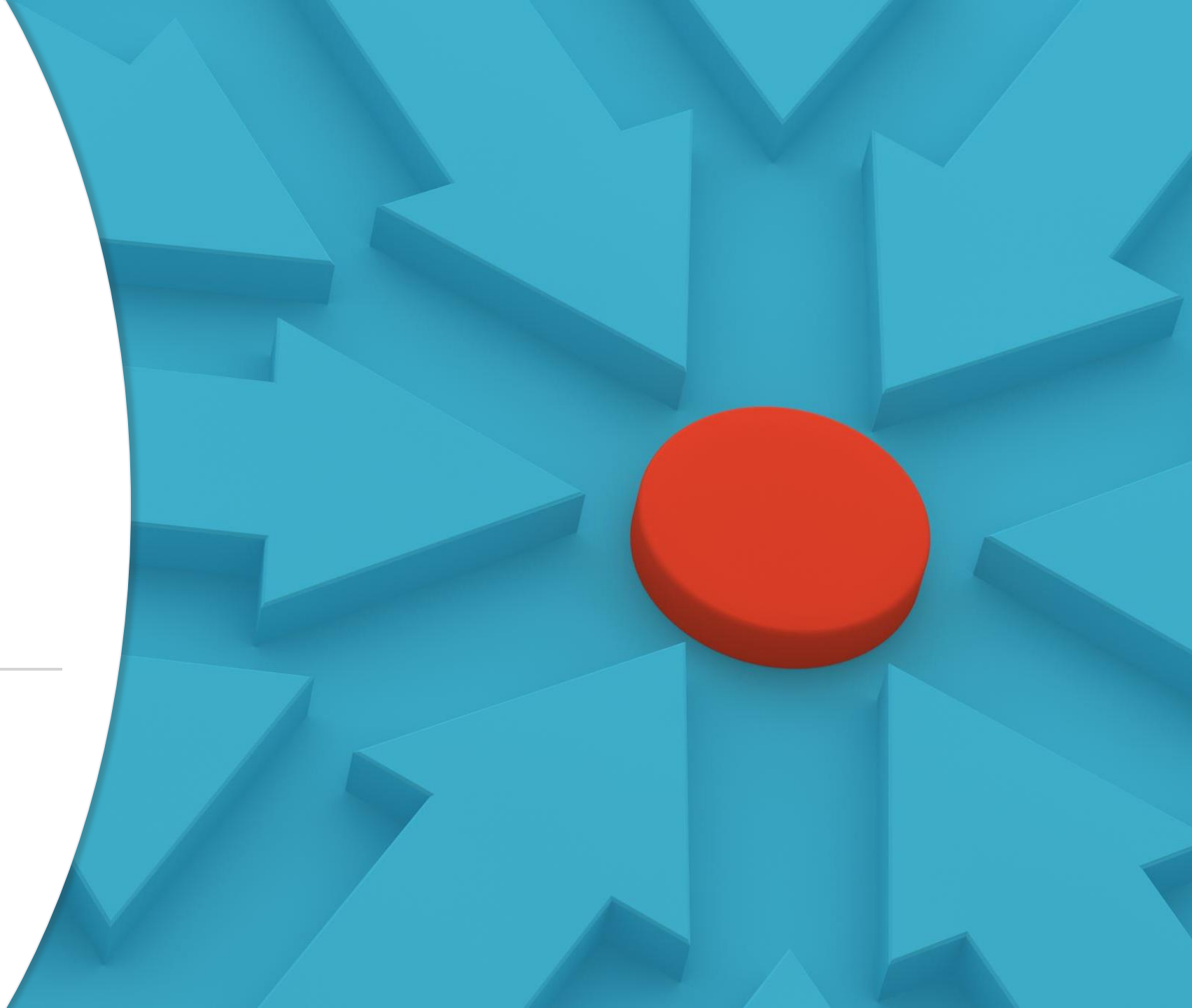
A brief introduction to

GAI Principle

Bo Yuan

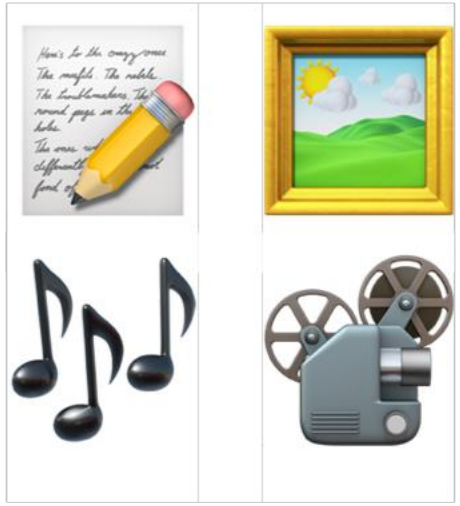


Outline

- Transformer
- LLM Training
- DeepSeek

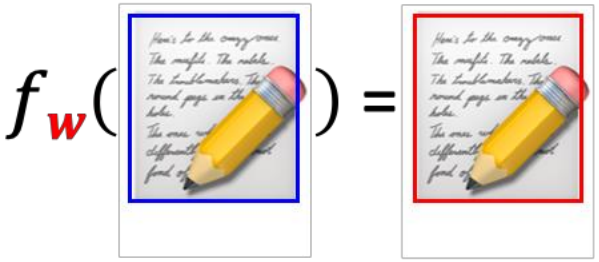
What is GAI?

GAI = Generative Artificial Intelligent



via

LLM = Large Language Model



Sequence to Sequence

Hardness of GAI

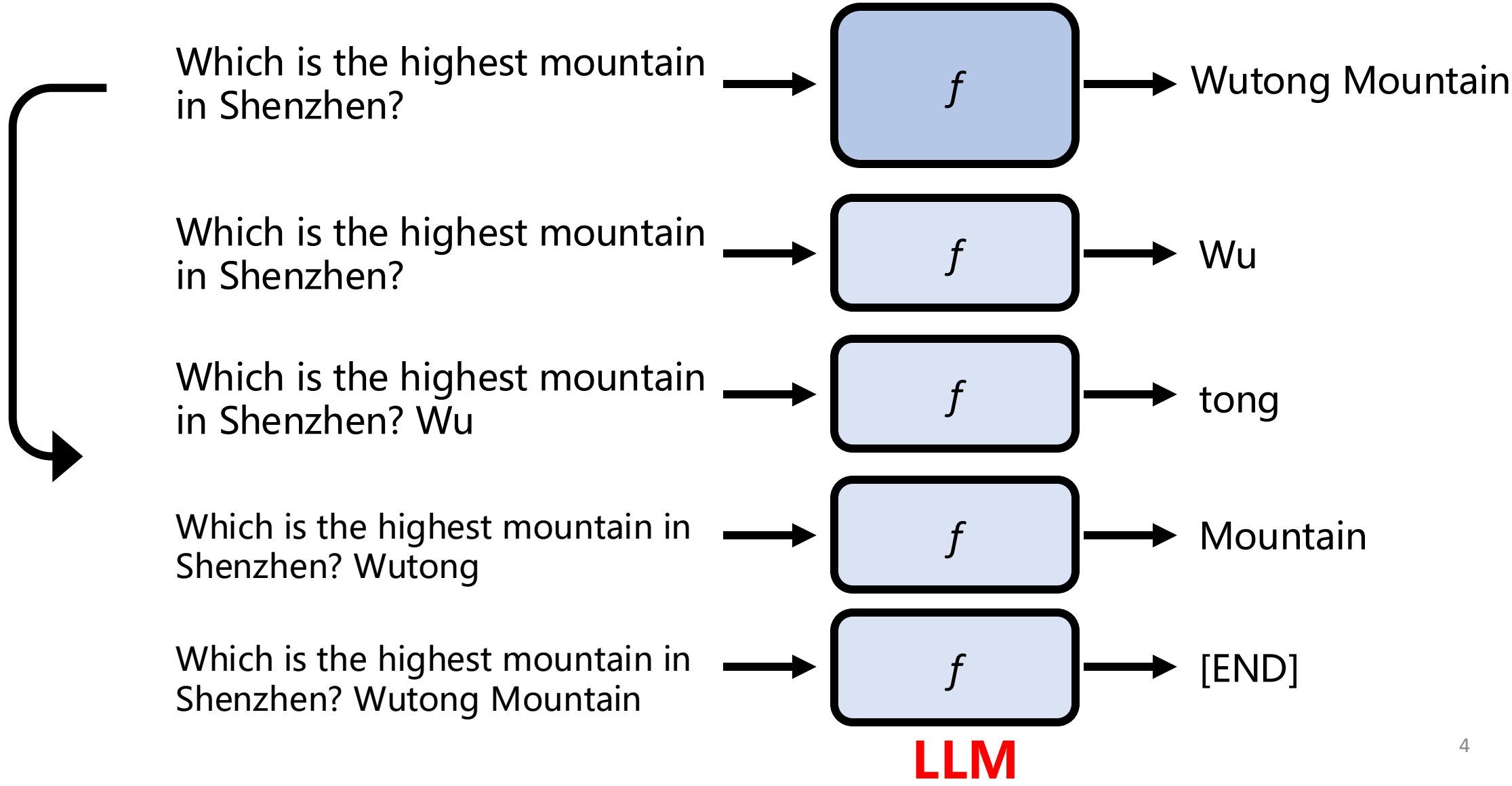
Q: Write a poem consisting of 100 words.

A:

Number of possible answers?

$1000 \times 1000 \times 1000 \times 1000 \ldots\ldots = 1000^{100} = 10^{300}$

LLM


Which is the highest mountain in Shenzhen?
f
Wutong Mountain
Which is the highest mountain in Shenzhen?
f
Wu
Which is the highest mountain in Shenzhen? Wu
f
tong
Which is the highest mountain in Shenzhen? Wutong
f
Mountain
Which is the highest mountain in Shenzhen? Wutong Mountain
f
[END]
LLM

Language Model

N-gram → MLP/ANN/FF → RNN → Transformer

Transformer

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

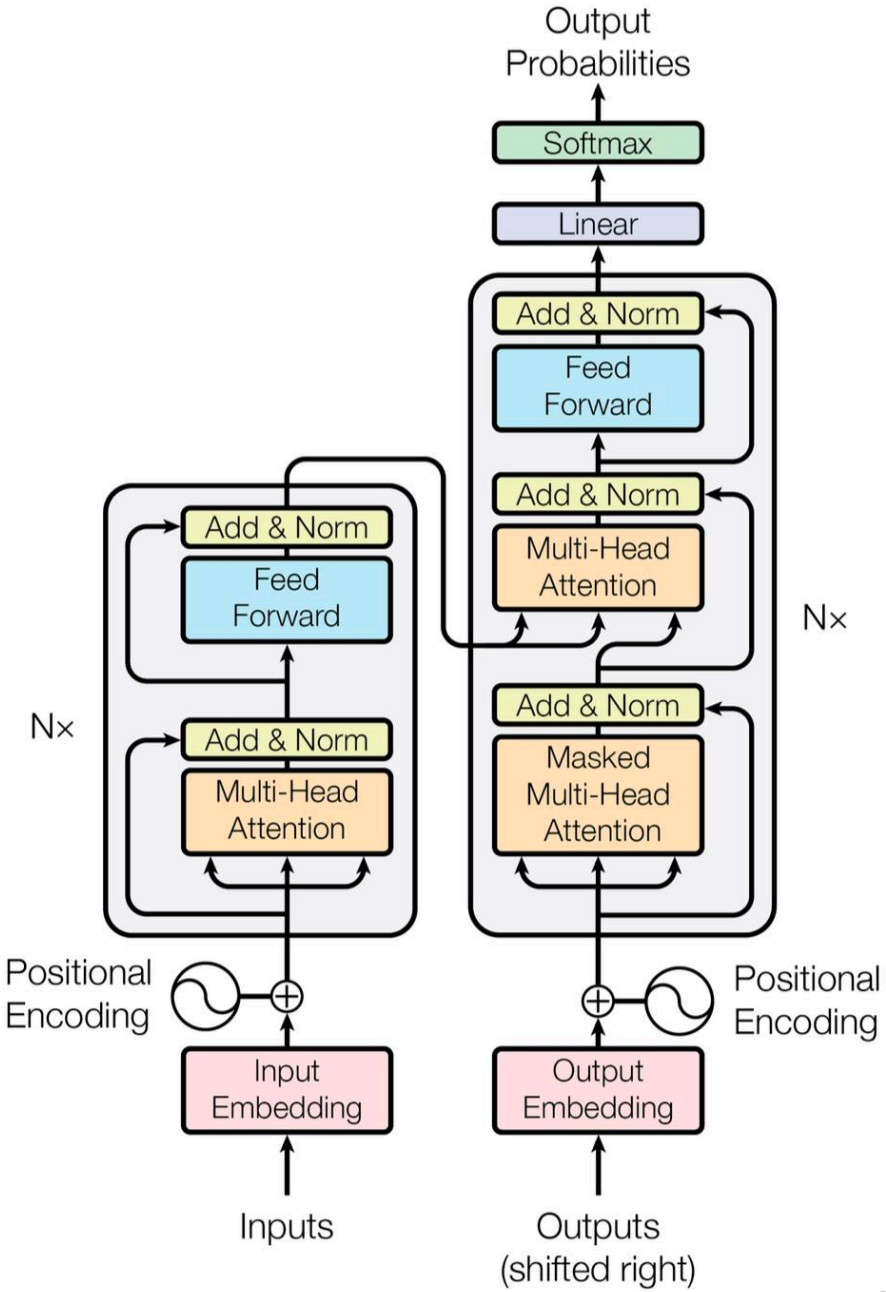
Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

