

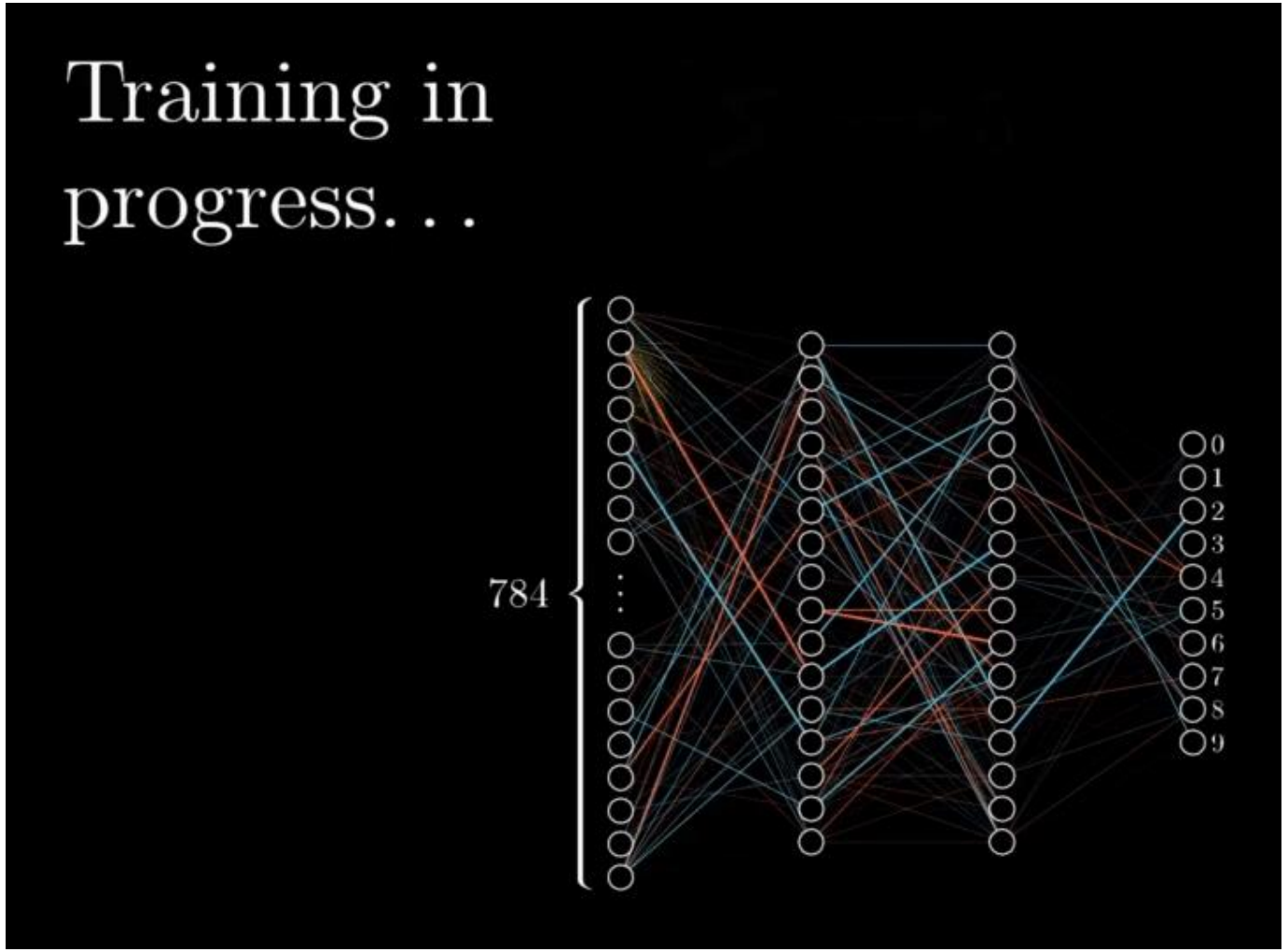
Efficient AI and Large Language Models

Baichuan Huang (TA in ML4IOT 2026)

Department of Electrical and Information Technology, Lund University, Sweden

baichuan.huang@eit.lth.se

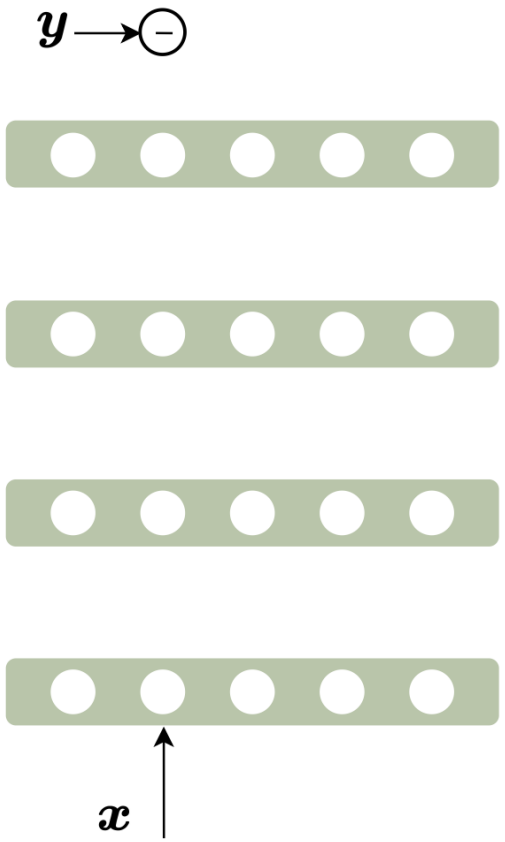
Backpropagation



https://www.youtube.com/watch?v=VkHfRKewkWw

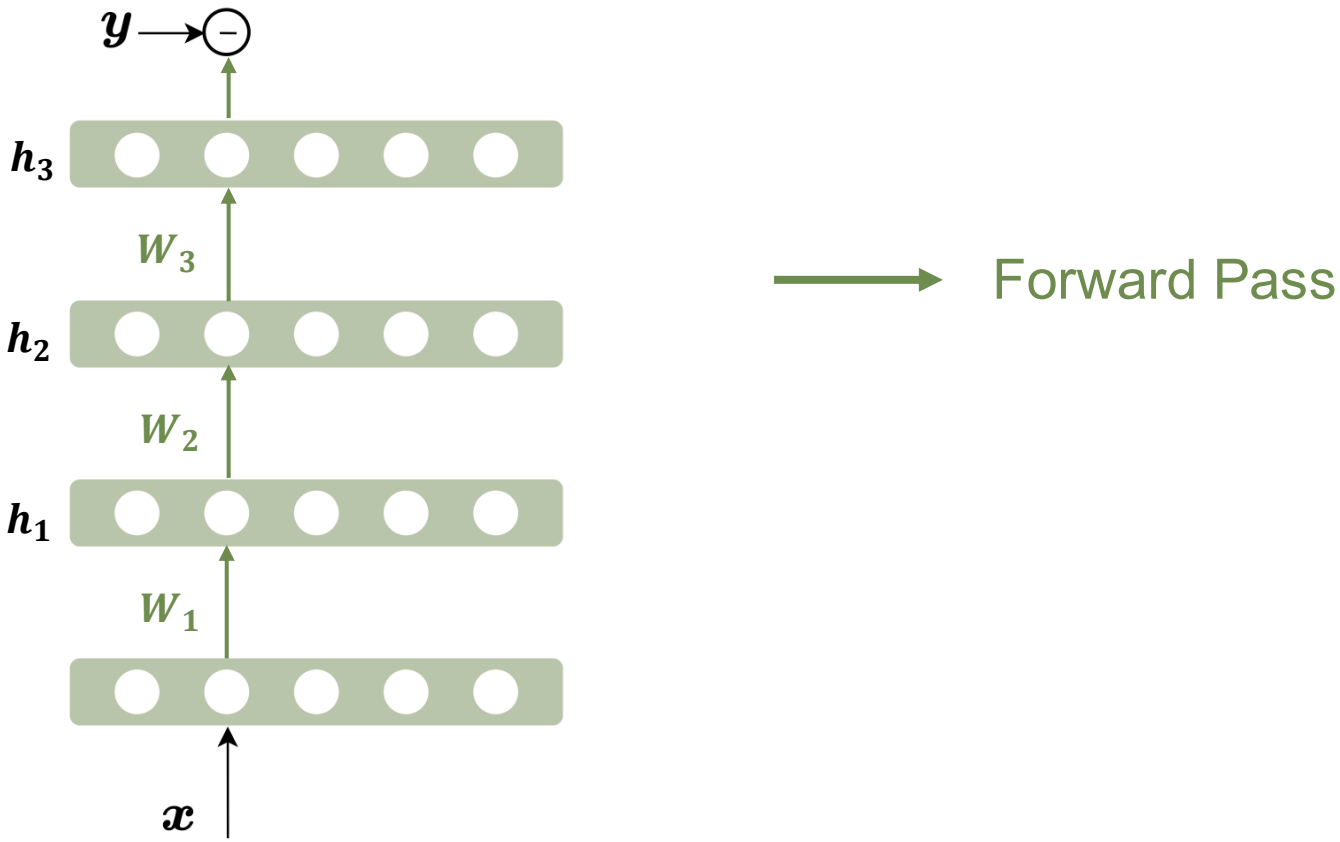
https://robodk.com/blog/robodks-virtual-assistant/neuralnetwork-training/

The Process of Backpropagation



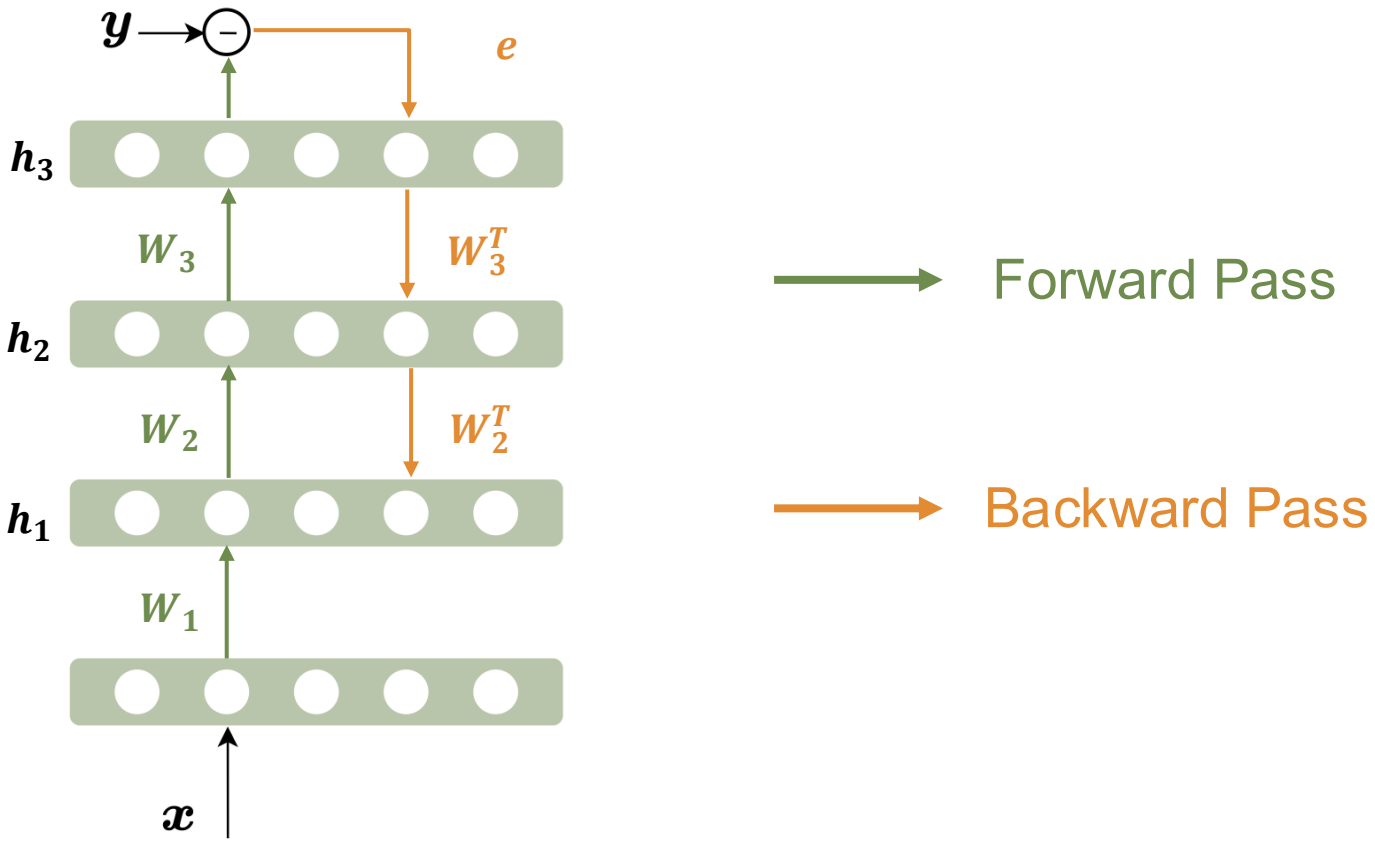
David E Rumelhart, Geoffrey E Hinton, et.al, “Learning representations by back-propagating errors,” Nature, 1986.

The Process of Backpropagation



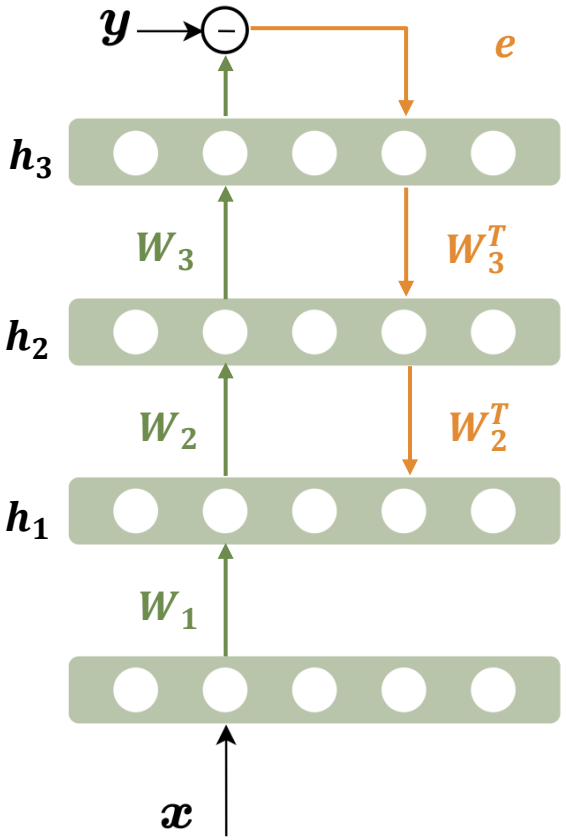
David E Rumelhart, Geoffrey E Hinton, et.al, “Learning representations by back-propagating errors,” Nature, 1986.

The Process of Backpropagation



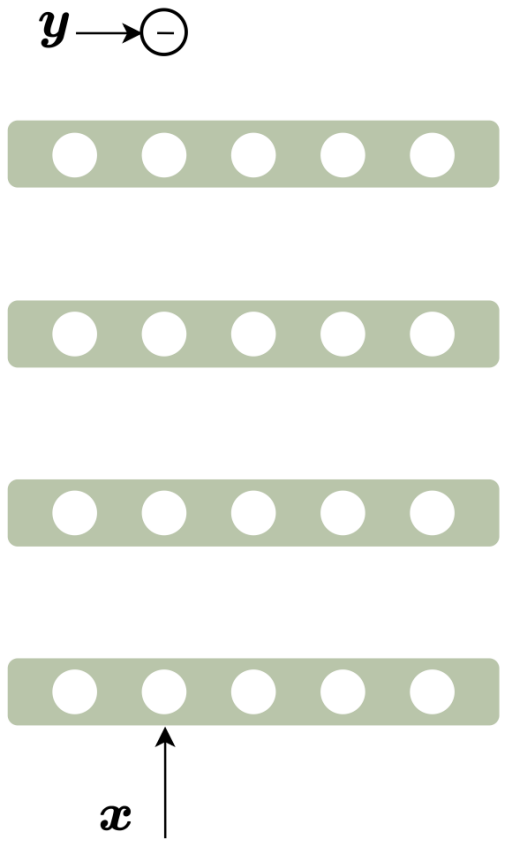
David E Rumelhart, Geoffrey E Hinton, et.al, “Learning representations by back-propagating errors,” Nature, 1986.

The Biological Implausibility of BP



David E Rumelhart, Geoffrey E Hinton, et.al, “Learning representations by back-propagating errors,” Nature, 1986.

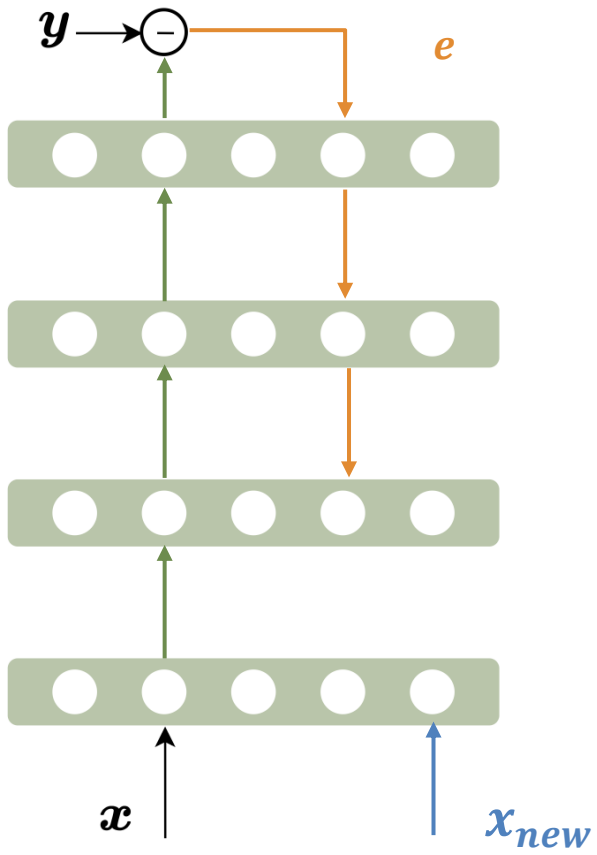
The Biological Implausibility of BP



Locking

David E Rumelhart, Geoffrey E Hinton, et.al, “Learning representations by back-propagating errors,” Nature, 1986.

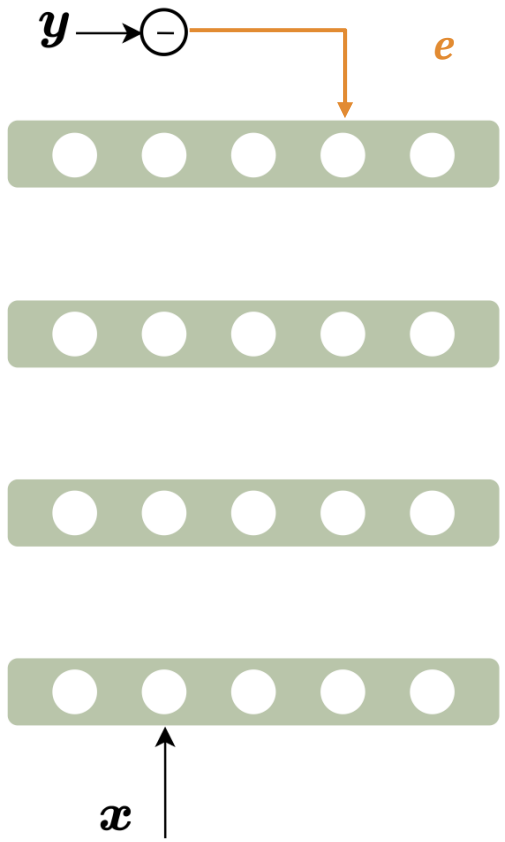
The Biological Implausibility of BP



Locking

David E Rumelhart, Geoffrey E Hinton, et.al, “Learning representations by back-propagating errors,” Nature, 1986.

The Biological Implausibility of BP

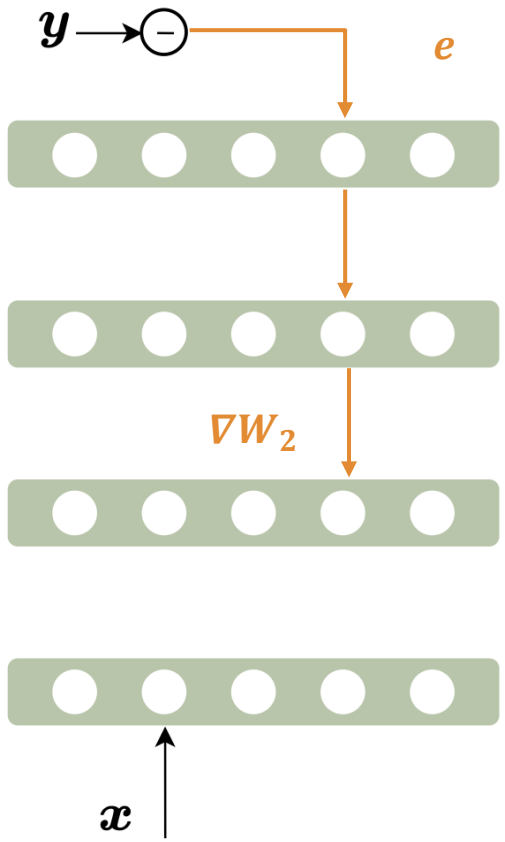


Locking

Non-Locality

David E Rumelhart, Geoffrey E Hinton, et.al, “Learning representations by back-propagating errors,” Nature, 1986.

The Biological Implausibility of BP

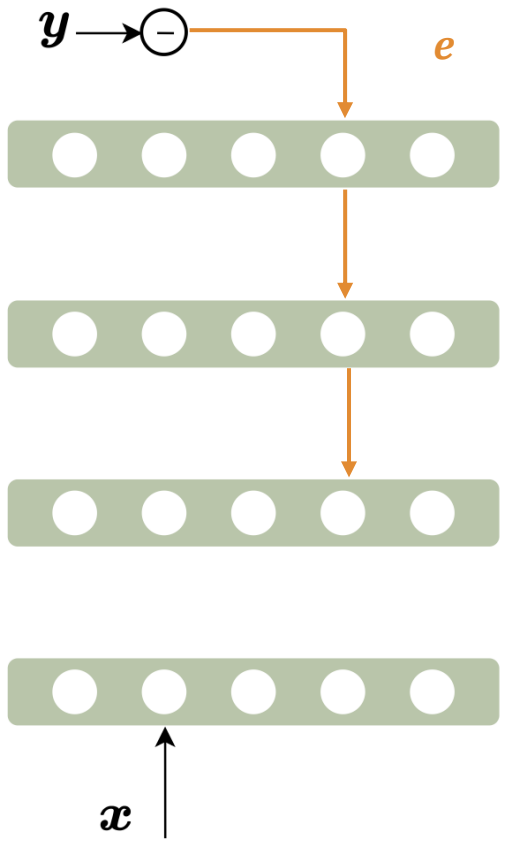


Locking

Non-Locality

David E Rumelhart, Geoffrey E Hinton, et.al, “Learning representations by back-propagating errors,” Nature, 1986.

The Biological Implausibility of BP



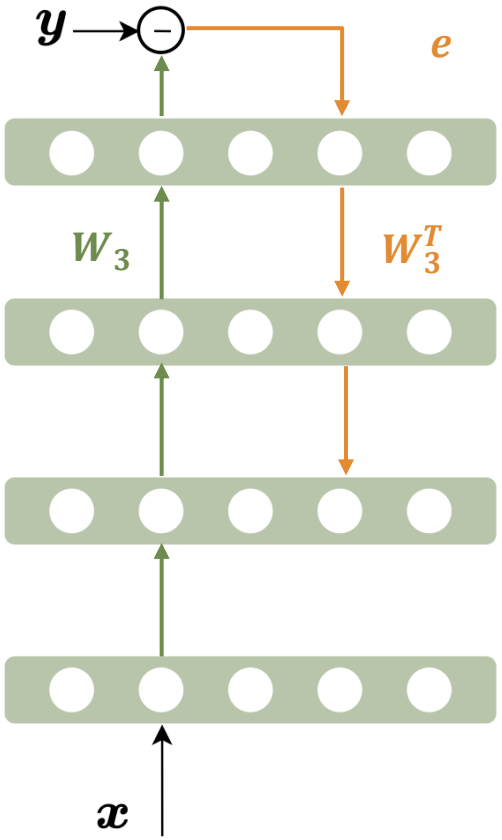
Locking

Non-Locality

Weight Transport

David E Rumelhart, Geoffrey E Hinton, et.al, “Learning representations by back-propagating errors,” Nature, 1986.

The Biological Implausibility of BP



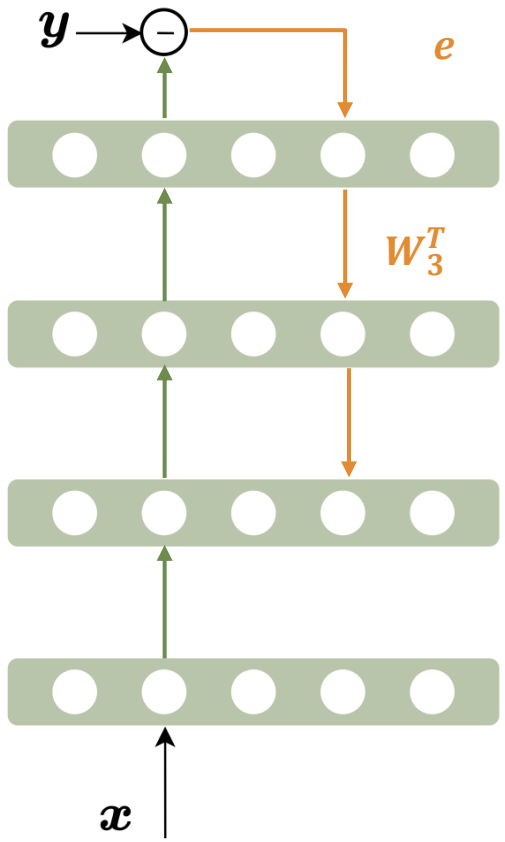
Locking

Non-Locality

Weight Transport

David E Rumelhart, Geoffrey E Hinton, et.al, “Learning representations by back-propagating errors,” Nature, 1986.

The Biological Implausibility of BP



Locking

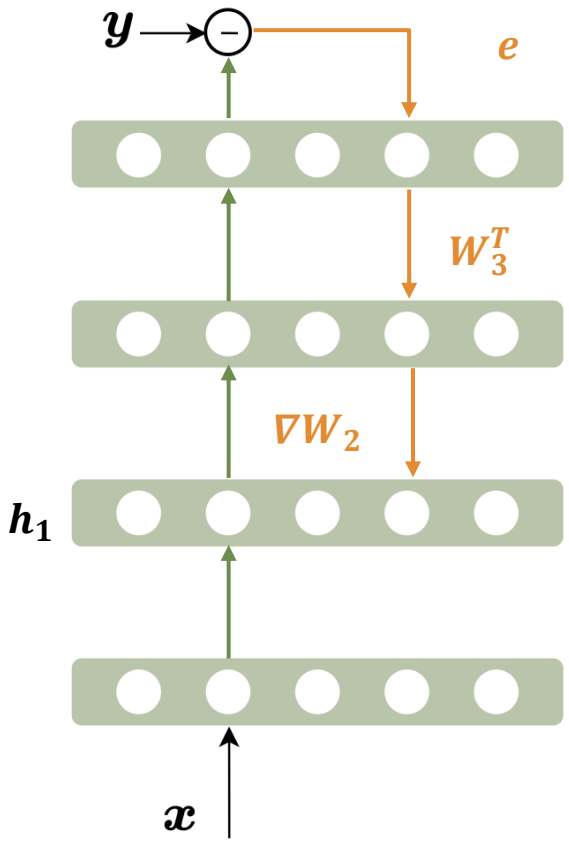
Non-Locality

Weight Transport

Frozen Activities

David E Rumelhart, Geoffrey E Hinton, et.al, “Learning representations by back-propagating errors,” Nature, 1986.

The Biological Implausibility of BP



Locking

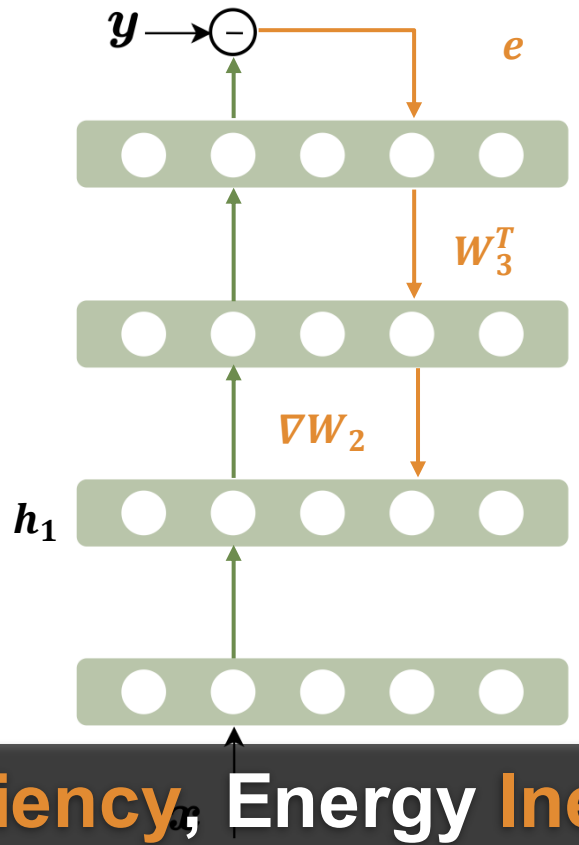
Non-Locality

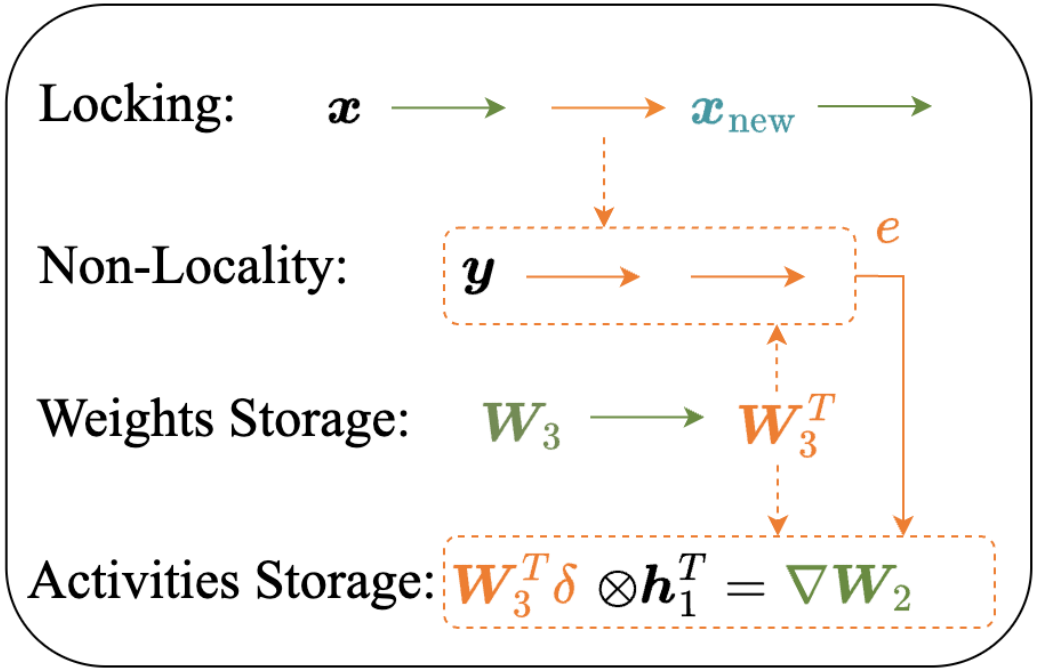
Weight Transport

Frozen Activities

David E Rumelhart, Geoffrey E Hinton, et.al, “Learning representations by back-propagating errors,” Nature, 1986.

The Biological Implausibility of BP





Memory Inefficiency, Energy Inefficiency, Computational Inefficiency

David E Rumelhart, Geoffrey E Hinton, et.al, “Learning representations by back-propagating errors,” Nature, 1986.

