examples of a task are given to the language model which improves its ability to reason.[5] CoT prompting can also be a zero-shot learning task by prepending text to the prompt that encourages a chain of thought (e.g. "Let's think step by step"), which may also improve the performance of a language model in multi-step reasoning problems.[6] The broad accessibility of these tools were driven by the publication of several open-source notebooks and community-led projects for image synthesis.[7]

A description for handling prompts reported that over 2,000 public prompts for around 170 datasets were available in February 2022.[8]

How would you come up with a solution for this problem?

The kid is throwing rocks at the window

The <subject>kid</subject> is throwing <object>rocks</object> at the <destination>window</destination>

Prompt Engineering

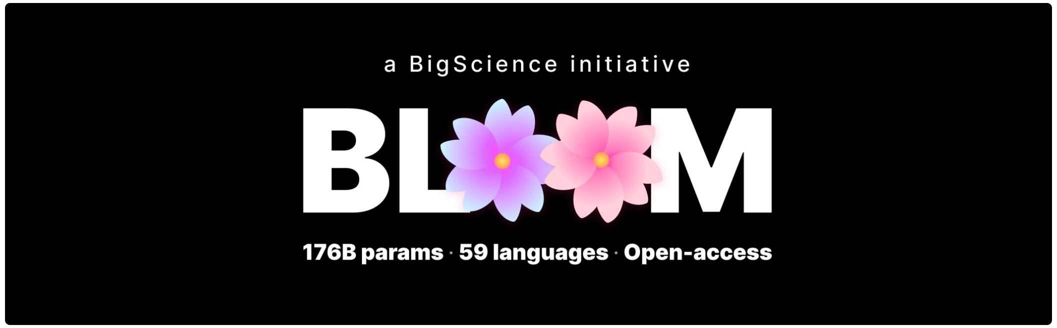
Input: The cat is throwing the ball into the ground
Output: The <subject>cat</subject> is throwing the <object>ball</object> into the <destination>ground</ground>

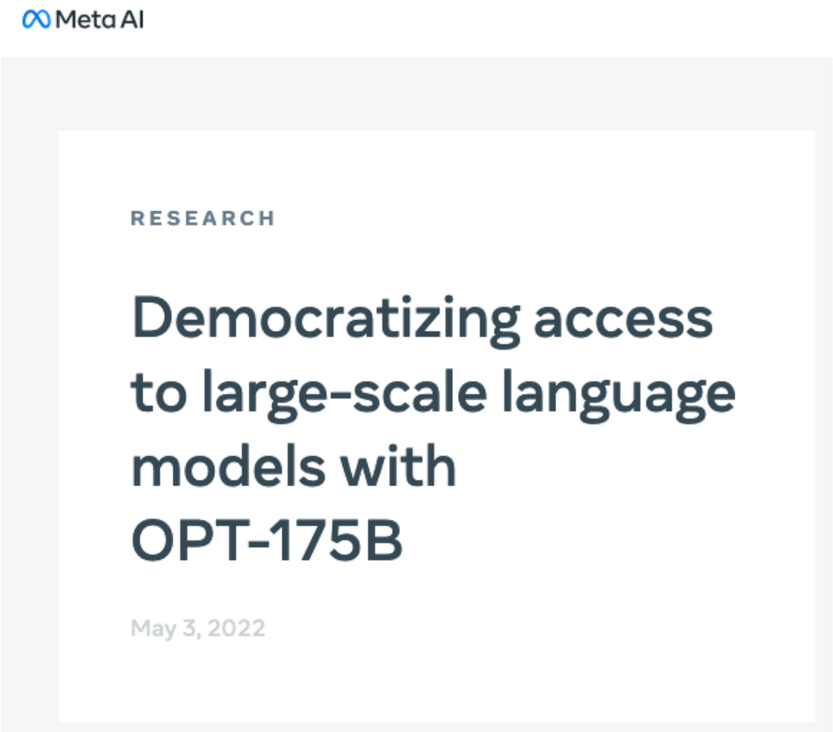
Input: The snake is being attacked by the wolf
Output: The <object>snake</object> is being attacked by the <actor>wolf</actor>

Input: The kid is throwing rocks at the window
Output:

Prompt Engineering

- Any Large Language Model (LLM) such as GPT-3 can be turned into a general purpose problem solver in this way.
- Obviously, it is not going to work well for every use case.
- Other Large Language Models trained at the scale of GPT-3 that are actually publicly available:
- BLOOM-176B and OPT-175B:

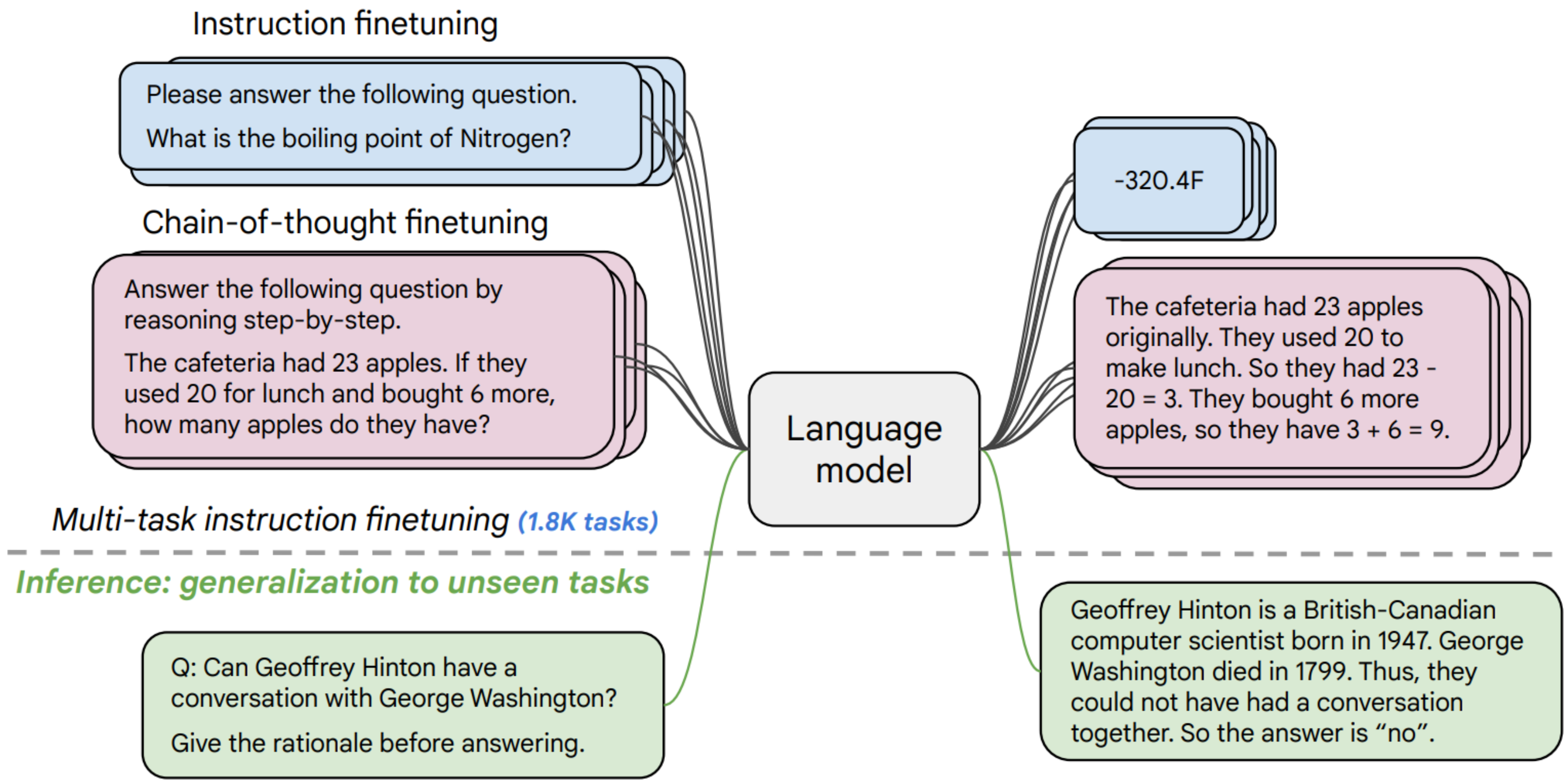




However these are still limited

- Predicting the next word can lead to intelligent behavior such as the one exemplified earlier however this still limited
- What makes some of the new LLMs special? ChatGPT (GPT-3.5, 3.5 Turbo, 4, 4-turbo), FLAN-T5, OPT-IML