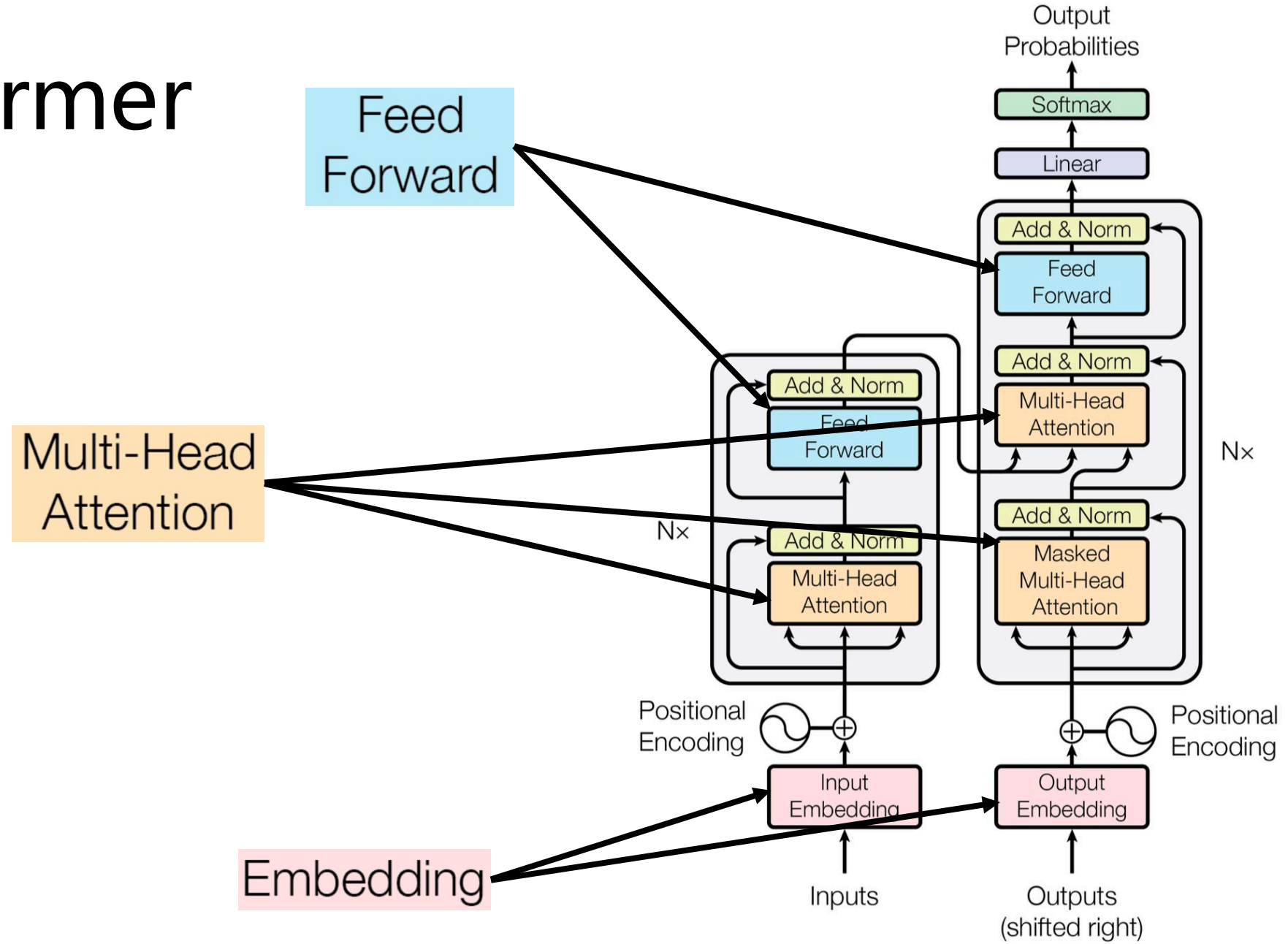
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

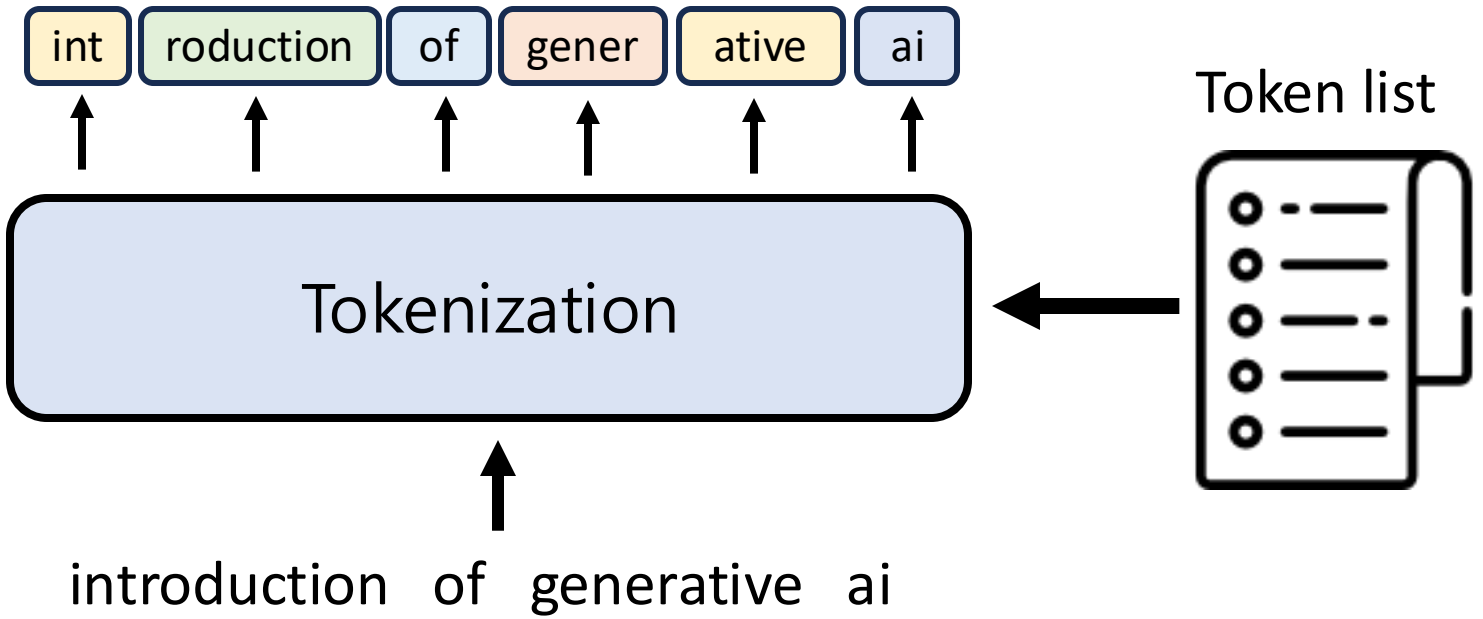
Illia Polosukhin* ‡ illia.polosukhin@gmail.com



Transformer

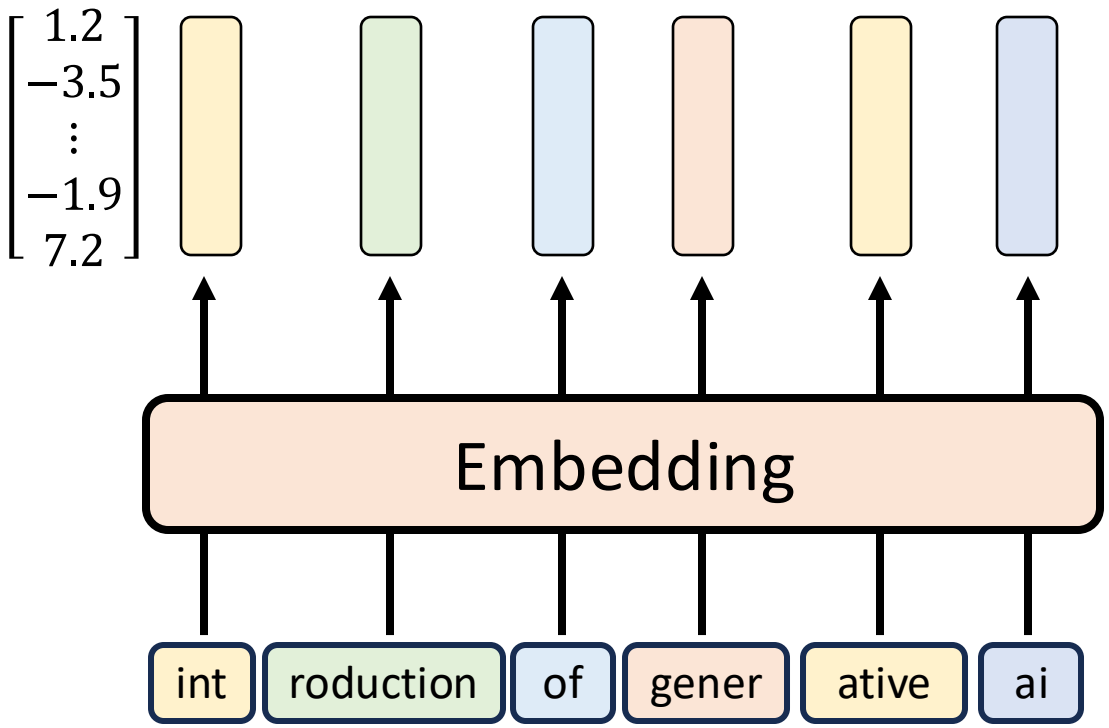


Tokenization (Pre-processing)