Biologically Plausible Alternatives



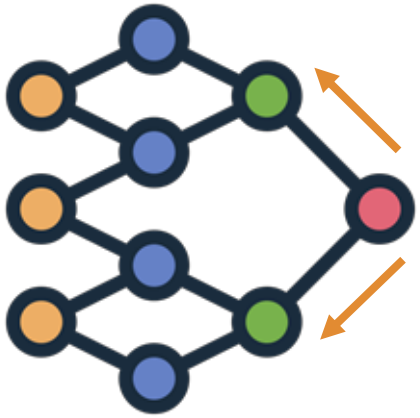
Human Brain
(~20 Watts)

G. Hinton. The forward-forward algorithm: Some preliminary investigations, 2022.

Biologically Plausible Alternatives



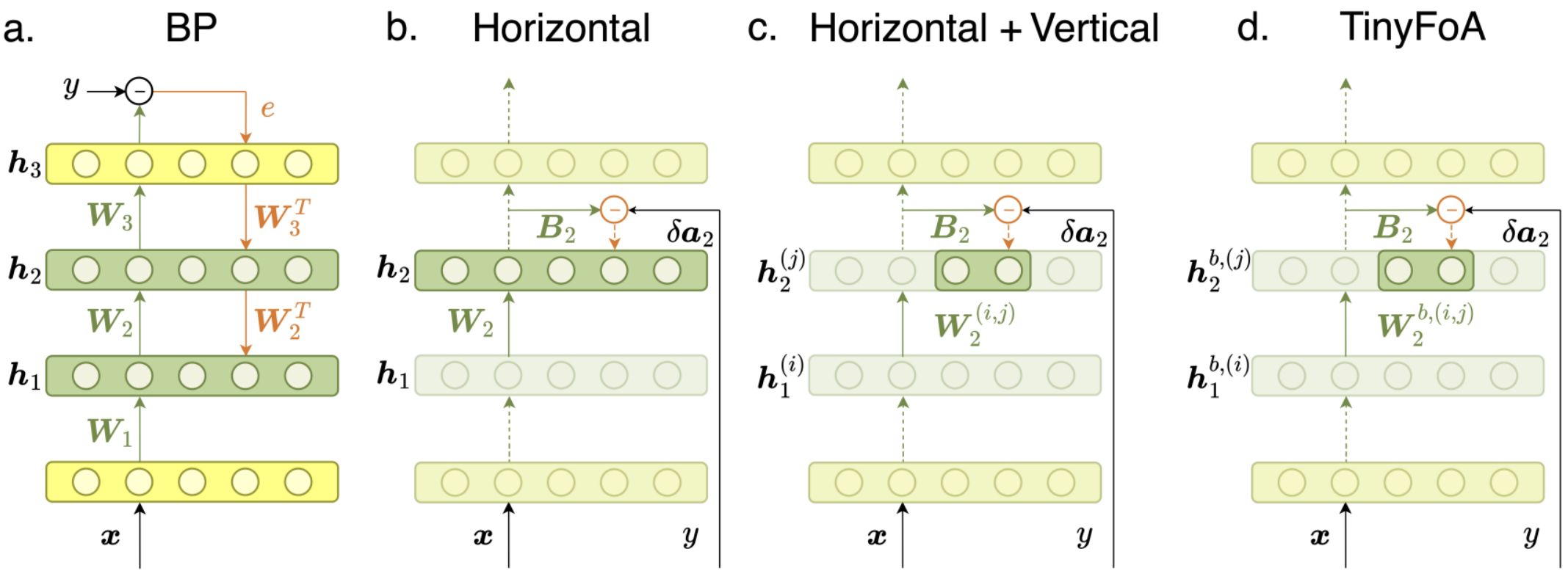
Human Brain
(~20 Watts)



Back-Propagation
(Bio-Implausible)

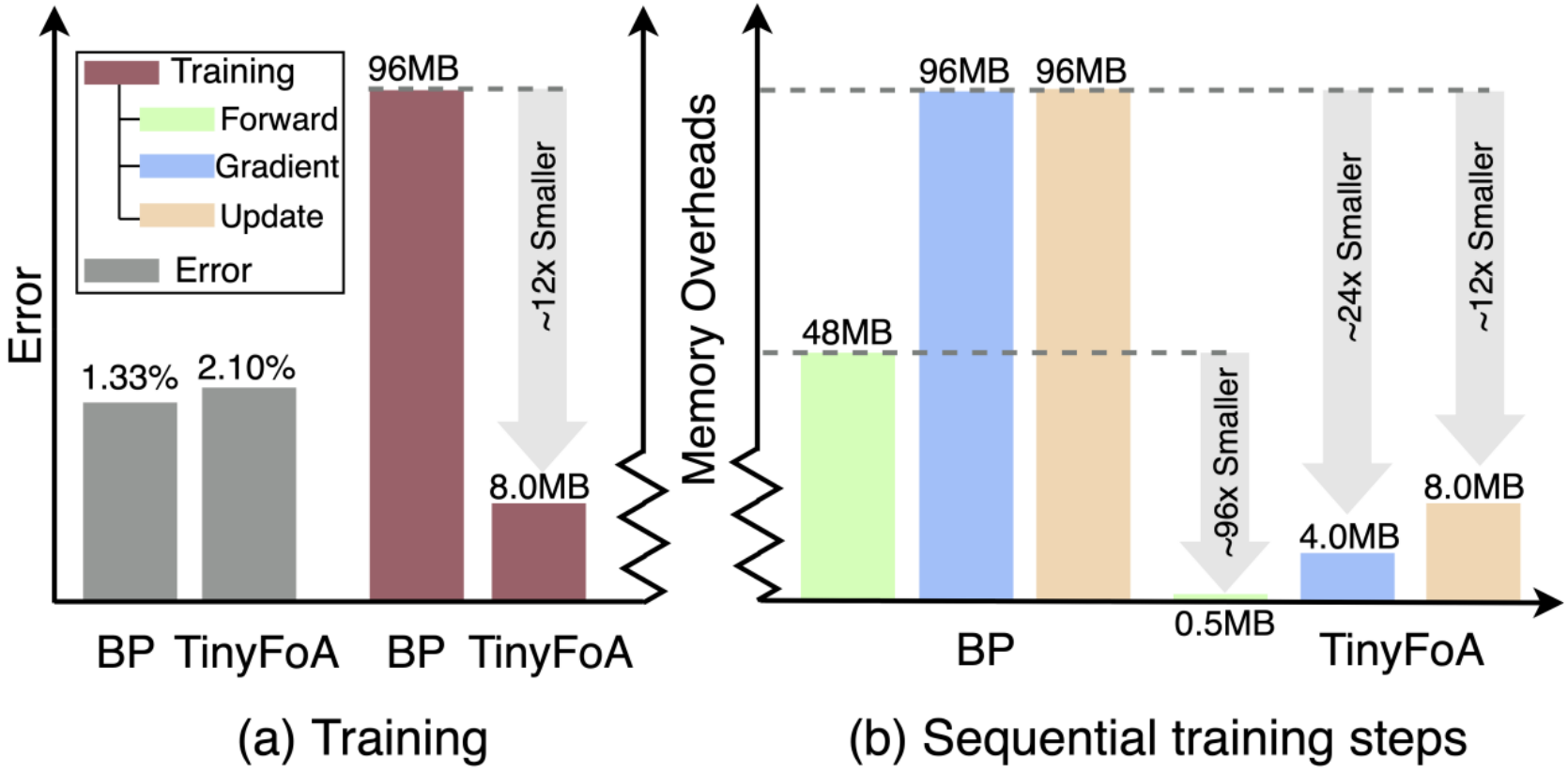
G. Hinton. The forward-forward algorithm: Some preliminary investigations, 2022.

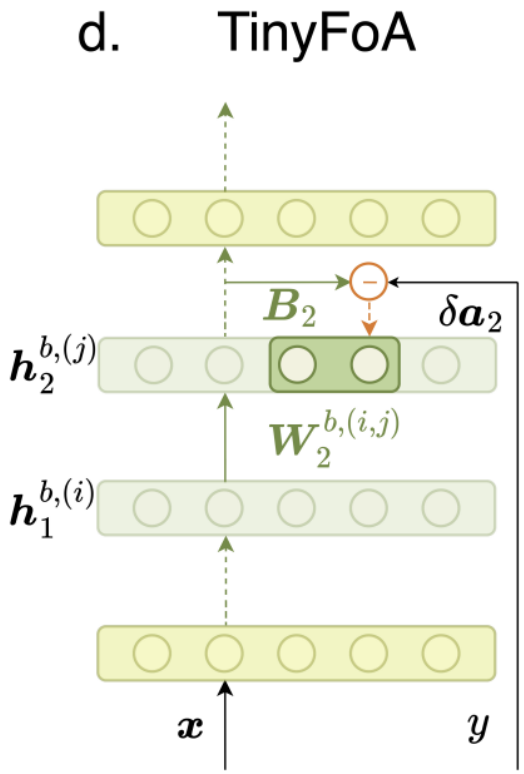
TinyFoA



Huang, Baichuan, and Amir Aminifar. "TinyFoA: Memory Efficient Forward-Only Algorithm for On-Device Learning." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 39. No. 16. 2025.

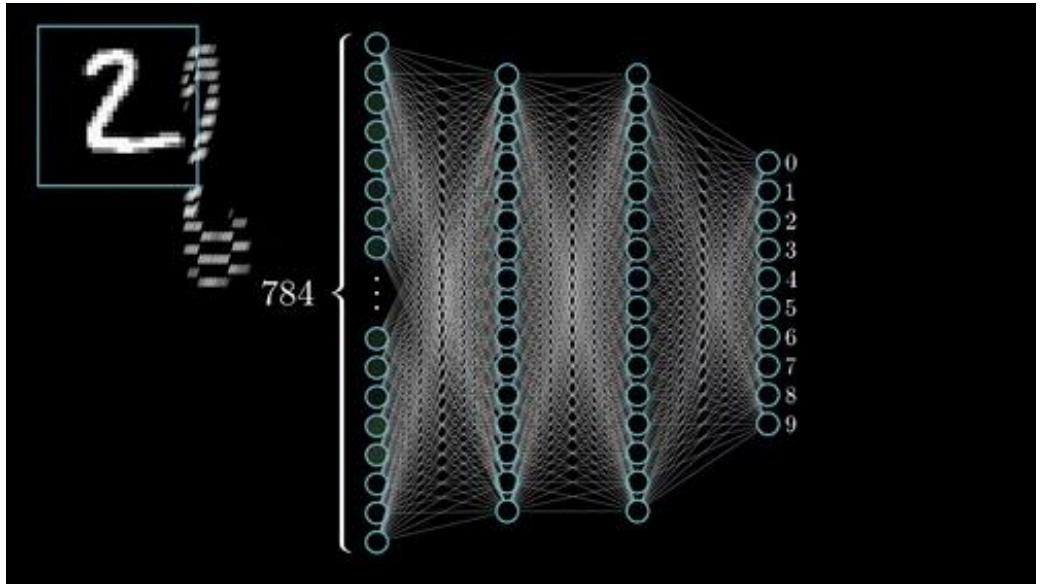
TinyFoA





Huang, Baichuan, and Amir Aminifar. "TinyFoA: Memory Efficient Forward-Only Algorithm for On-Device Learning." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 39. No. 16. 2025.

Quick Recap of Neural Network Layers in Deep Learning

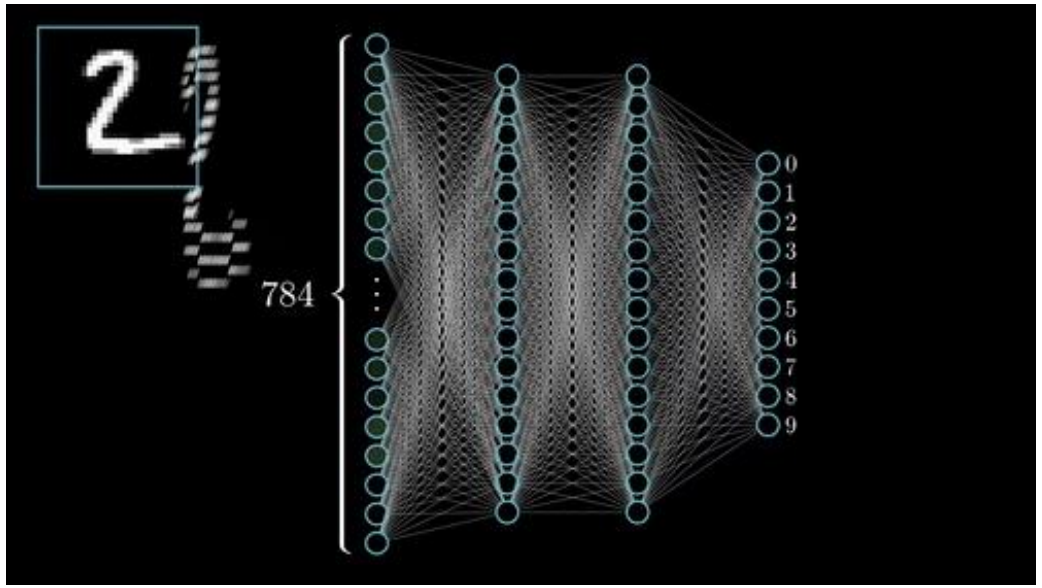


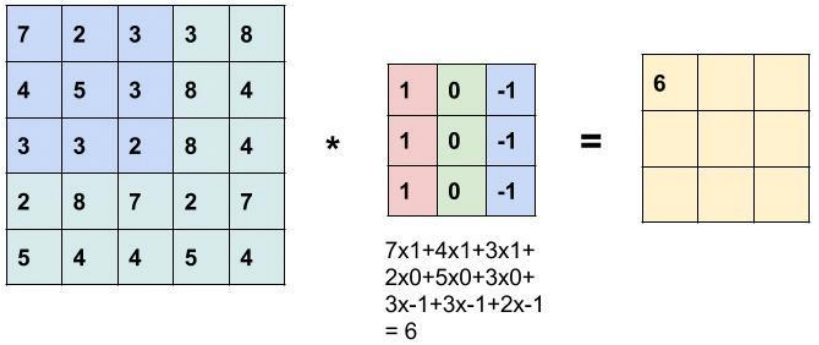
Fully Connected (FC)

$$y_j = \sum_i (W_{ij} \cdot x_i) + b_j$$

https://youtube.com/playlist?list=PLZHQObOWTQDNU6R1_67000Dx_ZCJB-3pi&si=2MkQ_kFeVhhb8yh0

Quick Recap of Neural Network Layers in Deep Learning





Fully Connected (FC)

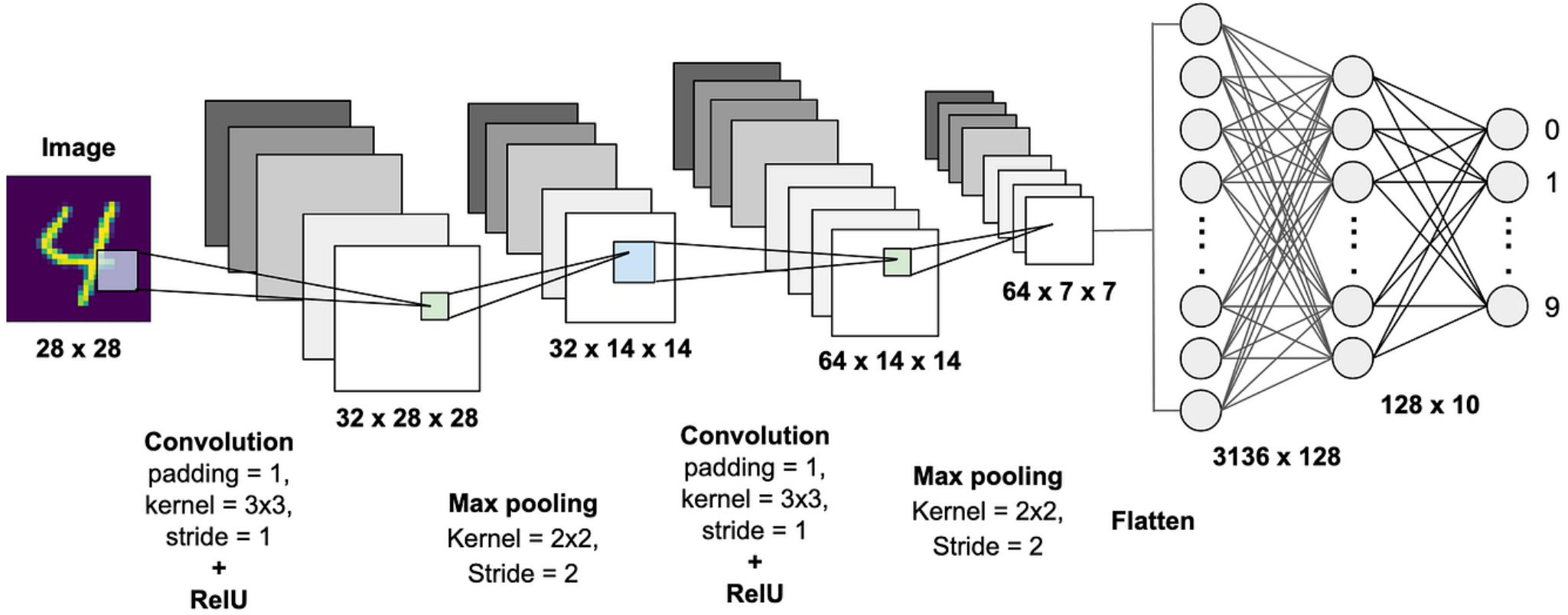
$$y_j = \sum_i (W_{ij} \cdot x_i) + b_j$$

Convolutional (CNN)

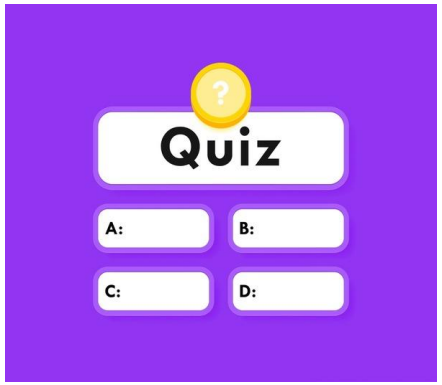
$$y_{ij} = \sum_{w=1}^{W} \sum_{h=1}^{H} x(i+m, j+n) \cdot W_{mn}$$

https://youtube.com/playlist?list=PLZHQObOWTQDNU6R1_67000Dx_ZCJB-3pi&si=2MkQ_kFeVhhb8yh0

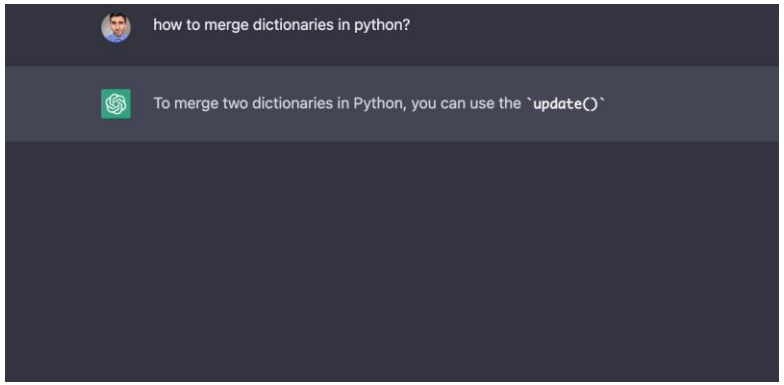
Real Application with FC and CNN



Large Language Models (LLMs)



Classification



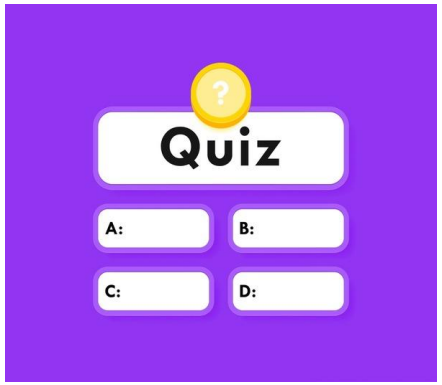
Generation

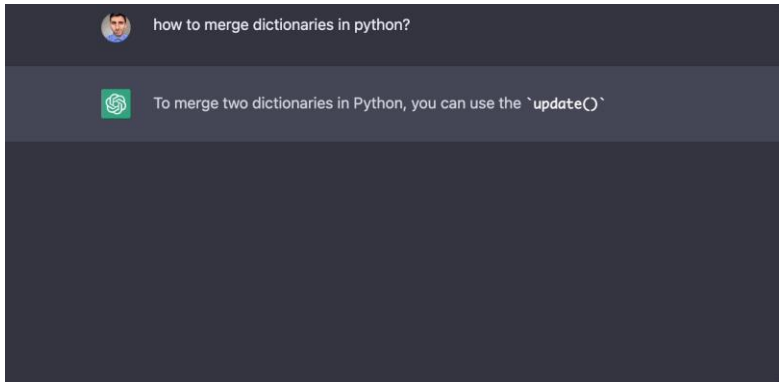


seq2seq

https://youtu.be/LPZh9BOjkQs?si=osdwVV73MbiIjff4

Large Language Models (LLMs)







Classification

Understanding

Generation

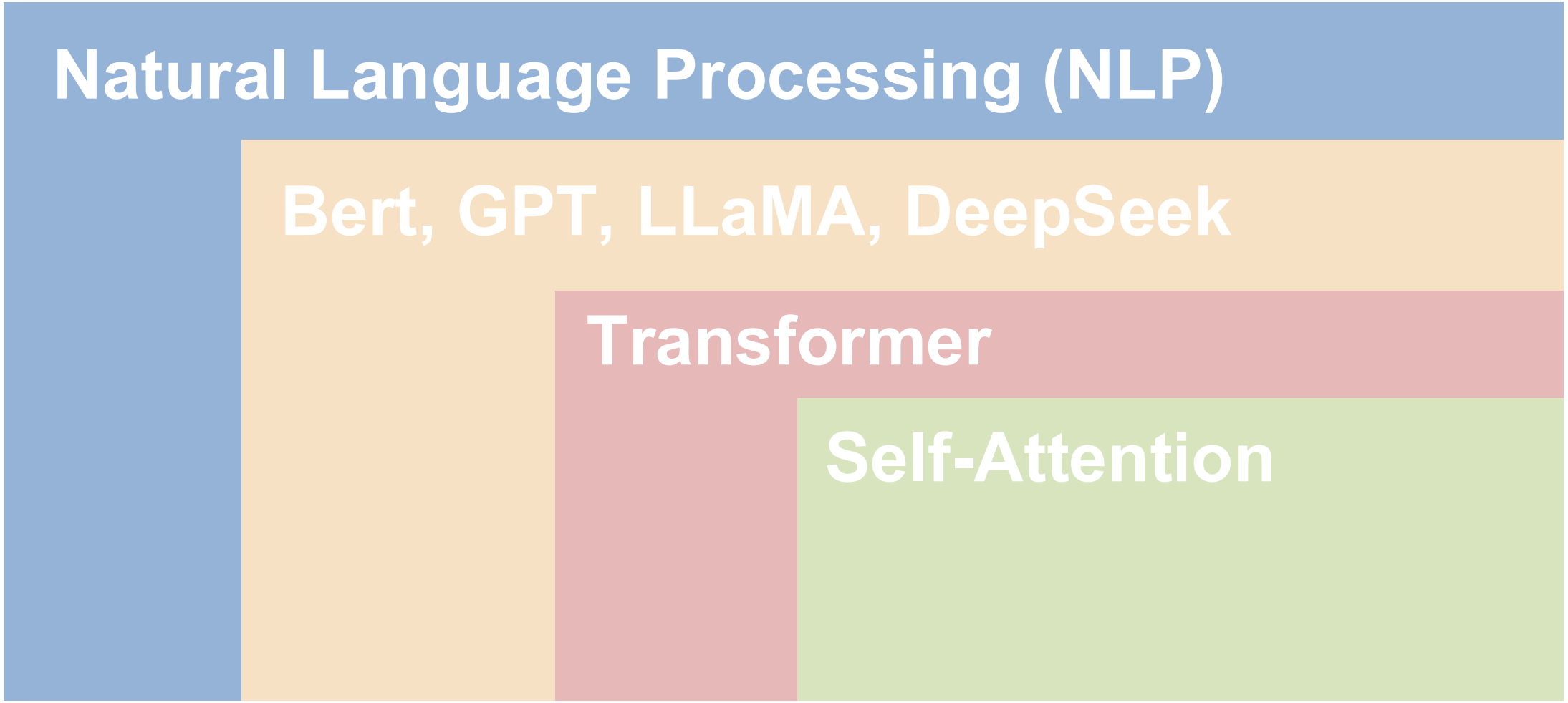
Dialogue/Coding

seq2seq

Translator

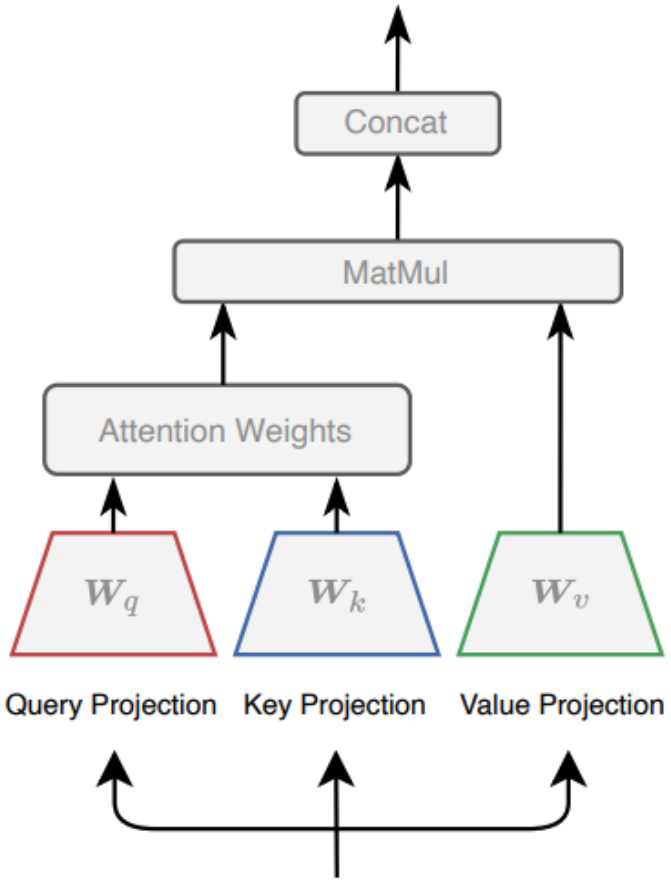
https://youtu.be/LPZh9BOjkQs?si=osdwVV73MbiIjff4

Domain Scope



Guo, Daya, et al. "Deepseek-r1 incentivizes reasoning in llms through reinforcement learning." *Nature* 645.8081 (2025): 633-638.

Self-Attention



$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Self-Attention (example)

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

I like football, but basketball more

https://sebastianraschka.com/blog/2023/self-attention-from-scratch.html

Self-Attention (example)

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

I like football, but basketball more

Token
I
like
football
,
but
basketball
more

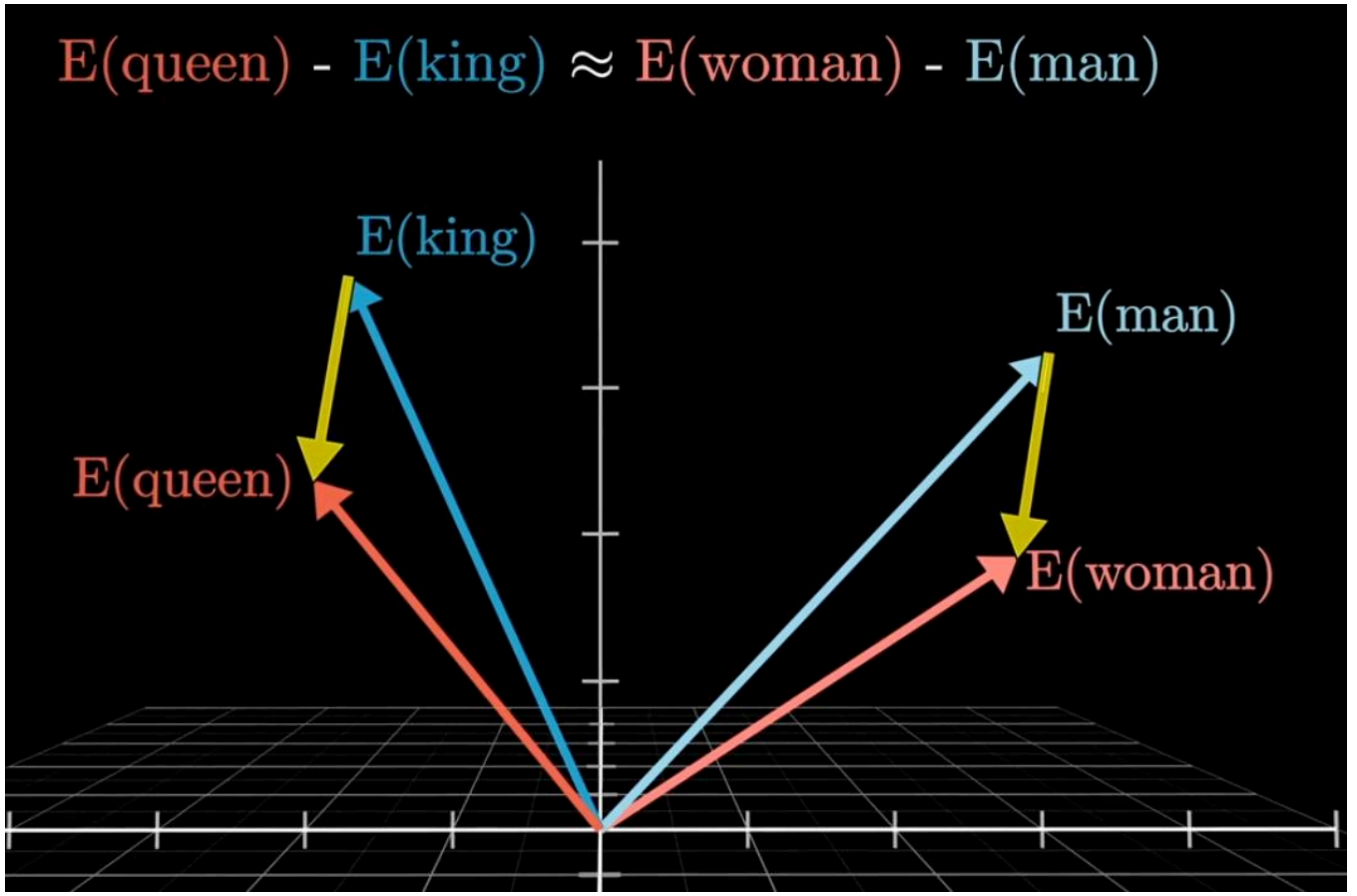
https://sebastianraschka.com/blog/2023/self-attention-from-scratch.html

Self-Attention (example)

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

I like football, but basketball more

Token	Embedding
I	$[0.1, 0.0]$
like	$[0.9, 0.1]$
football	$[0.8, 0.9]$
,	$[0.0, 0.0]$
but	$[0.2, 0.1]$
basketball	$[0.9, 0.8]$
more	$[0.4, 0.2]$



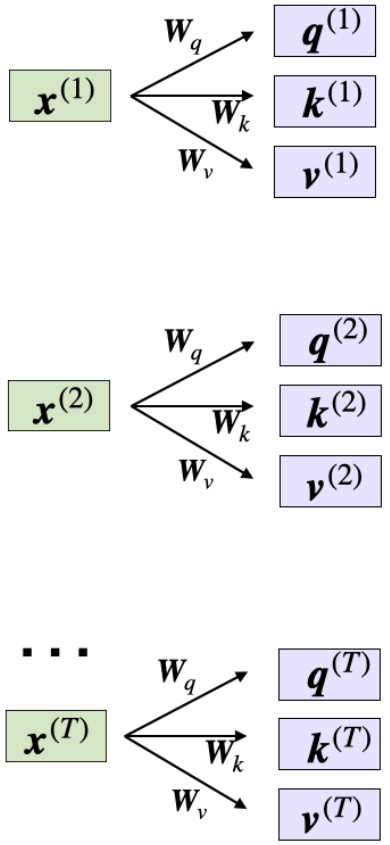
https://sebastianraschka.com/blog/2023/self-attention-from-scratch.html

Self-Attention (example)

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

I like football, but basketball more

Token	Embedding
I	$[0.1, 0.0]$
like	$[0.9, 0.1]$
football	$[0.8, 0.9]$
,	$[0.0, 0.0]$
but	$[0.2, 0.1]$
basketball	$[0.9, 0.8]$
more	$[0.4, 0.2]$



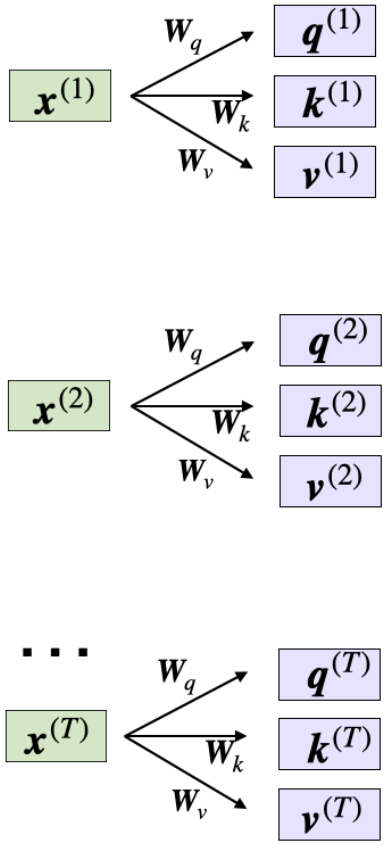
https://sebastianraschka.com/blog/2023/self-attention-from-scratch.html

Self-Attention (example)

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

I like football, but basketball more

Token	Embedding
I	$[0.1, 0.0]$
like	$[0.9, 0.1]$
football	$[0.8, 0.9]$
,	$[0.0, 0.0]$
but	$[0.2, 0.1]$
basketball	$[0.9, 0.8]$
more	$[0.4, 0.2]$



Query
What am I looking for?

Key
What is this token about?