Instruction Tuning (e.g. FLAN-T5 by Google)



https://arxiv.org/pdf/2210.11416.pdf

FLAN-T5

Model input (Disambiguation QA)

Q: In the following sentences, explain the antecedent of the pronoun (which thing the pronoun refers to), or state that it is ambiguous.

Sentence: The reporter and the chef will discuss their favorite dishes.

Options:
(A) They will discuss the reporter's favorite dishes
(B) They will discuss the chef's favorite dishes
(C) Ambiguous

A: Let's think step by step.

Before instruction finetuning

The reporter and the chef will discuss their favorite dishes.
The reporter and the chef will discuss the reporter's favorite dishes.
The reporter and the chef will discuss the chef's favorite dishes.
The reporter and the chef will discuss the reporter's and the chef's favorite dishes.

✖ **(doesn't answer question)**

FLAN-T5

Model input (Disambiguation QA)

Q: In the following sentences, explain the antecedent of the pronoun (which thing the pronoun refers to), or state that it is ambiguous.

Sentence: The reporter and the chef will discuss their favorite dishes.

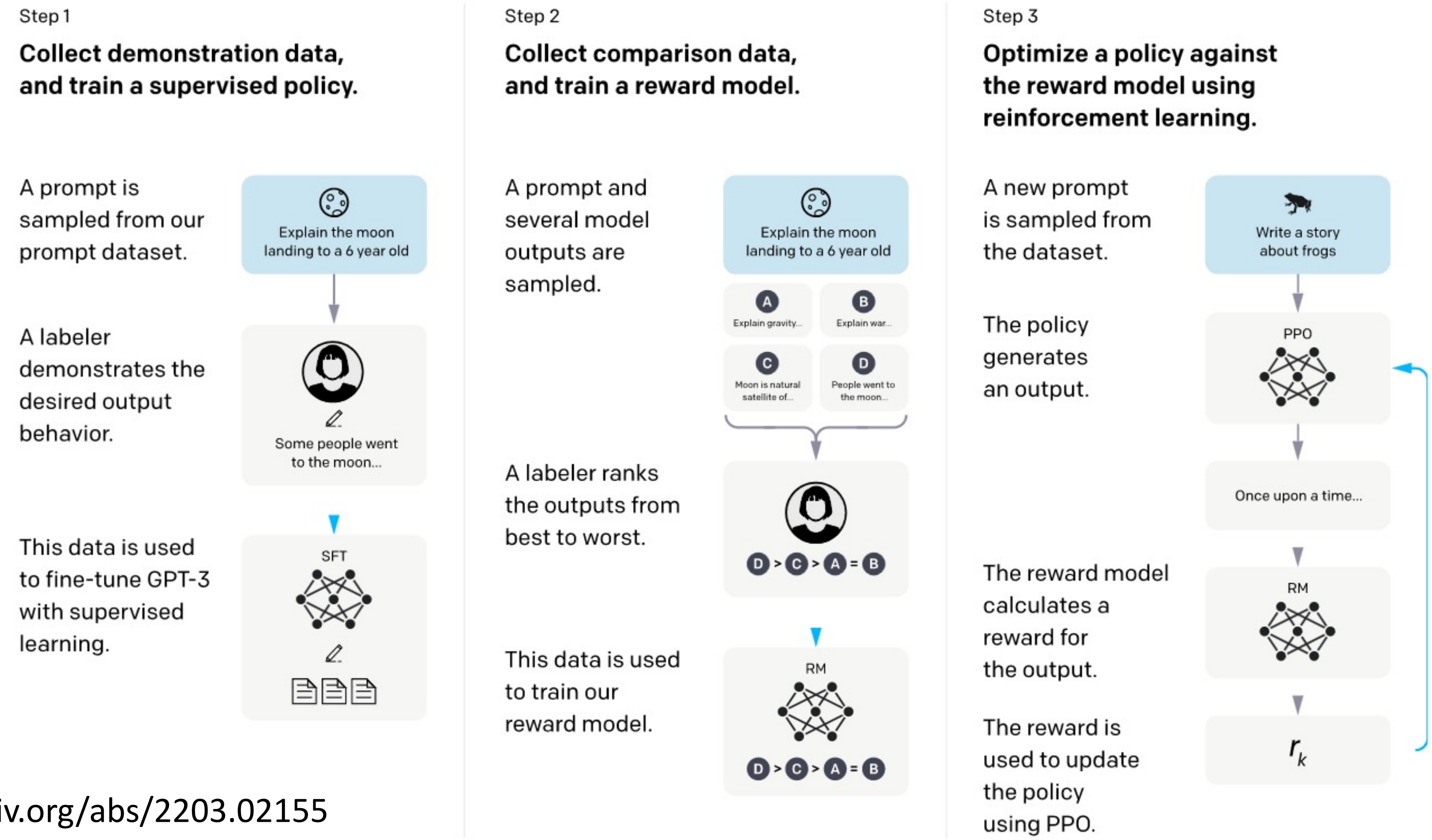
Options:
(A) They will discuss the reporter's favorite dishes
(B) They will discuss the chef's favorite dishes
(C) Ambiguous

A: Let's think step by step.

After instruction finetuning

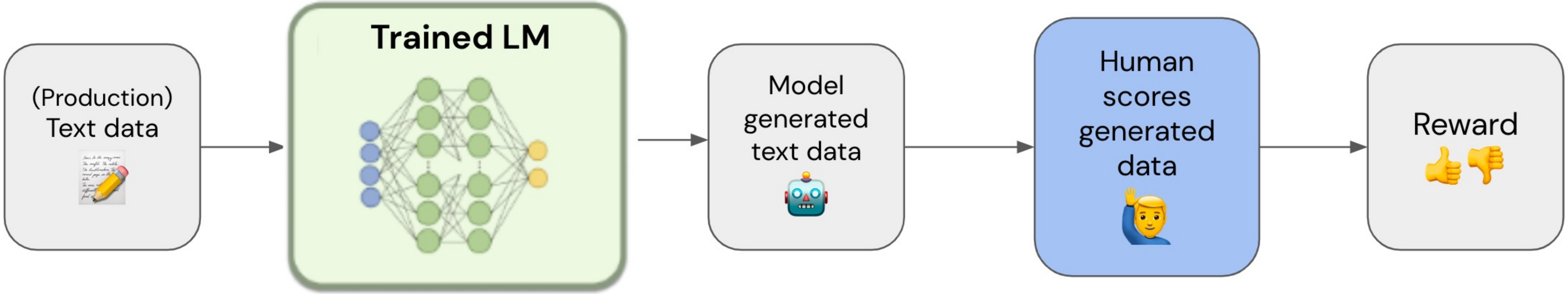
The reporter and the chef will discuss their favorite dishes does not indicate whose favorite dishes they will discuss. So, the answer is (C).