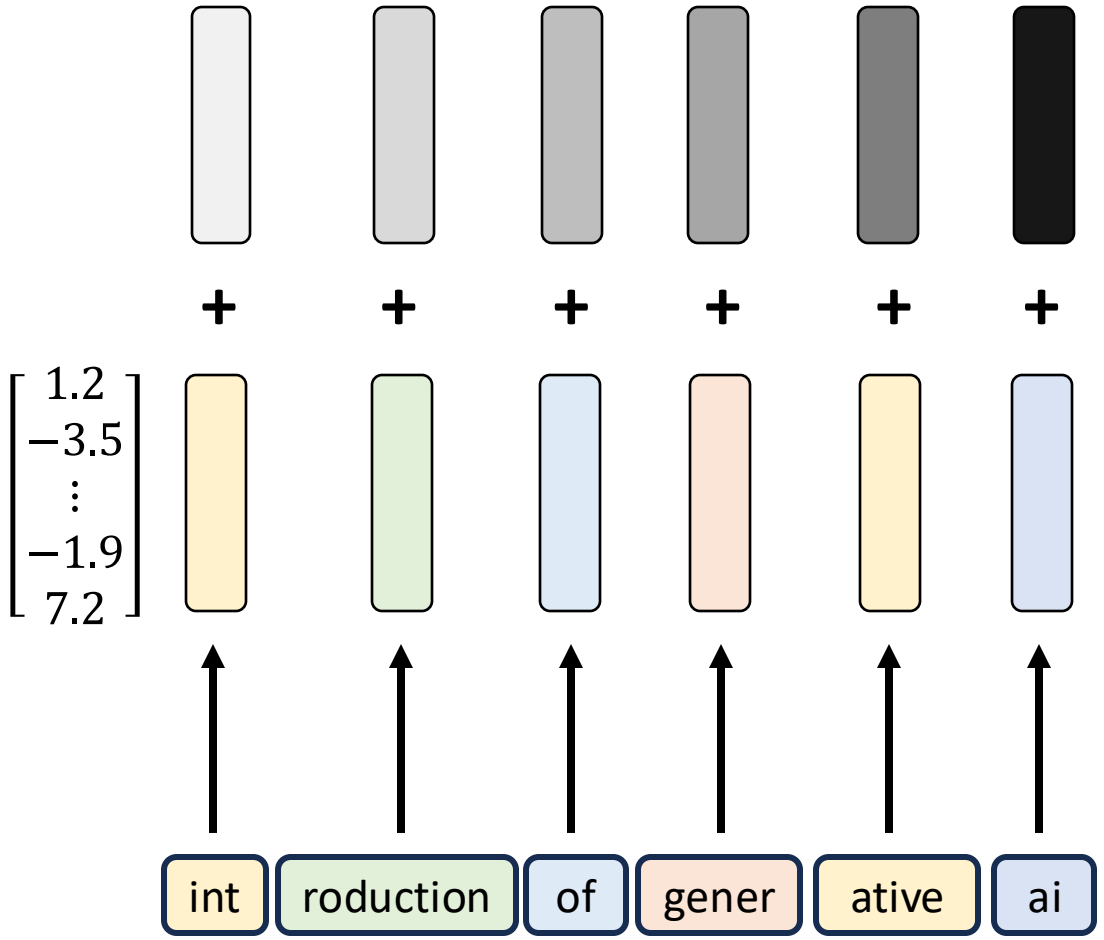


https://lunary.ai/openai-tokenizer

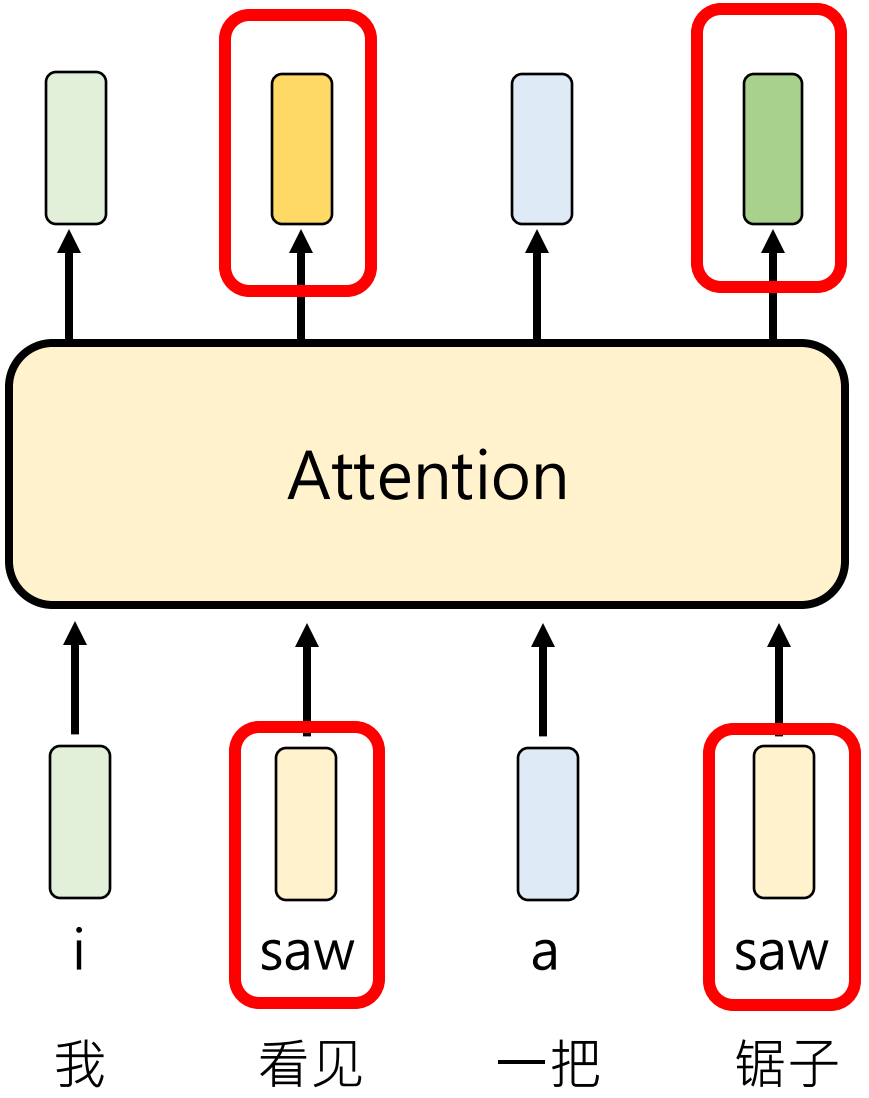
Embedding (Semantic)



Embedding (Positional)



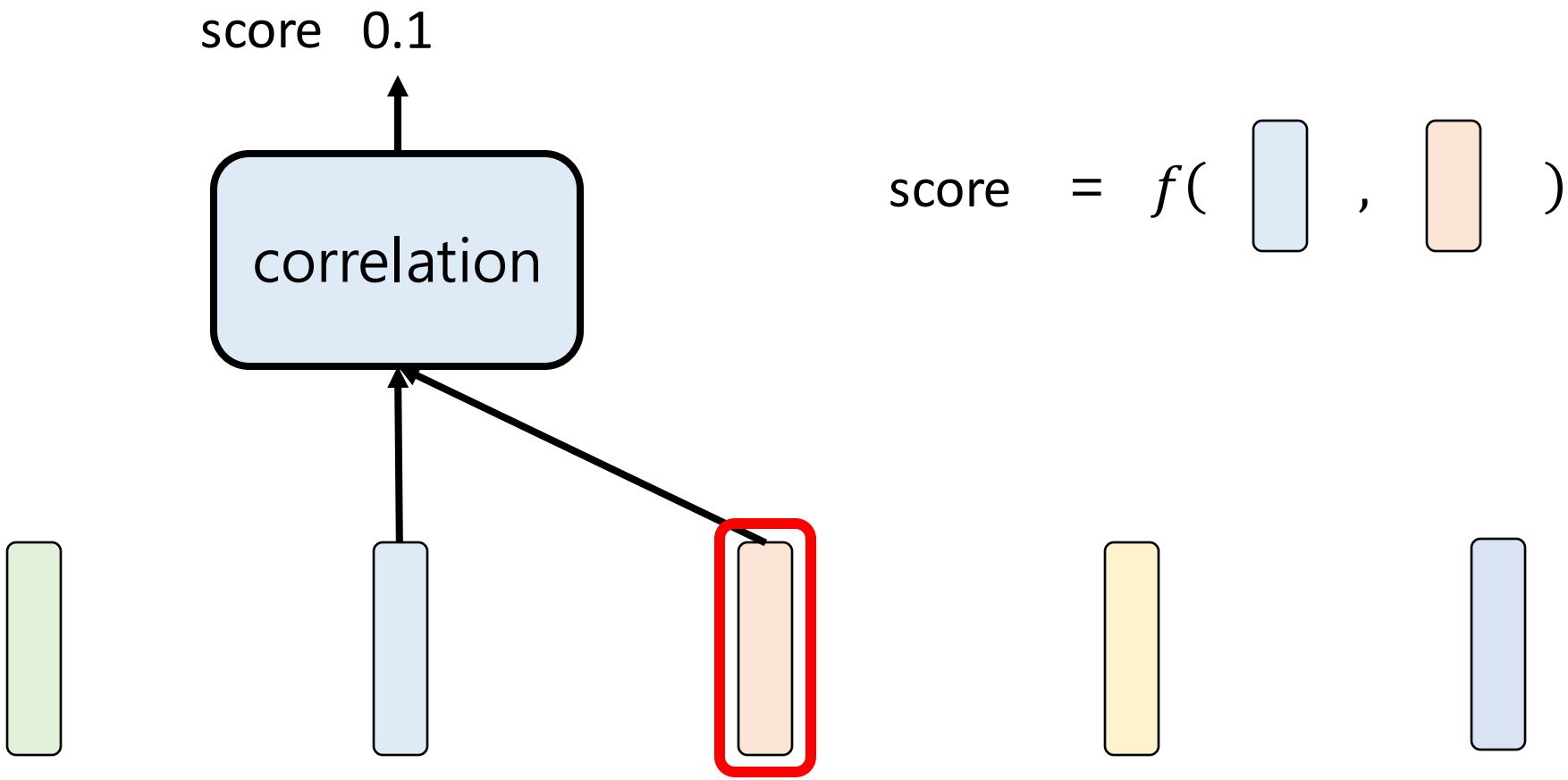
Attention (Context)