Value
Here is the actual information

https://sebastianraschka.com/blog/2023/self-attention-from-scratch.html

Self-Attention (if $W_{q,k,v} = I$, We pick Q=like)

$$\text{softmax}(\frac{QK^T}{\sqrt{d_k}})$$

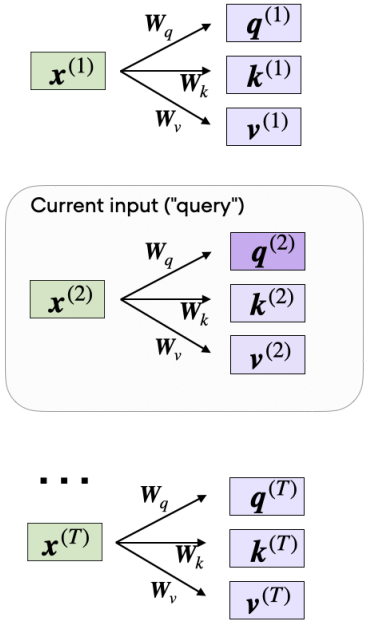
Token	Embedding
I	$[0.1, 0.0]$
like	$[0.9, 0.1]$
football	$[0.8, 0.9]$
,	$[0.0, 0.0]$
and	$[0.2, 0.1]$
basketball	$[0.9, 0.8]$
more	$[0.4, 0.2]$

https://sebastianraschka.com/blog/2023/self-attention-from-scratch.html

Self-Attention (if $W_{q,k,v} = I$, We pick Q=like)

$$\text{softmax}(\frac{QK^T}{\sqrt{d_k}})$$

Token	Embedding /Q/K/V
I	$[0.1, 0.0]$
like	$[0.9, 0.1]$
football	$[0.8, 0.9]$
,	$[0.0, 0.0]$
and	$[0.2, 0.1]$
basketball	$[0.9, 0.8]$
more	$[0.4, 0.2]$

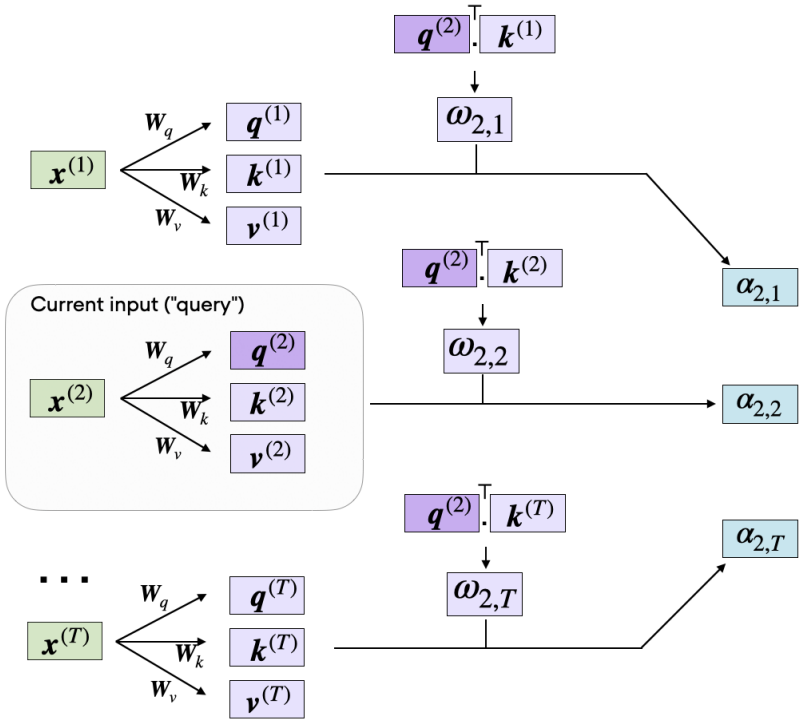


https://sebastianraschka.com/blog/2023/self-attention-from-scratch.html

Self-Attention (if $W_{q,k,v} = I$, We pick Q=like)

$$\text{softmax}(\frac{QK^T}{\sqrt{d_k}})$$

Token	Embedding /Q/K/V	$Q \cdot K^T$
I	$[0.9, 0.1]$	$\cdot [0.1, 0.0]^T = 0.09$
like	$[0.9, 0.1]$	
football	$[0.8, 0.9]$	
,	$[0.0, 0.0]$	
and	$[0.2, 0.1]$	
basketball	$[0.9, 0.8]$	
more	$[0.4, 0.2]$	

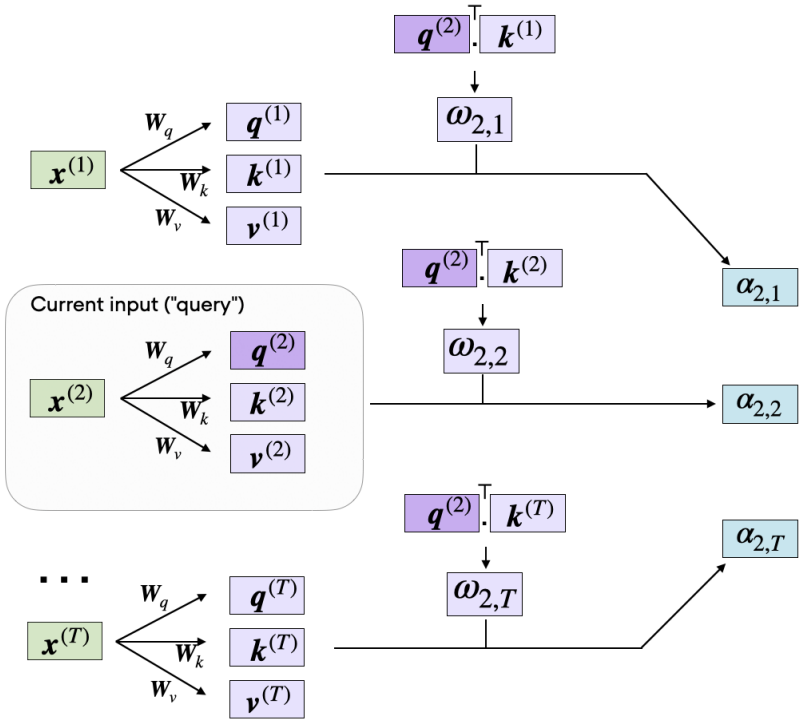


https://sebastianraschka.com/blog/2023/self-attention-from-scratch.html

Self-Attention (if $W_{q,k,v} = I$, We pick Q=like)

$$\text{softmax}(\frac{QK^T}{\sqrt{d_k}})$$

Token	Embedding /Q/K/V	$Q \cdot K^T$
I	$[0.9, 0.1] \cdot [0.1, 0.0]^T$	$= 0.09$
like	$[0.9, 0.1] \cdot [0.9, 0.1]^T$	0.82
football	$[0.9, 0.1] \cdot [0.8, 0.9]^T$	0.81
,	$[0.9, 0.1] \cdot [0.0, 0.0]^T$	0.00
and	$[0.9, 0.1] \cdot [0.2, 0.1]^T$	0.19
basketball	$[0.9, 0.1] \cdot [0.9, 0.8]^T$	0.89
more	$[0.9, 0.1] \cdot [0.4, 0.2]^T$	0.38

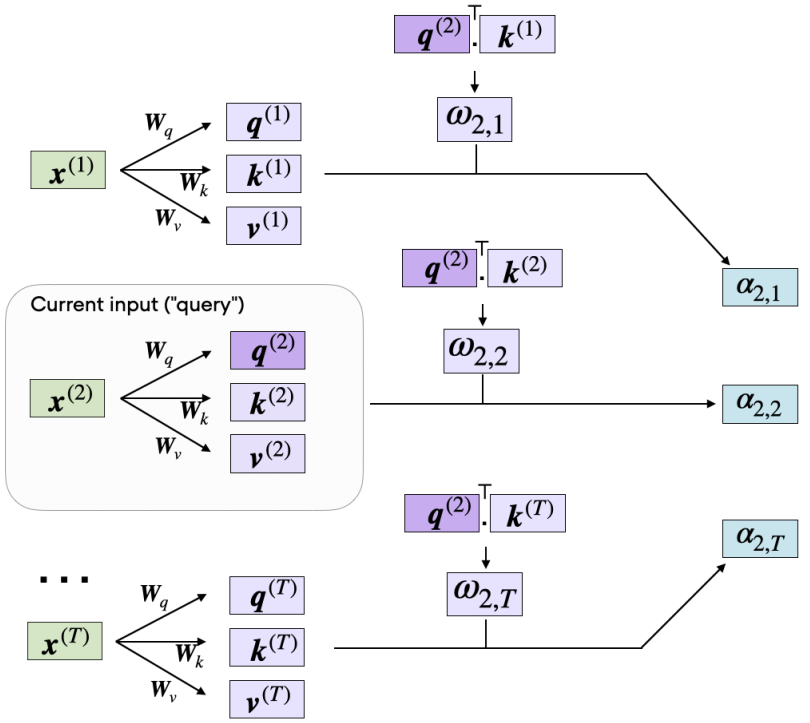


https://sebastianraschka.com/blog/2023/self-attention-from-scratch.html

Self-Attention (if $W_{q,k,v} = I$, We pick Q=like)

$$\text{softmax}(\frac{QK^T}{\sqrt{d_k}})$$

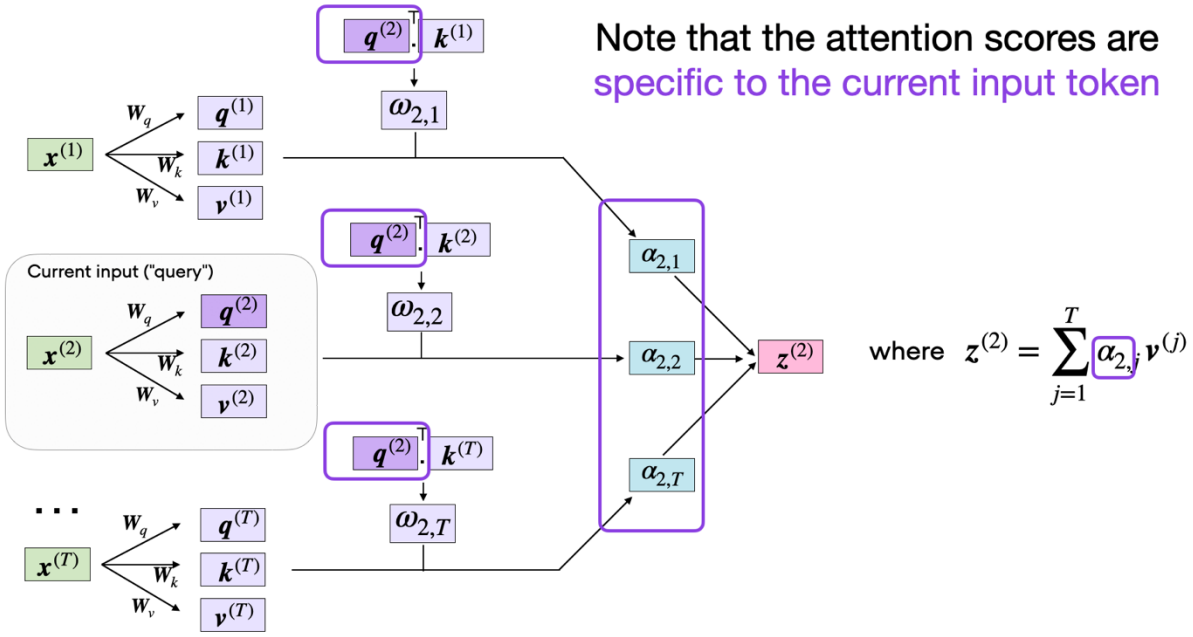
Token	Embedding /Q/K/V	$Q \cdot K^T$	softmax
I	$[0.9, 0.1] \cdot [0.1, 0.0]^T$	0.09	0.09
like	$[0.9, 0.1] \cdot [0.9, 0.1]^T$	0.82	0.19
football	$[0.9, 0.1] \cdot [0.8, 0.9]^T$	0.81	0.19
,	$[0.9, 0.1] \cdot [0.0, 0.0]^T$	0.00	0.08
and	$[0.9, 0.1] \cdot [0.2, 0.1]^T$	0.19	0.10
basketball	$[0.9, 0.1] \cdot [0.9, 0.8]^T$	0.89	0.20
more	$[0.9, 0.1] \cdot [0.4, 0.2]^T$	0.38	0.12



https://sebastianraschka.com/blog/2023/self-attention-from-scratch.html

Self-Attention

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

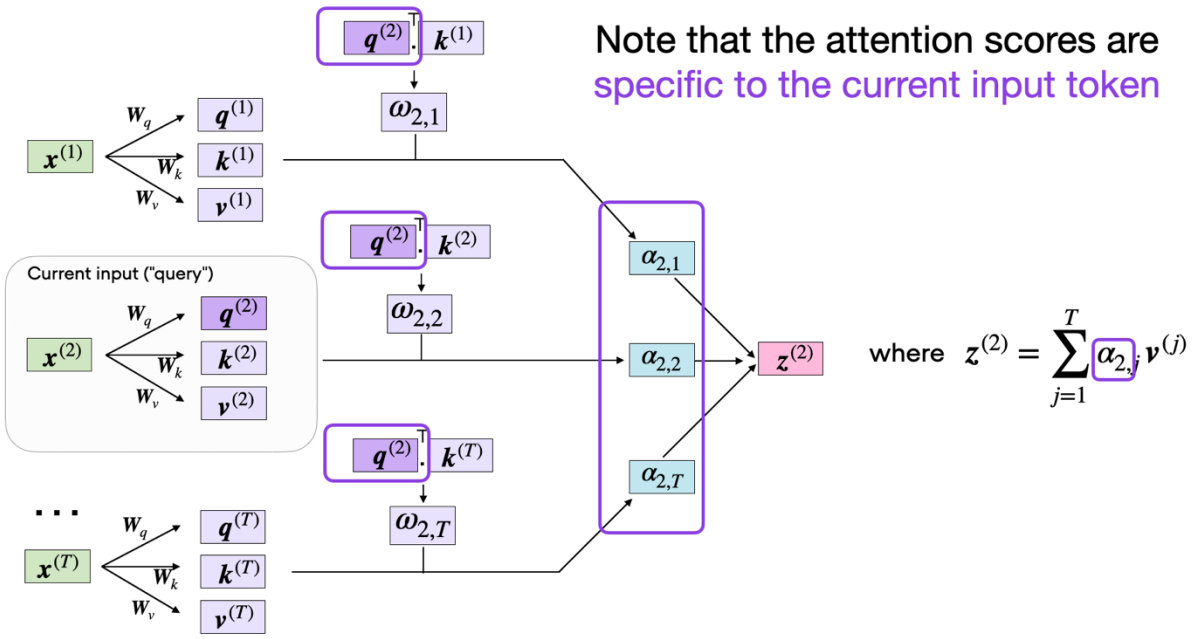


Specific Query

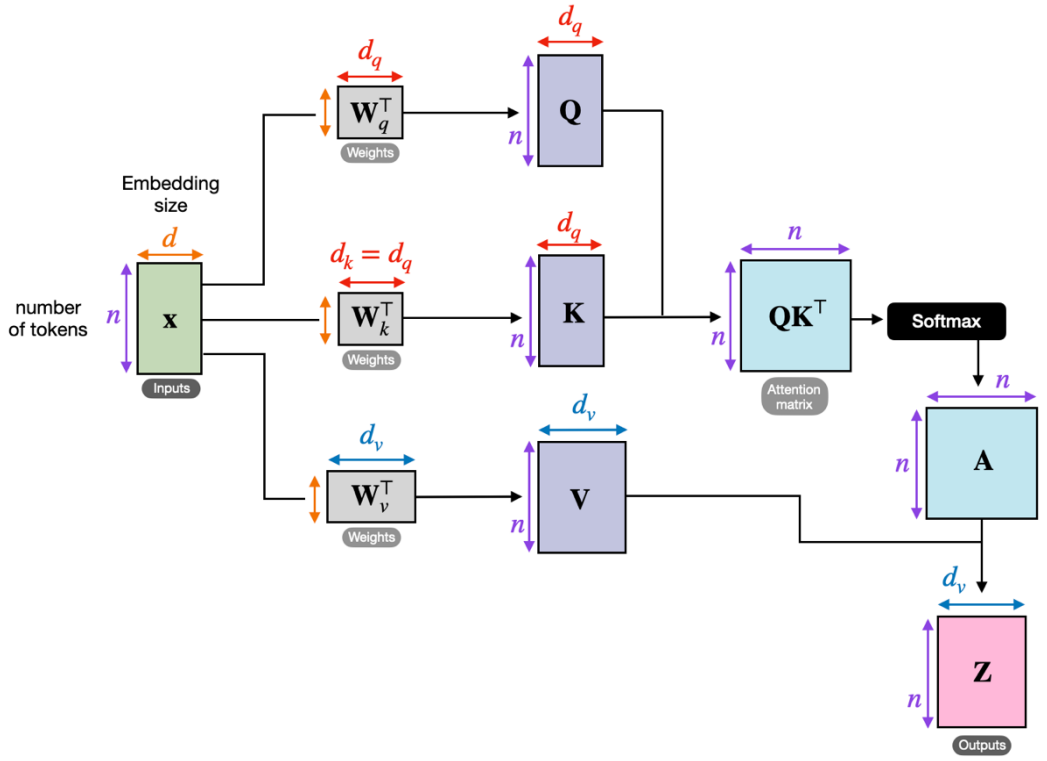
https://sebastianraschka.com/blog/2023/self-attention-from-scratch.html

Self-Attention

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$



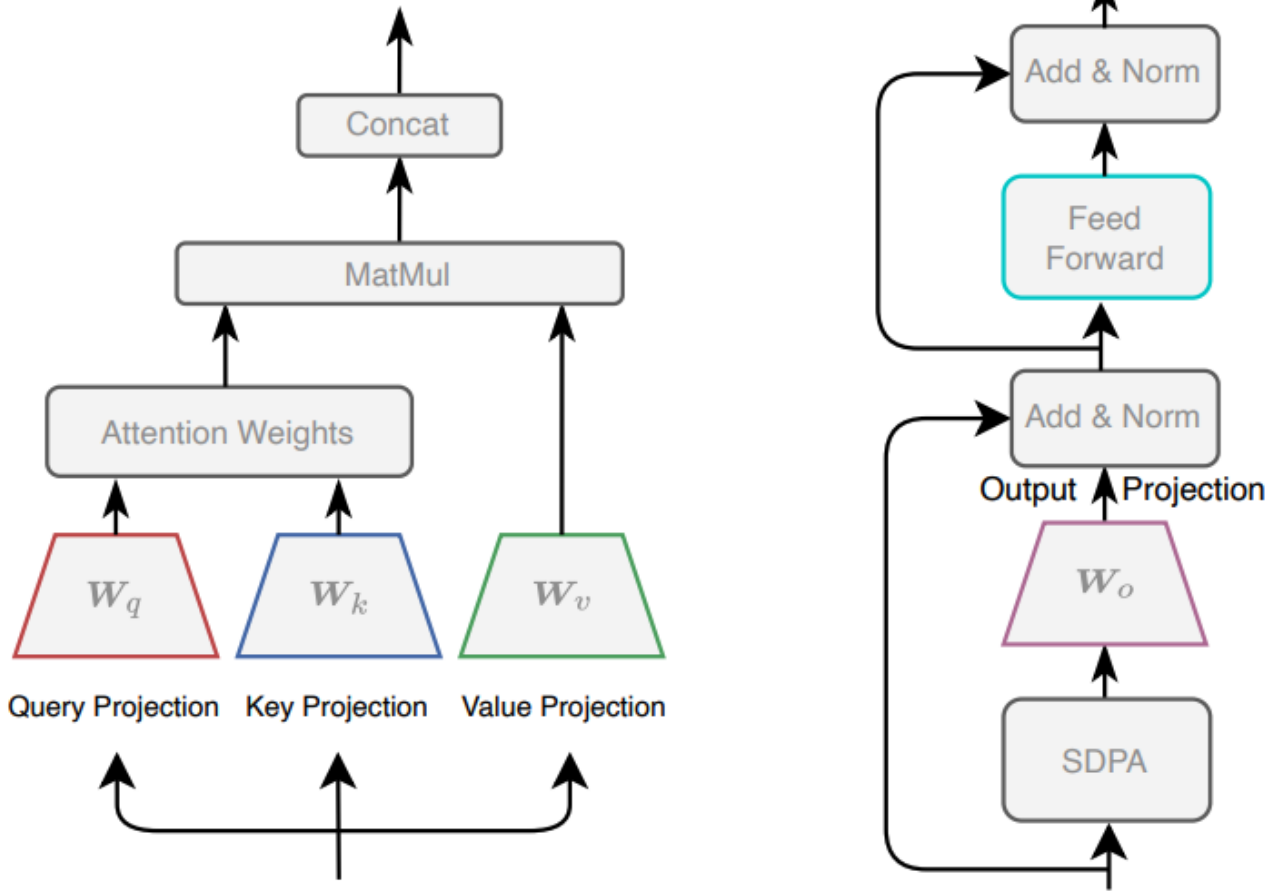
Specific Query



Parallel All Tokens

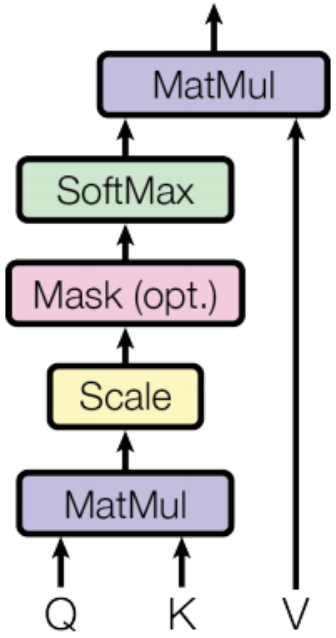
https://sebastianraschka.com/blog/2023/self-attention-from-scratch.html

One Transformer Layer



Attention to Transformer

Self-Attention

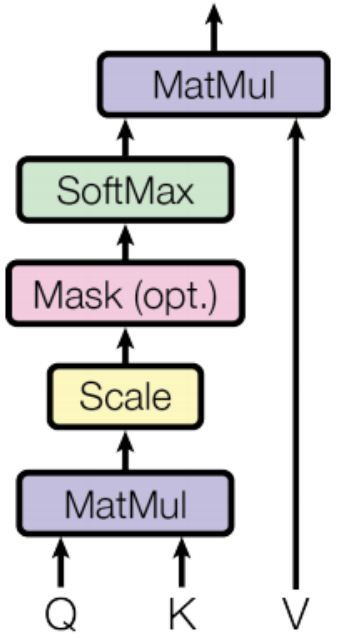


$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).
https://poloclub.github.io/transformer-explainer/

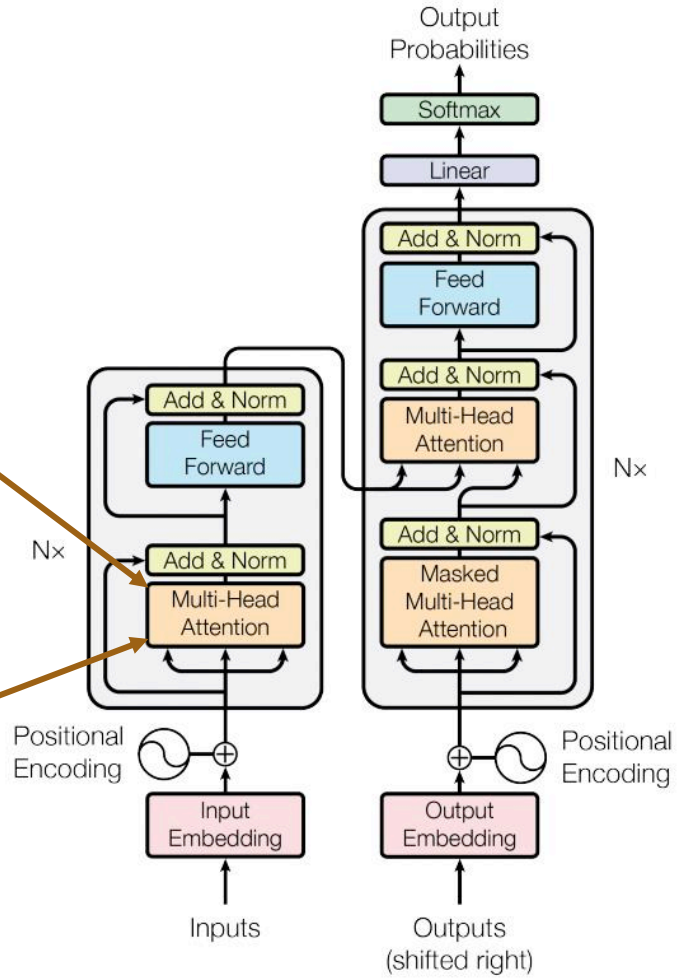
Attention to Transformer

Self-Attention



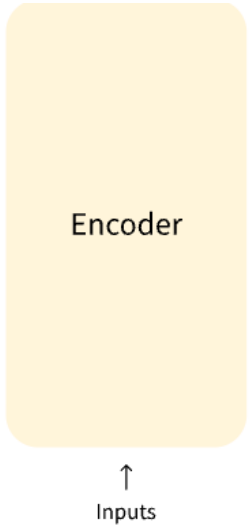
$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Transformer

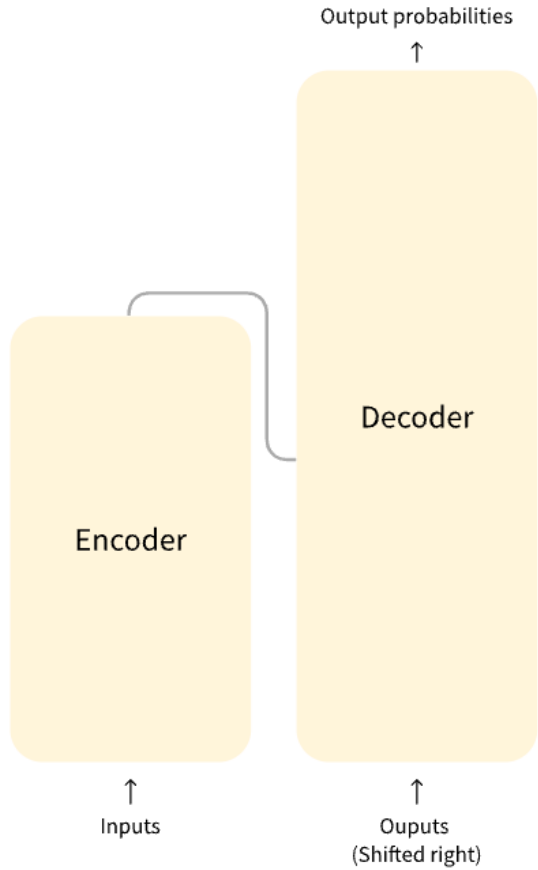


Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).
https://poloclub.github.io/transformer-explainer/

LLM Architectures







Classification

Bert, RoBERTa

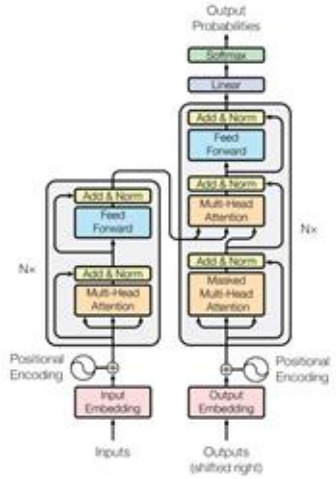
Generation

GPT, LLaMA

seq2seq

T5

Environmental Impact of Training Transformer

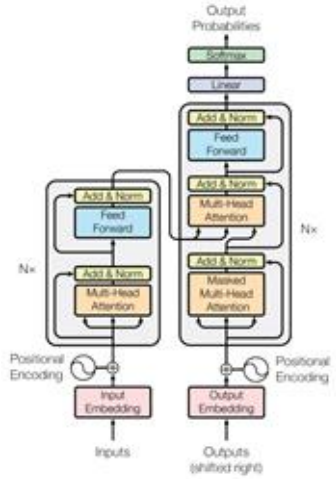


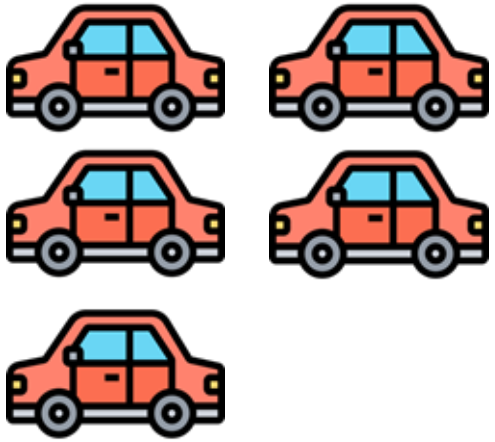
Training Transformer (Strubell E. 2020)

CO_2 626,155 lbs

Strubell E, et al. Energy and policy considerations for modern deep learning research. AAAI, 2020.
Vaswani A. Attention is all you need. NeurIPS, 2017.
https://www.forbes.com/sites/robtoews/2020/06/17/deep-learnings-climate-change-problem/

Environmental Impact of Training Transformer





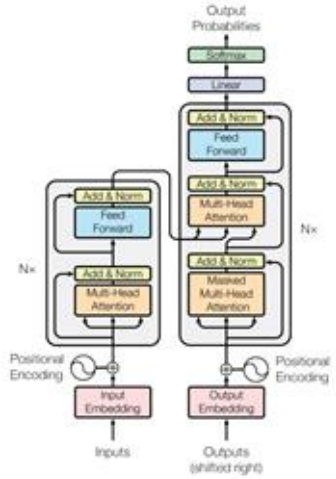
Training Transformer (Strubell E. 2020)

Total Lifetime of a Car

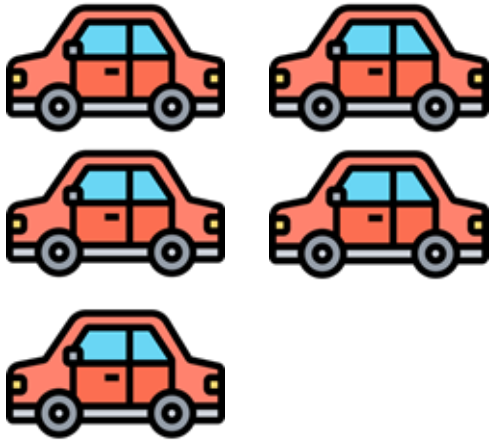
CO_2 626,155 lbs = 5×126,000 lbs

Strubell E, et al. Energy and policy considerations for modern deep learning research. AAAI, 2020.
Vaswani A. Attention is all you need. NeurIPS, 2017.
https://www.forbes.com/sites/robtoews/2020/06/17/deep-learnings-climate-change-problem/

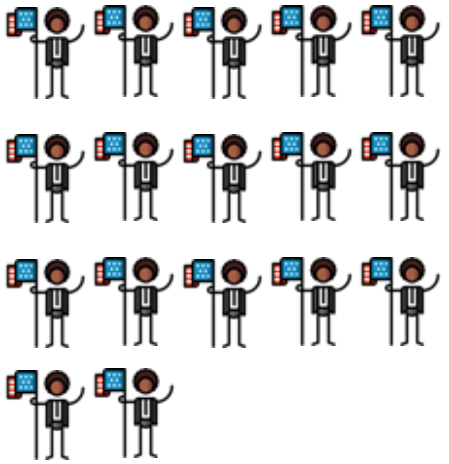
Environmental Impact of Training Transformer



Training Transformer (Strubell E. 2020)



Total Lifetime of a Car

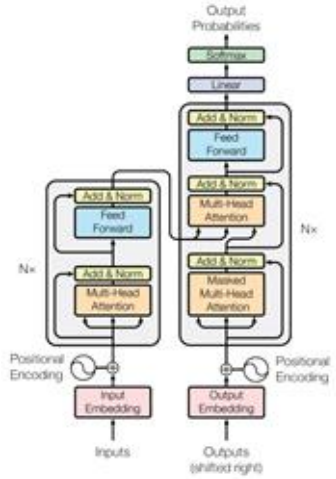


Average American in a Year

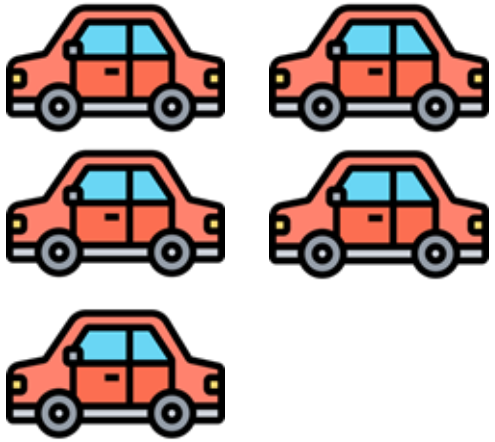
CO_2 626,155 lbs = 5×126,000 lbs = 17×36,156 lbs

Strubell E, et al. Energy and policy considerations for modern deep learning research. AAAI, 2020.
Vaswani A. Attention is all you need. NeurIPS, 2017.
https://www.forbes.com/sites/robtoews/2020/06/17/deep-learnings-climate-change-problem/

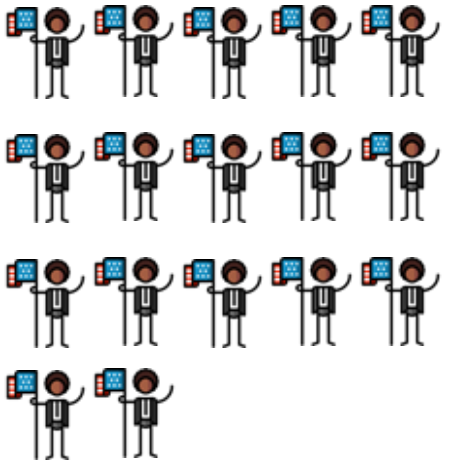
Environmental Impact of Training Transformer



Training Transformer (Strubell E. 2020)



Total Lifetime of a Car



Average American in a Year

CO_2 626,155 lbs = 5×126,000 lbs = 17×36,156 lbs

The computational resources needed to produce a best-in-class AI model has on average doubled every 3.4 months.