InstructGPT (ChatGPT)



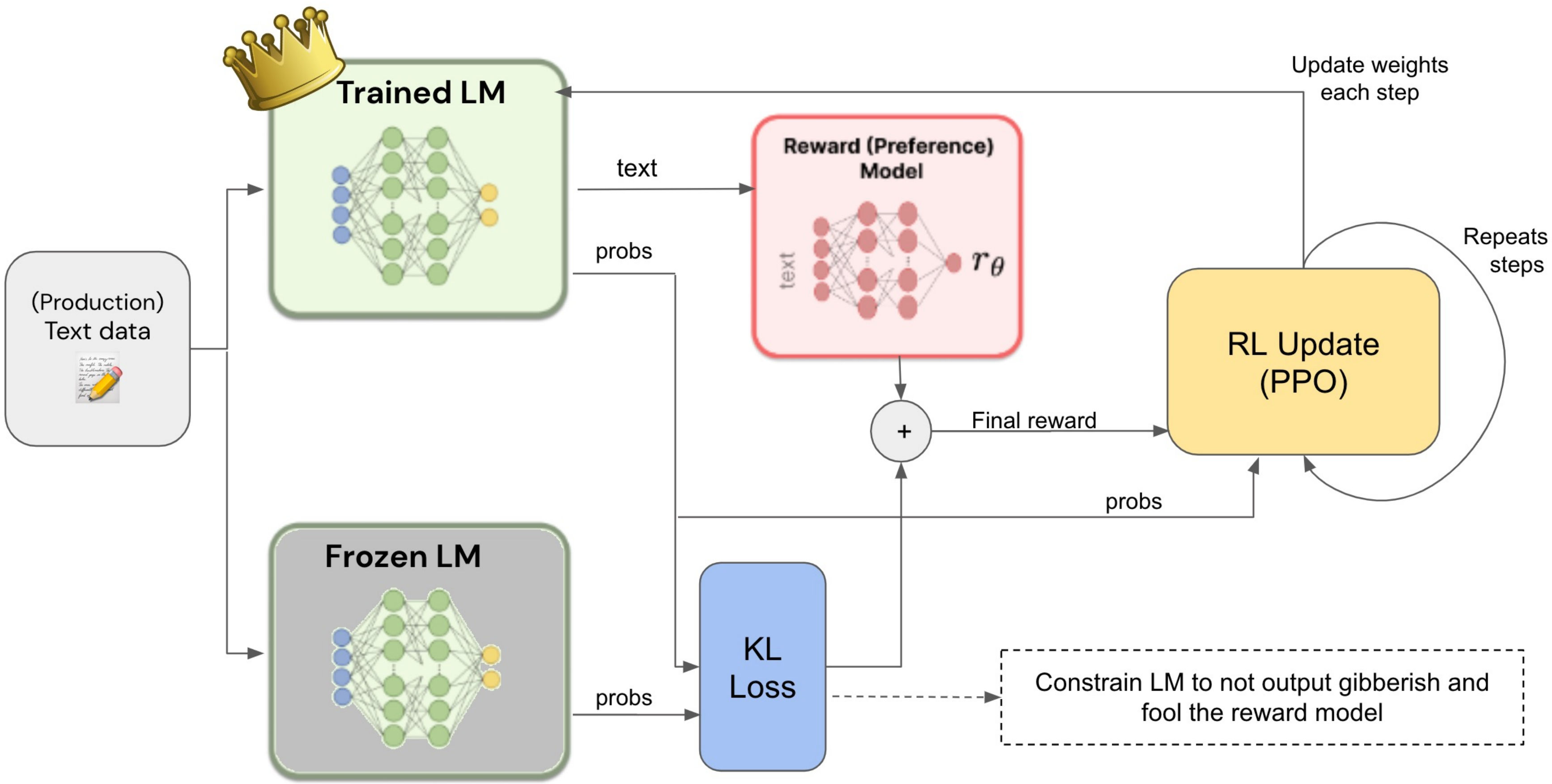
https://arxiv.org/abs/2203.02155

Step by Step: Train a Reward Model that learns from Human Ratings e.g. from 1 to 5



https://gist.github.com/JoaoLages/c6f2dfd13d2484aa8bb0b2d567fbf093

Step by Step: Train the LM to generate text that gets high reward but still produces stuff that makes sense



https://gist.github.com/JoaoLages/c6f2dfd13d2484aa8bb0b2d567fbf093

Recommended Slide Deck

Natural Language Processing with Deep Learning

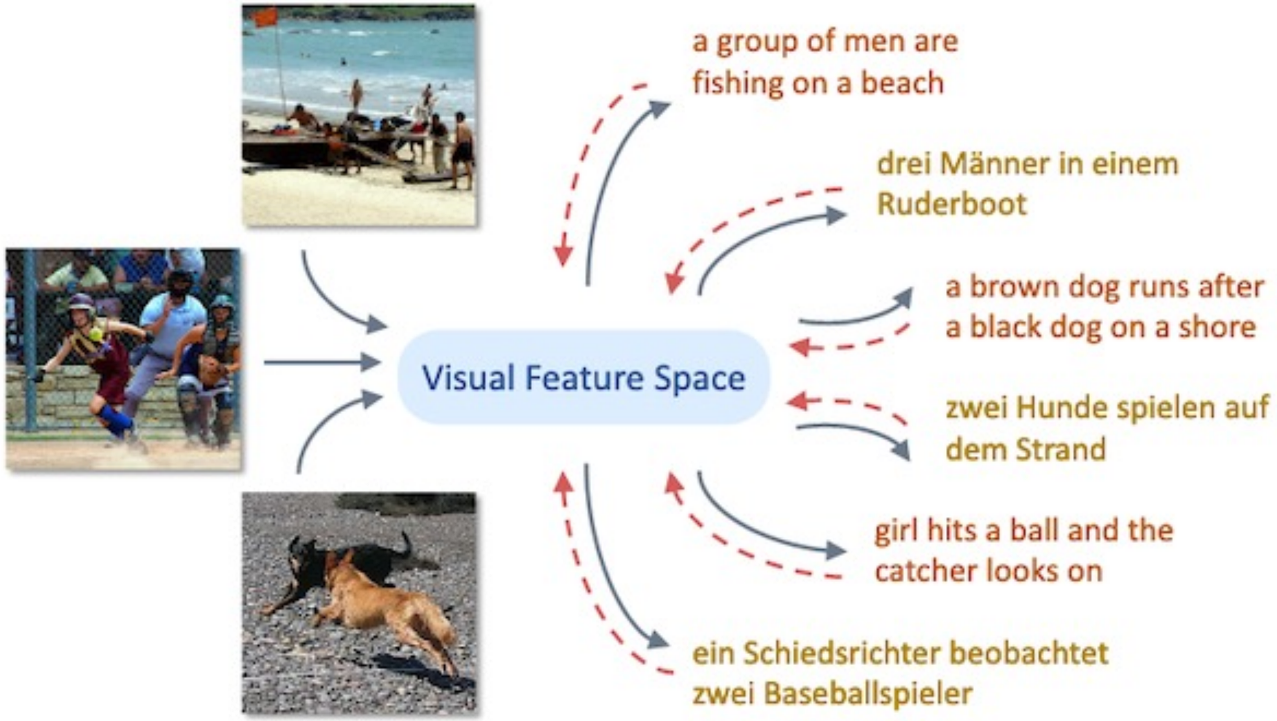
CS224N/Ling284

Jesse Mu

Lecture 11: Prompting, Instruction Finetuning, and RLHF

http://web.stanford.edu/class/cs224n/slides/cs224n-2023-lecture11-prompting-rlhf.pdf