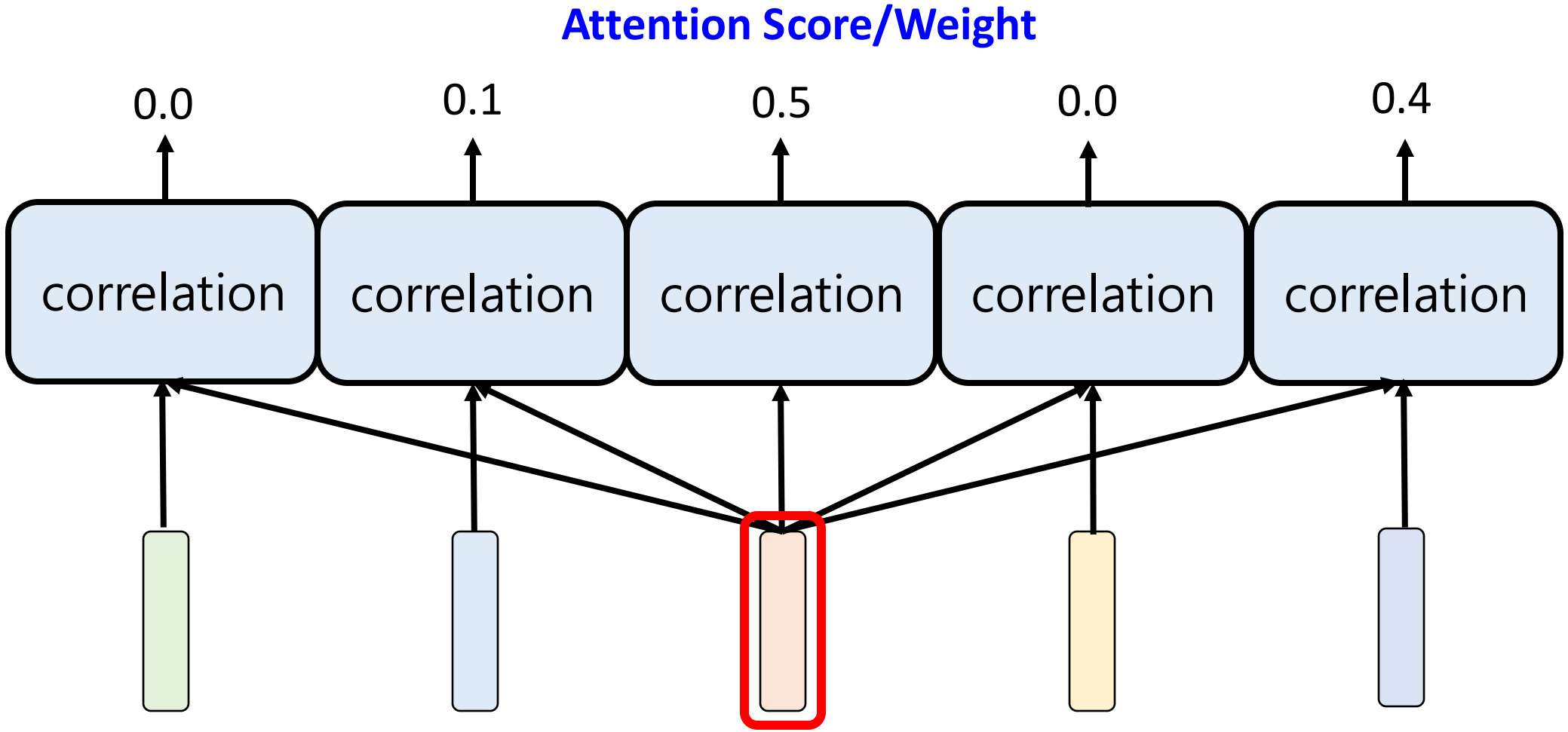
苹 果 电 脑

苹 果 好 吃

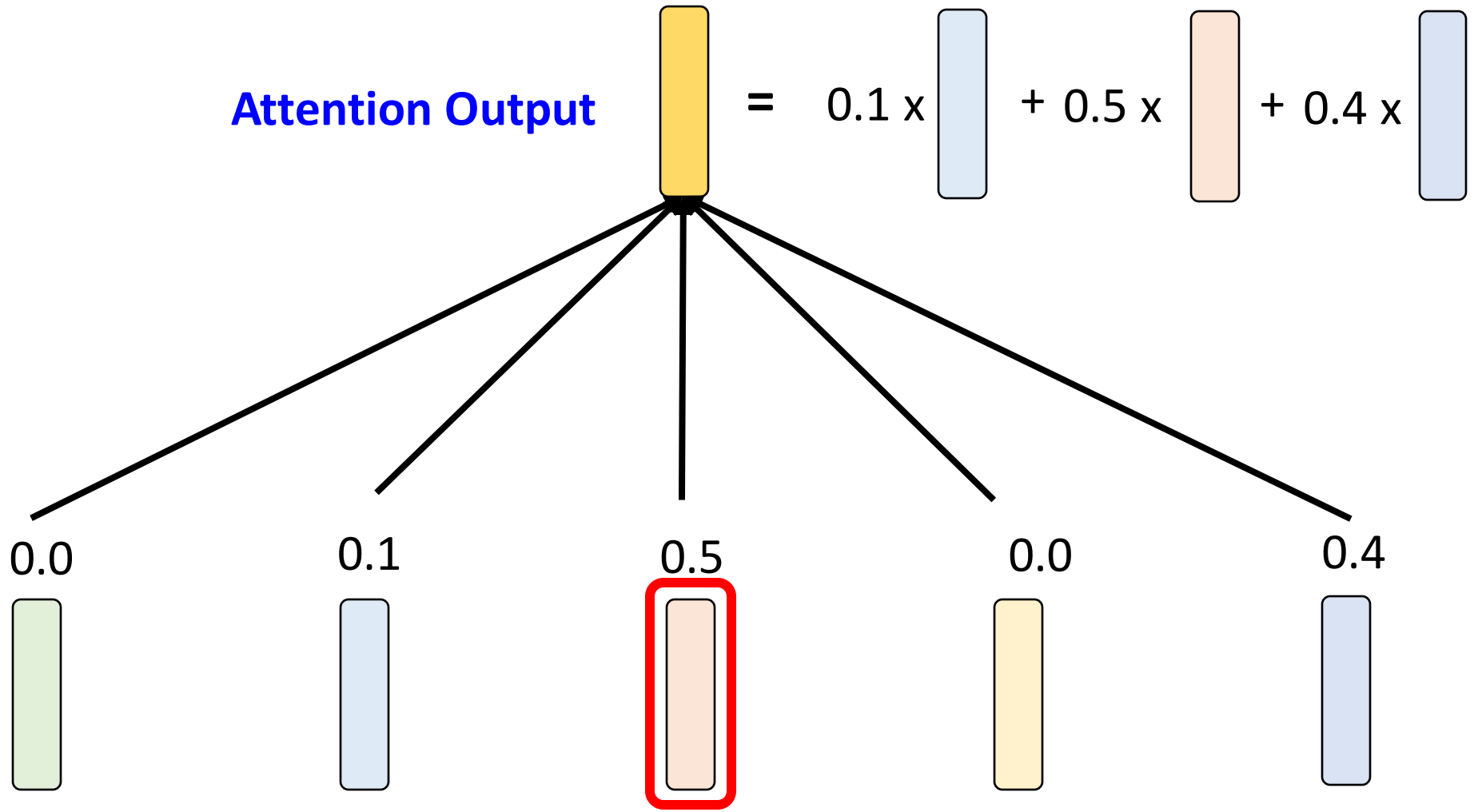
Attention (Context)



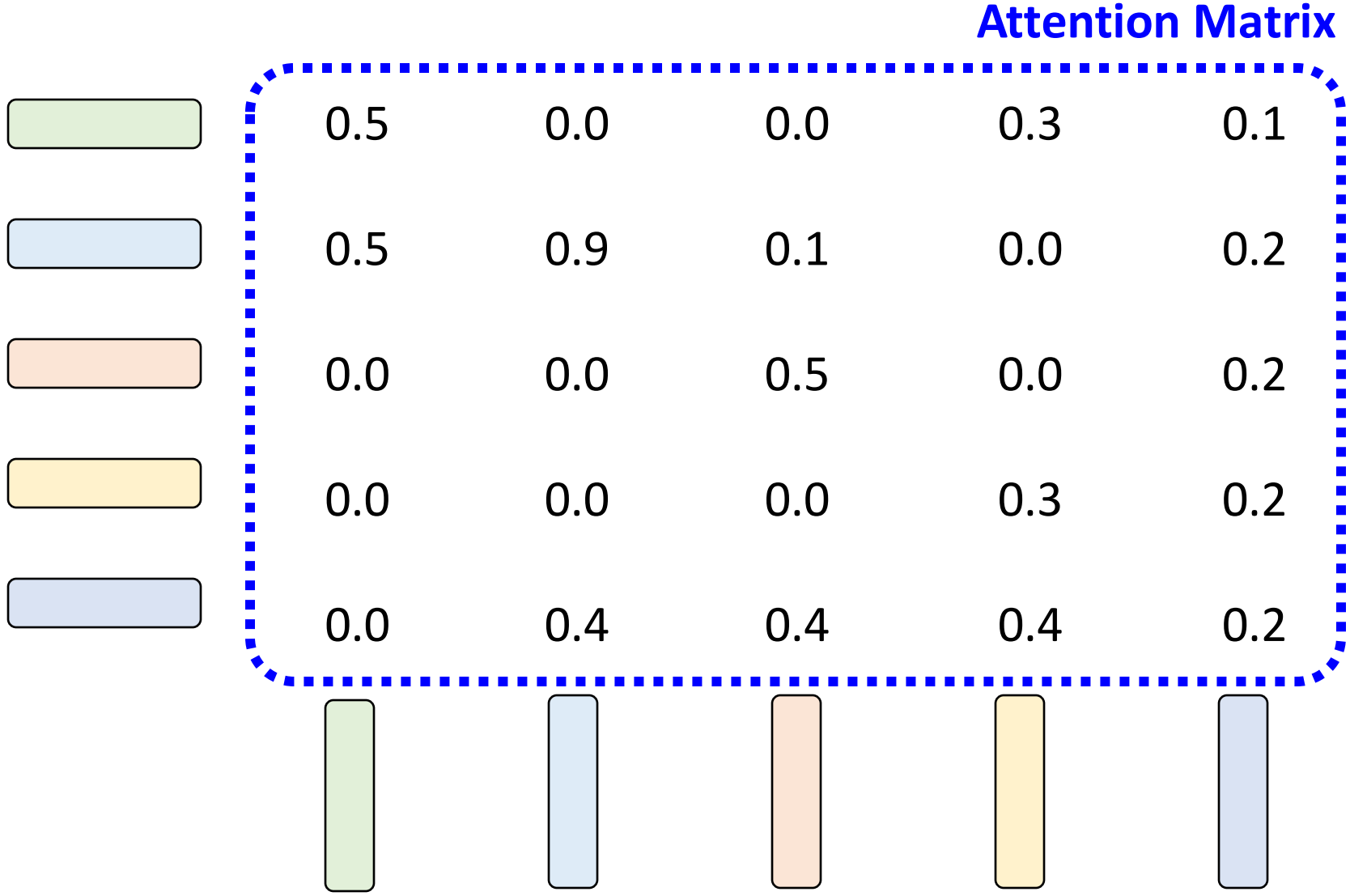
Attention (Context)



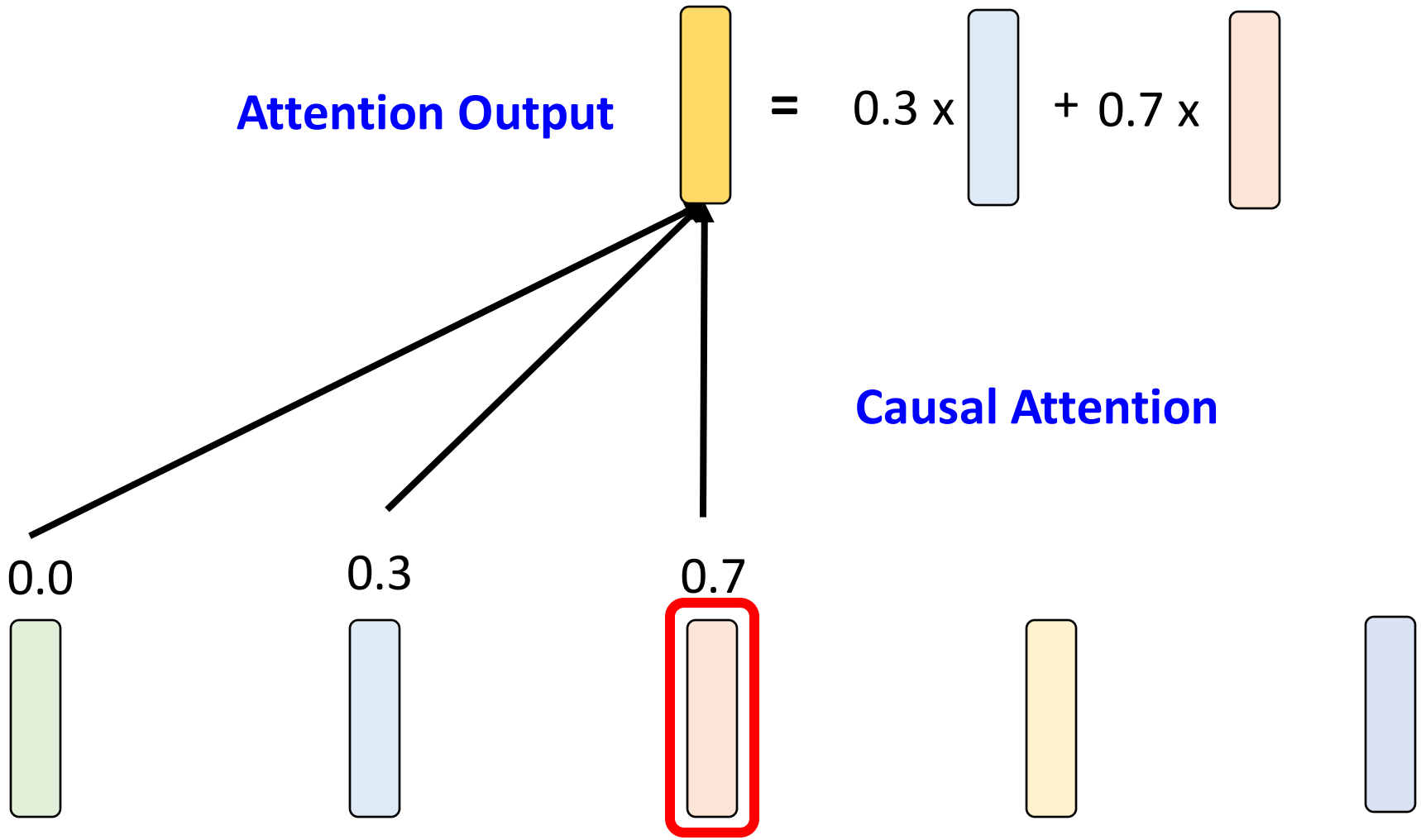
Attention (Context)