Strubell E, et al. Energy and policy considerations for modern deep learning research. AAAI, 2020.
Vaswani A. Attention is all you need. NeurIPS, 2017.
https://www.forbes.com/sites/robtoews/2020/06/17/deep-learnings-climate-change-problem/

Energy Consumption of Training LLMs



GPT-3



GPT-4

D. Patterson, et al. Carbon emissions and large neural network training, 2021.
https://tinyml.substack.com/p/the-carbon-impact-of-large-language
Data sources: U.S. Energy Information Administration, Electric Power Research Institute (EPRI)

Energy Consumption of Training LLMs

	GPT-3		GPT-4
CO_2	1,216,950 lbs	×13	15,238,333 lbs

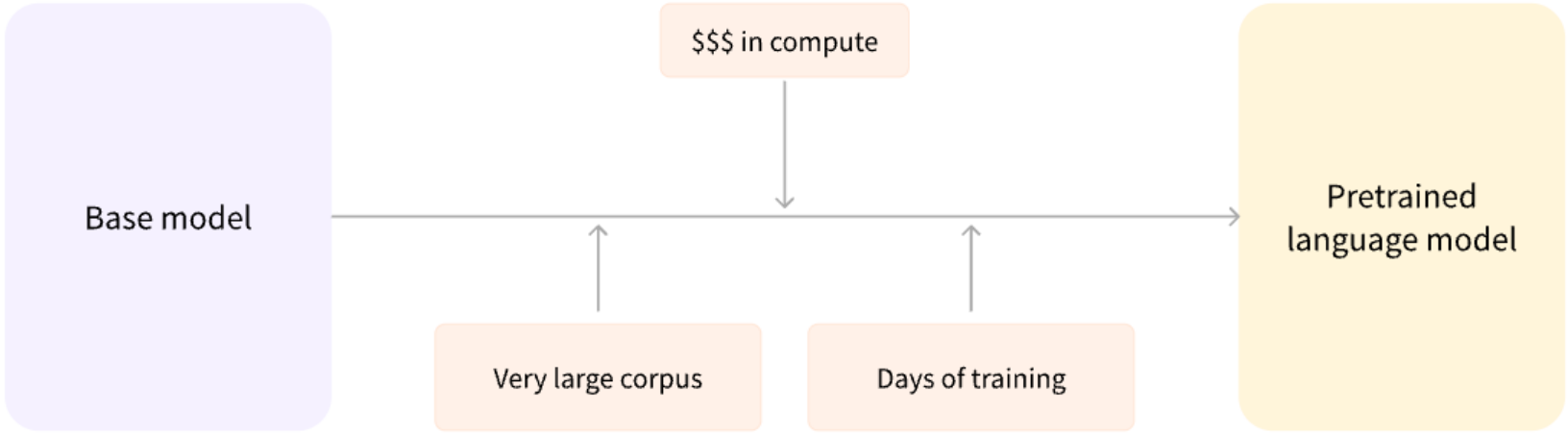
D. Patterson, et al. Carbon emissions and large neural network training, 2021.
https://tinyml.substack.com/p/the-carbon-impact-of-large-language
Data sources: U.S. Energy Information Administration, Electric Power Research Institute (EPRI)

Energy Consumption of Training LLMs

	GPT-3		GPT-4
CO_2	1,216,950 lbs	×13	15,238,333 lbs
	1,287 Megawatt-Hour	×48	62,318 Megawatt-Hour

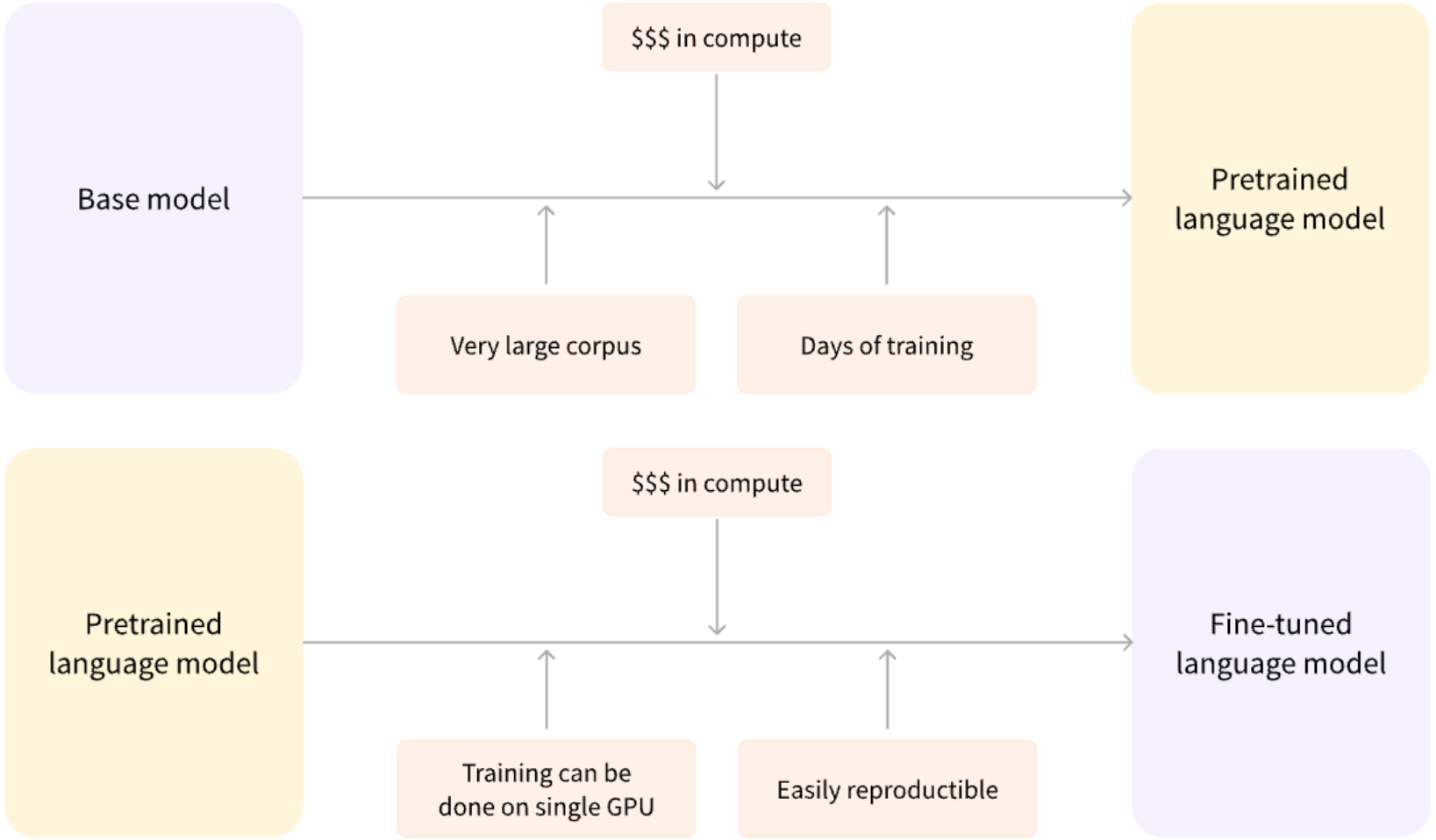
D. Patterson, et al. Carbon emissions and large neural network training, 2021.
https://tinyml.substack.com/p/the-carbon-impact-of-large-language
Data sources: U.S. Energy Information Administration, Electric Power Research Institute (EPRI)

Pre-Trained LLMs to Task Adaptation



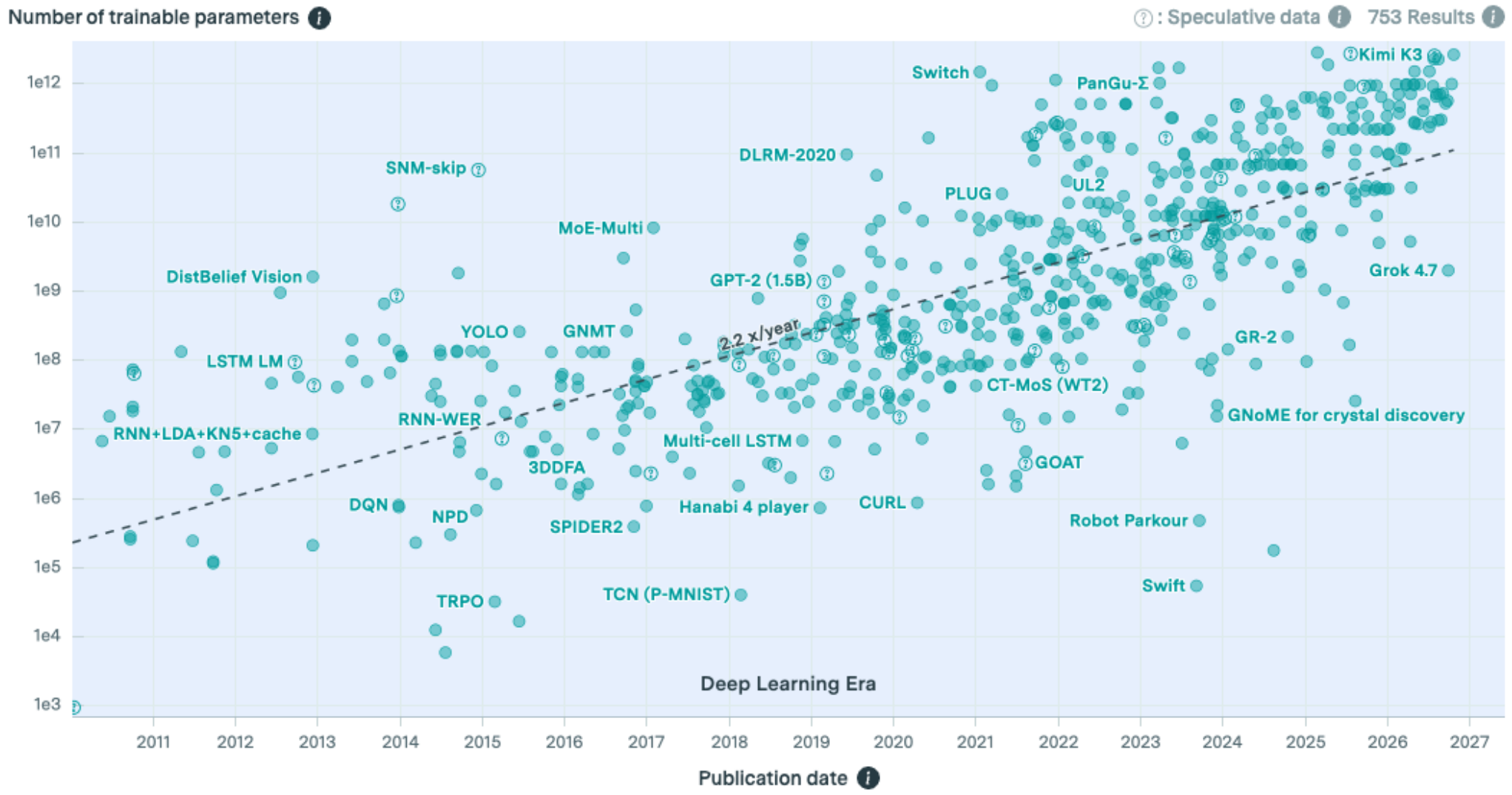
https://huggingface.co/learn/llm-course/en/chapter1/4

Pre-Trained LLMs to Task Adaptation



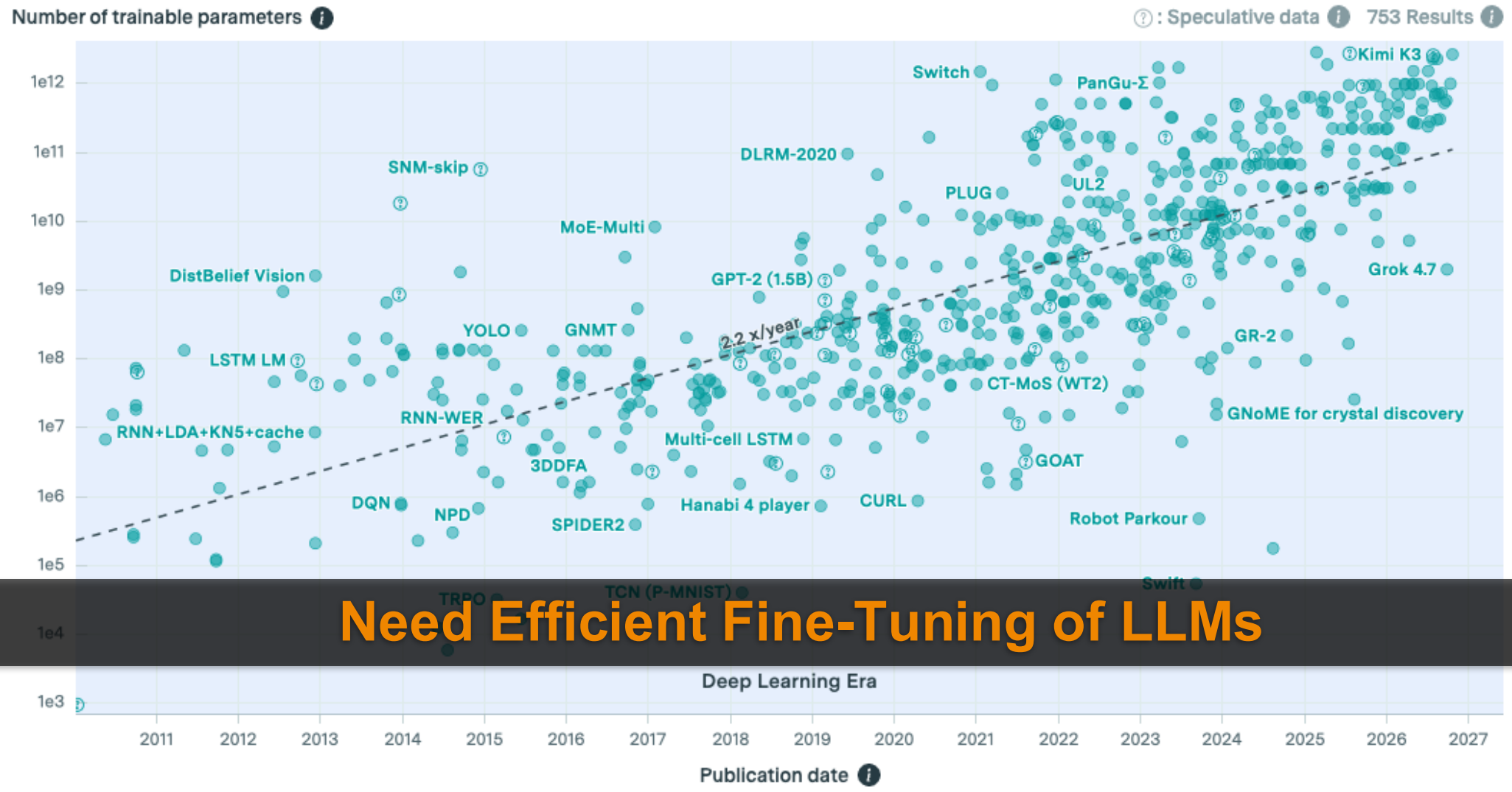
https://huggingface.co/learn/llm-course/en/chapter1/4

Rapid Advancements



Epoch AI. ”data on ai models”. URL https://epoch.ai/data/ai-models. Accessed: 04 Nov 2026.

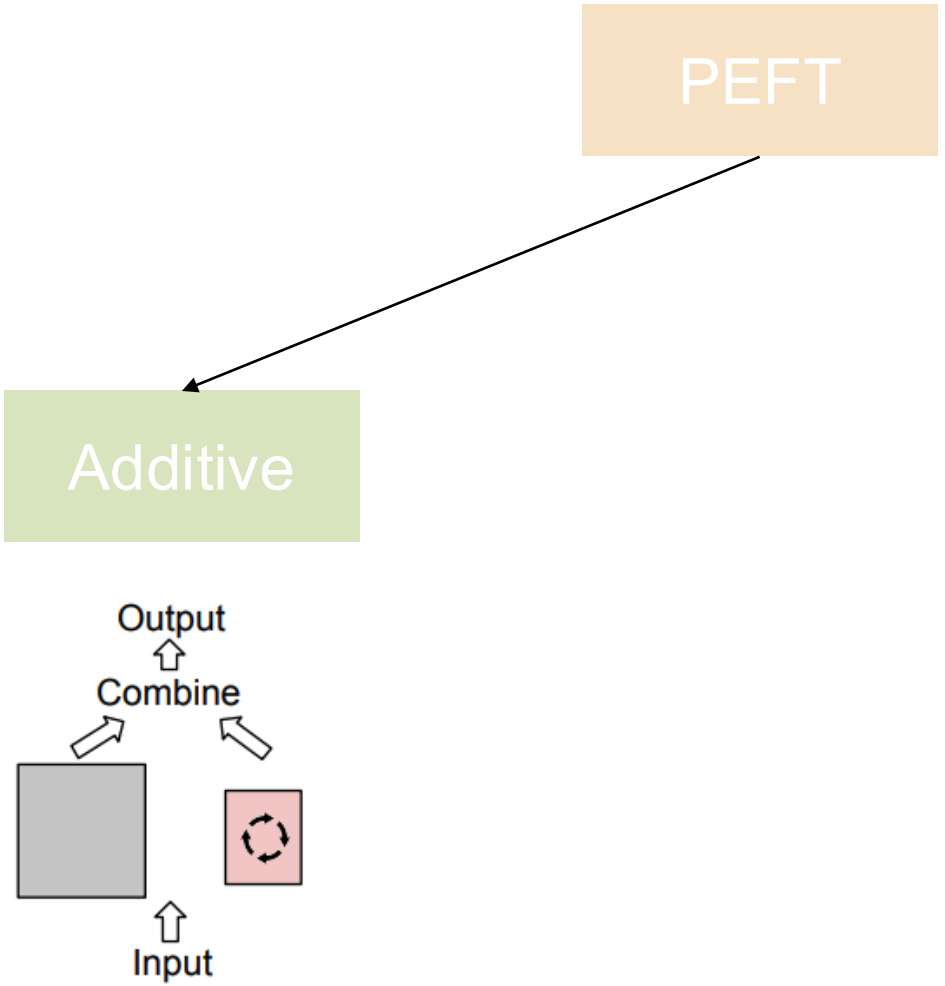
Rapid Advancements



Need Efficient Fine-Tuning of LLMs

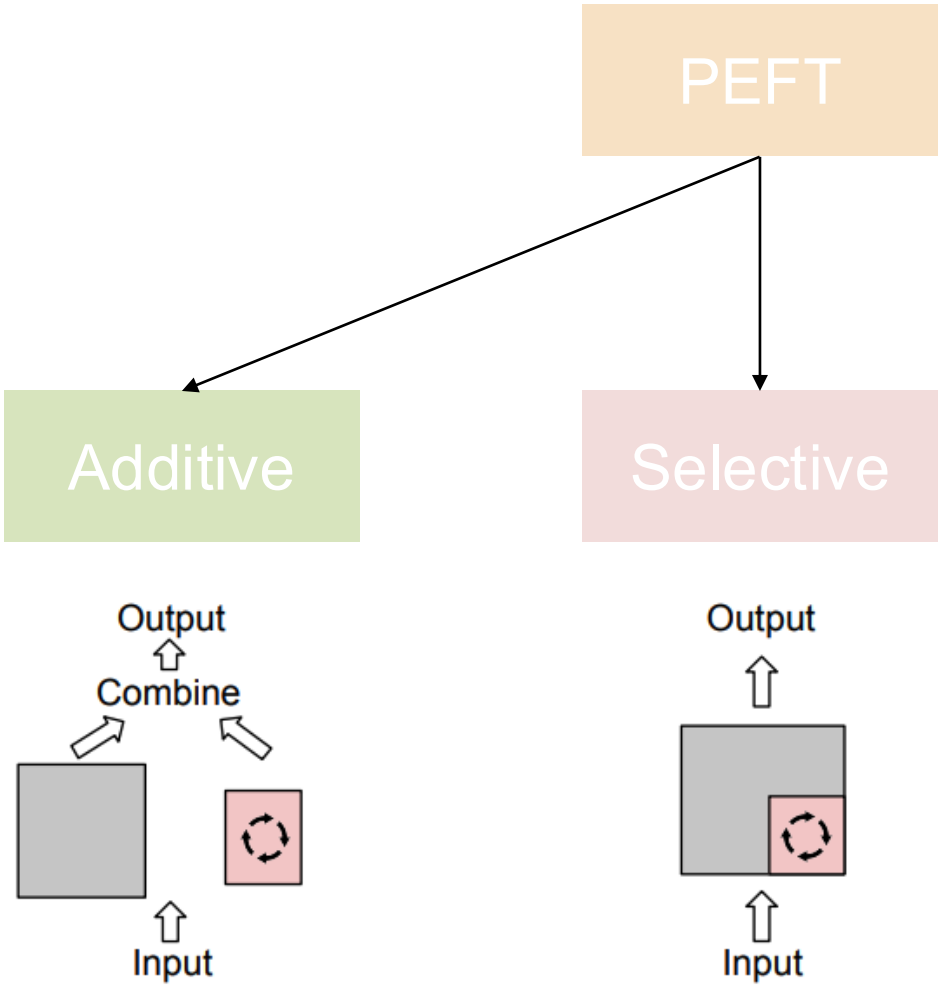
Epoch AI. ”data on ai models”. URL https://epoch.ai/data/ai-models. Accessed: 04 Nov 2026.

Parameter-Efficient Fine-Tuning (PEFT)



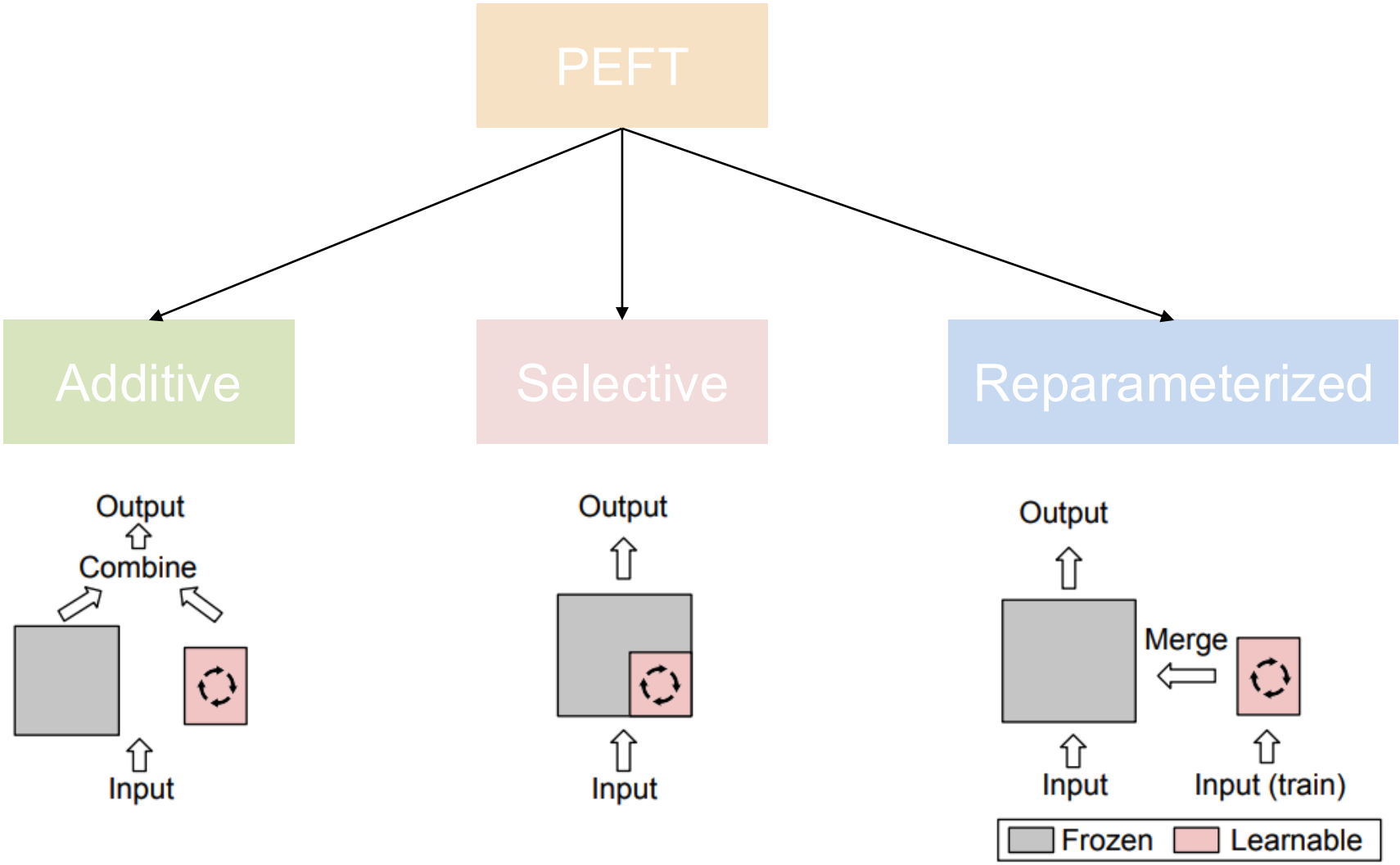
Han, Zeyu, et al. "Parameter-Efficient Fine-Tuning for Large Models: A Comprehensive Survey." *Transactions on Machine Learning Research*. 2024

Parameter-Efficient Fine-Tuning (PEFT)



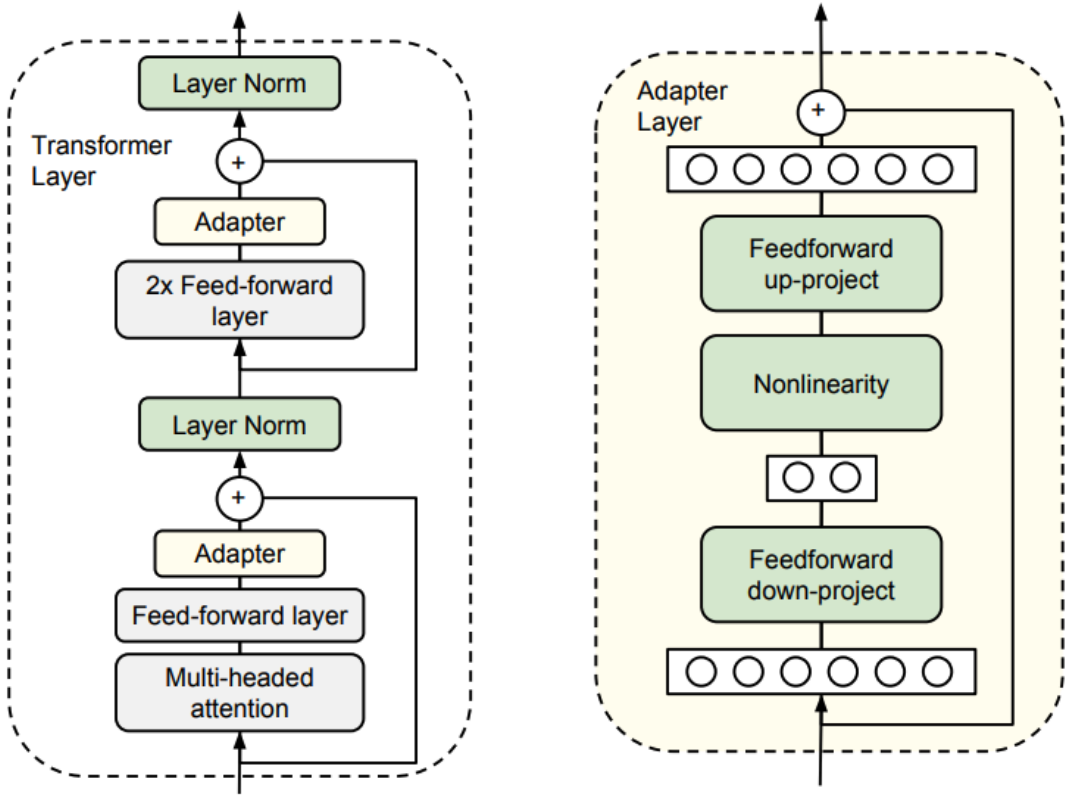
Han, Zeyu, et al. "Parameter-Efficient Fine-Tuning for Large Models: A Comprehensive Survey." *Transactions on Machine Learning Research*. 2024

Parameter-Efficient Fine-Tuning (PEFT)



Han, Zeyu, et al. "Parameter-Efficient Fine-Tuning for Large Models: A Comprehensive Survey." *Transactions on Machine Learning Research*. 2024

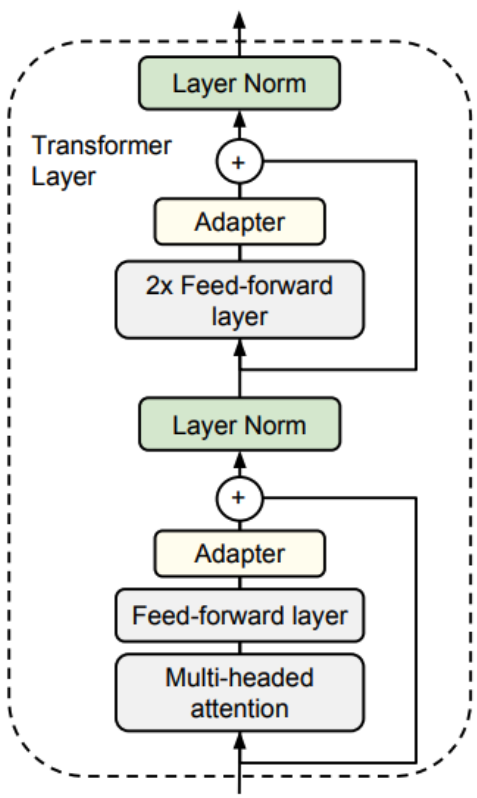
PEFT-Additive

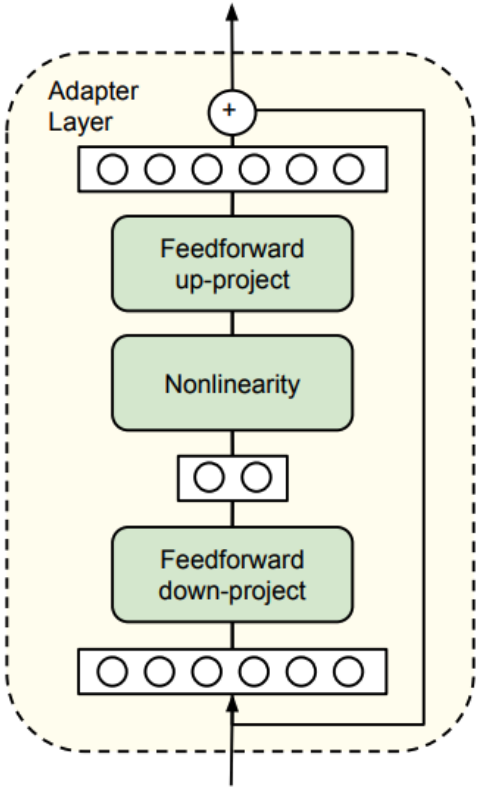


Adapter-based

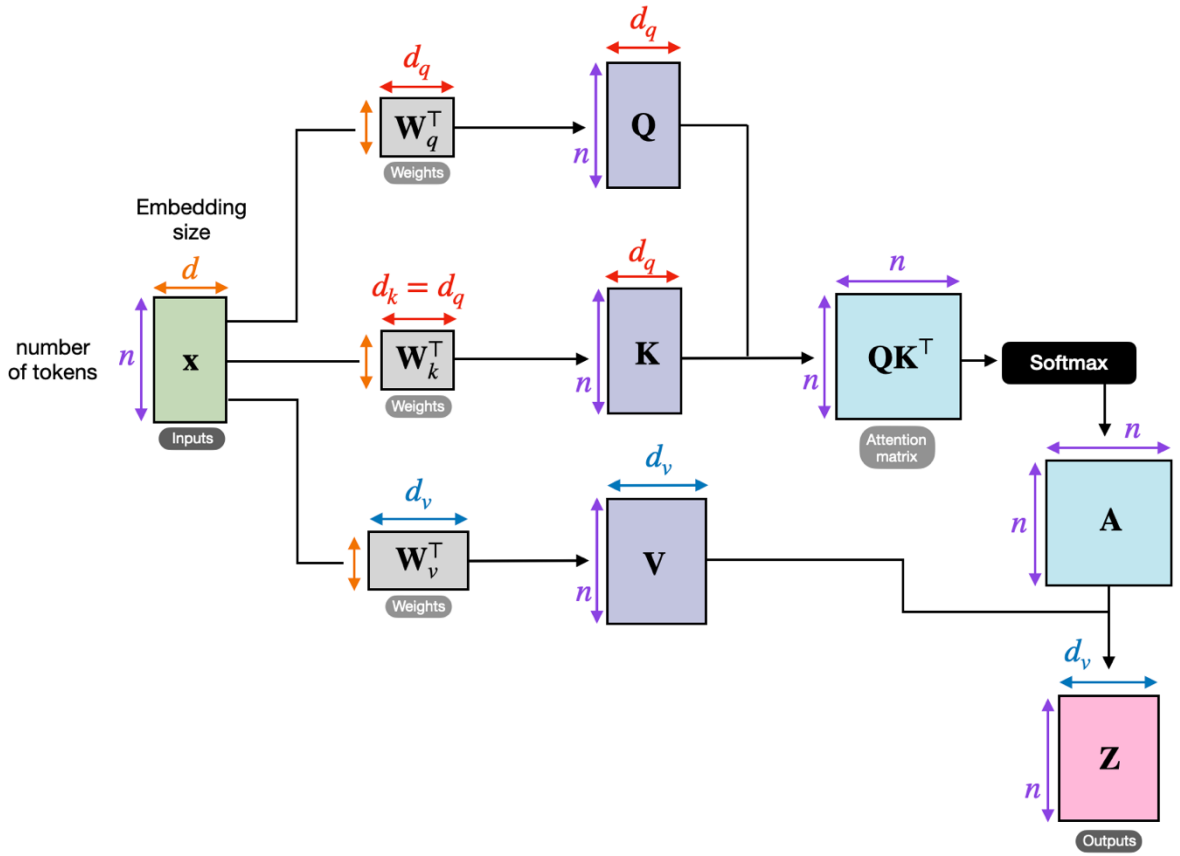
Houlsby, Neil, et al. "Parameter-efficient transfer learning for NLP." *International conference on machine learning*. PMLR, 2019.