Next Step: Multimodality



Multimodal Few-Shot Learning with Frozen Language Models

Maria Tsimpoukelli*
DeepMind
mrts@deepmind.com

Jacob Menick*
DeepMind
University College London
jmenick@deepmind.com

Serkan Cabi*
DeepMind
cabi@deepmind.com

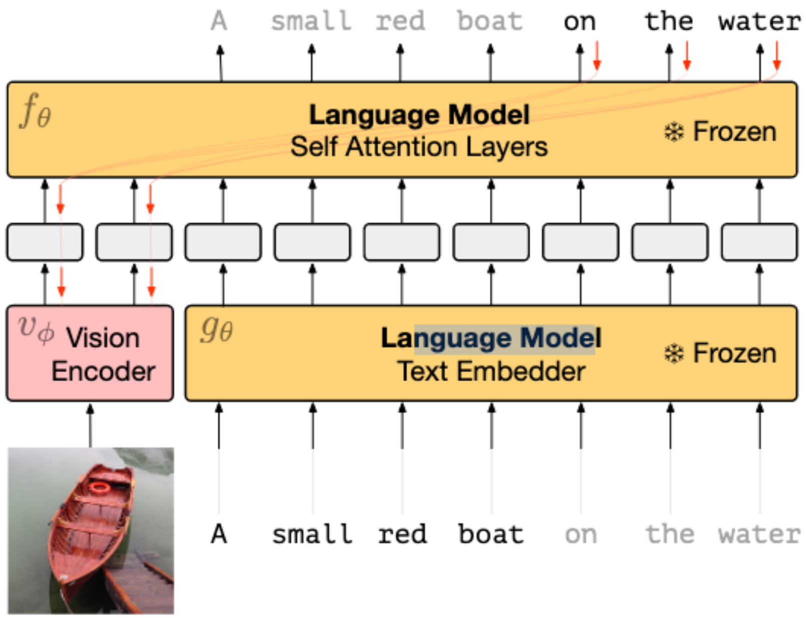
S. M. Ali Eslami
DeepMind
aeslami@deepmind.com

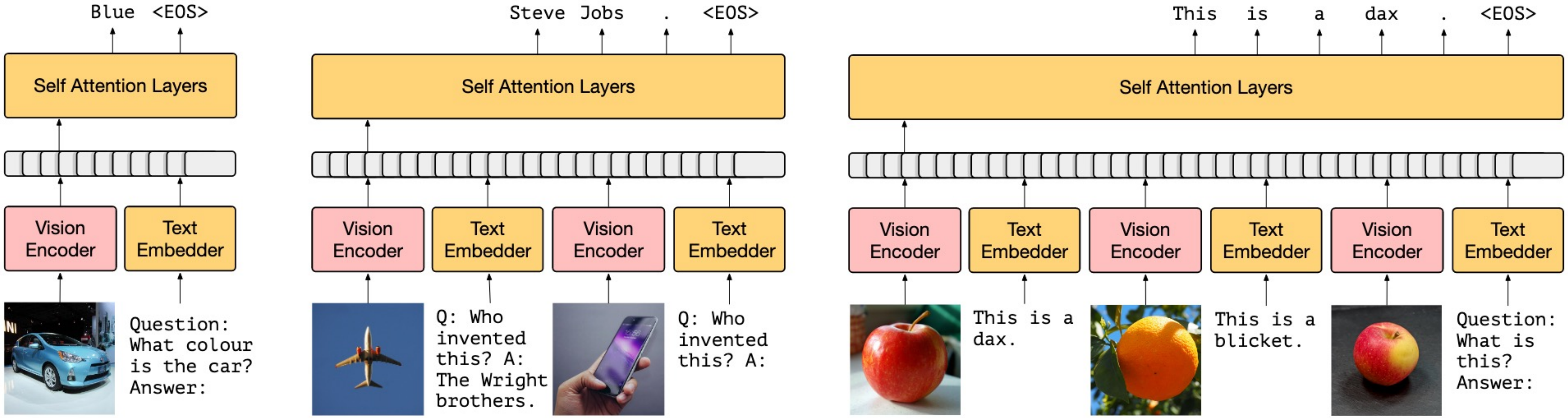
Oriol Vinyals
DeepMind
vinyals@deepmind.com

Felix Hill
DeepMind
felixhill@deepmind.com

NeurIPS 2021

Training:





(a) **0-shot VQA**

(b) **1-shot outside-knowledge VQA**

(c) **Few-shot image classification**

Flamingo

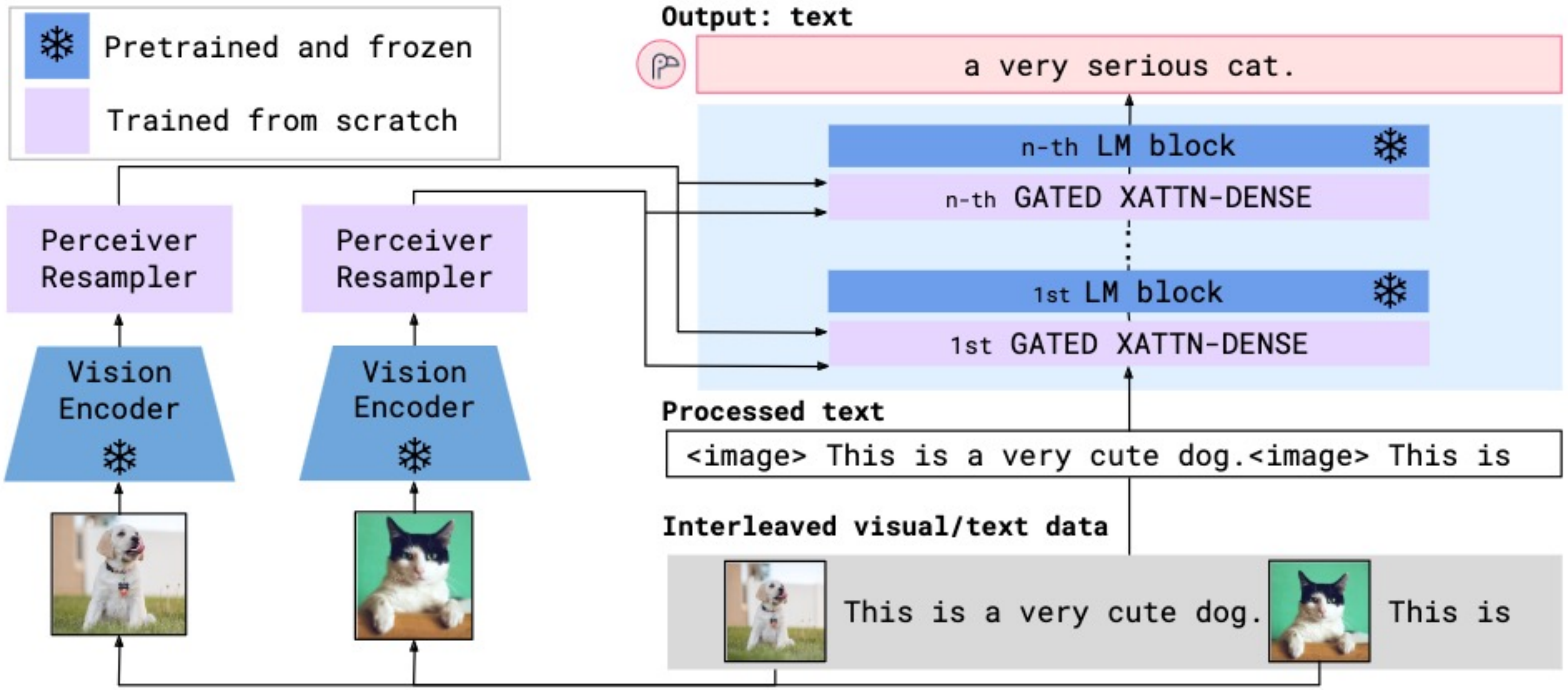
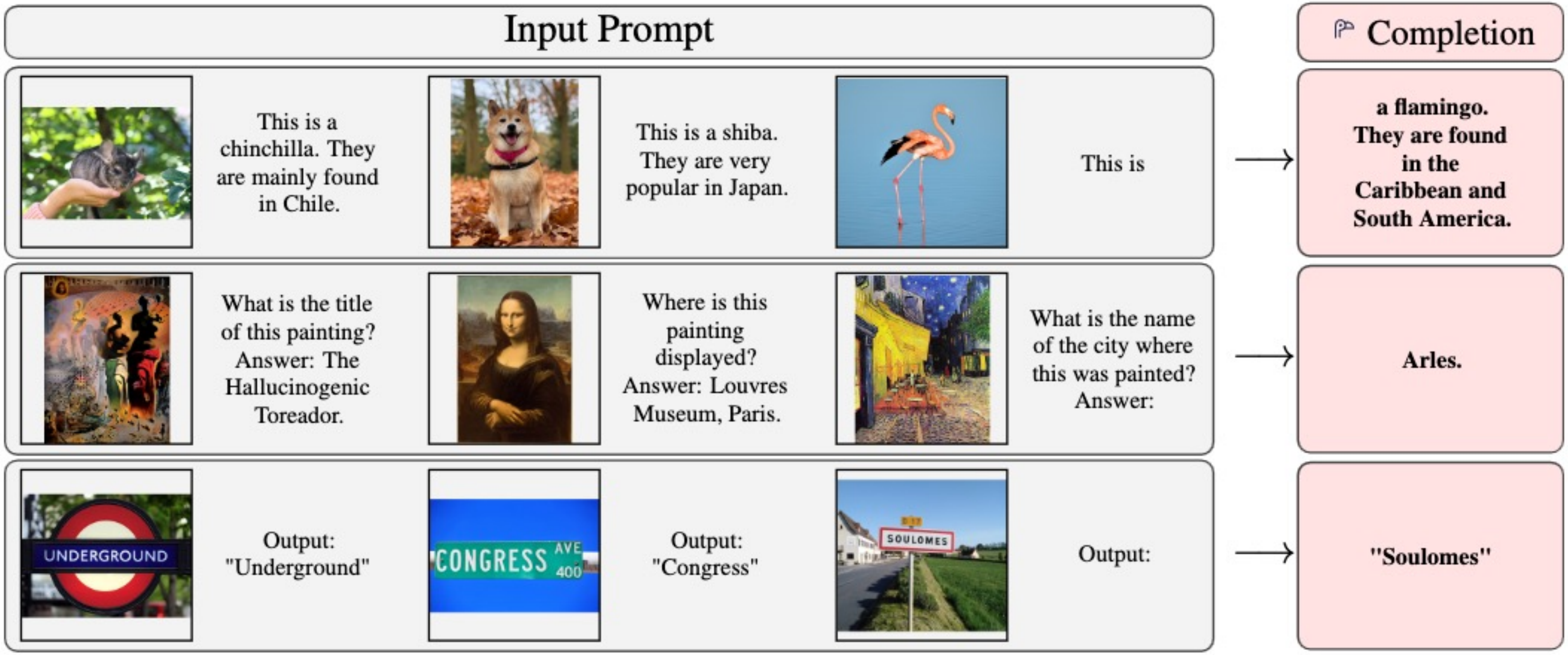


Figure 3: **Flamingo architecture overview.** Flamingo is a family of visual language models (VLMs) that take as input visual data interleaved with text and produce free-form text as output.