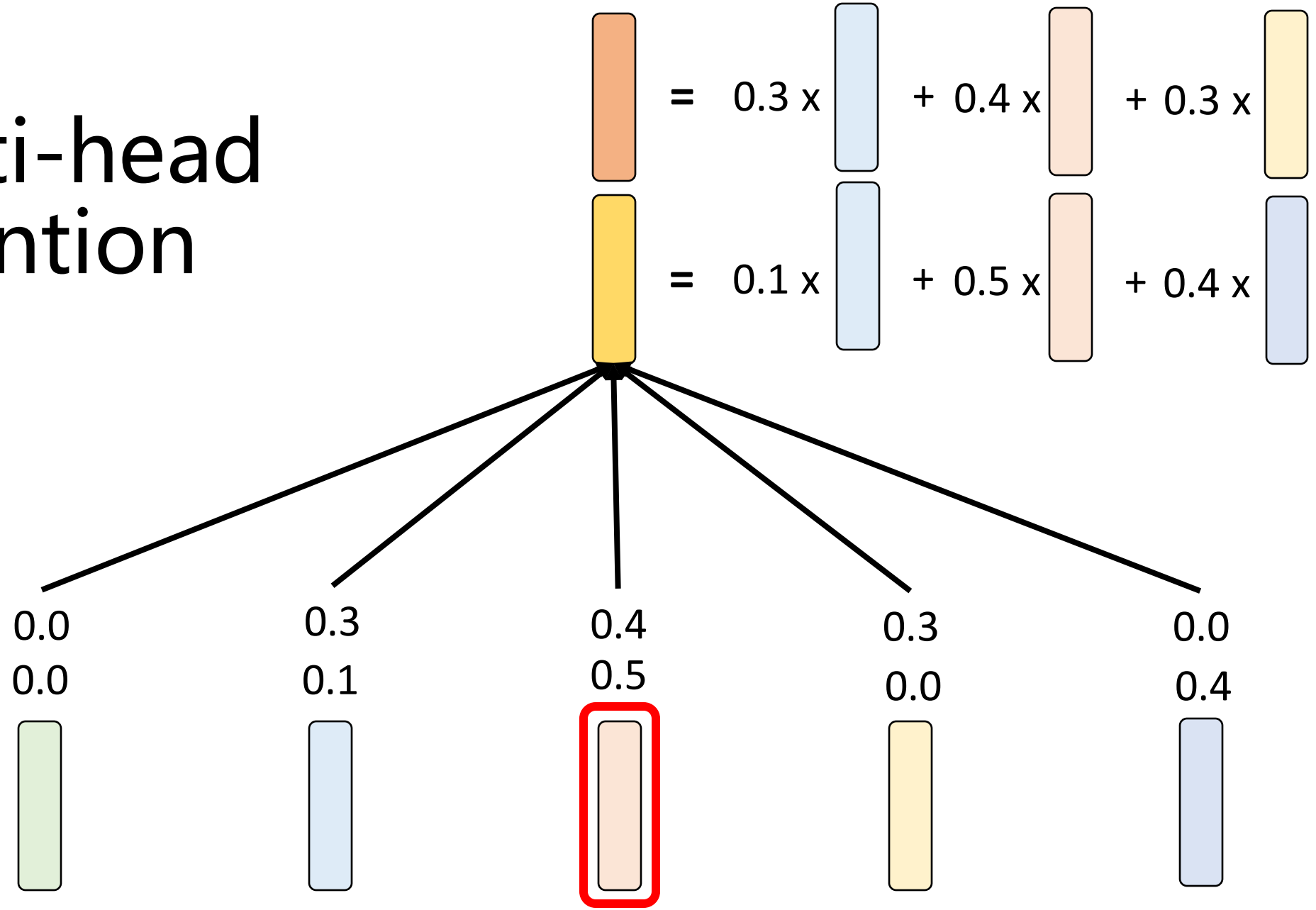
Attention (Context)



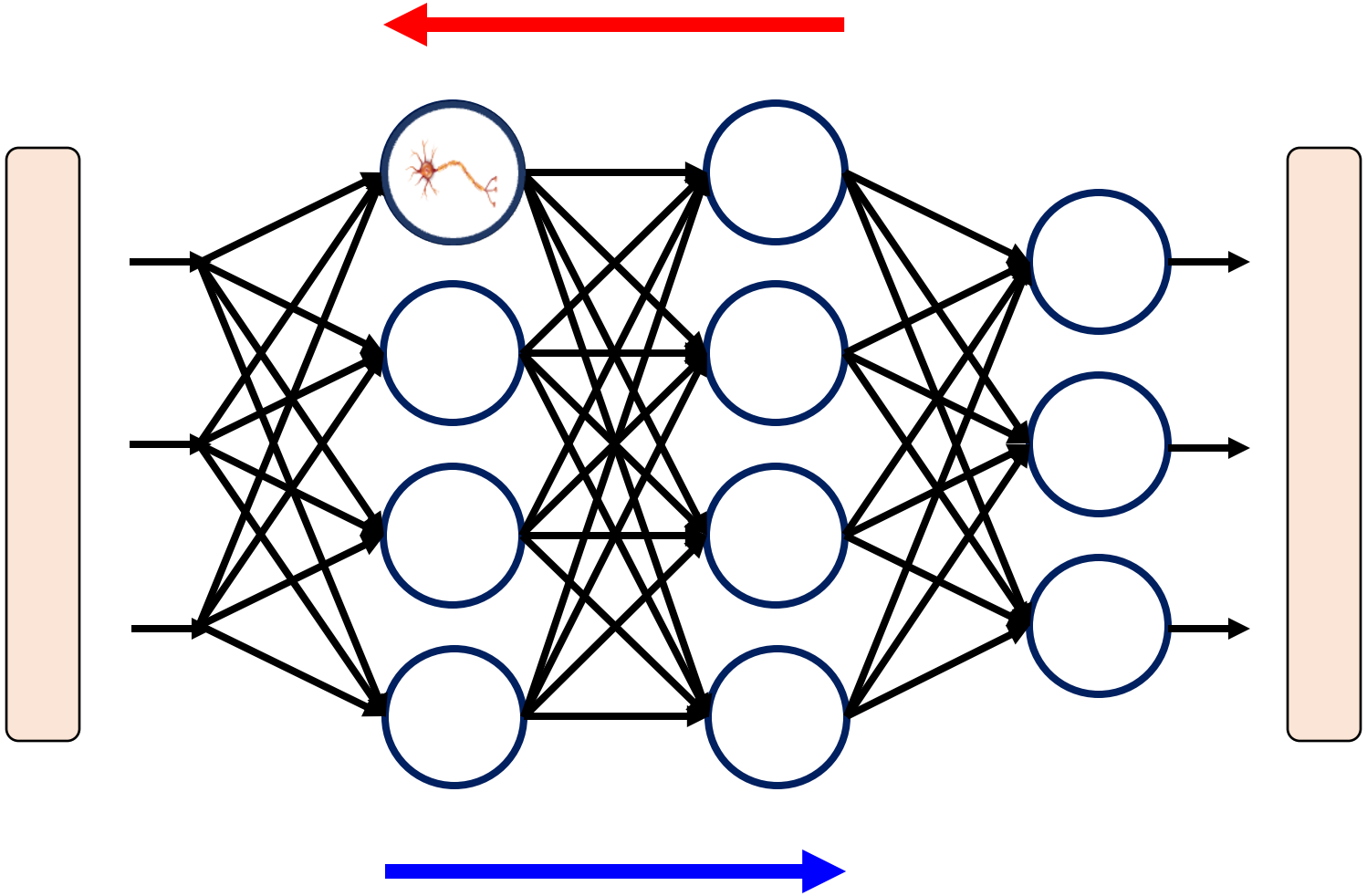
Attention (Context)