PEFT-Additive





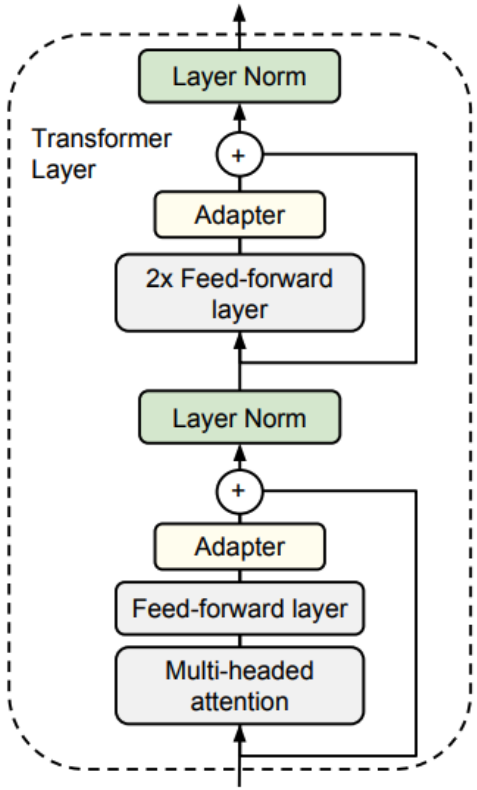
Adapter-based

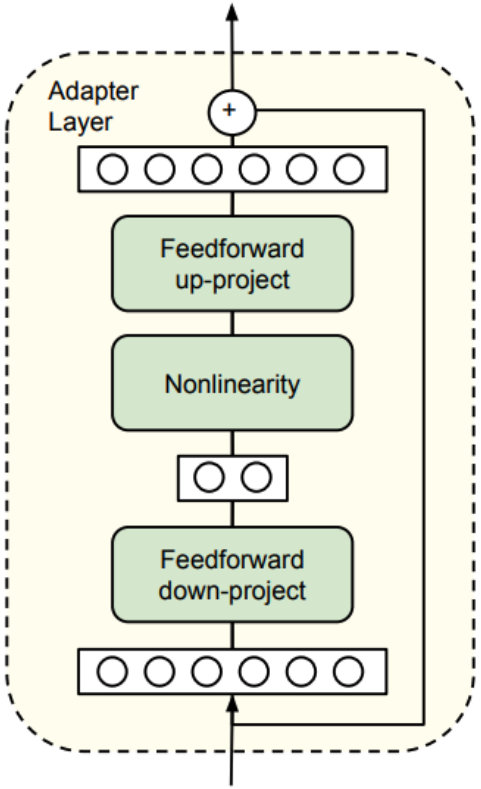


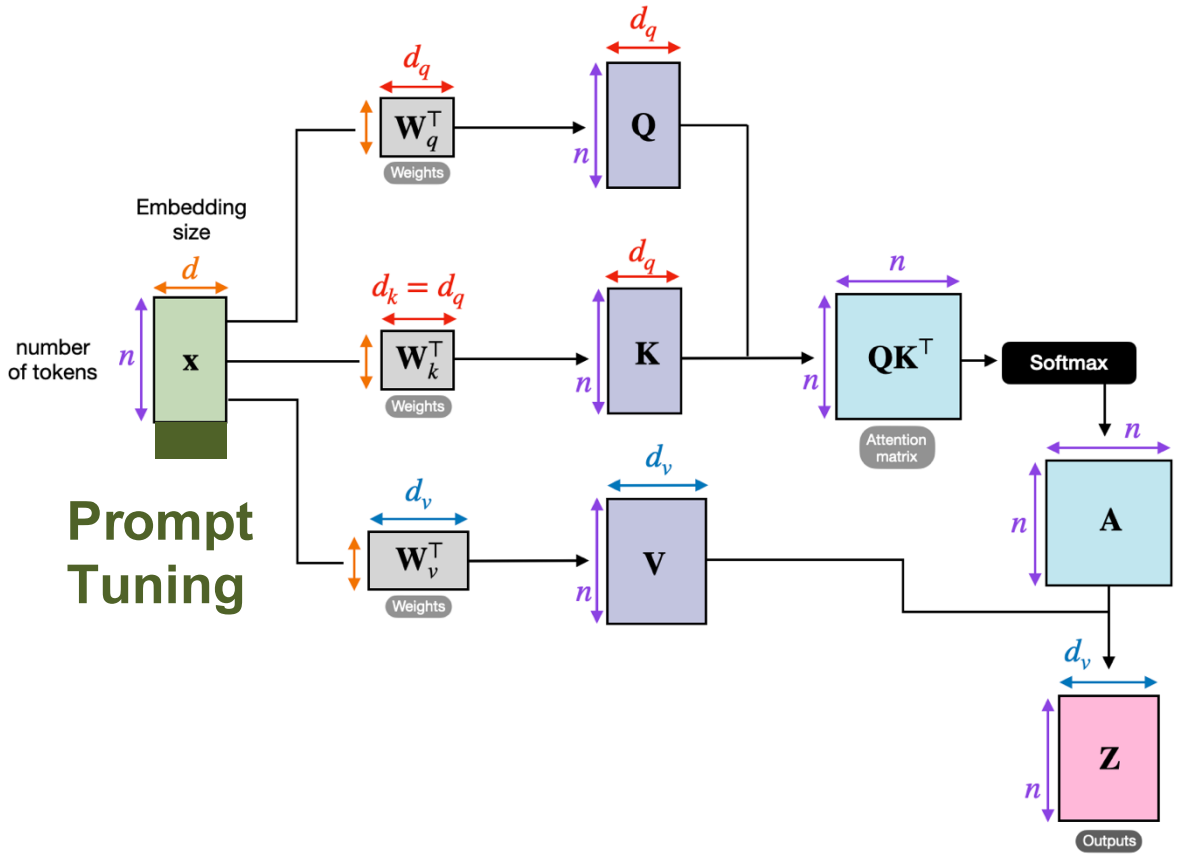
Prompt-based

Houlsby, Neil, et al. "Parameter-efficient transfer learning for NLP." *International conference on machine learning*. PMLR, 2019.

PEFT-Additive





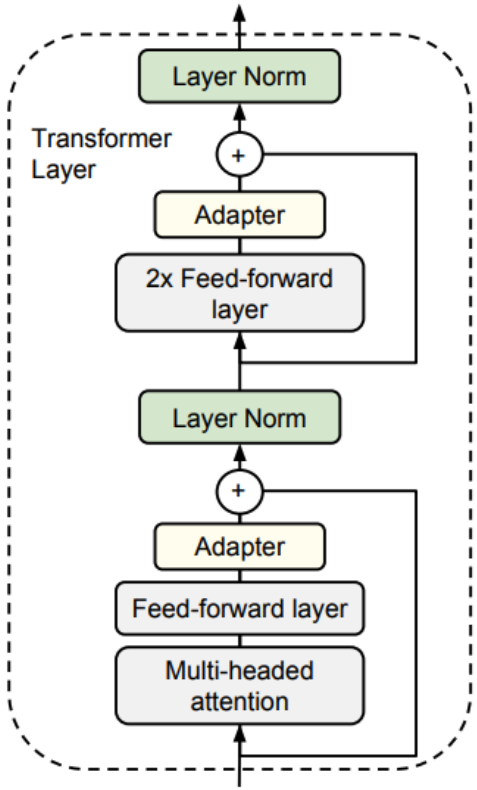


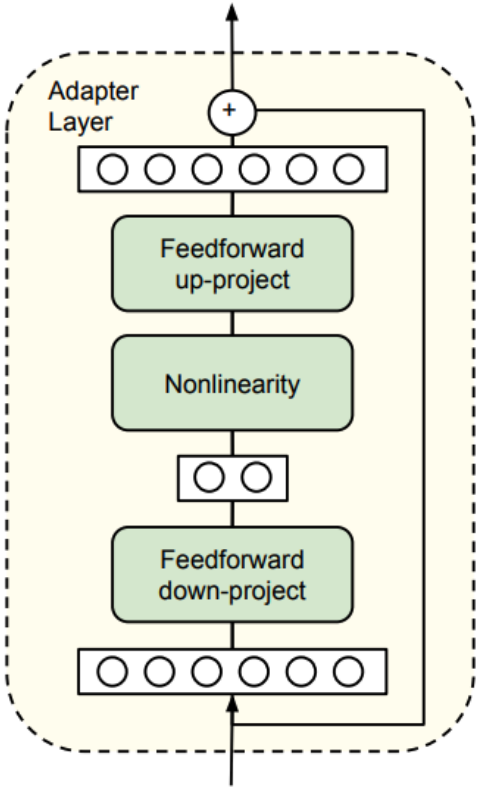
Adapter-based

Prompt-based

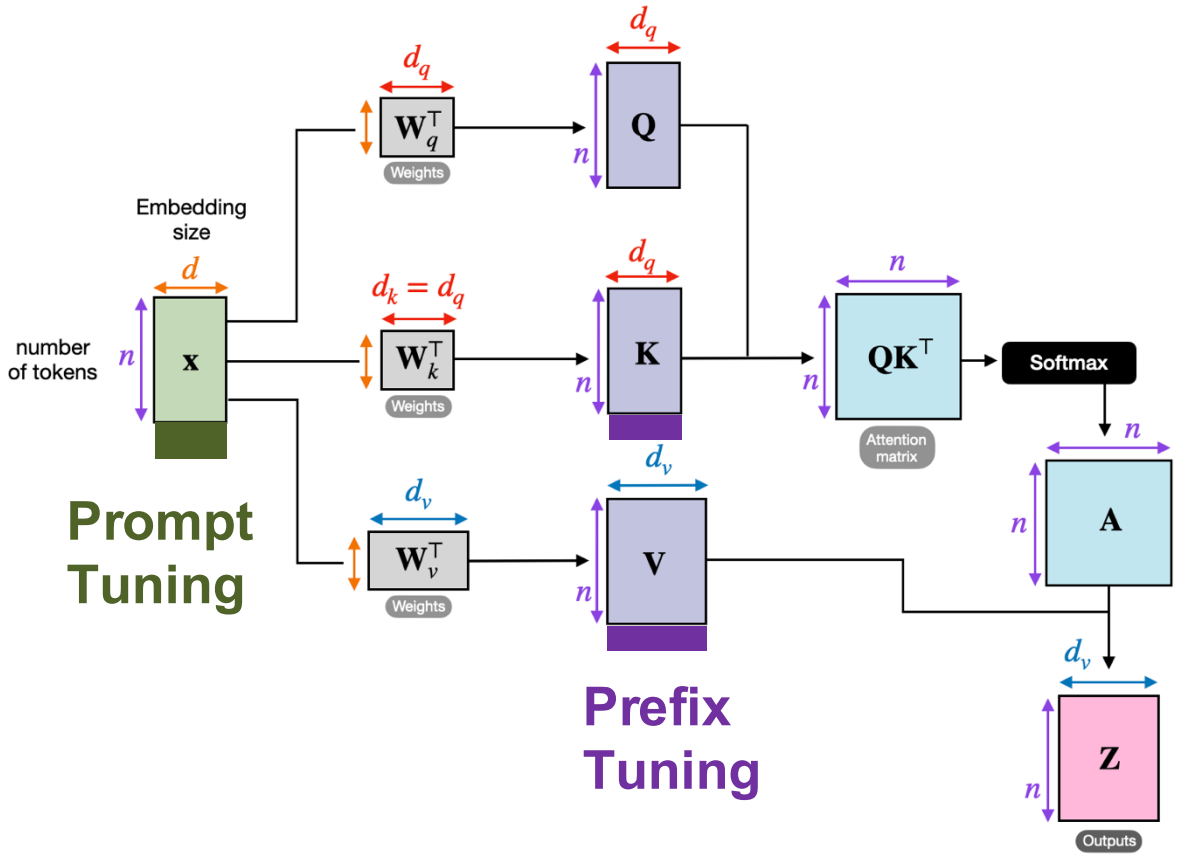
Houlsby, Neil, et al. "Parameter-efficient transfer learning for NLP." *International conference on machine learning*. PMLR, 2019.

PEFT-Additive





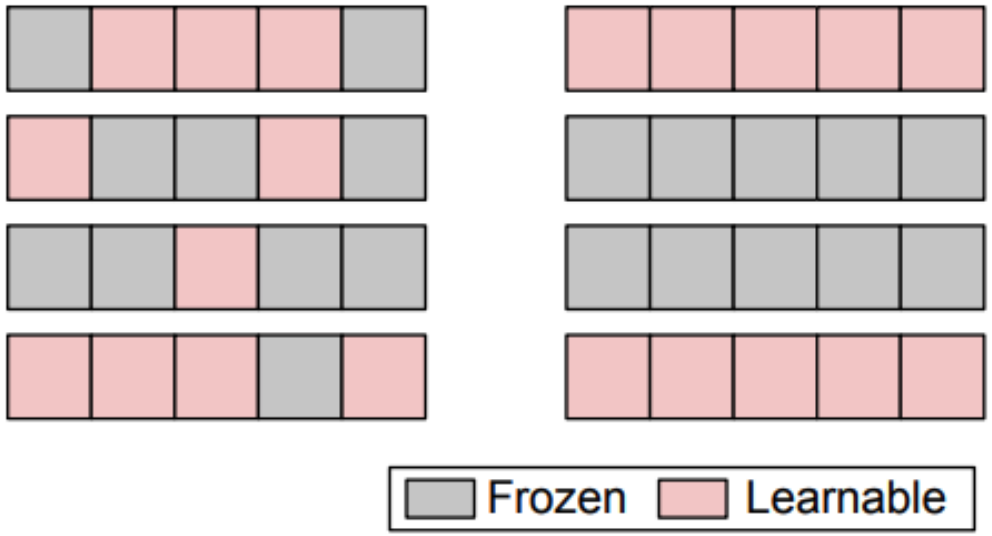
Adapter-based



Prompt-based

Houlsby, Neil, et al. "Parameter-efficient transfer learning for NLP." *International conference on machine learning*. PMLR, 2019.

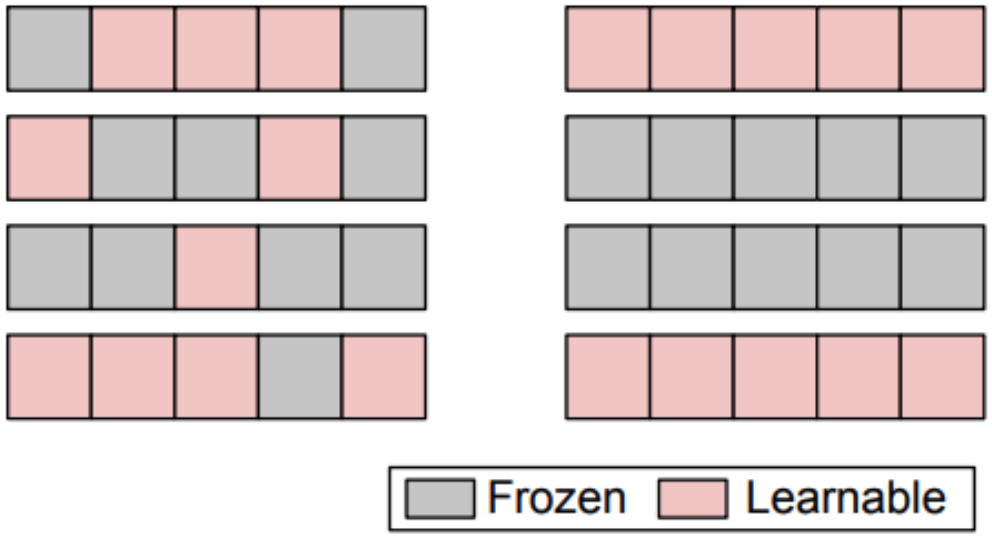
PEFT-Selective



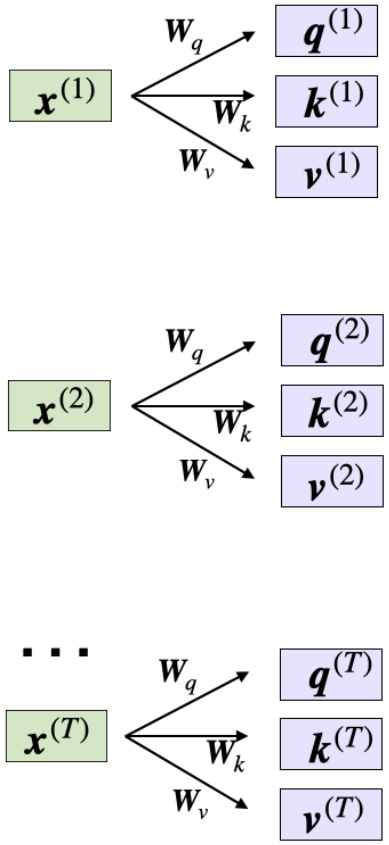
Unstructual **Structual**

Huang, Baichuan, Ananth Balashankar, and Amir Aminifar. "BEFT: Bias-Efficient Fine-Tuning of Language Models." (2025).

PEFT-Selective



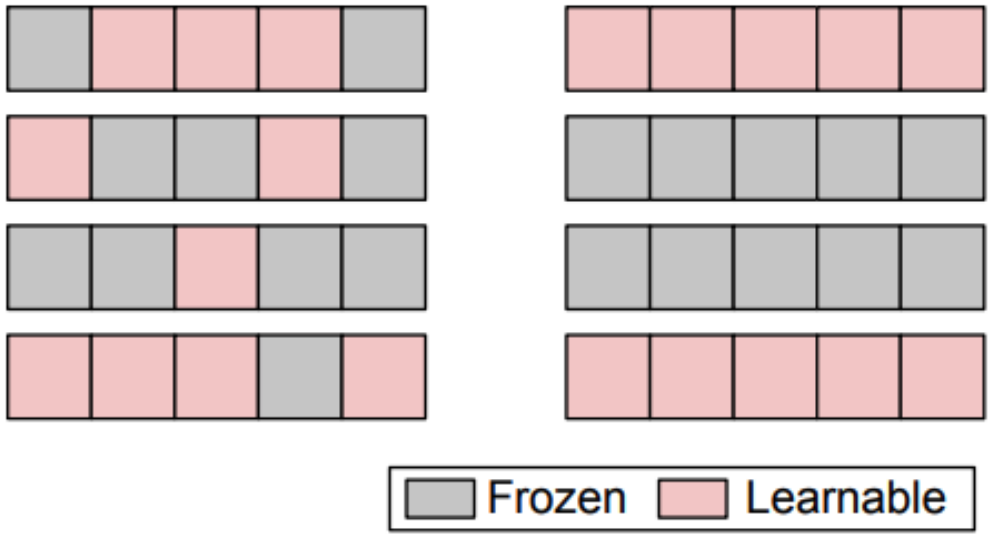
Unstructual **Structual**



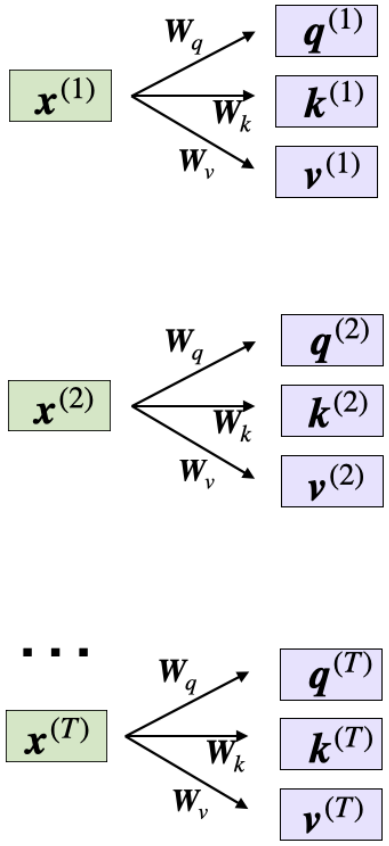
Example of structual

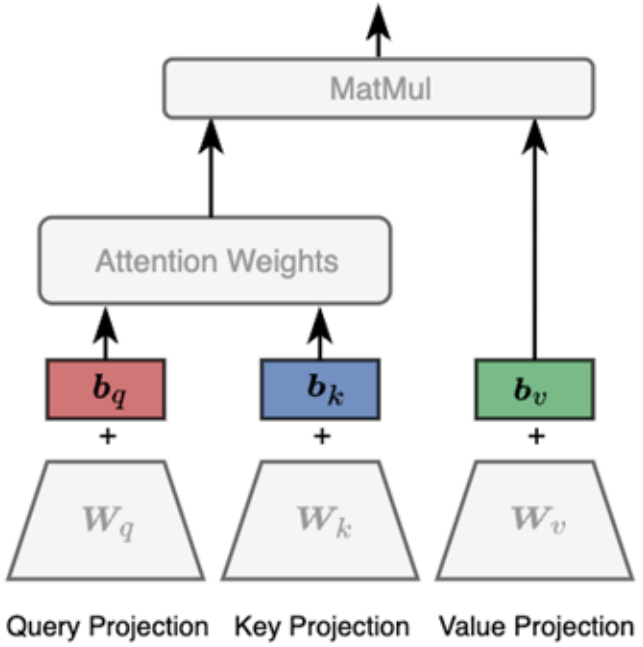
Huang, Baichuan, Ananth Balashankar, and Amir Aminifar. "BEFT: Bias-Efficient Fine-Tuning of Language Models." (2025).

PEFT-Selective



Unstructual **Structual**

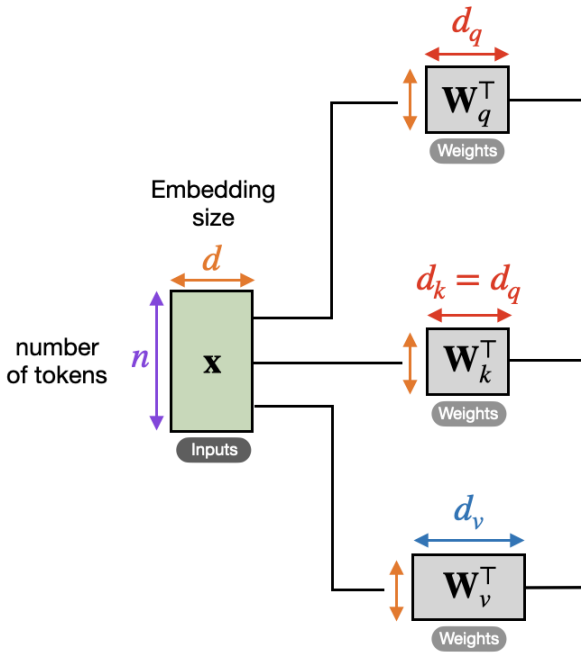




Example of structual

Huang, Baichuan, Ananth Balashankar, and Amir Aminifar. "BEFT: Bias-Efficient Fine-Tuning of Language Models." (2025).

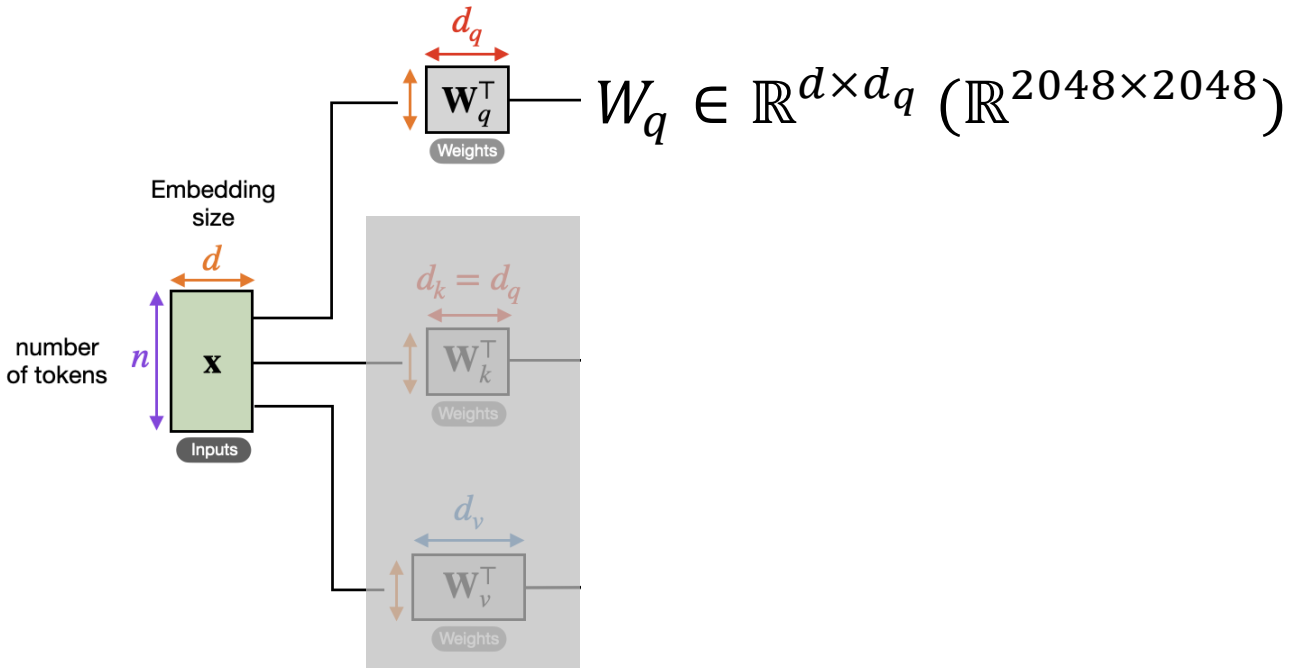
PEFT-Reparameterized



Low-Rank Decomposition

Hu, Edward J., et al. "Lora: Low-rank adaptation of large language models." *ICLR* 1.2 (2022): 3.

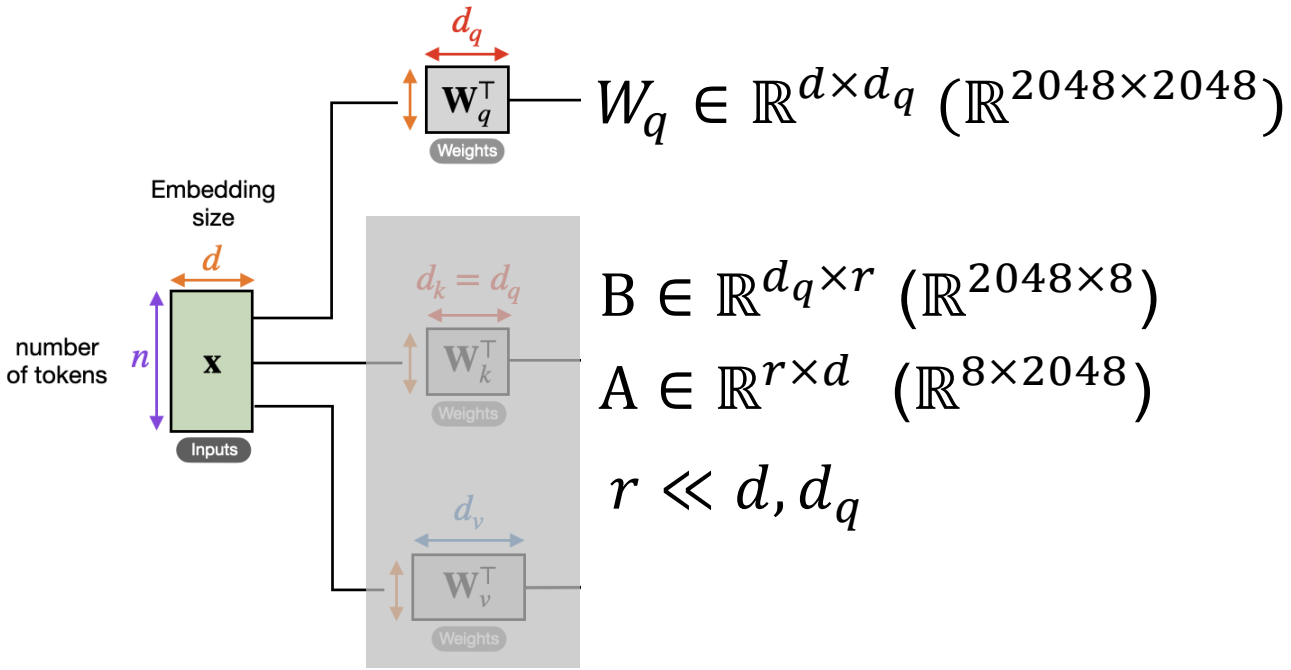
PEFT-Reparameterized



Low-Rank Decomposition

Hu, Edward J., et al. "Lora: Low-rank adaptation of large language models." *ICLR* 1.2 (2022): 3.

PEFT-Reparameterized



$$W_q \in \mathbb{R}^{d \times d_q} \; (\mathbb{R}^{2048 \times 2048})$$

$$\mathrm{B} \in \mathbb{R}^{d_q \times r} \; (\mathbb{R}^{2048 \times 8})$$

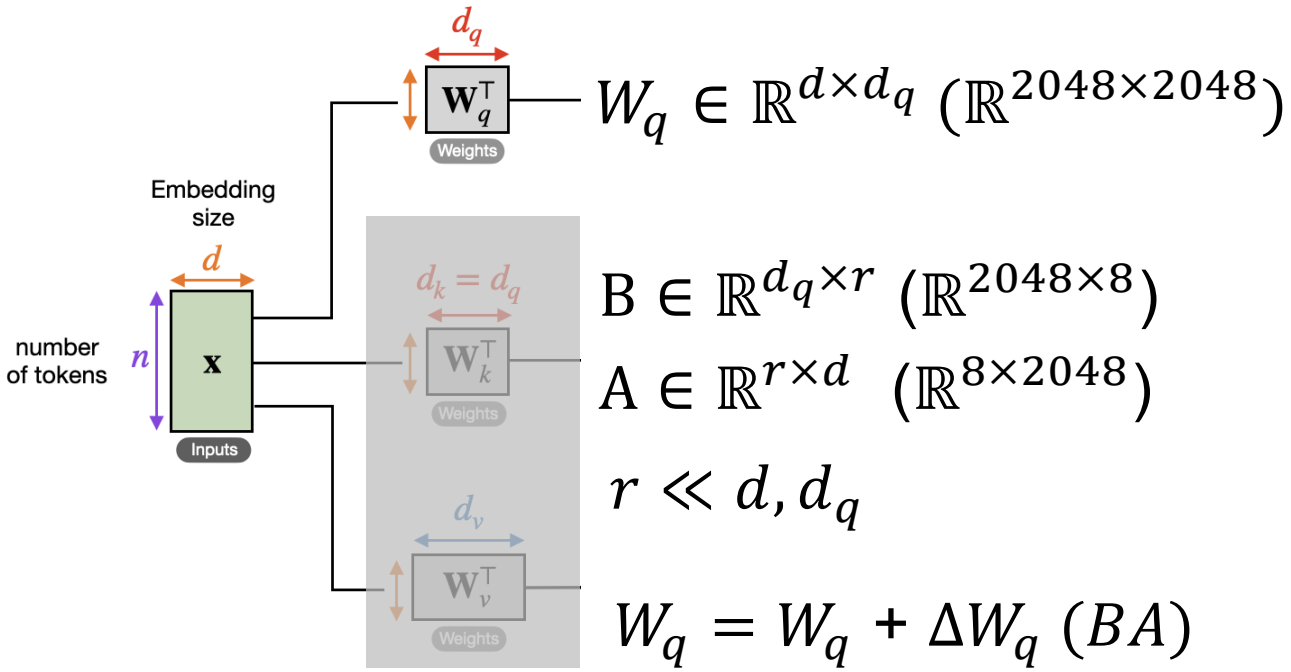
$$\mathrm{A} \in \mathbb{R}^{r \times d} \; (\mathbb{R}^{8 \times 2048})$$

$$r \ll d, d_q$$

Low-Rank Decomposition

Hu, Edward J., et al. "Lora: Low-rank adaptation of large language models." *ICLR* 1.2 (2022): 3.

PEFT-Reparameterized



$$W_q \in \mathbb{R}^{d \times d_q} \; (\mathbb{R}^{2048 \times 2048})$$

$$\mathrm{B} \in \mathbb{R}^{d_q \times r} \; (\mathbb{R}^{2048 \times 8})$$

$$\mathrm{A} \in \mathbb{R}^{r \times d} \; (\mathbb{R}^{8 \times 2048})$$

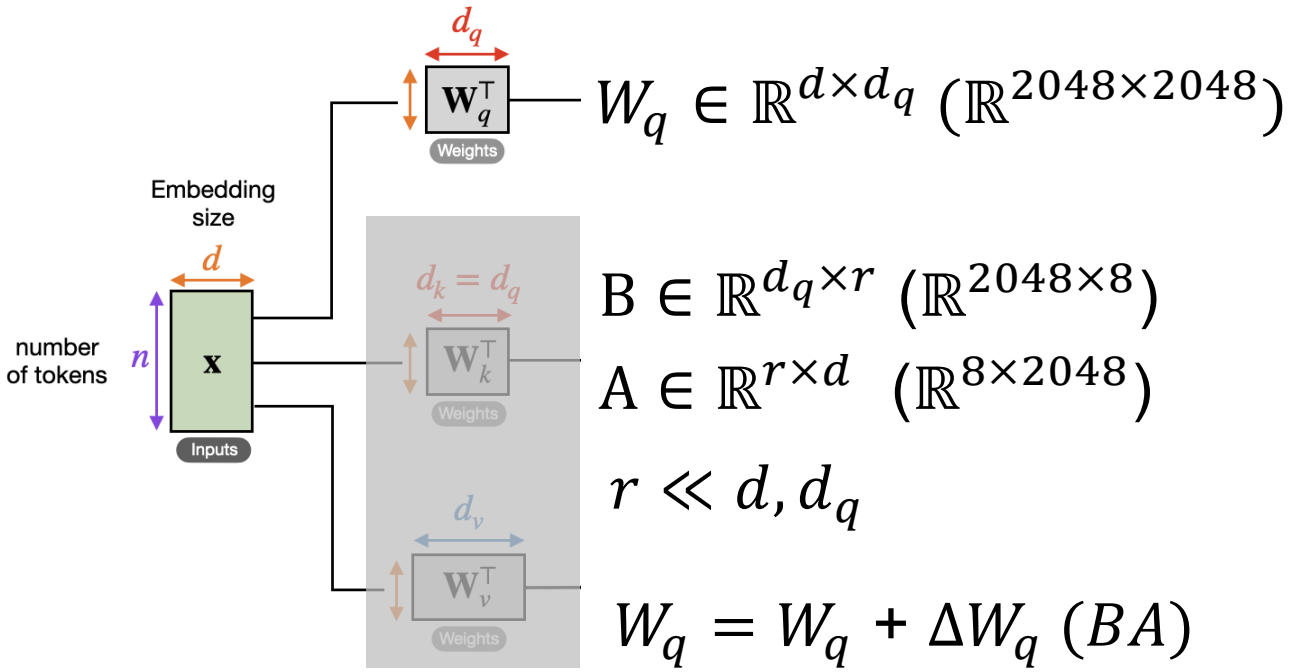
$$r \ll d, d_q$$

$$W_q = W_q + \Delta W_q \; (BA)$$

Low-Rank Decomposition

Hu, Edward J., et al. "Lora: Low-rank adaptation of large language models." *ICLR* 1.2 (2022): 3.