https://arxiv.org/pdf/2204.14198.pdf

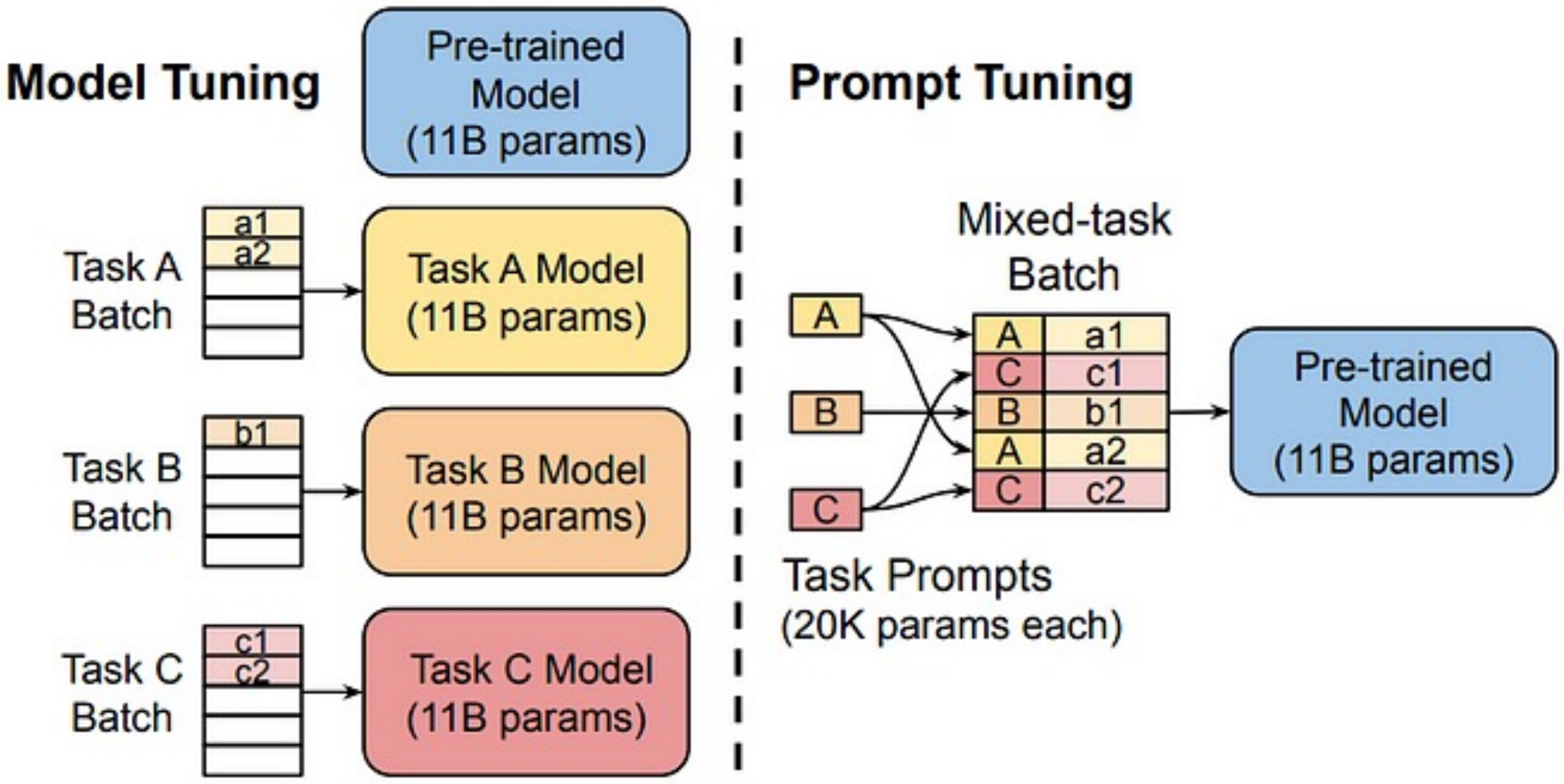
Flamingo



https://arxiv.org/pdf/2204.14198.pdf

LLM Efficient Model Finetuning

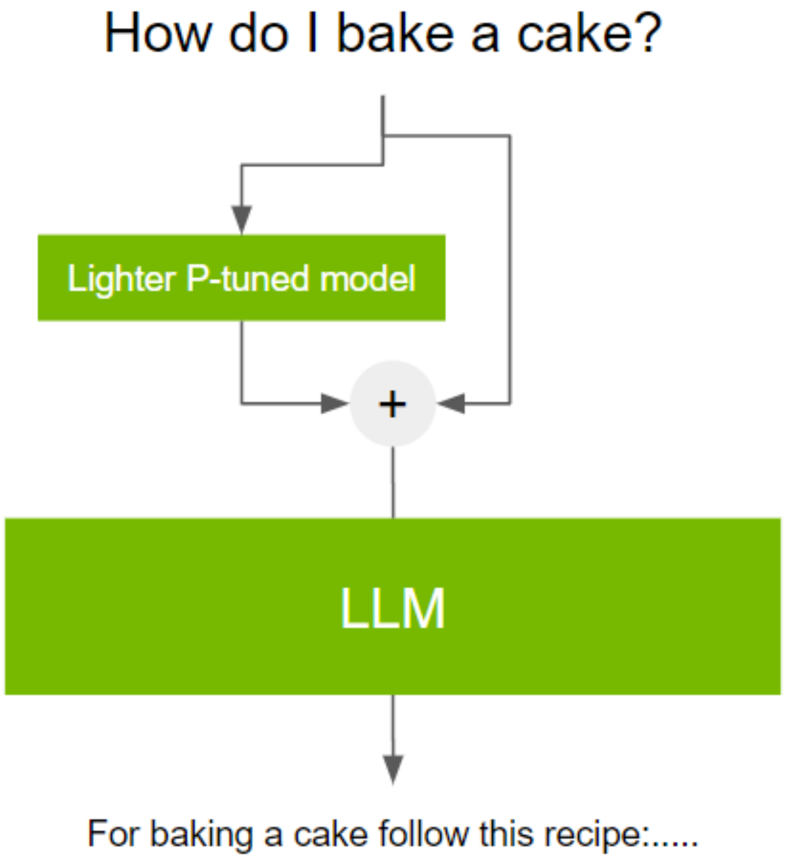
- Prompt Tuning



https://medium.com/@shahshreyansh20/prompt-tuning-a-powerful-technique-for-adapting-llms-to-new-tasks-6d6fd9b83557

LLM Efficient Model Finetuning

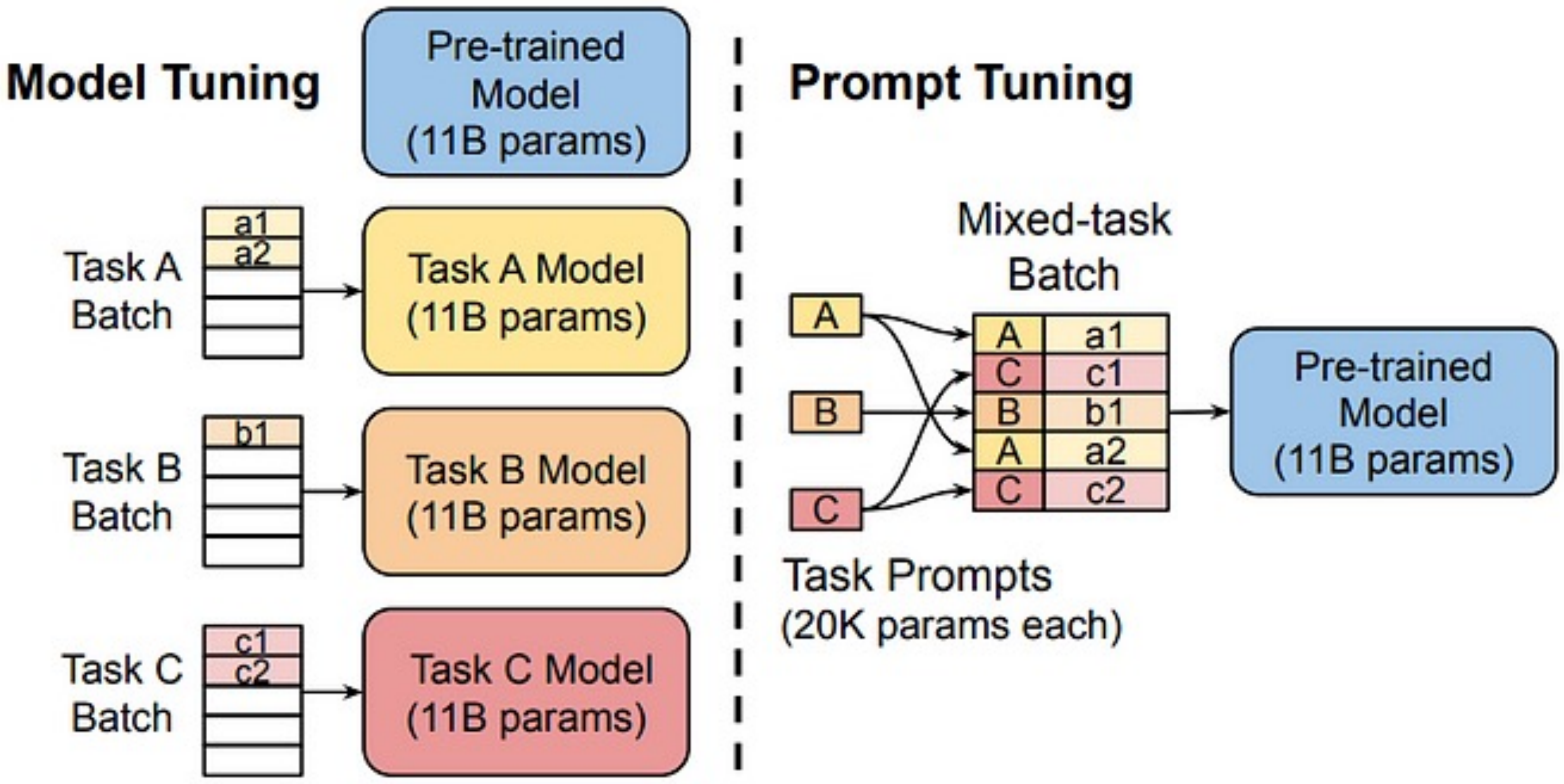
- Prompt Tuning



https://medium.com/@shahshreyansh20/prompt-tuning-a-powerful-technique-for-adapting-llms-to-new-tasks-6d6fd9b83557