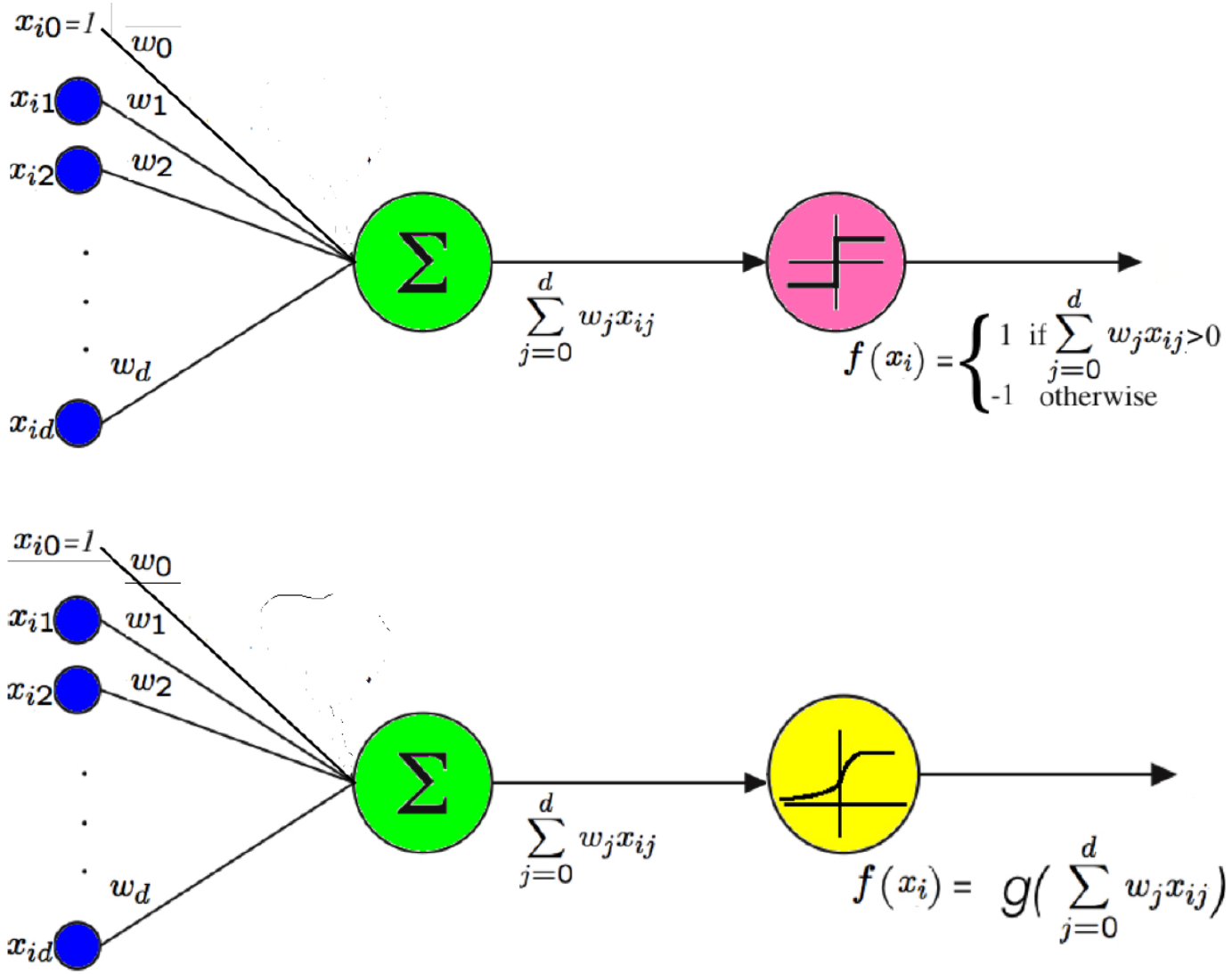
Multi-head Attention



Feedforward-Backpropagation



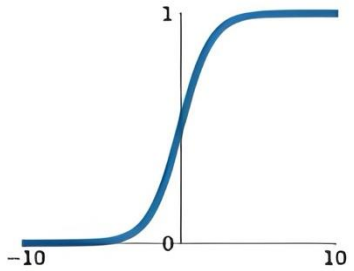
Perceptron


$x_{i0}=1$
w_0
x_{i1}
w_1
x_{i2}
w_2
w_d
x_{id}
Σ
$\sum_{j=0}^{d} w_j x_{ij}$
$f(x_i) = \begin{cases} 1 & \text{if } \sum_{j=0}^{d} w_j x_{ij} > 0 \\ -1 & \text{otherwise} \end{cases}$
$x_{i0}=1$
w_0
x_{i1}
w_1
x_{i2}
w_2
w_d
x_{id}
Σ
$\sum_{j=0}^{d} w_j x_{ij}$
$f(x_i) = g(\sum_{j=0}^{d} w_j x_{ij})$