PEFT-Reparameterized



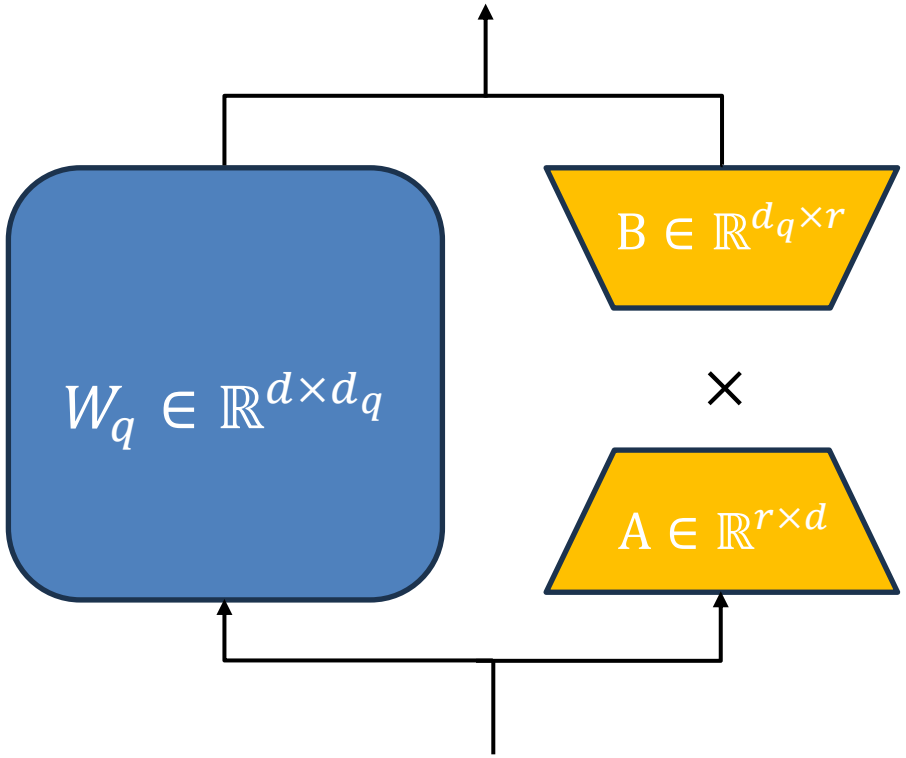
$W_q \in \mathbb{R}^{d \times d_q}$ $(\mathbb{R}^{2048 \times 2048})$

$\mathrm{B} \in \mathbb{R}^{d_q \times r}$ $(\mathbb{R}^{2048 \times 8})$

$\mathrm{A} \in \mathbb{R}^{r \times d}$ $(\mathbb{R}^{8 \times 2048})$

$r \ll d, d_q$

$W_q = W_q + \Delta W_q \; (BA)$

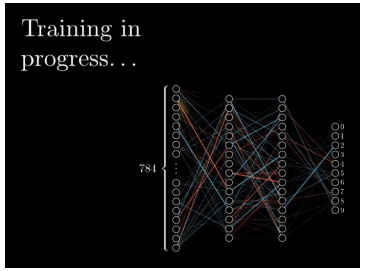


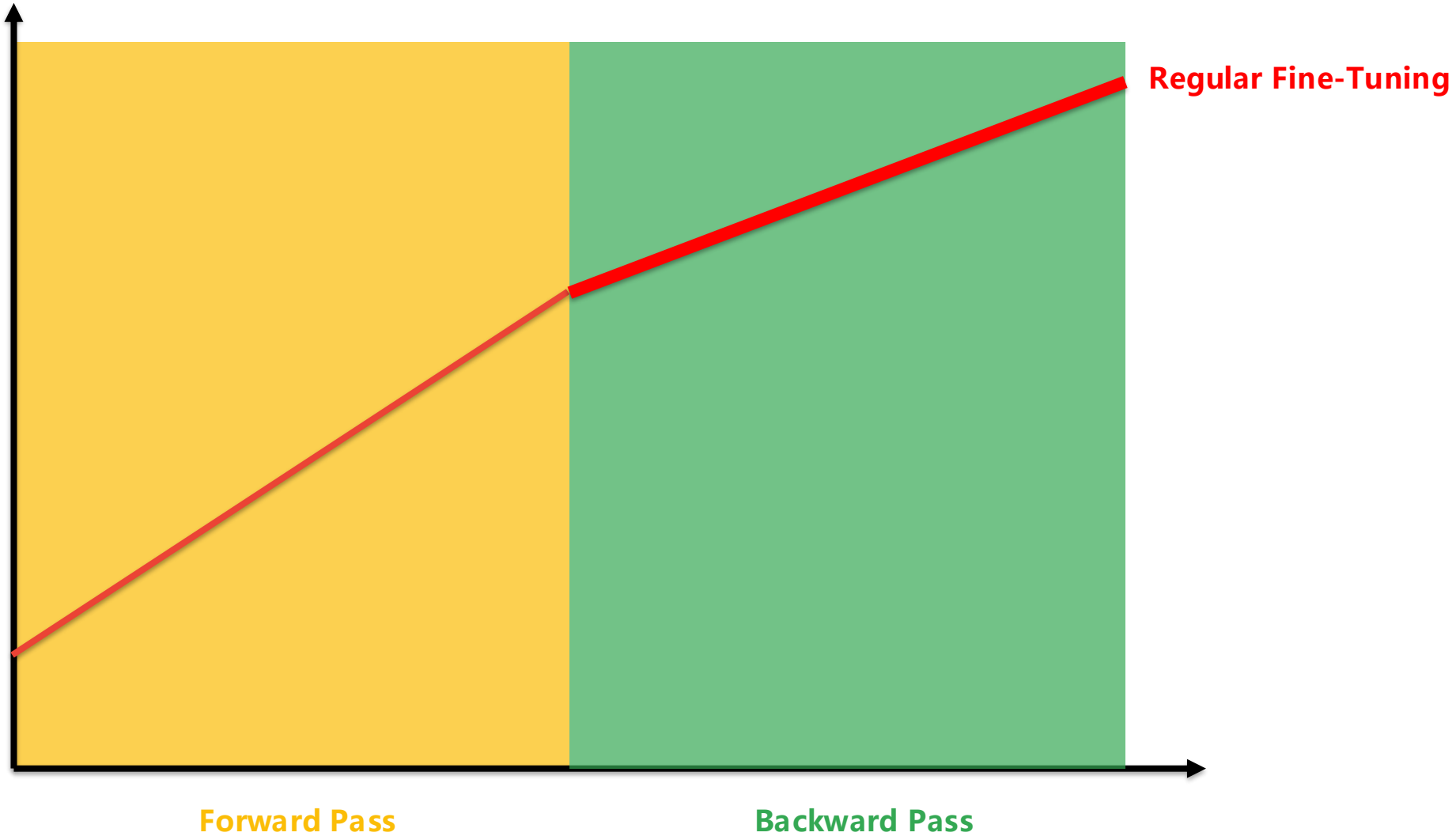
Low-Rank Decomposition

LoRA

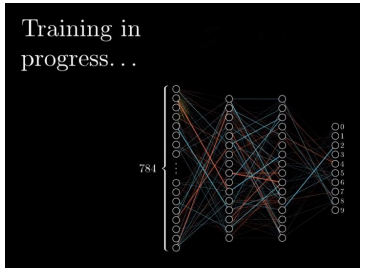
Hu, Edward J., et al. "Lora: Low-rank adaptation of large language models." *ICLR* 1.2 (2022): 3.

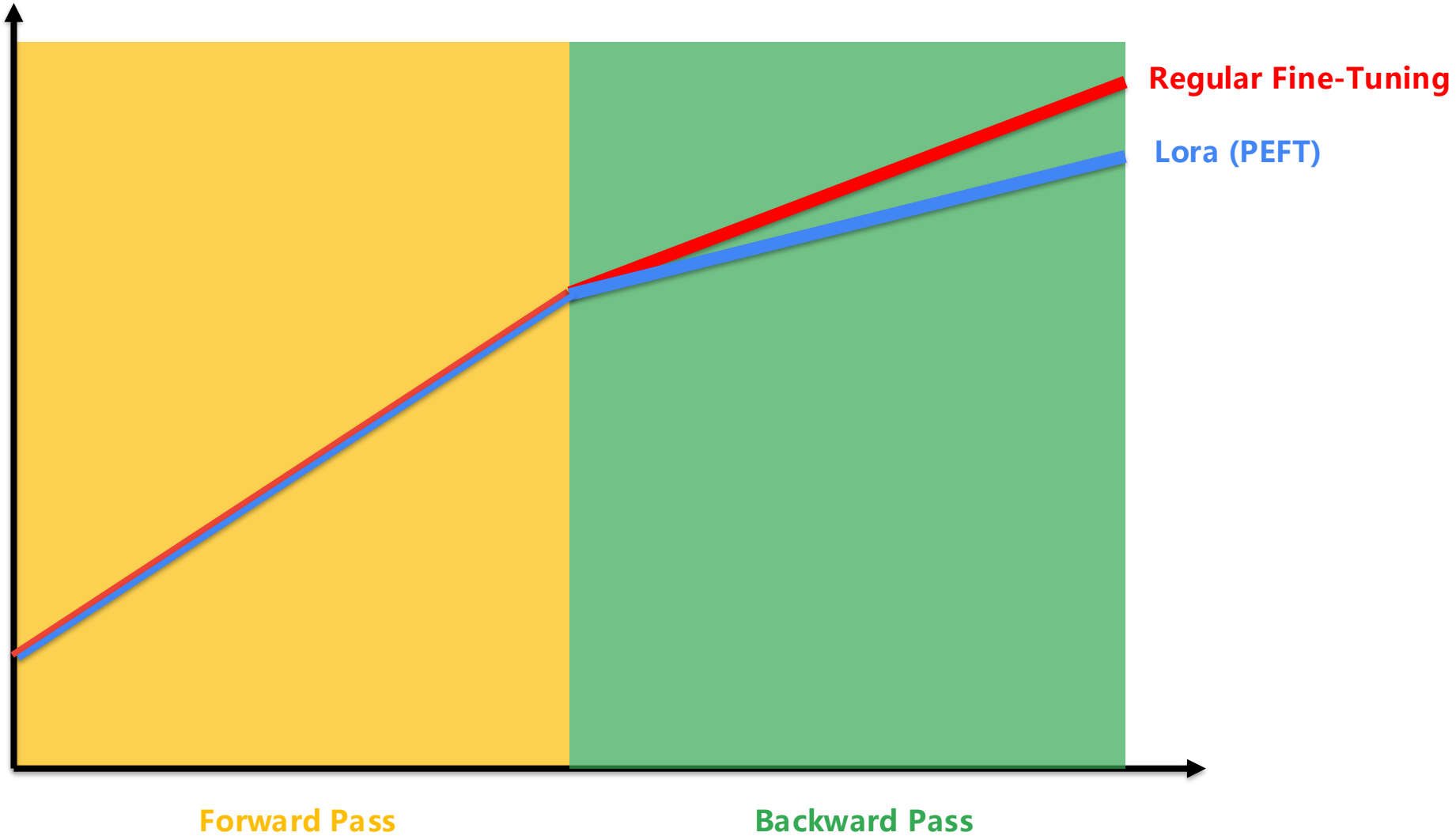
Parameter Efficiency Reduces Training Memory?



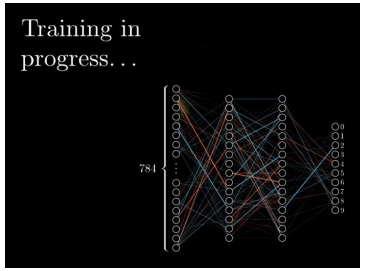


Parameter Efficiency Reduces Training Memory?





Parameter Efficiency Reduces Training Memory?



GPU Memory

Regular Fine-Tuning

Lora (PEFT)

...

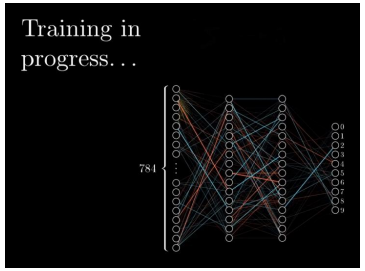
Prefix-Tuning (PEFT)

Forward Pass

Backward Pass

Dynamic Training Process
(Theoratically)

Parameter Efficiency Reduces Training Memory?



GPU Memory

Regular Fine-Tuning

Lora (PEFT)

...

Prefix-Tuning (PEFT)

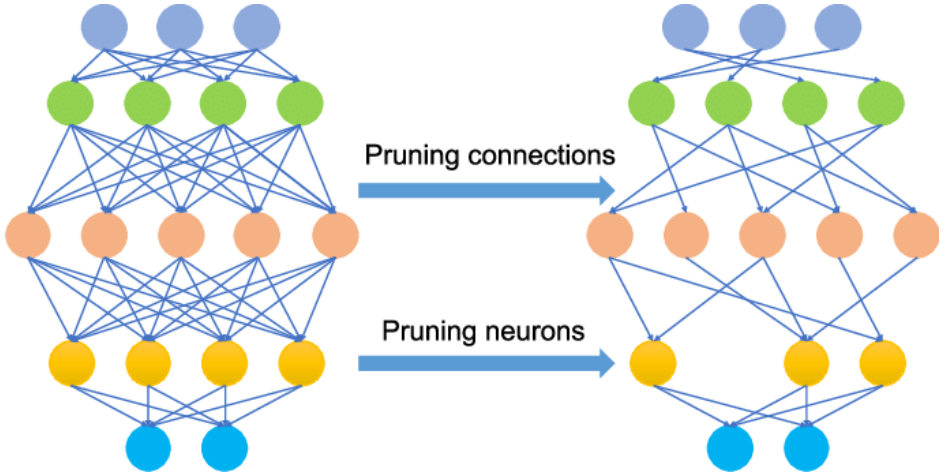
In-Place Update

Forward Pass

Backward Pass

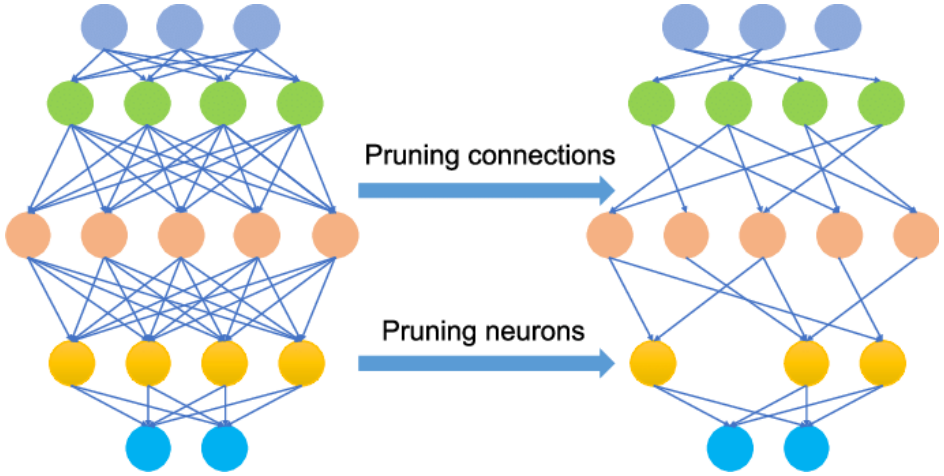
Dynamic Training Process
(Theoratically)

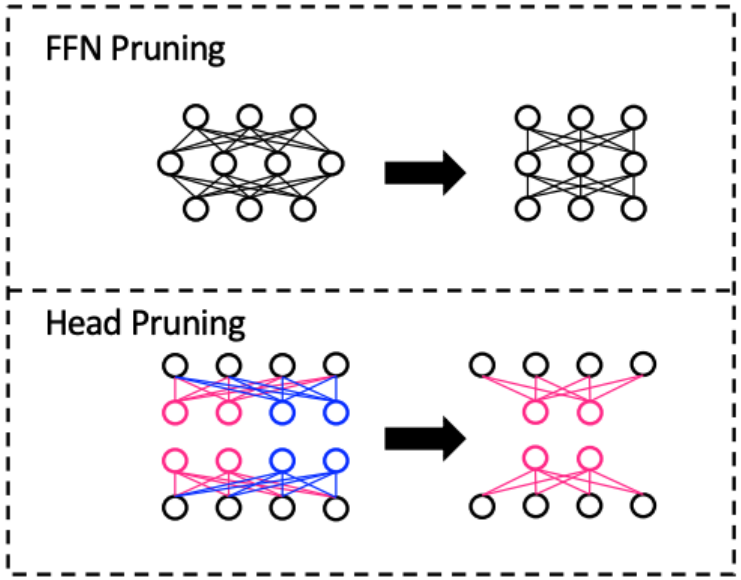
Memory Efficient Fine-Tuning-Pruning



Han, Song, Huizi Mao, and William J. Dally. "Deep Compression: Compressing Deep Neural Network with Pruning, Trained Quantization and Huffman Coding." *ICLR*. 2016.
Gu, Naibin, et al. "Light-PEFT: Lightening Parameter-Efficient Fine-Tuning via Early Pruning." *Findings of the Association for Computational Linguistics ACL 2024*. 2024.
Wang, Yuxin, et al. "CFSP: An Efficient Structured Pruning Framework for LLMs with Coarse-to-Fine Activation Information." *Proceedings of the 31st International Conference on Computational Linguistics*. 2025.

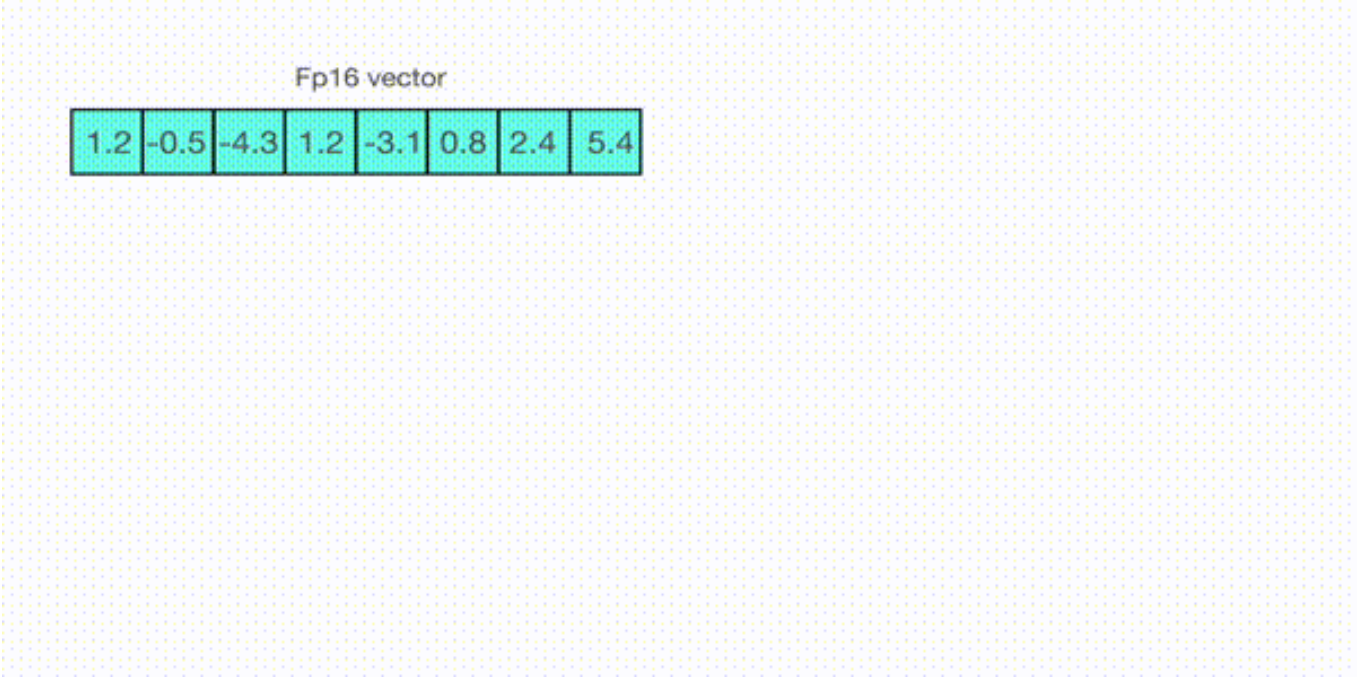
Memory Efficient Fine-Tuning-Pruning





Han, Song, Huizi Mao, and William J. Dally. "Deep Compression: Compressing Deep Neural Network with Pruning, Trained Quantization and Huffman Coding." *ICLR*. 2016.
Gu, Naibin, et al. "Light-PEFT: Lightening Parameter-Efficient Fine-Tuning via Early Pruning." *Findings of the Association for Computational Linguistics ACL 2024*. 2024.
Wang, Yuxin, et al. "CFSP: An Efficient Structured Pruning Framework for LLMs with Coarse-to-Fine Activation Information." *Proceedings of the 31st International Conference on Computational Linguistics*. 2025.

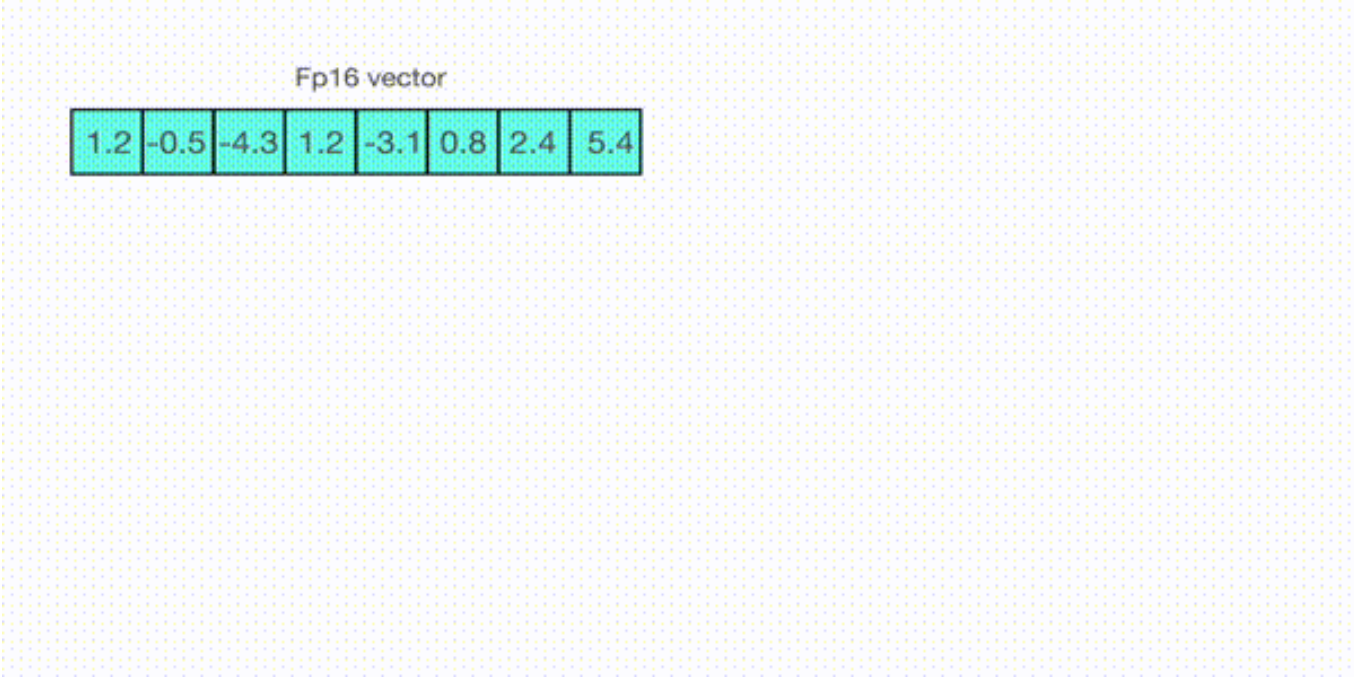
Memory Efficient Fine-Tuning-Quantization



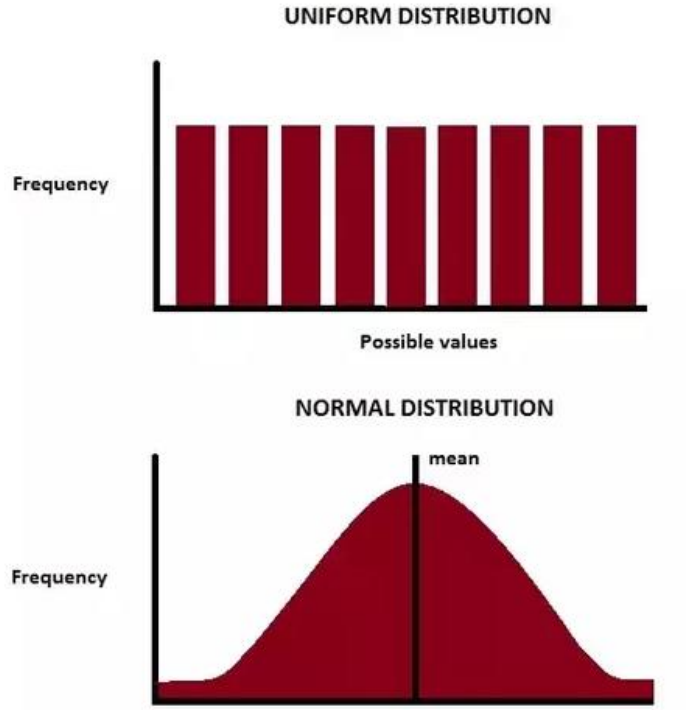
Quantization

Dettmers, Tim, et al. "Qlora: Efficient finetuning of quantized llms." *Advances in neural information processing systems* 36 (2023): 10088-10115.
https://huggingface.co/blog/hf-bitsandbytes-integration

Memory Efficient Fine-Tuning-Quantization



Quantization

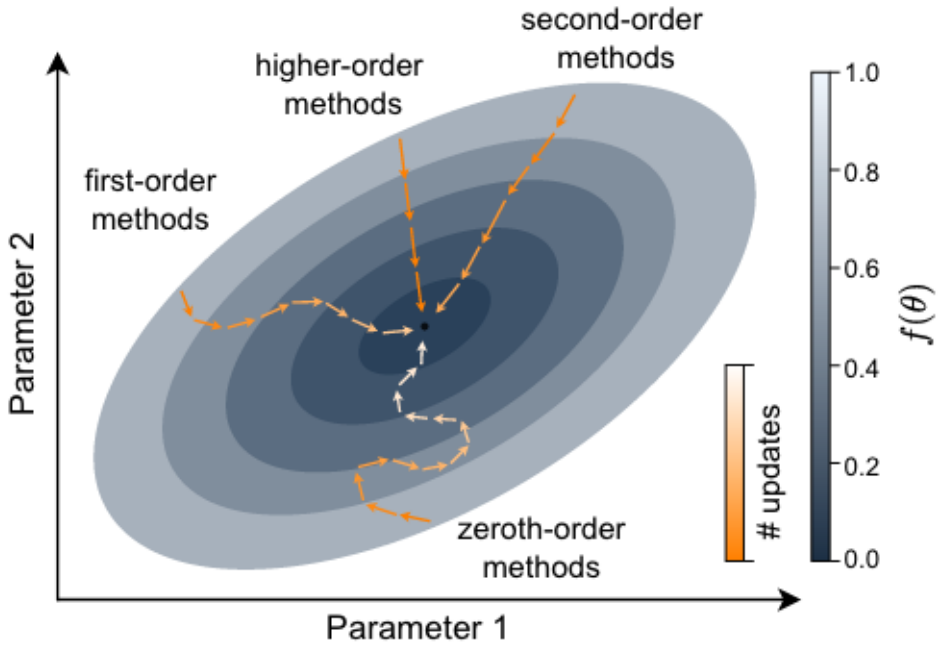


4-bit NormalFloat (NF4)

QLoRA

Dettmers, Tim, et al. "Qlora: Efficient finetuning of quantized llms." *Advances in neural information processing systems* 36 (2023): 10088-10115.
https://huggingface.co/blog/hf-bitsandbytes-integration

Memory Efficient Fine-Tuning-Zeroth-Order Gradient

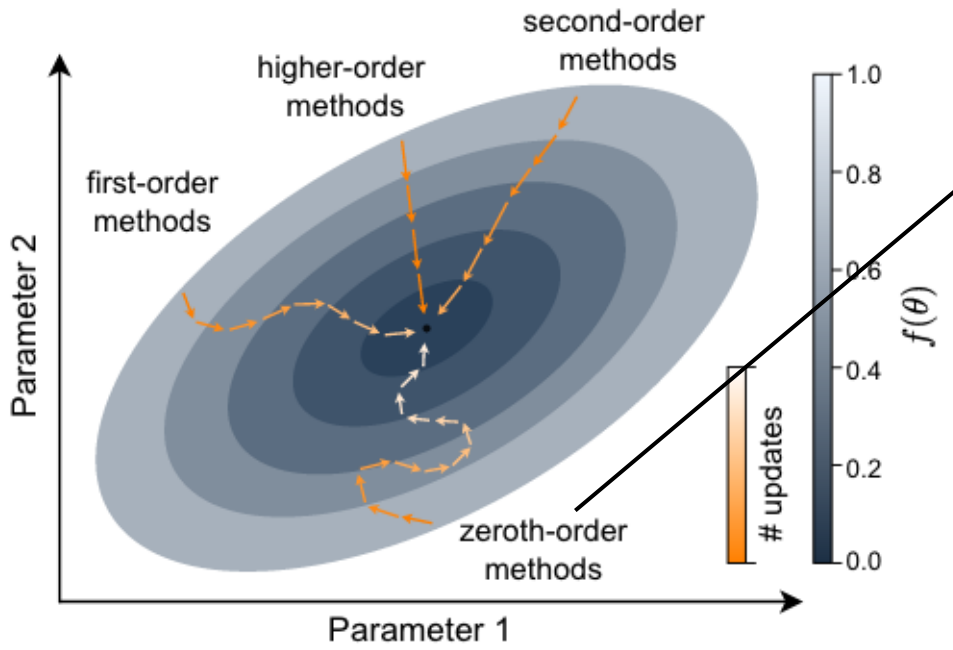


Zhang, Yihua, et al. "Revisiting zeroth-order optimization for memory-efficient LLM fine-tuning: a benchmark." *Proceedings of the 41st International Conference on Machine Learning*. 2024.

Fernández, Jesús García, Nasir Ahmad, and Marcel van Gerven. "A Unified Perspective on Optimization in Machine Learning and Neuroscience: From Gradient Descent to Neural Adaptation." *arXiv preprint arXiv:2510.18812* (2025).

https://sites.google.com/view/zo-tutorial-aaai-2024/

Memory Efficient Fine-Tuning-Zeroth-Order Gradient



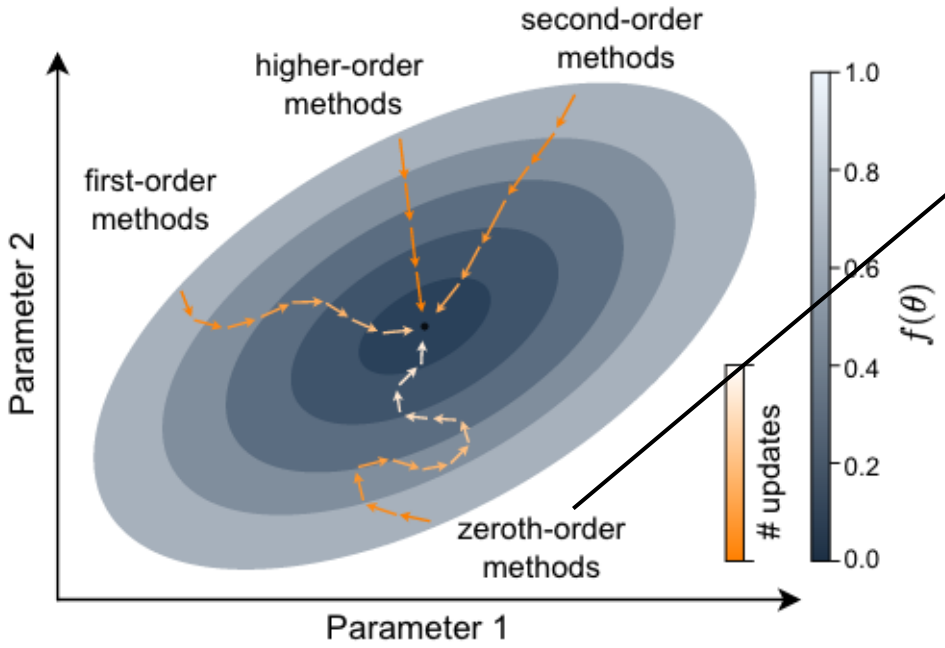
The gradient can be approximated by random gradient estimation

Zhang, Yihua, et al. "Revisiting zeroth-order optimization for memory-efficient LLM fine-tuning: a benchmark." *Proceedings of the 41st International Conference on Machine Learning*. 2024.