LLM Efficient Model Finetuning

- Prompt Tuning

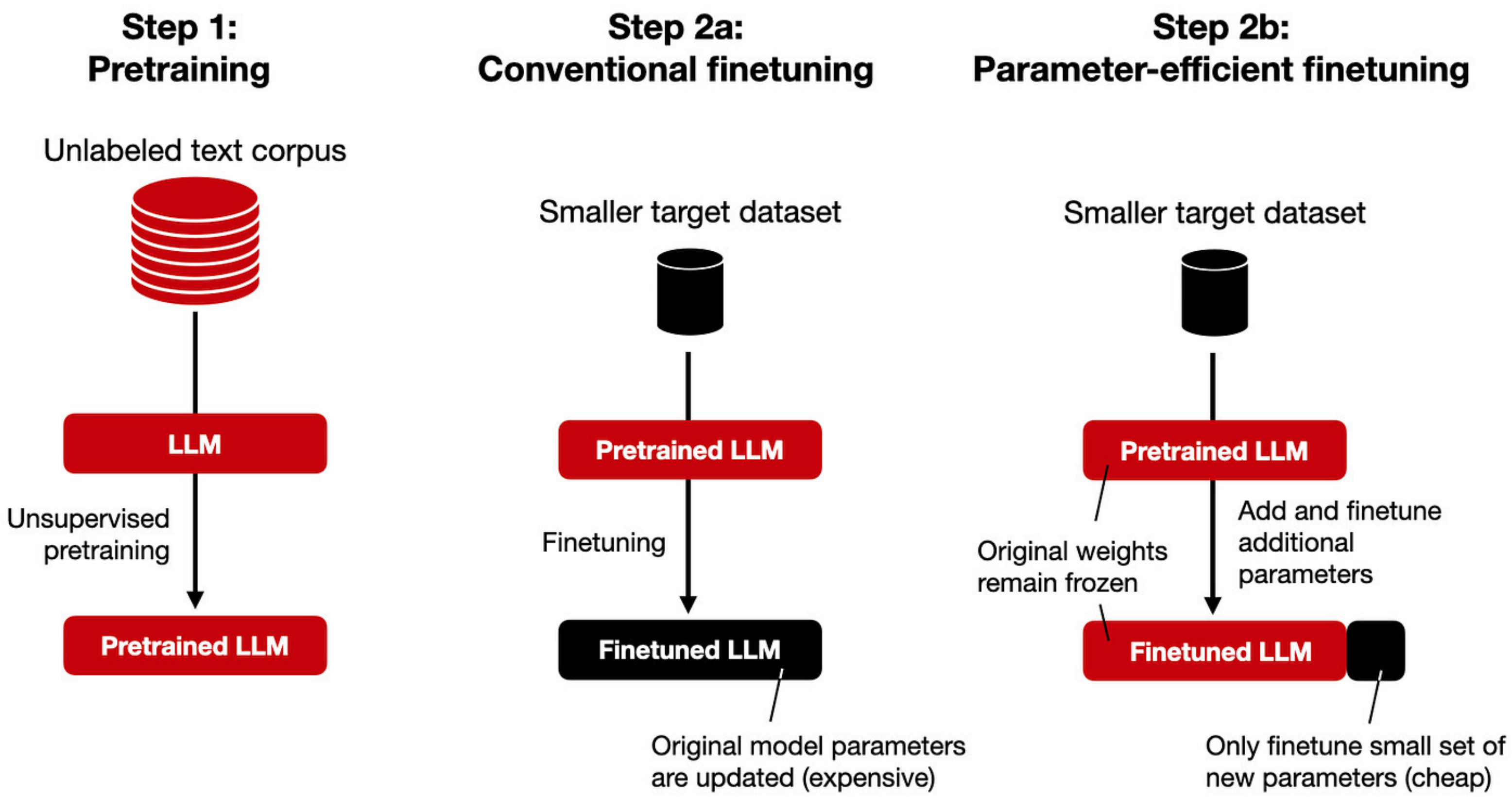


https://medium.com/@shahshreyansh20/prompt-tuning-a-powerful-technique-for-adapting-llms-to-new-tasks-6d6fd9b83557

LLM Efficient Model Finetuning

- Prompt Tuning
- Parameter Efficient Finetuning (PEFT)