Activation Functions

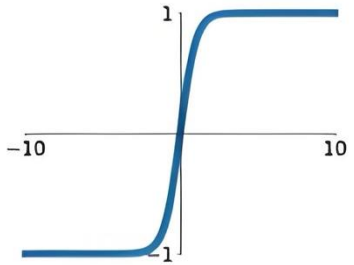
Sigmoid

$$\sigma(x) = \frac{1}{1+e^{-x}}$$



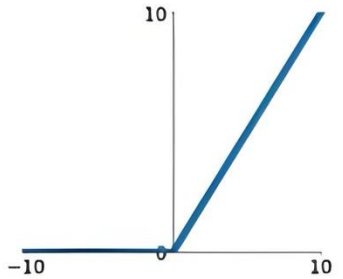
tanh

$$\tanh(x)$$



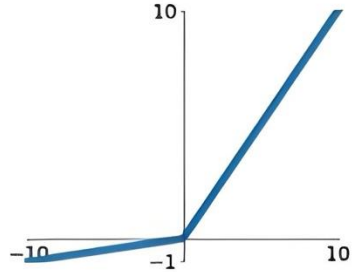
ReLU

$$\max(0, x)$$



Leaky ReLU

$$\max(0.1x, x)$$

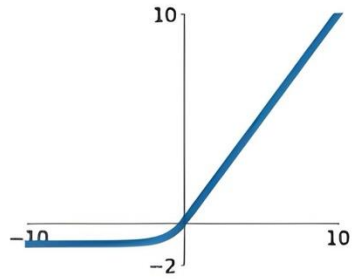


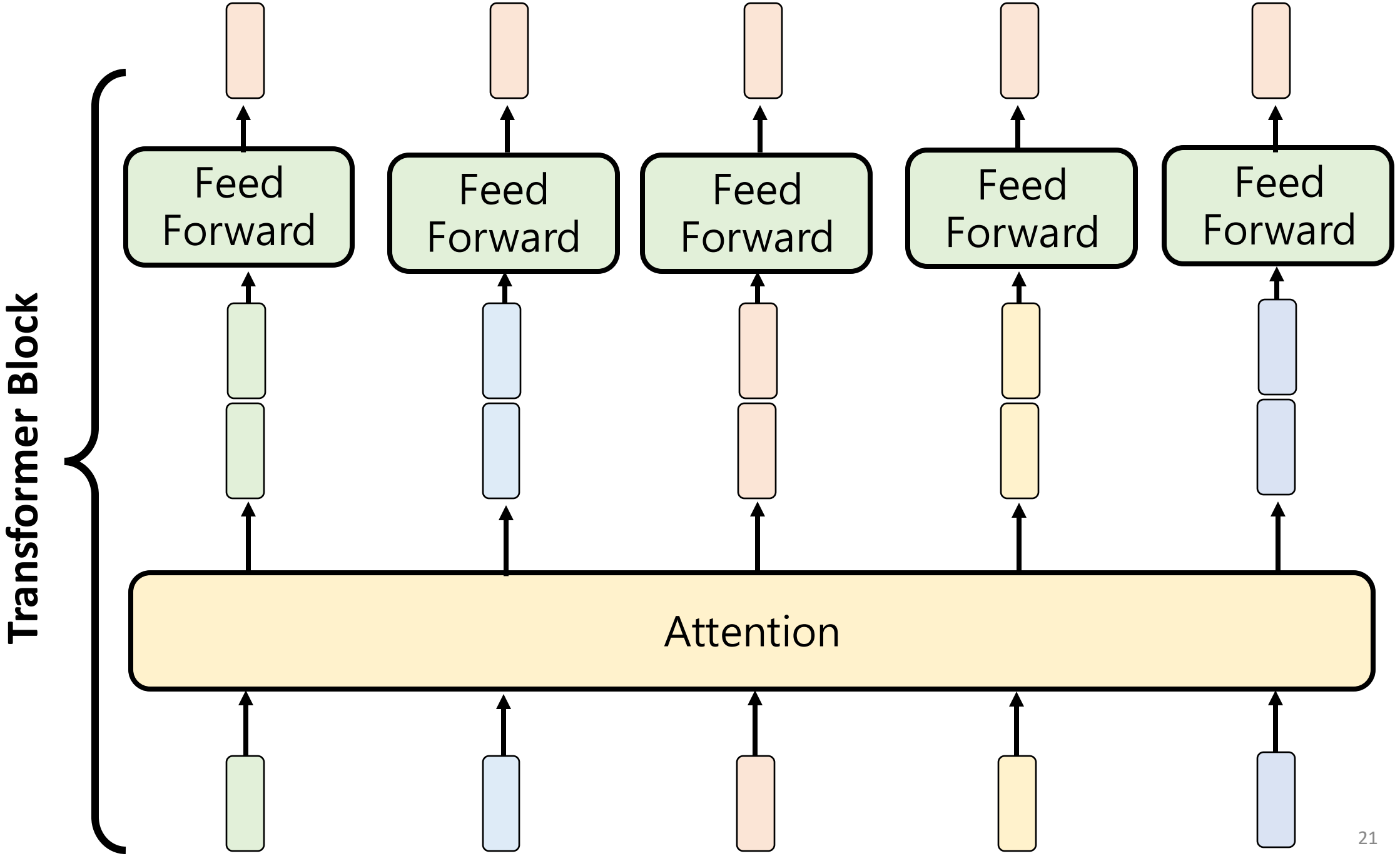
Maxout

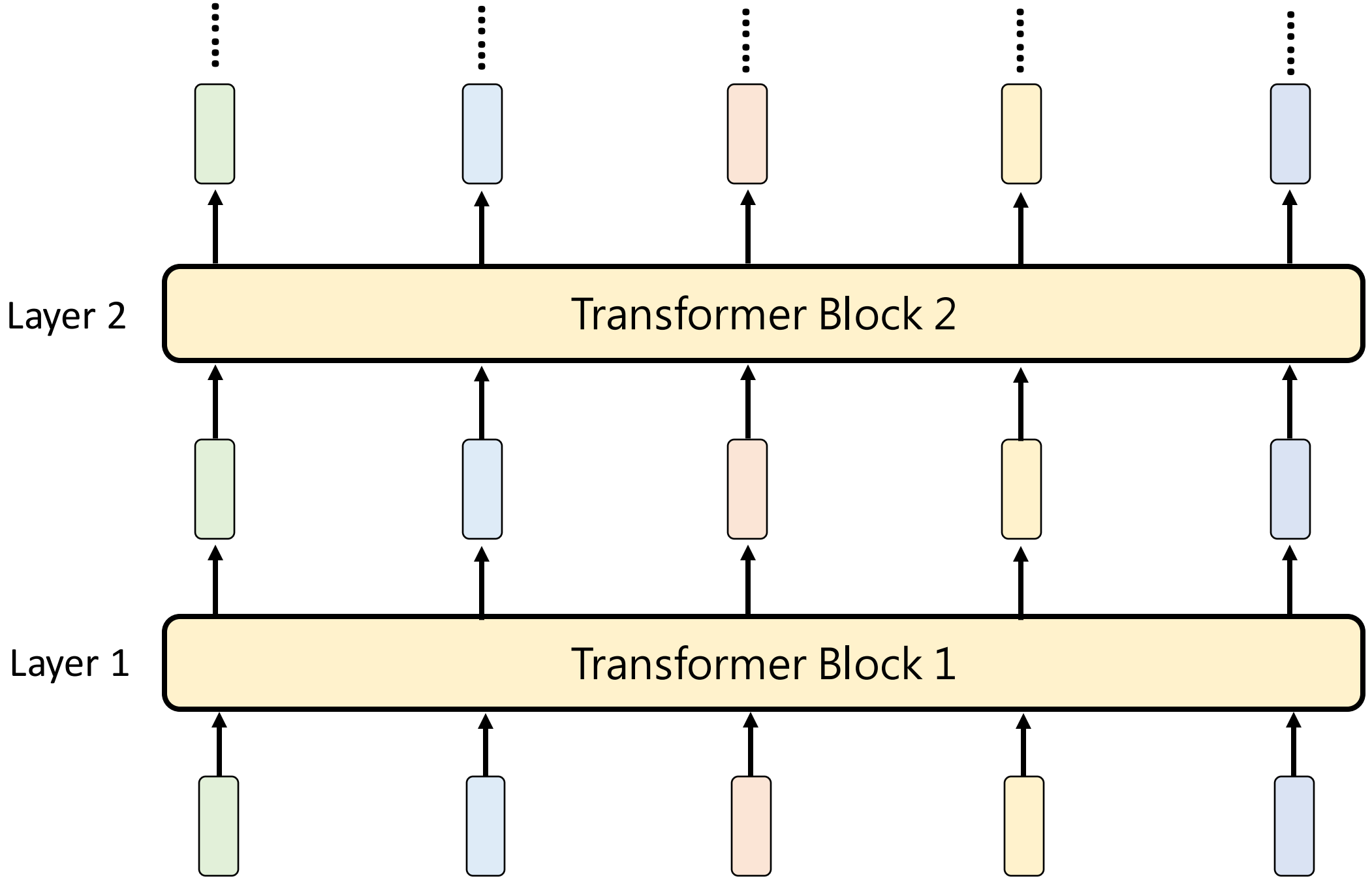
$$\max(w_1^T x + b_1, w_2^T x + b_2)$$

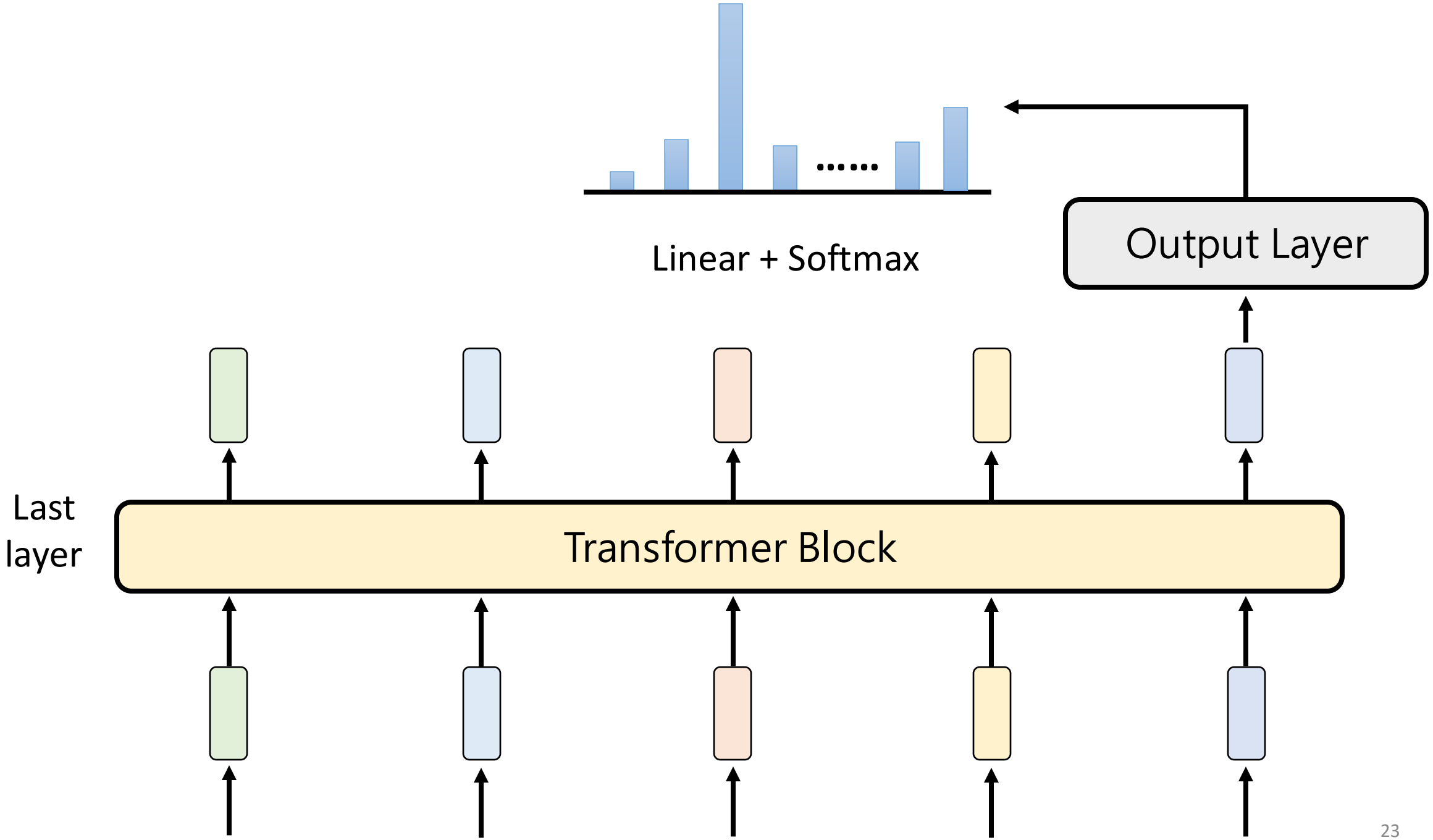
ELU

$$\begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$$

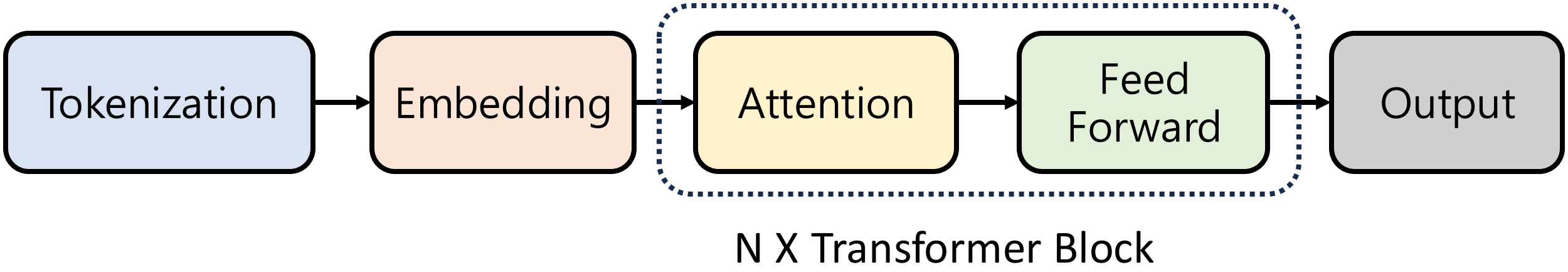



Transformer Block
Feed Forward
Feed Forward
Feed Forward
Feed Forward
Feed Forward
Attention

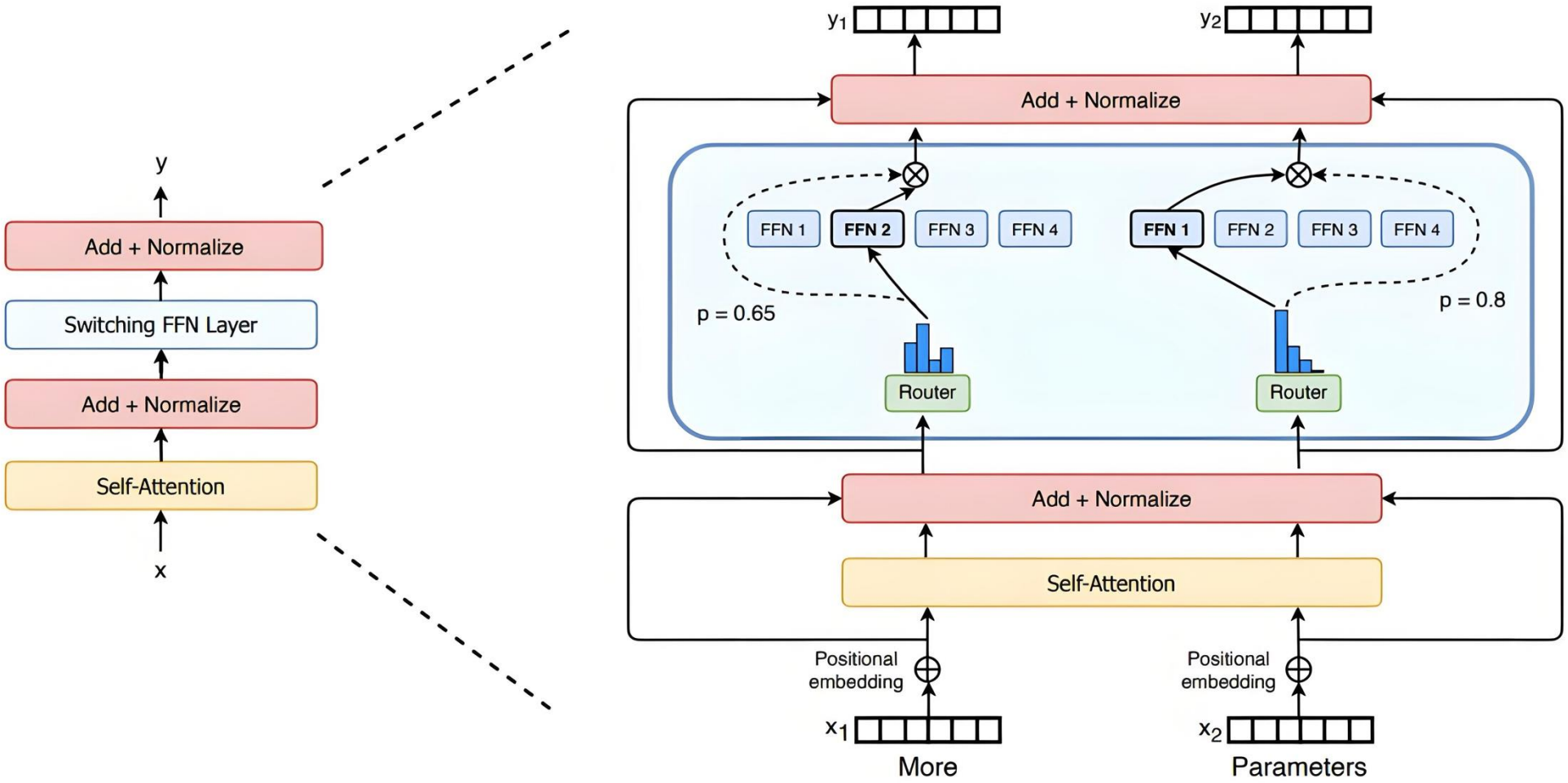

Layer 2
Transformer Block 2
Layer 1
Transformer Block 1


Linear + Softmax
Output Layer
Last layer
Transformer Block

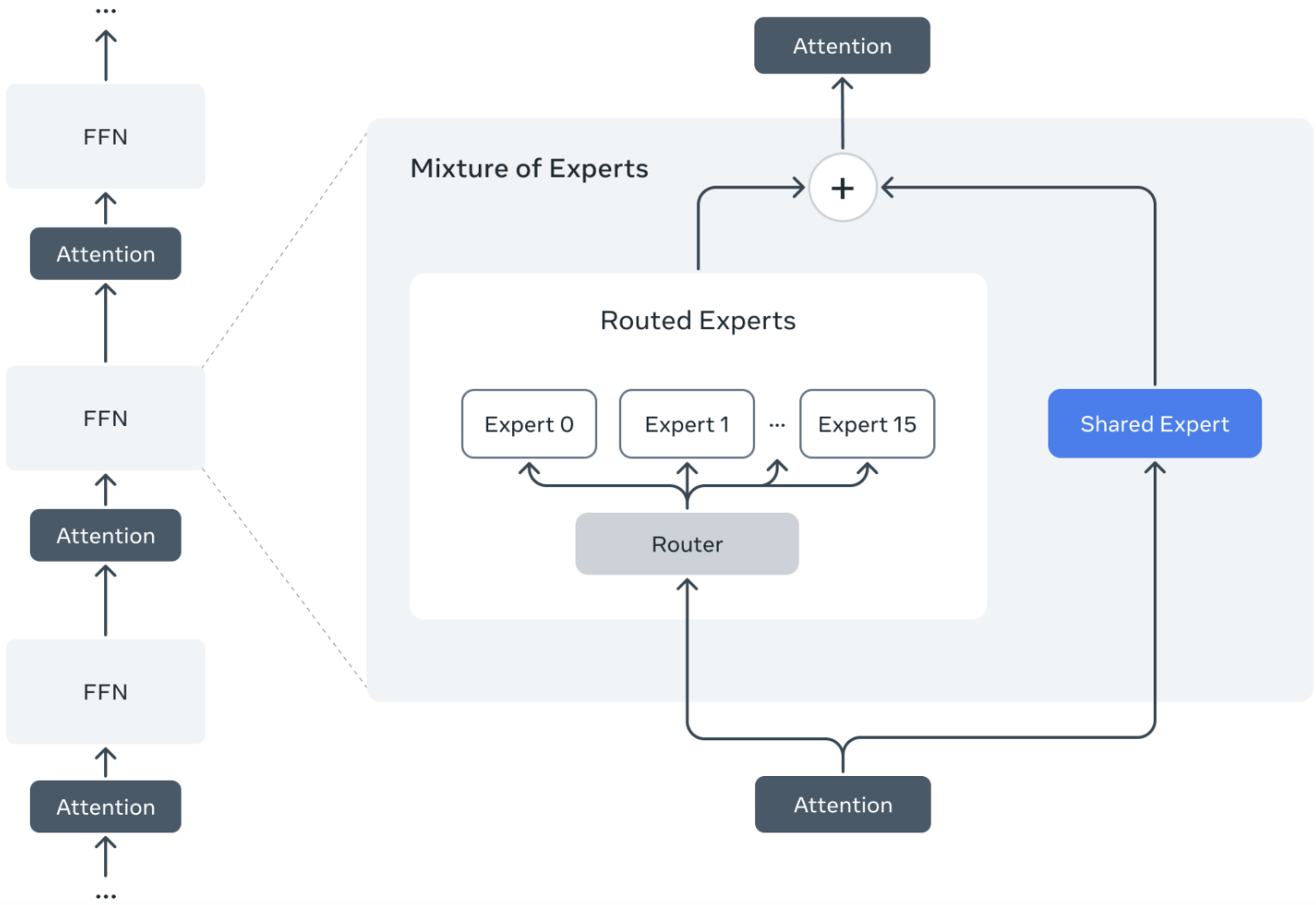
Summary: Transformer



Mixture of Experts (MoE)