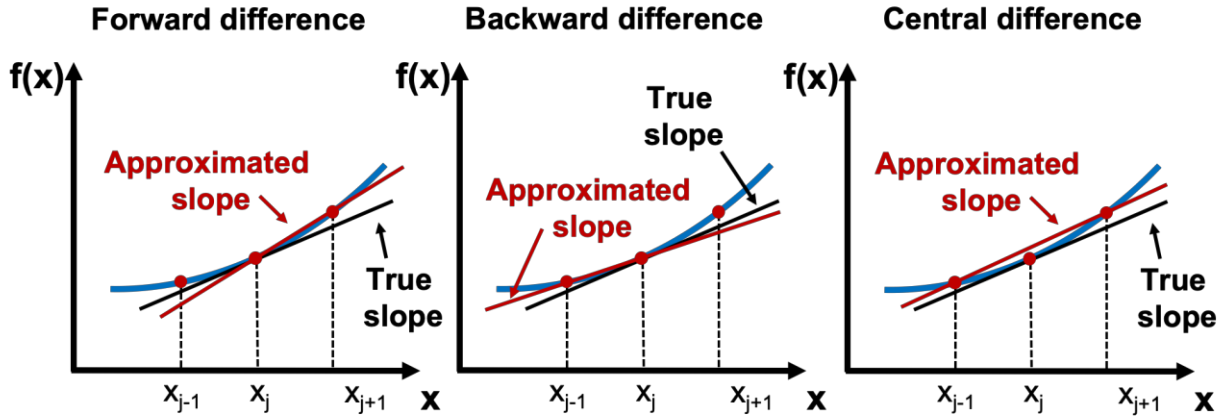
Fernández, Jesús García, Nasir Ahmad, and Marcel van Gerven. "A Unified Perspective on Optimization in Machine Learning and Neuroscience: From Gradient Descent to Neural Adaptation." *arXiv preprint arXiv:2510.18812* (2025).

https://sites.google.com/view/zo-tutorial-aaai-2024/

Memory Efficient Fine-Tuning-Zeroth-Order Gradient



The gradient can be approximated by random gradient estimation

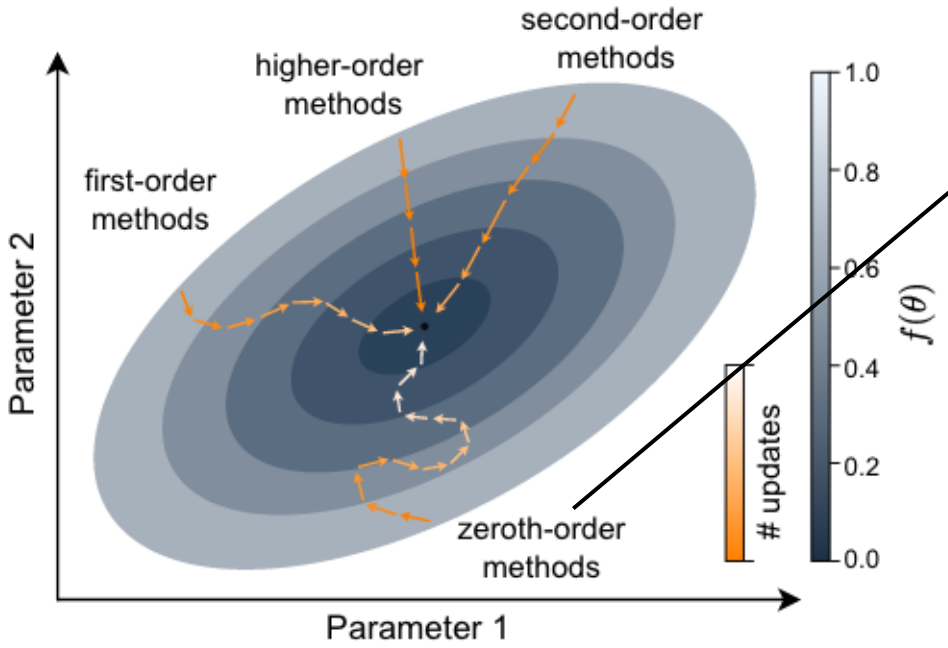


Zhang, Yihua, et al. "Revisiting zeroth-order optimization for memory-efficient LLM fine-tuning: a benchmark." *Proceedings of the 41st International Conference on Machine Learning*. 2024.

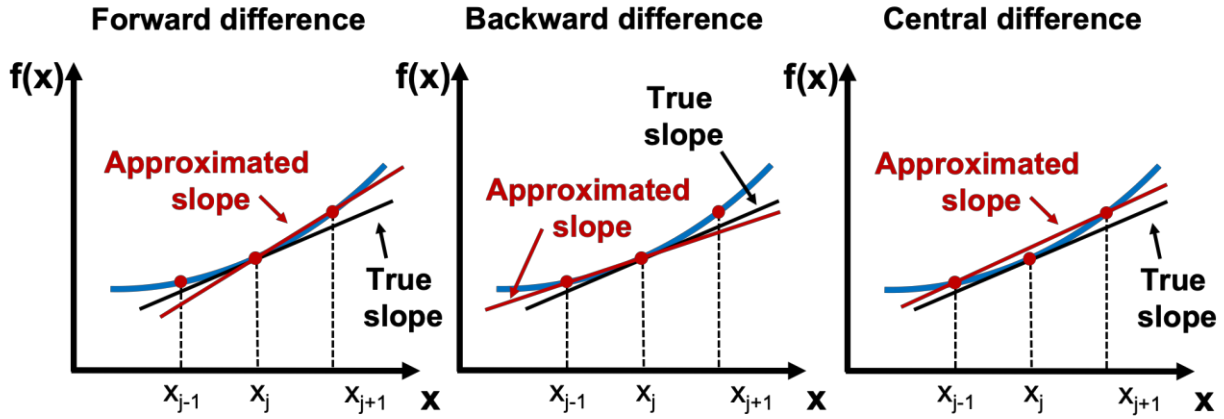
Fernández, Jesús García, Nasir Ahmad, and Marcel van Gerven. "A Unified Perspective on Optimization in Machine Learning and Neuroscience: From Gradient Descent to Neural Adaptation." *arXiv preprint arXiv:2510.18812* (2025).

https://sites.google.com/view/zo-tutorial-aaai-2024/

Memory Efficient Fine-Tuning-Zeroth-Order Gradient



The gradient can be approximated by random gradient estimation



$$\nabla f(\theta) \approx \frac{f(\theta + h\xi) - f(\theta - h\xi)}{2}(h\xi)^{-1}$$

Zhang, Yihua, et al. "Revisiting zeroth-order optimization for memory-efficient LLM fine-tuning: a benchmark." *Proceedings of the 41st International Conference on Machine Learning*. 2024.

Fernández, Jesús García, Nasir Ahmad, and Marcel van Gerven. "A Unified Perspective on Optimization in Machine Learning and Neuroscience: From Gradient Descent to Neural Adaptation." *arXiv preprint arXiv:2510.18812* (2025).

https://sites.google.com/view/zo-tutorial-aaai-2024/

Without Resources for Any Fine-Tuning

Prompt Engineering

```
Classify the sentiment of the following sentence as positive or negative:
"I love this movie!"
```

Ram, Ori, et al. "In-context retrieval-augmented language models." *Transactions of the Association for Computational Linguistics* 11, 2023.

Without Resources for Any Fine-Tuning

LUND UNIVERSITY

Prompt Engineering

```
Classify the sentiment of the following sentence as positive or negative:
"I love this movie!"
```

In-Context Learning

```
Review: "It was amazing!" → Label: Positive
Review: "Too boring." → Label: Negative
Review: "I loved the actors!" → Label:
```

Ram, Ori, et al. "In-context retrieval-augmented language models." *Transactions of the Association for Computational Linguistics* 11, 2023.

Without Resources for Any Fine-Tuning

LUND UNIVERSITY

Prompt Engineering

```
Classify the sentiment of the following sentence as positive or negative:
"I love this movie!"
```

In-Context Learning

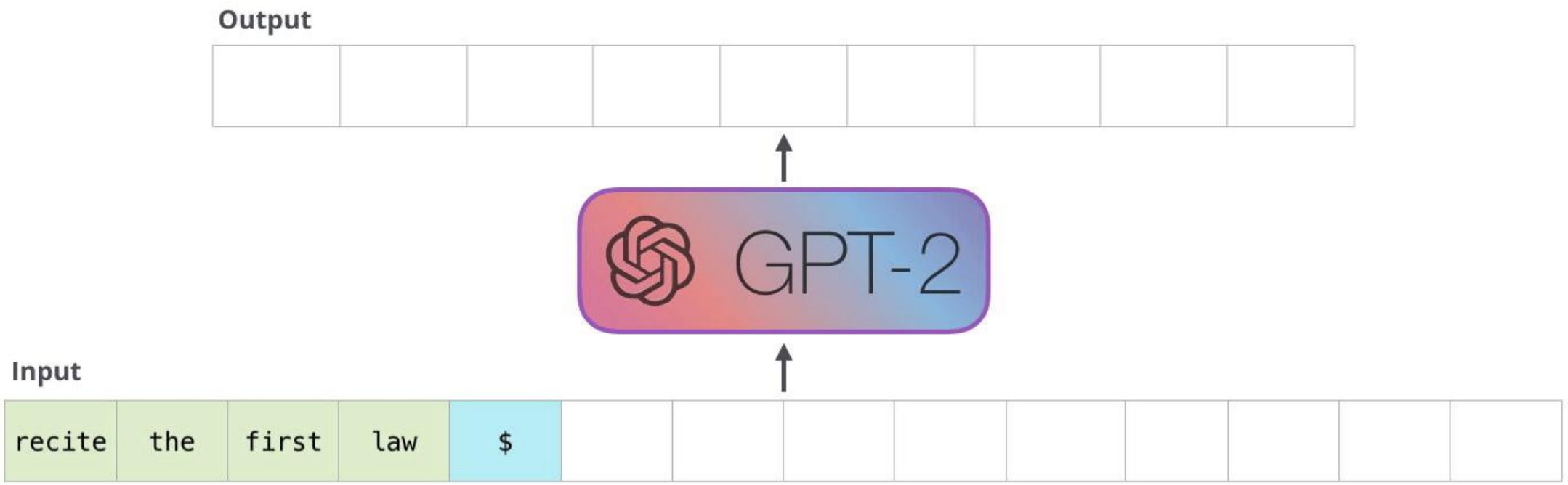
```
Review: "It was amazing!" → Label: Positive
Review: "Too boring." → Label: Negative
Review: "I loved the actors!" → Label:
```

Retrieval-Augmented Generation (RAG)

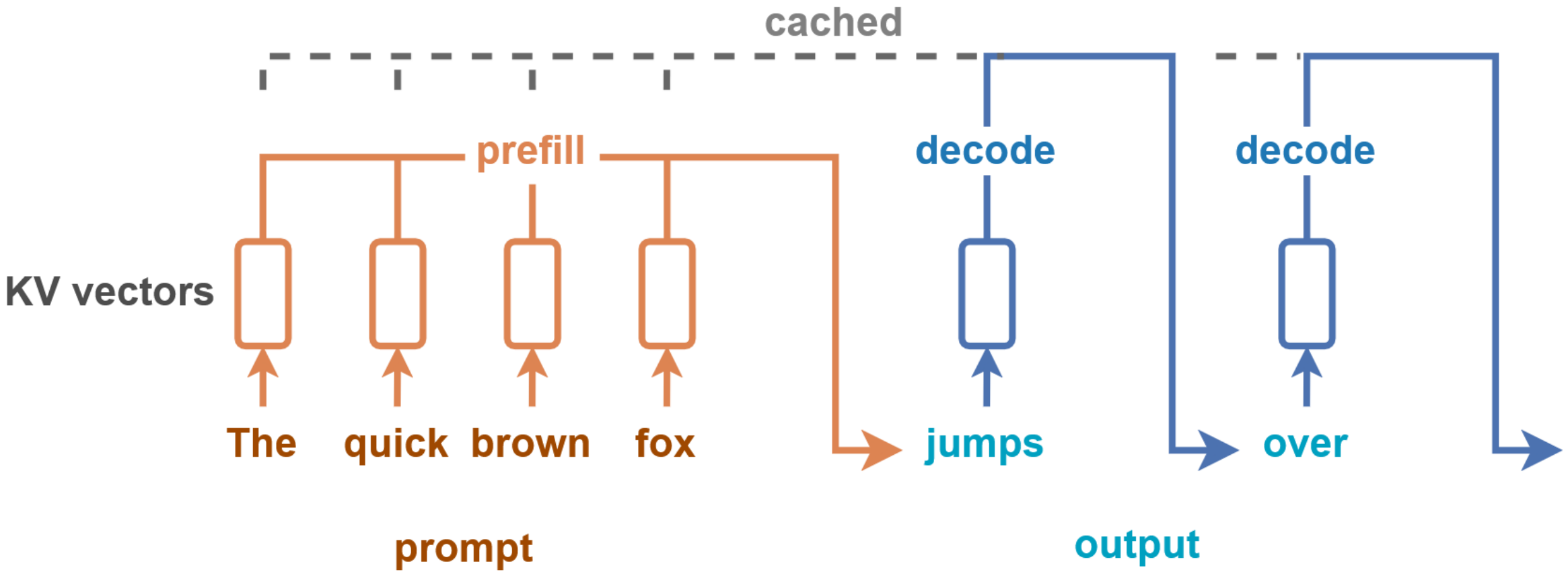
```
Query: What is photosynthesis?
↓
Retrieved: "Photosynthesis is the process by which green plants..."
↓
LLM: "Photosynthesis is the process used by plants..."
```

Ram, Ori, et al. "In-context retrieval-augmented language models." *Transactions of the Association for Computational Linguistics* 11, 2023.

Inference for LLMS (Generation Task)

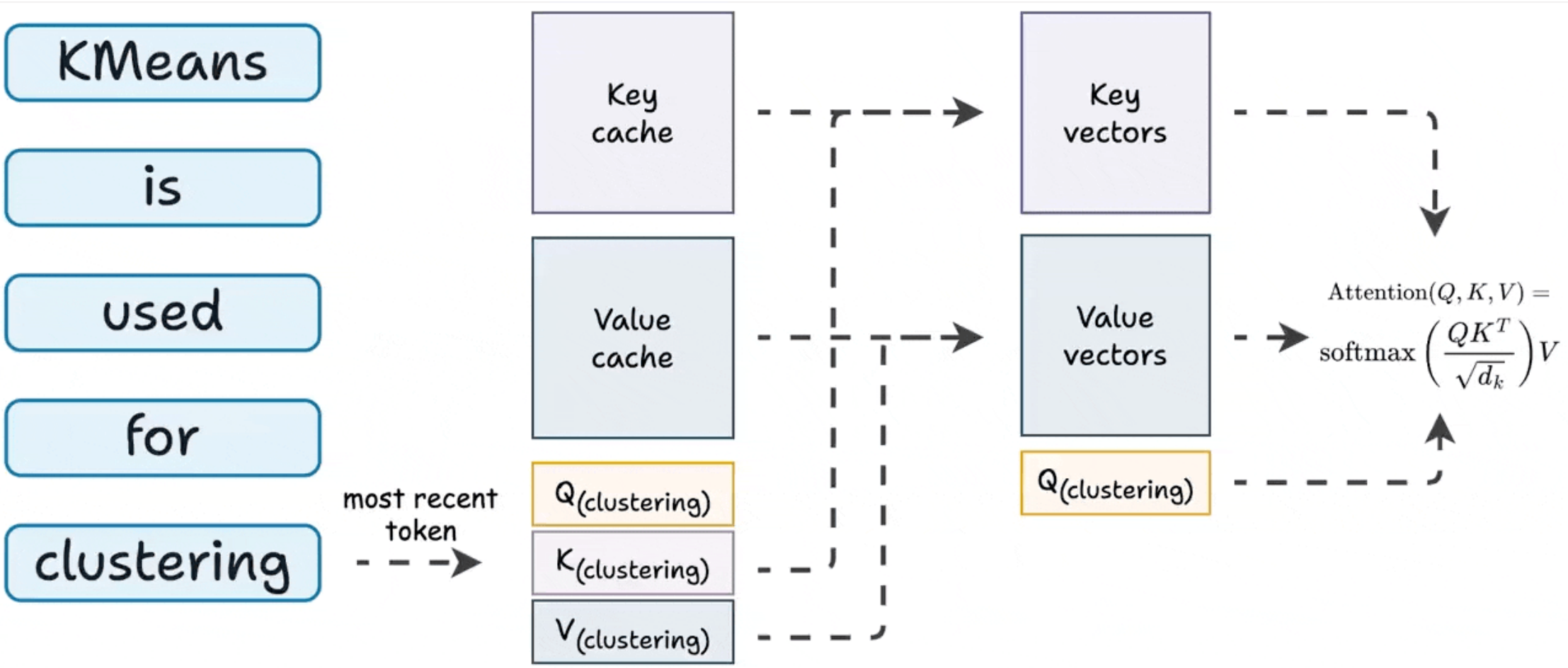


LLM Inference Steps



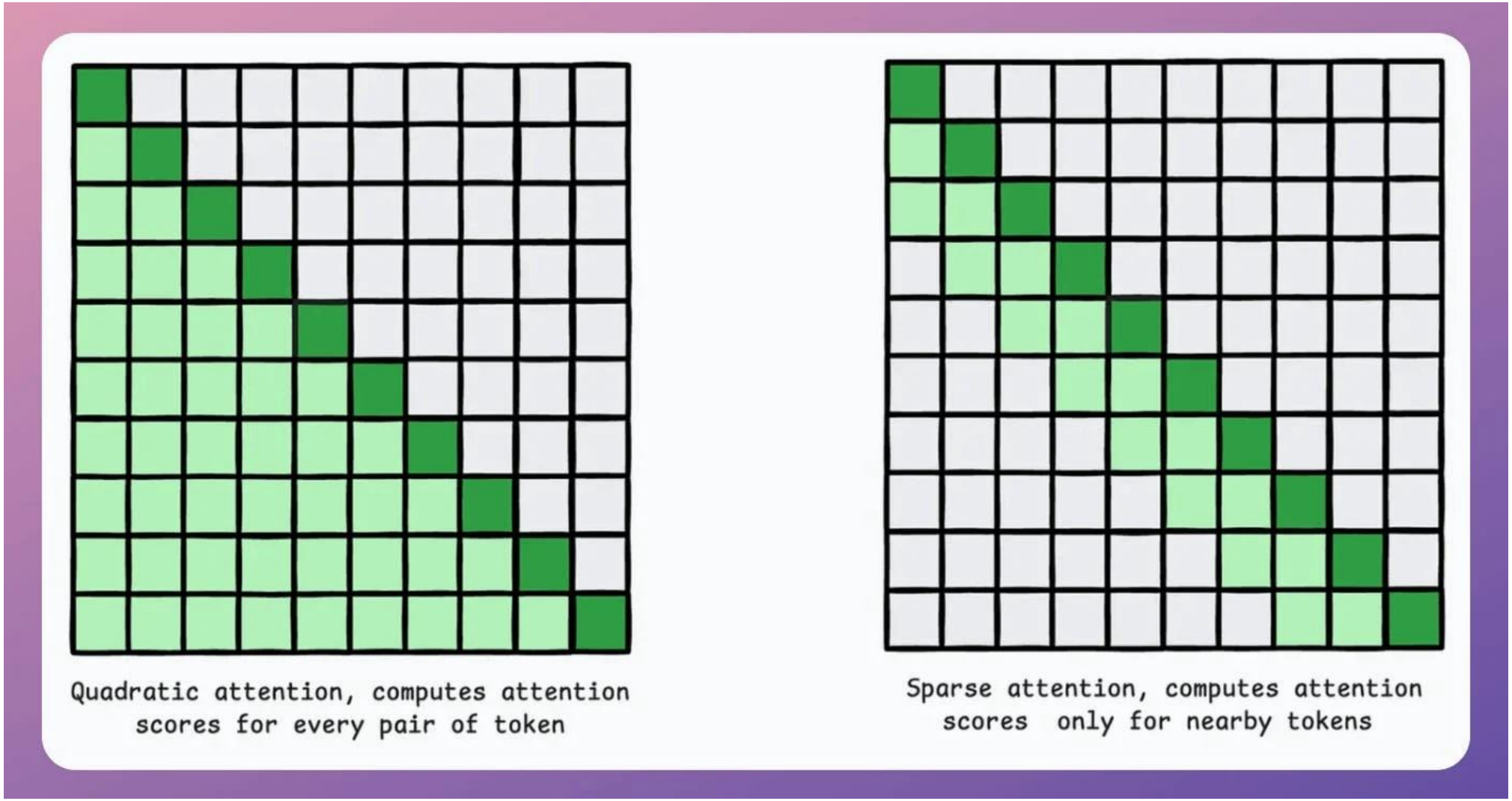
https://huggingface.co/blog/tngtech/llm-performance-prefill-decode-concurrent-requests

Efficient Inference for LLMs-KV Cache



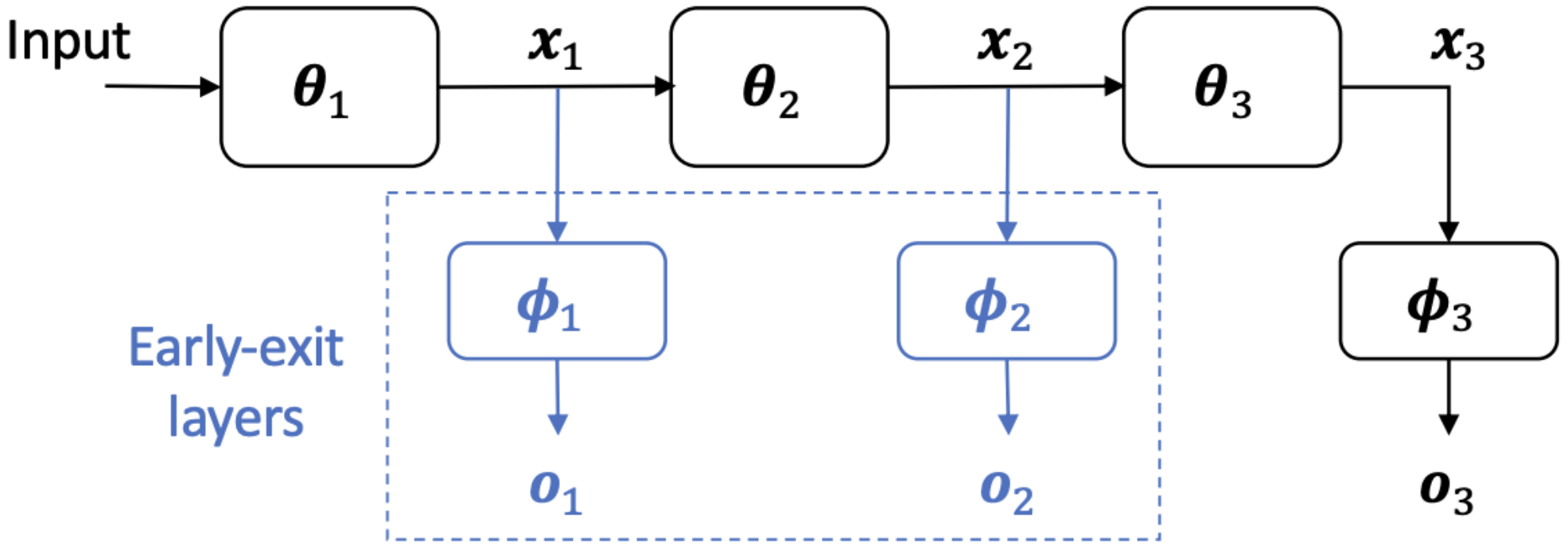
https://blog.dailydoseofds.com/p/kv-caching-in-llms-explained-visually

Efficient Inference for LLMs-Sparse Attention



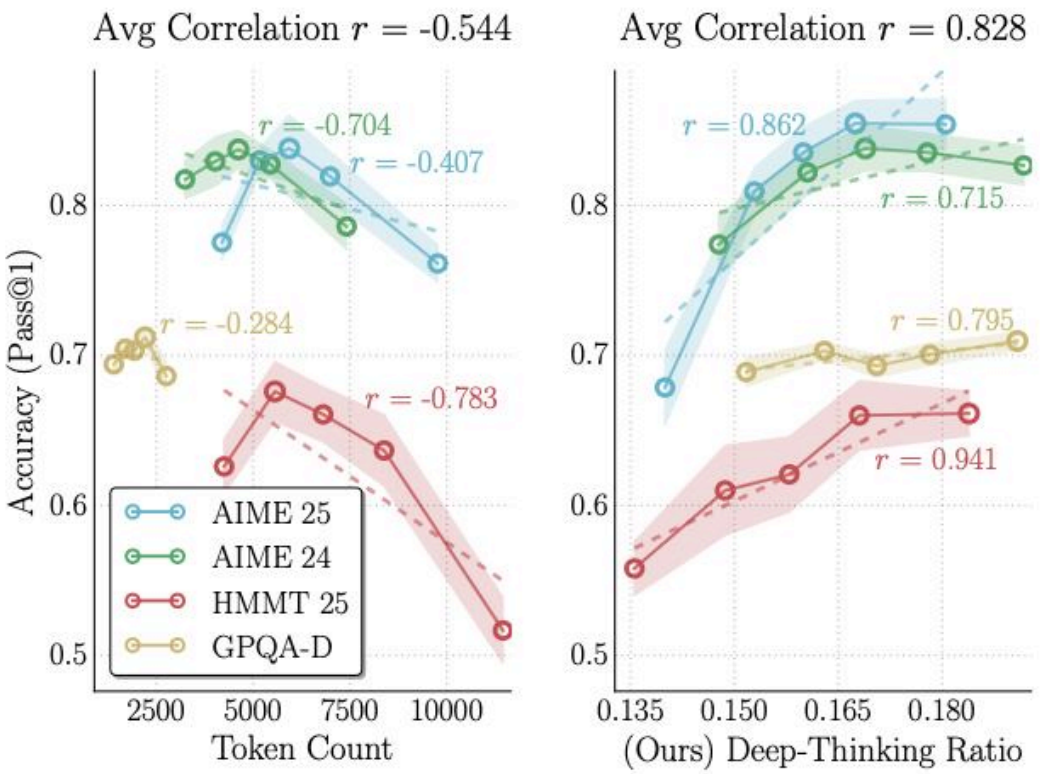
https://www.dailydoseofds.com/p/techniques-to-extend-context-length-of-llms/

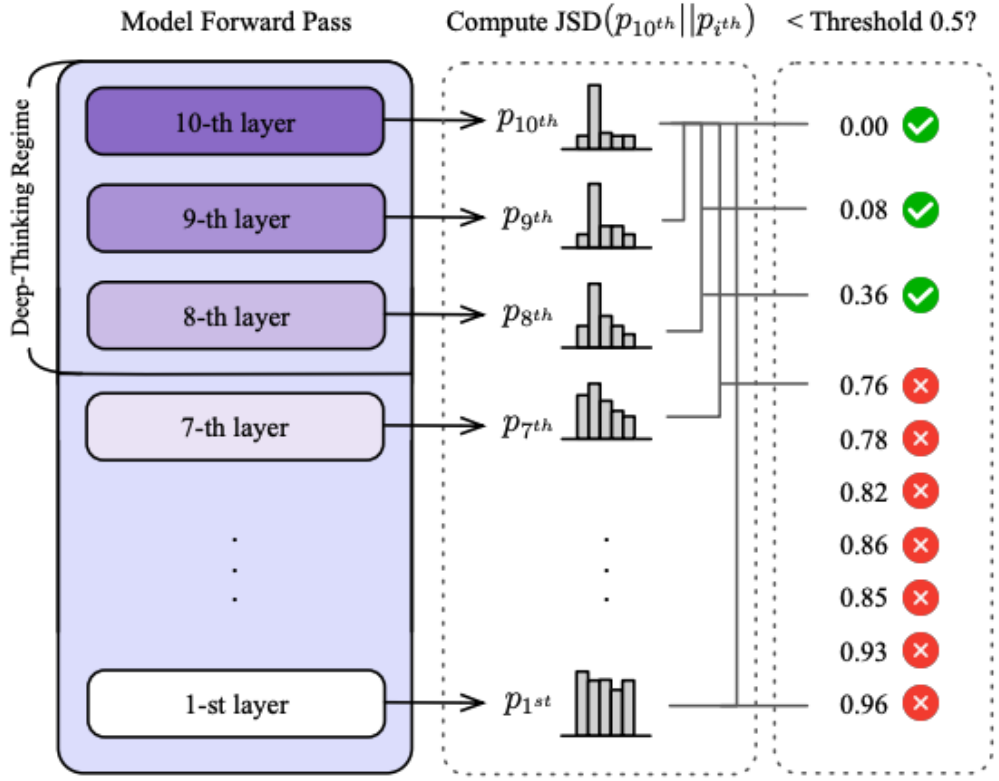
Efficient Inference for LLMs-Early Existing



Chen, Yanxi, et al. "EE-LLM: Large-Scale Training and Inference of Early-Exit Large Language Models with 3D Parallelism." *ICML*. 2024.

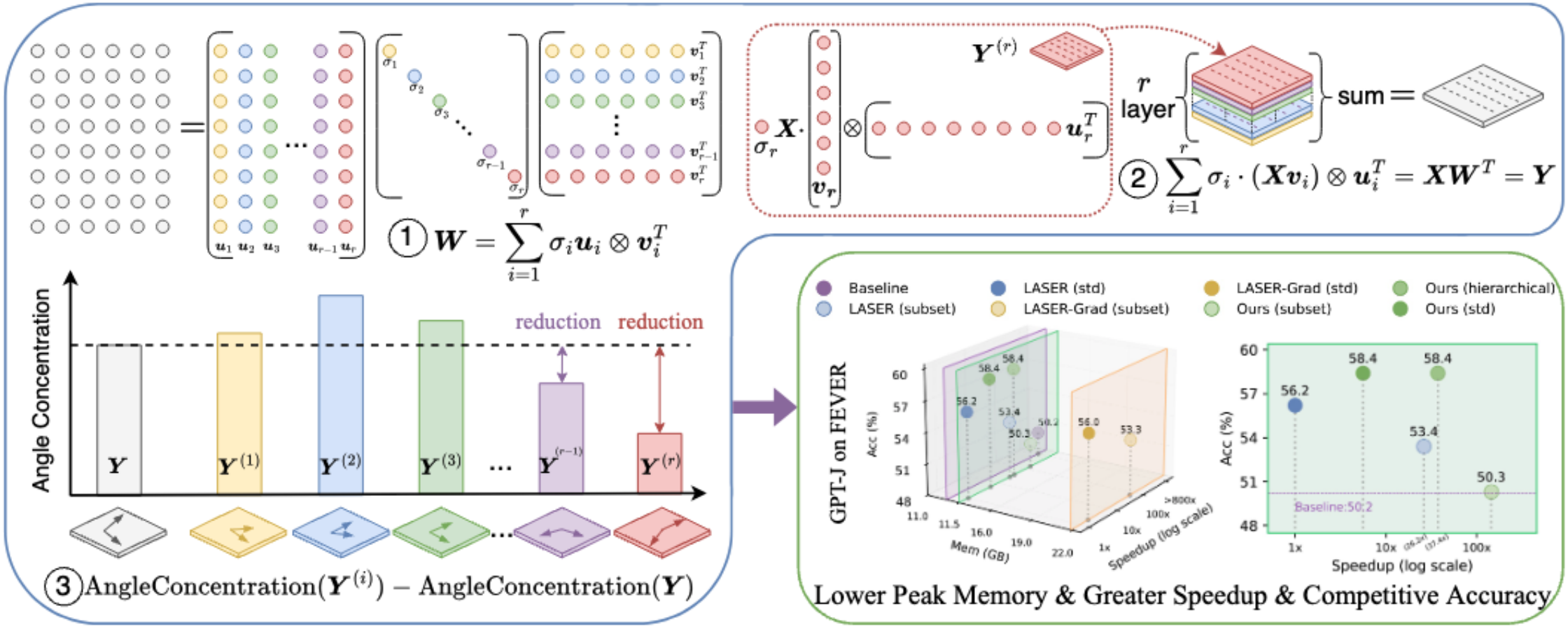
Early-Existing





Chen, Wei-Lin, et al. "Think Deep, Not Just Long: Measuring LLM Reasoning Effort via Deep-Thinking Tokens." *ICML 2026*

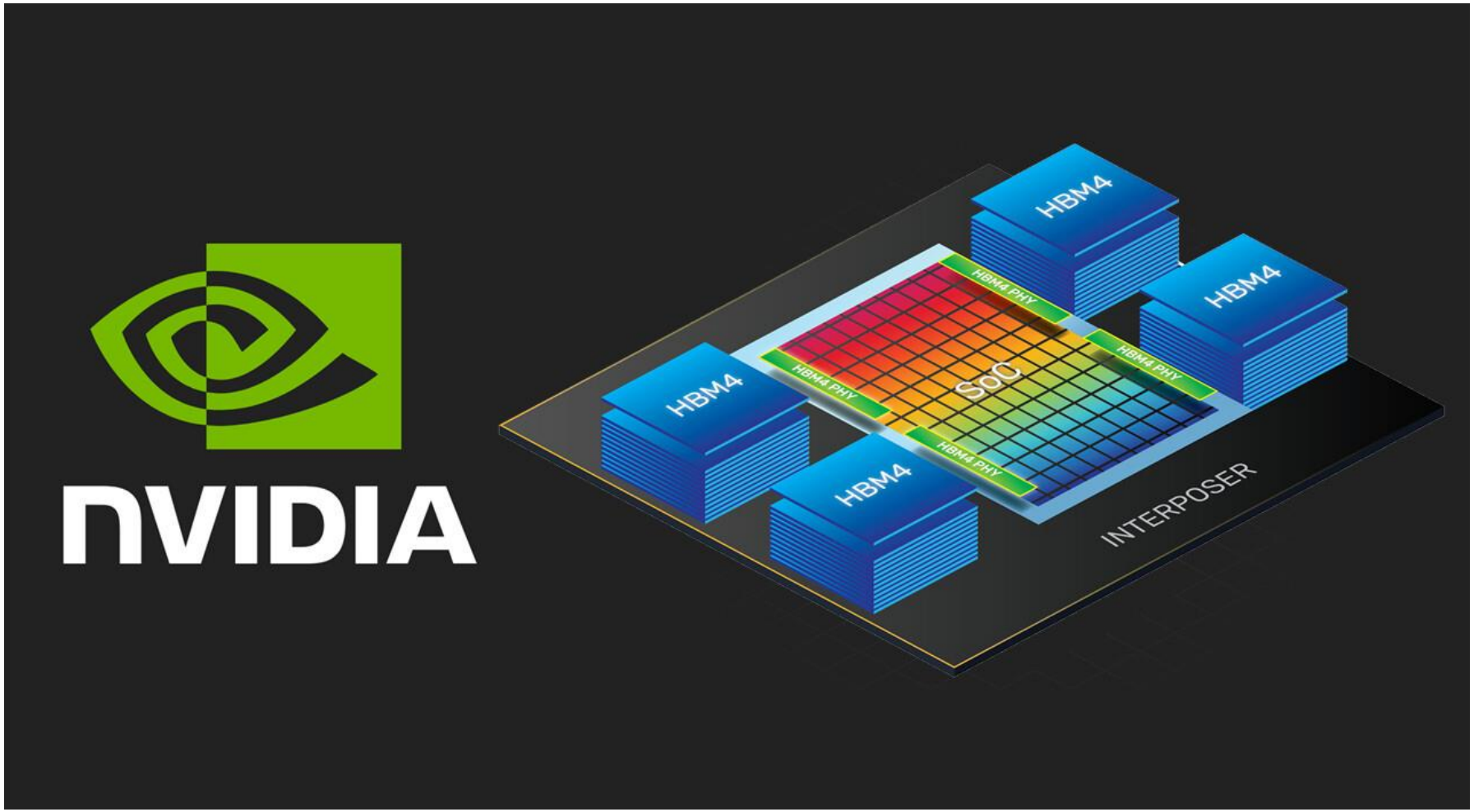
Efficient Adaptation During Inference



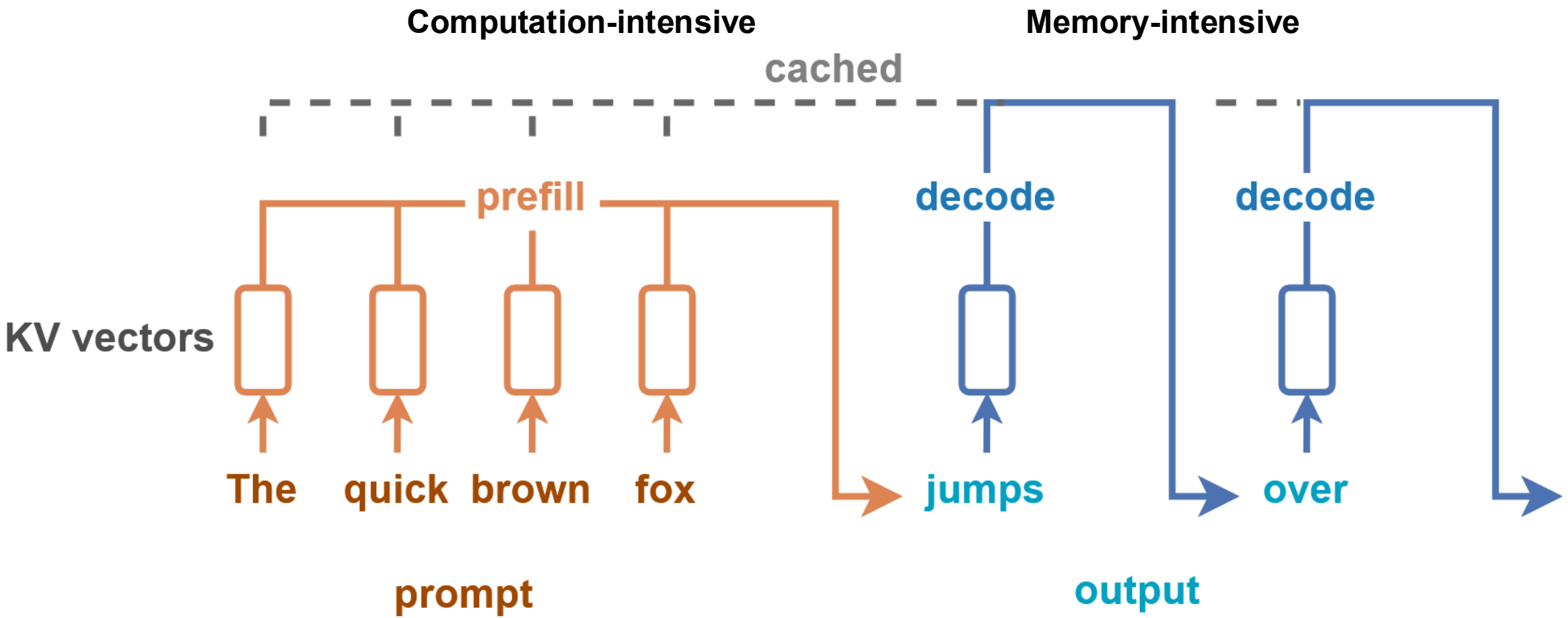
Our Forward-Only Search **Results**

Baichuan Huang, Ananth Balashankar, Amir Aminifar, “Efficient LLM Adaptation with Forward-Only Passes,” The 40th Annual Conference on Neural Information Processing Systems (NeurIPS), Australia, Dec. 2026.