https://github.com/huggingface/peft

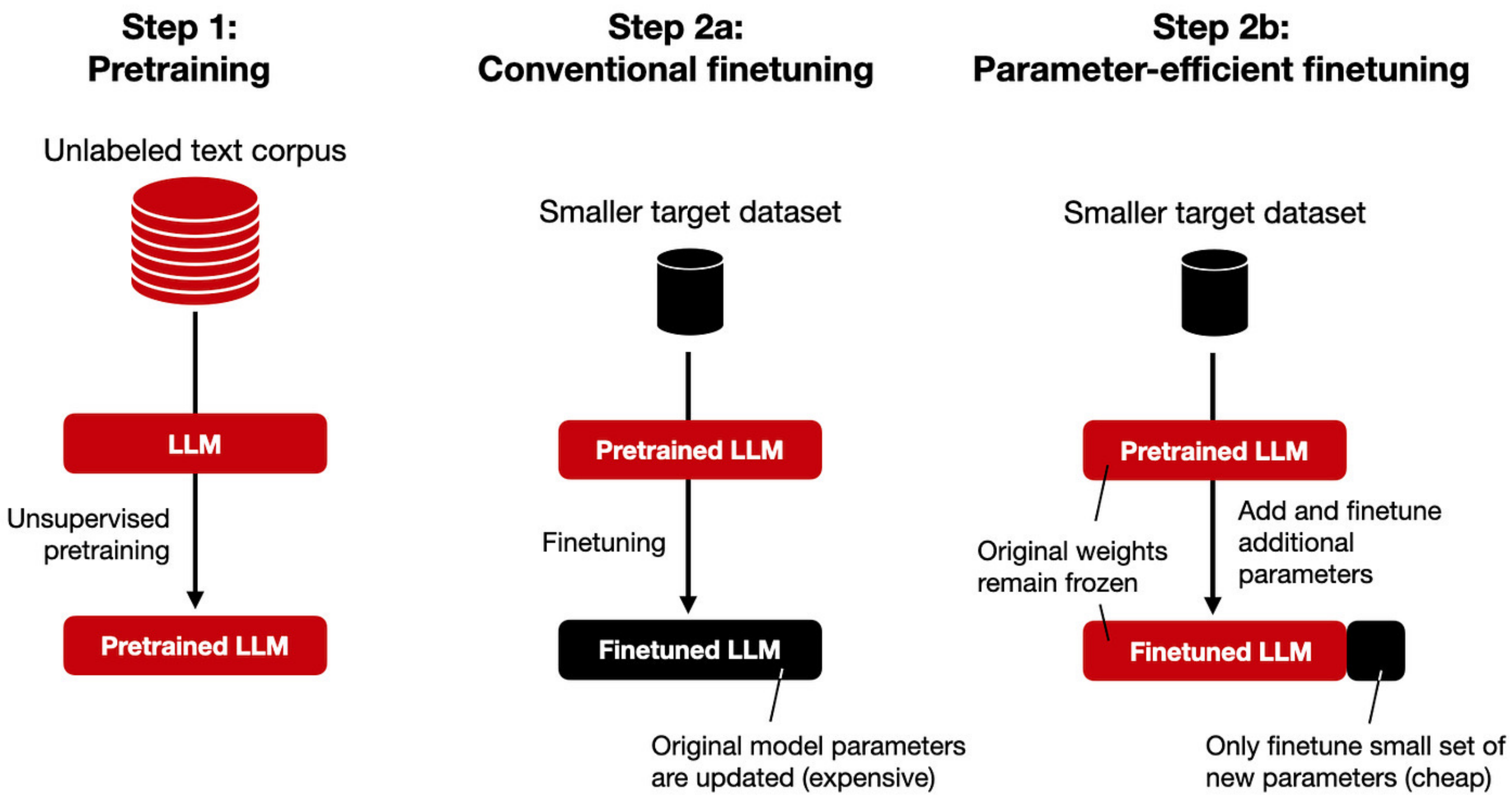


https://magazine.sebastianraschka.com/p/finetuning-llms-with-adapters

LLM Efficient Model Finetuning

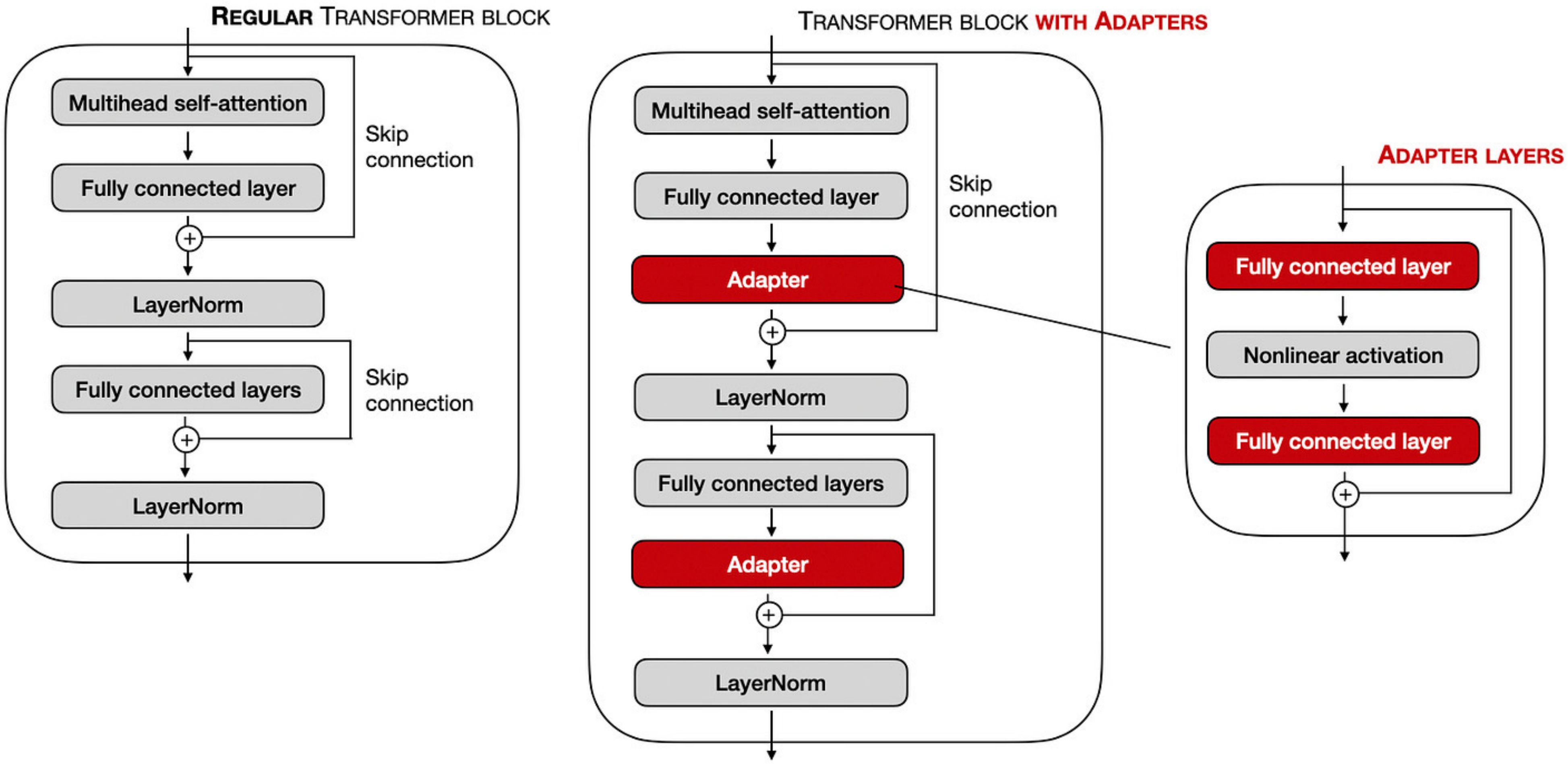
- Prompt Tuning
- Parameter Efficient Finetuning (PEFT)

https://github.com/huggingface/peft



https://magazine.sebastianraschka.com/p/finetuning-llms-with-adapters

LLM Efficient Model Finetuning: Adapters



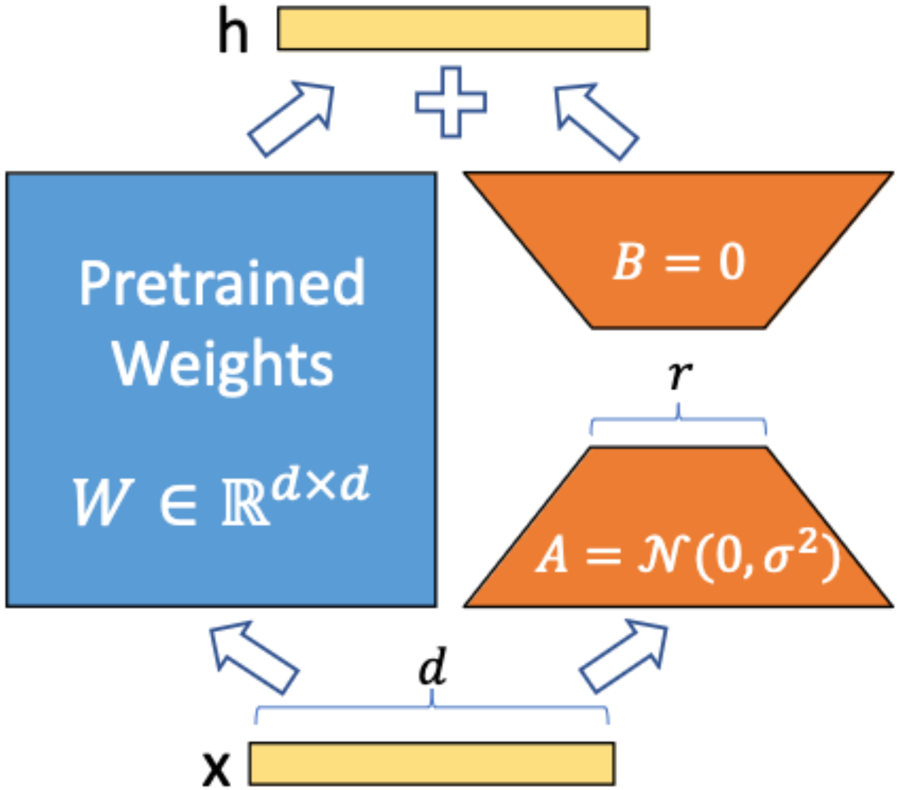
https://magazine.sebastianraschka.com/p/finetuning-llms-with-adapters

LLM Efficient Model Finetuning: LoRA

https://github.com/huggingface/peft

- LoRA: Low Rank Adaptation

$$h = W_0 x + \Delta W x = W_0 x + BAx$$



LoRA: Low-Rank Adaptation of Large Language Models

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen

Questions?