AI Hardware-NVIDIA GPU



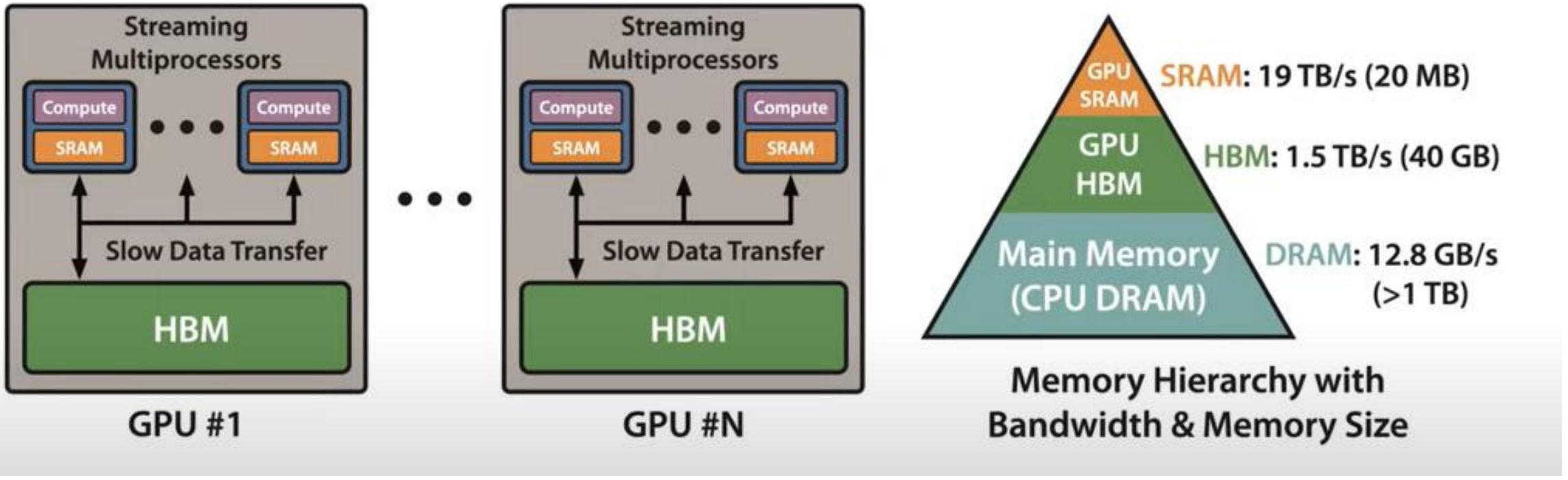
LLM Serving System



https://huggingface.co/blog/tngtech/llm-performance-prefill-decode-concurrent-requests

Efficient Inference for LLMs-Flash Attention

Background: GPU Compute Model & Memory Hierarchy



Dao, Tri, et al. "Flashattention: Fast and memory-efficient exact attention with io-awareness." *Advances in neural information processing systems* 35 (2022): 16344-16359.

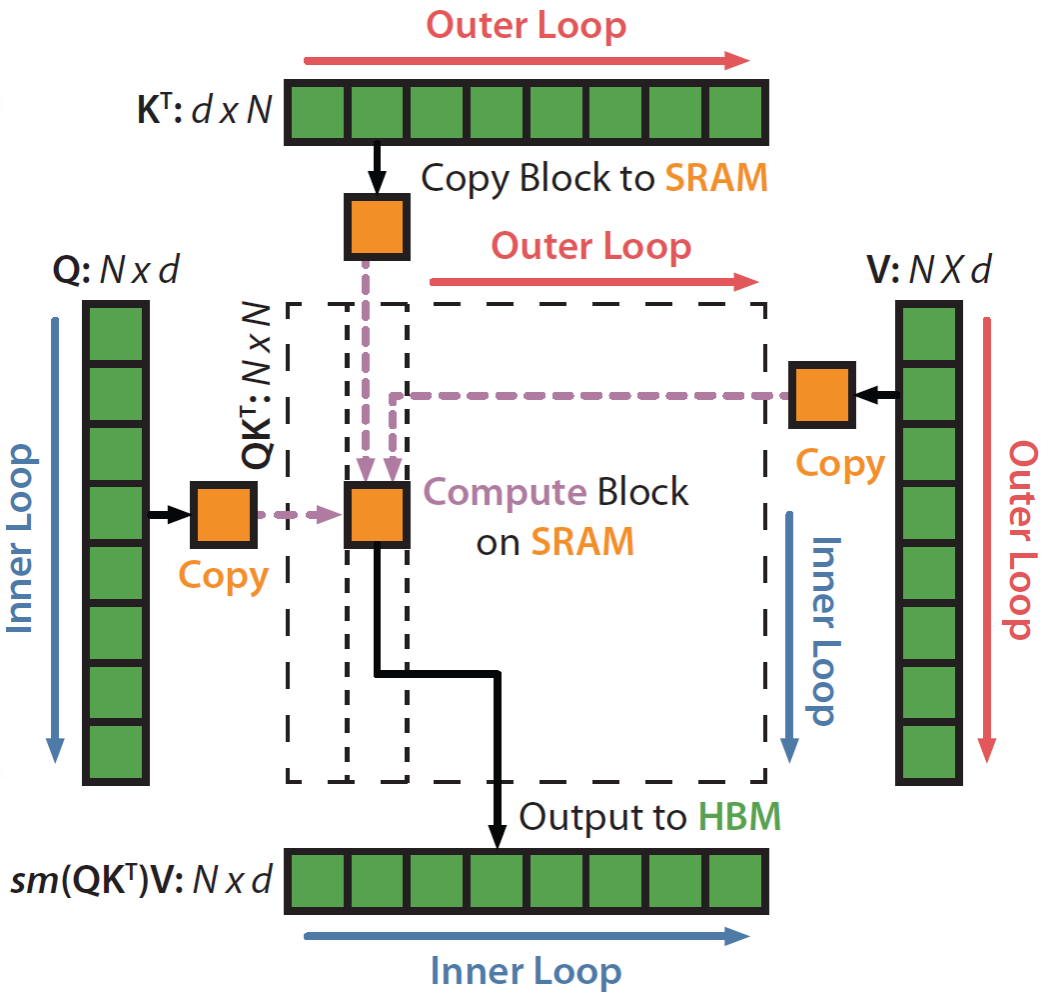
Efficient Inference for LLMs-Flash Attention

Vanilla Attention	Flash Attention
1. Matmul_op (Q,K) a. Read Q,K to SRAM b. Compute matmul A=QxK c. Write A to HBM 2. Mask_op a. Read A to SRAM b. Mask A into A' c. Write A' to HBM 3. Softmax_op a. Read A' to SRAM b. Softmax A' into A'' c. Write A'' to HBM	1. Read Q,K to SRAM 2. Compute A = QxK 3. Mask A into A' 4. Softmax A' into A'' 5. Write A'' to HBM

Dao, Tri, et al. "Flashattention: Fast and memory-efficient exact attention with io-awareness." *Advances in neural information processing systems* 35 (2022): 16344-16359.

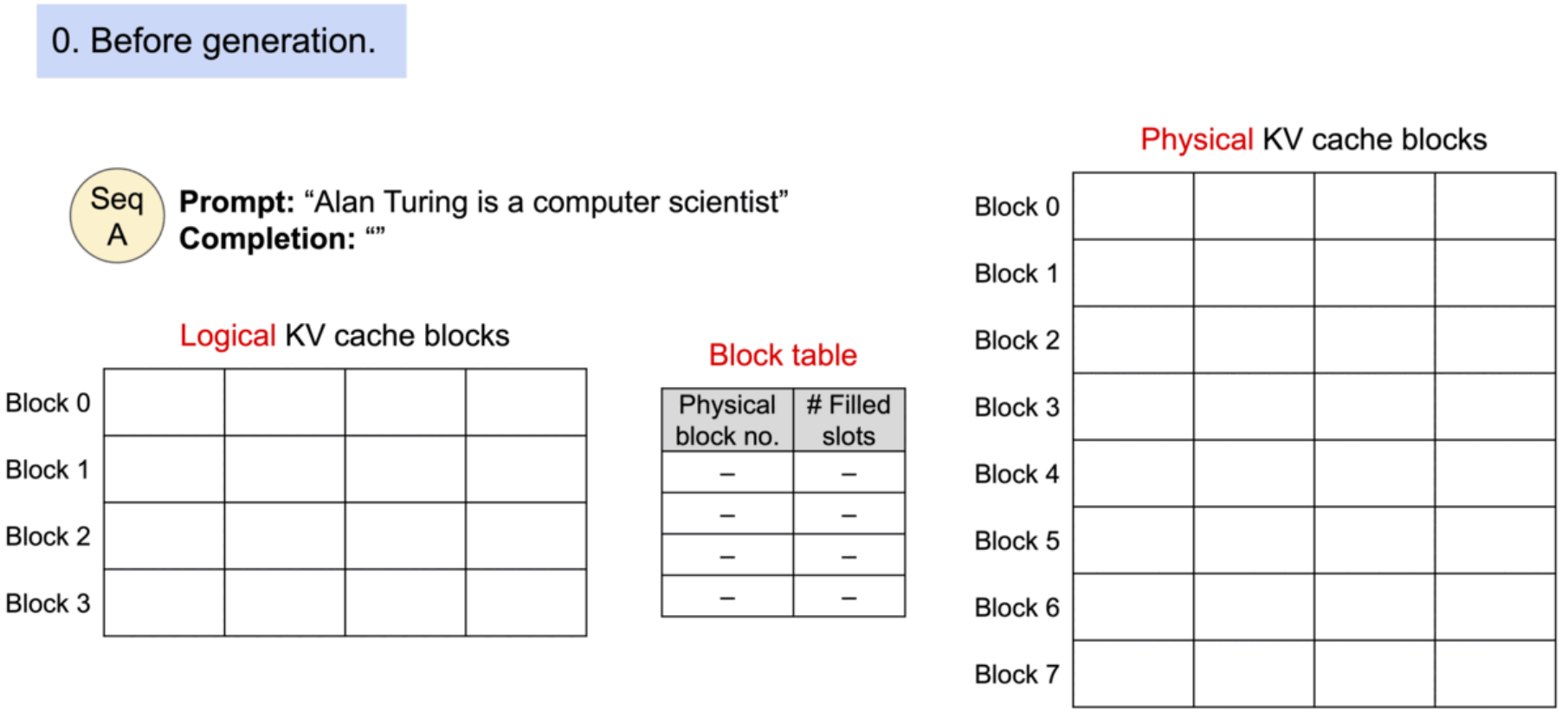
Efficient Inference for LLMs-Flash Attention

Vanilla Attention	Flash Attention
1. Matmul_op (Q,K) a. Read Q,K to SRAM b. Compute matmul A=QxK c. Write A to HBM 2. Mask_op a. Read A to SRAM b. Mask A into A' c. Write A' to HBM 3. Softmax_op a. Read A' to SRAM b. Softmax A' into A'' c. Write A'' to HBM	1. Read Q,K to SRAM 2. Compute A = QxK 3. Mask A into A' 4. Softmax A' into A'' 5. Write A'' to HBM



Dao, Tri, et al. "Flashattention: Fast and memory-efficient exact attention with io-awareness." *Advances in neural information processing systems* 35 (2022): 16344-16359.

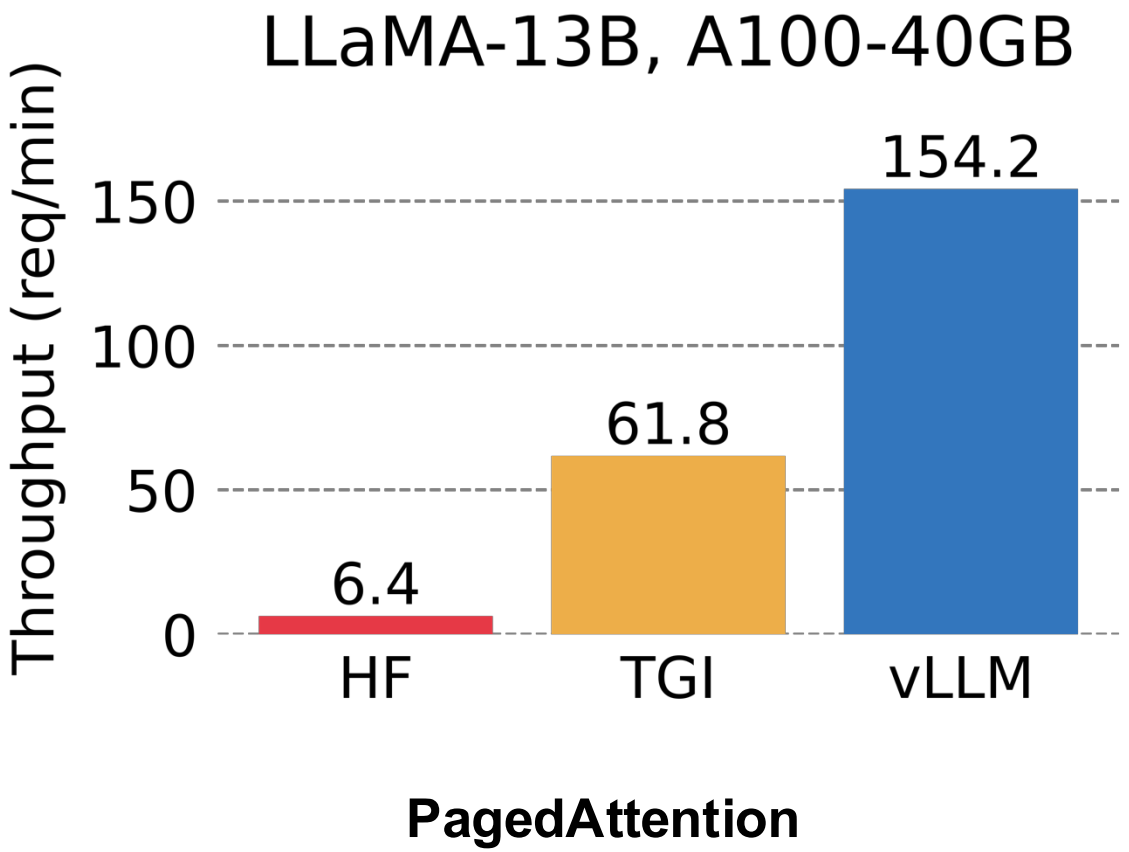
LLM Serving System-vLLM



PagedAttention

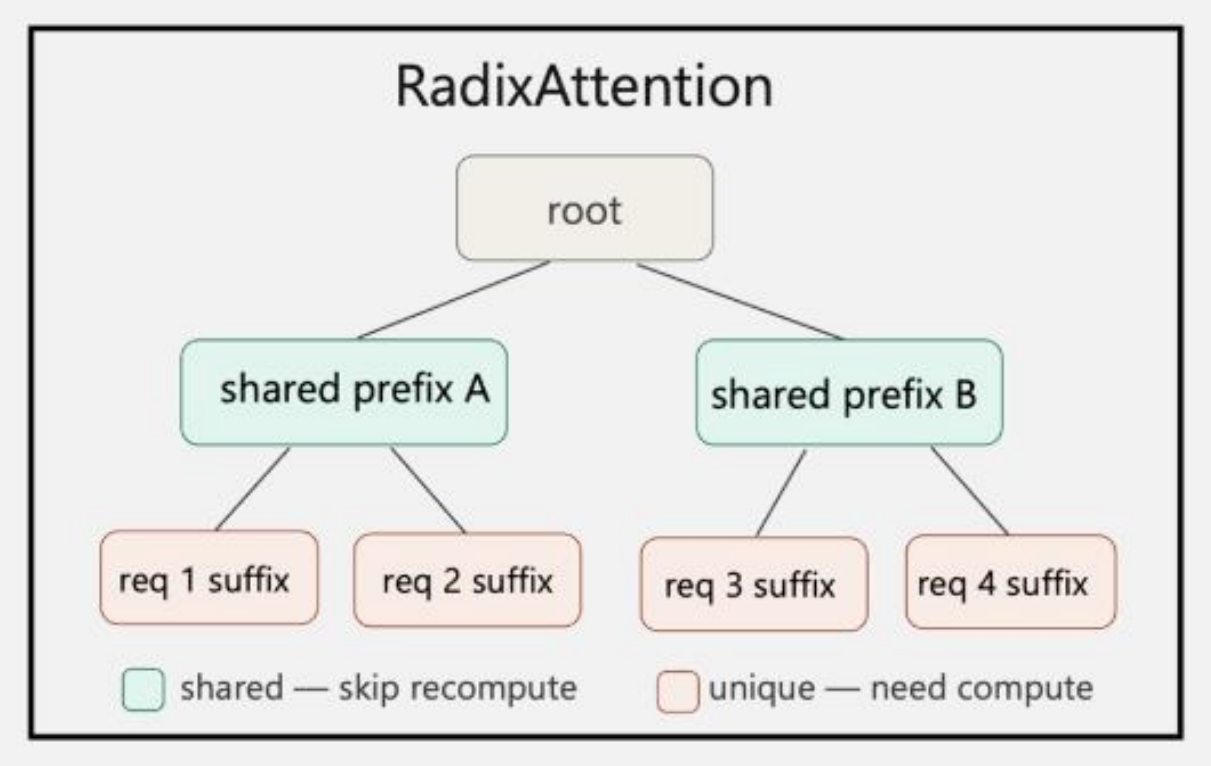
Kwon, Woosuk, et al. "Efficient memory management for large language model serving with pagedattention." SOSP 2023.

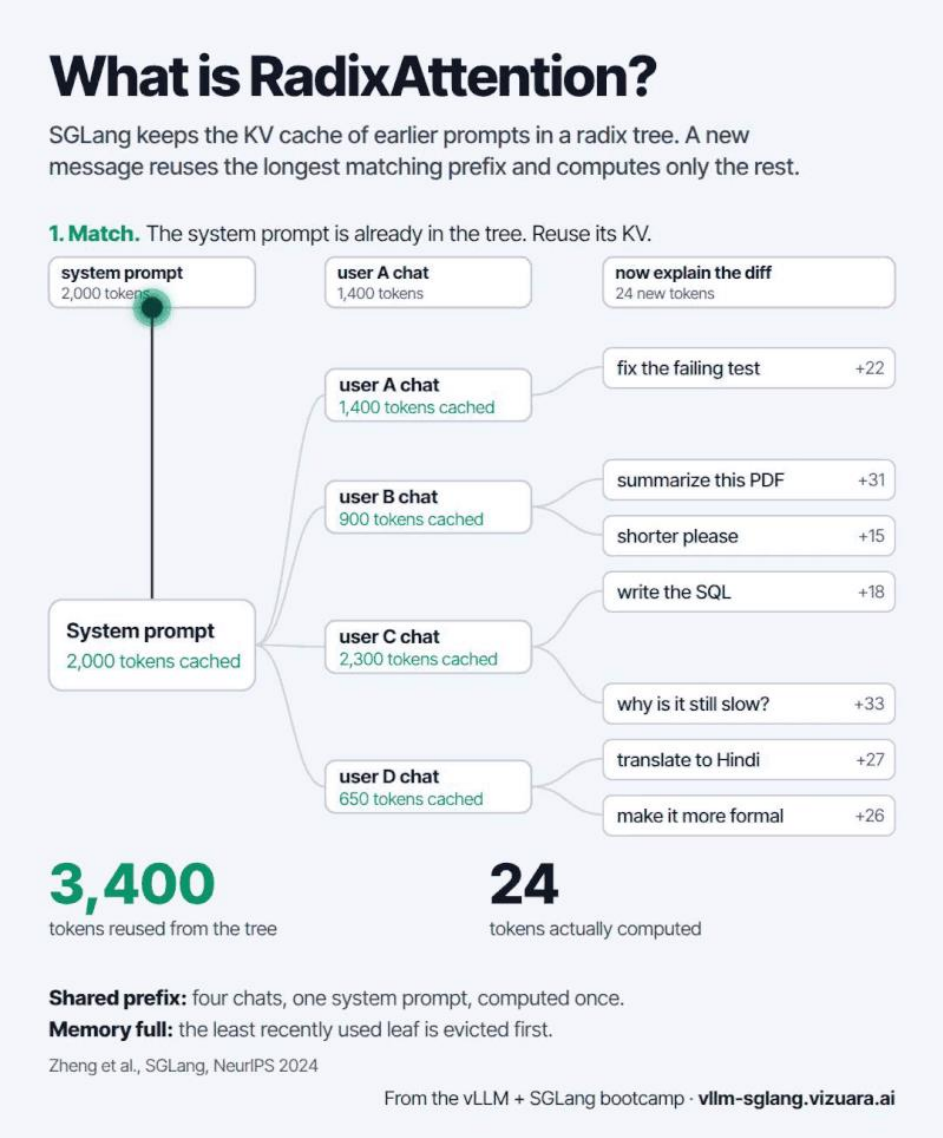
LLM Serving System-vLLM



Kwon, Woosuk, et al. "Efficient memory management for large language model serving with pagedattention." SOSP 2023.

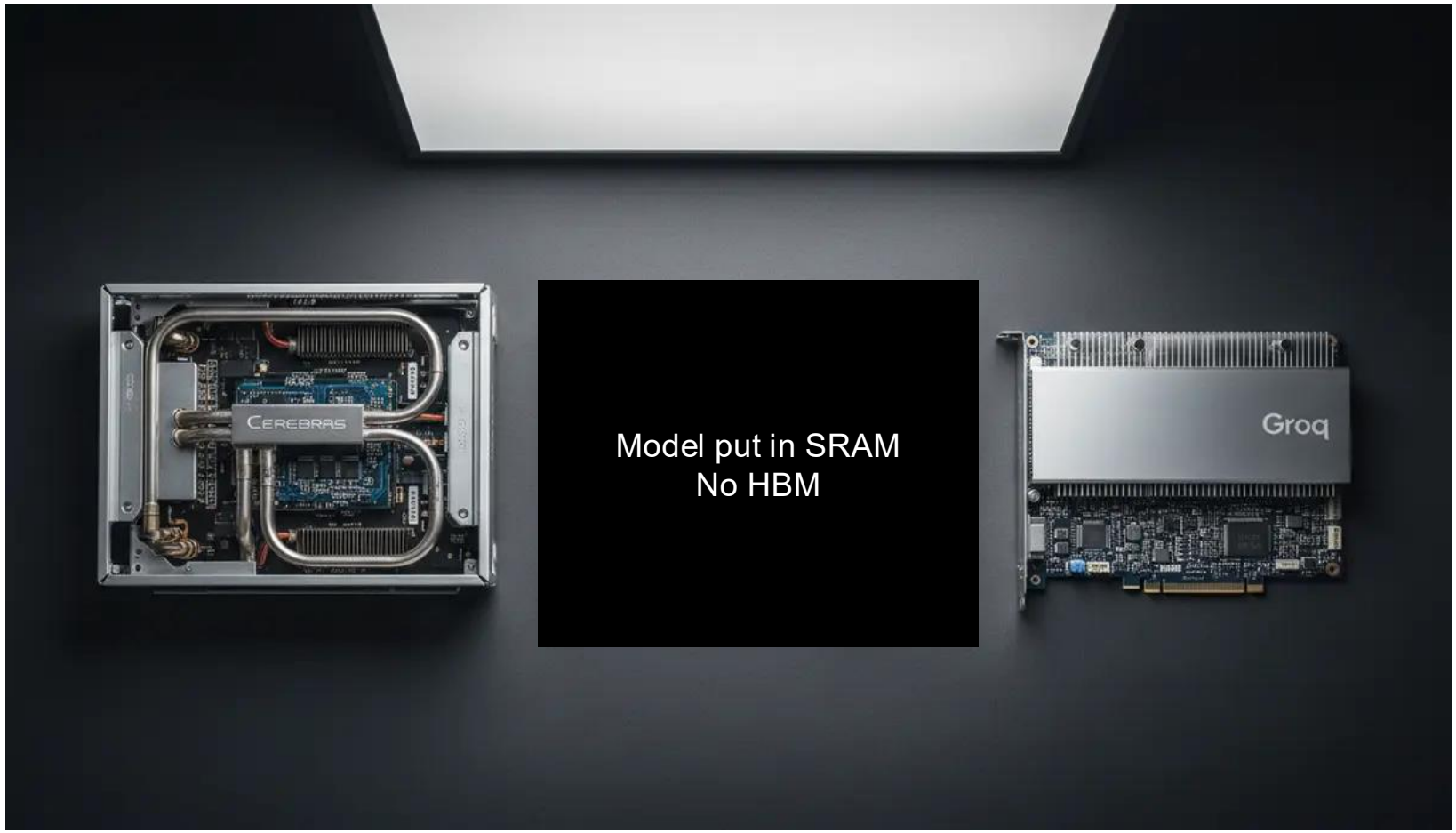
LLM Serving System-SGLang



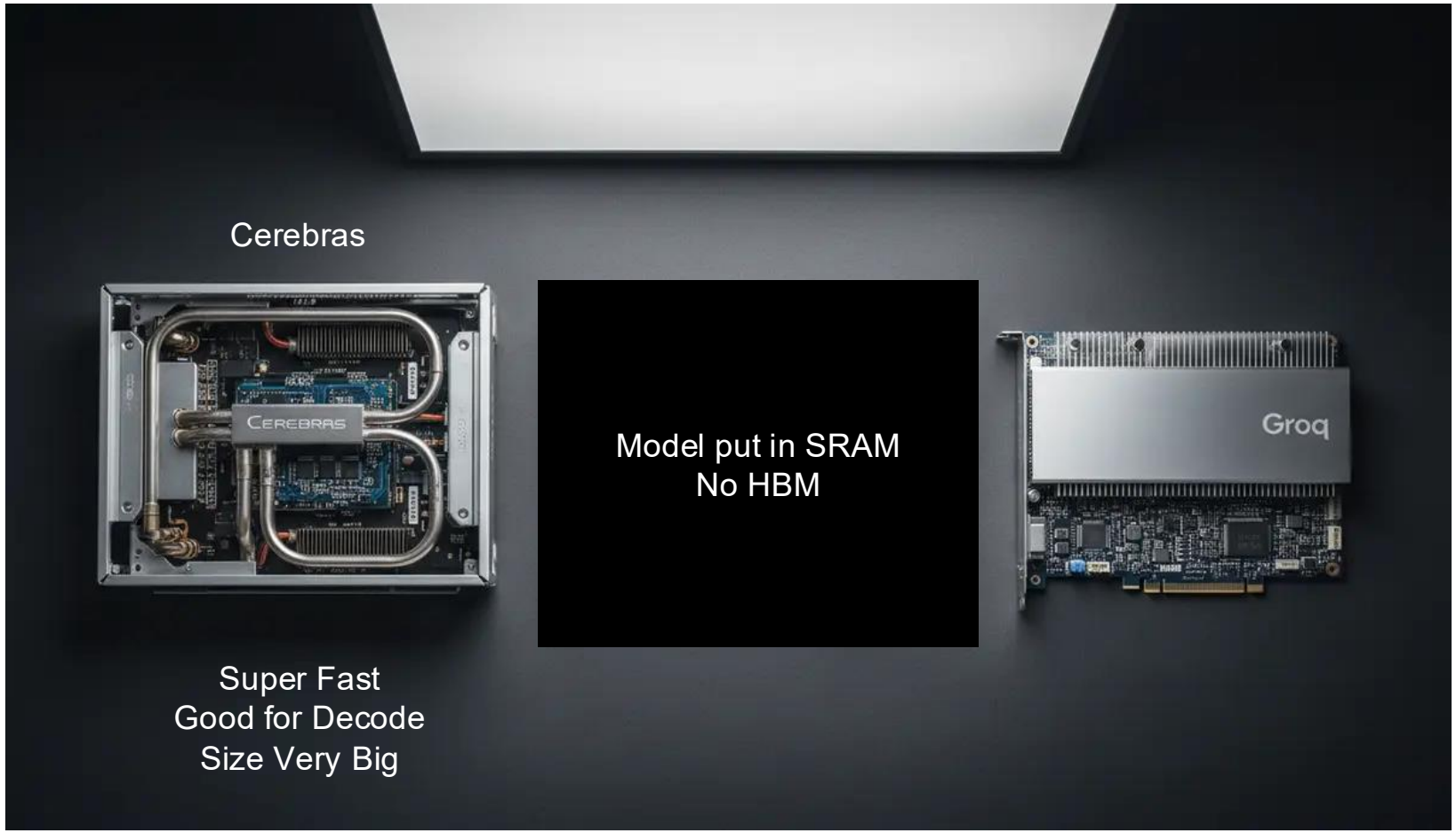


Zheng, Lianmin, et al. "Sglang: Efficient execution of structured language model programs." *Advances in neural information processing systems* 37 (2024).
https://lnkd.in/p/egeyaSGV
https://developer.arm.com/community/arm-community-blogs/b/ai-blog/posts/bringing-sglang-high-performance-llm-inference-to-arm-neoverse

AI Hardware



AI Hardware


Cerebras
CEREBRAS
Model put in SRAM
No HBM
Groq
Super Fast
Good for Decode
Size Very Big

AI Hardware

Cerebras



Super Fast
Good for Decode
Size Very Big

Model put in SRAM
No HBM

Groq



LPU
Static Scheduling
Compiler Pre-Execution
Reduce runtime contention and latency fluctuations