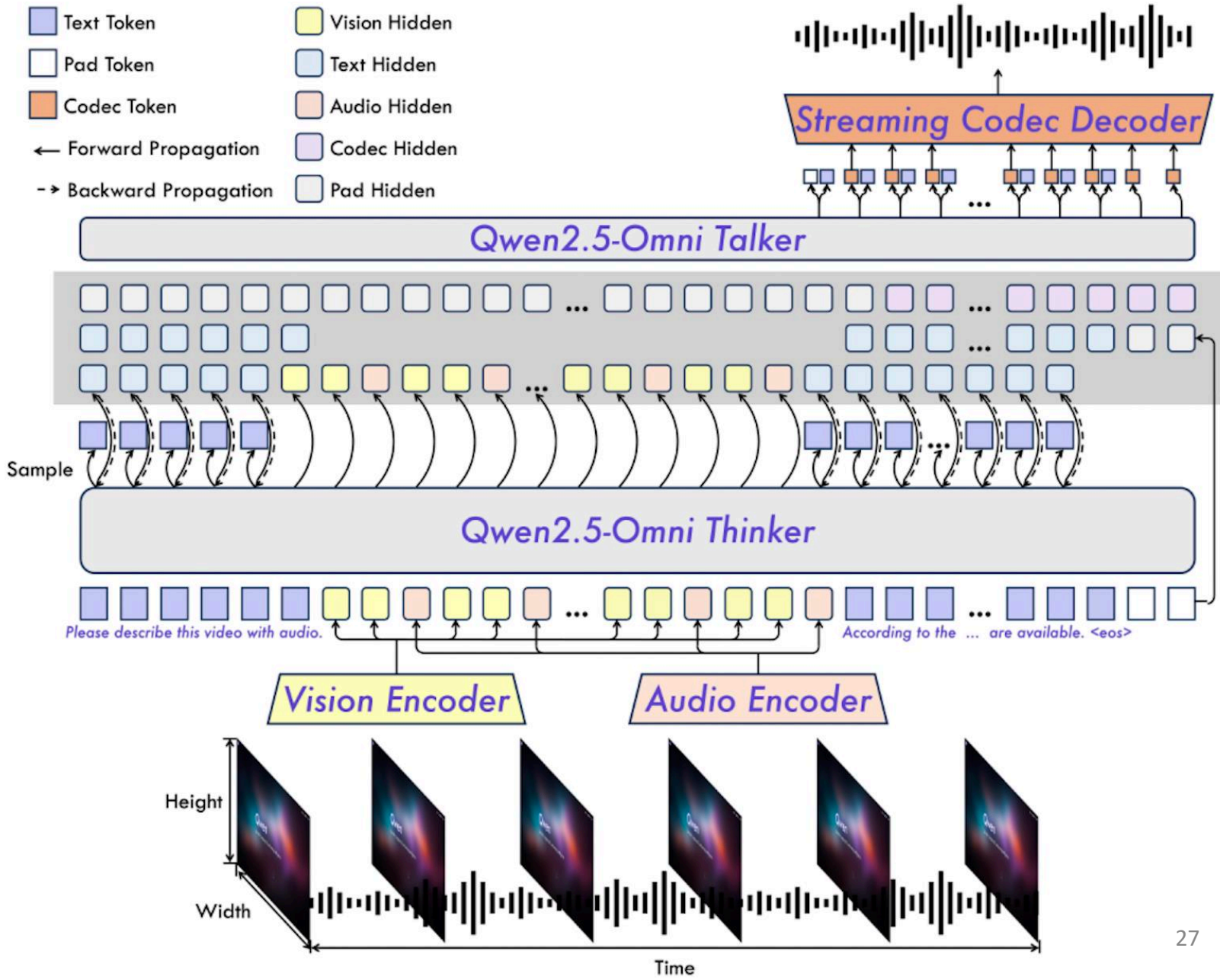


Mixture of Experts (MoE)



MM-LLM

Qwen2.5-Omni



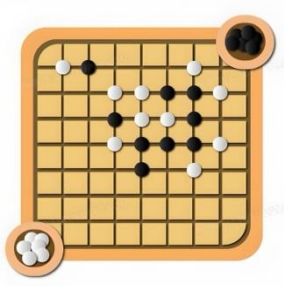
Outline

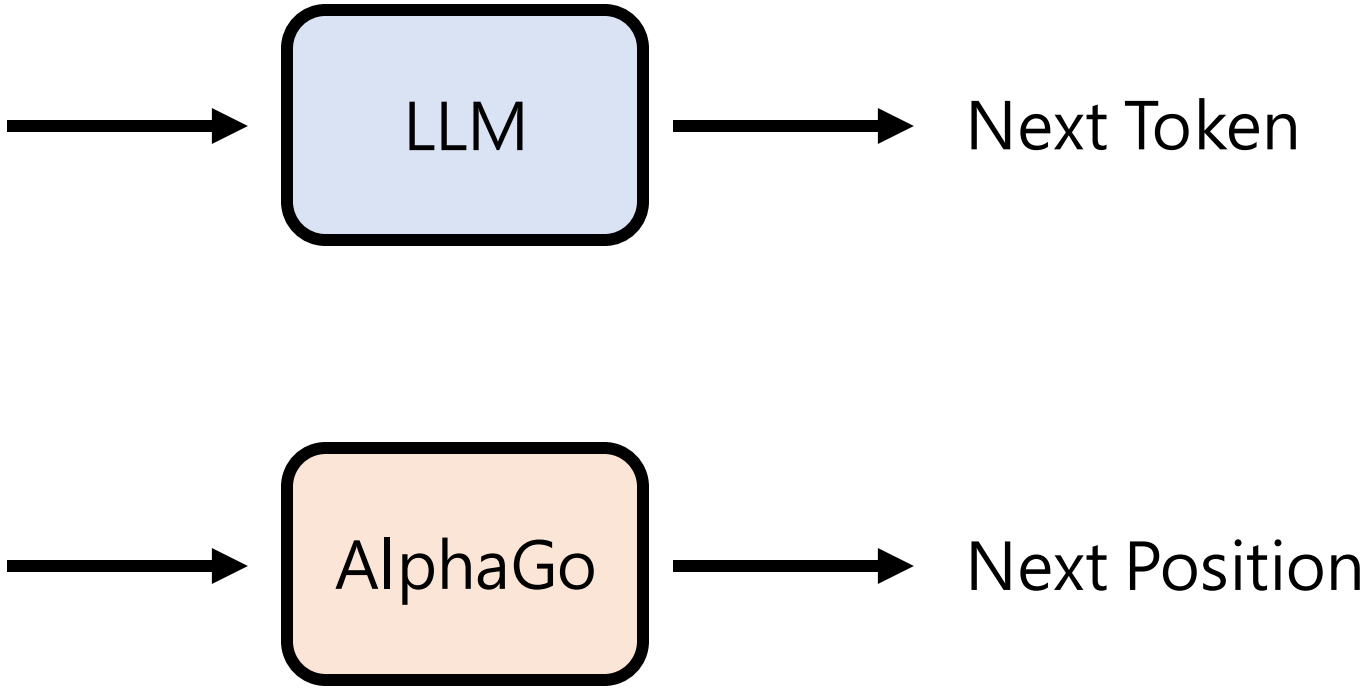
Transformer

LLM Training

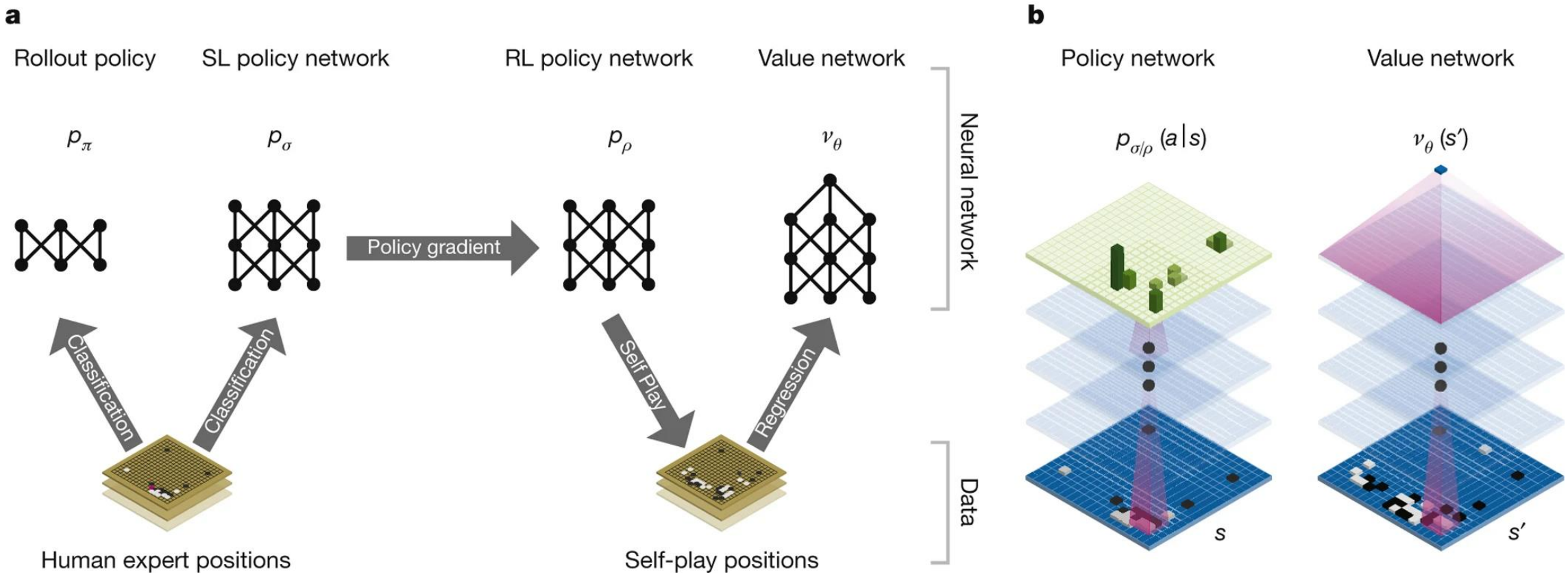
DeepSeek

LLM vs AlphaGo





AlphaGo

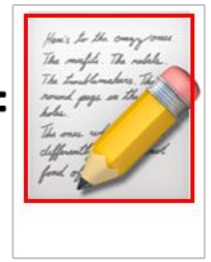

a
Rollout policy
SL policy network
RL policy network
Value network
p_π
p_σ
p_ρ
v_θ
Policy gradient
Classification
Classification
Self Play
Regression
Neural network
Data
Human expert positions
Self-play positions
b
Policy network
Value network
$p_{\sigma/\rho}(a|s)$
$v_\theta(s')$
s
s'

LLM Training

- Foundation Model
  - ① LLM Pre-train by Self-supervised Learning

- Alignment
  - ② LLM Fine-tuning by Supervised Learning
  - ③ LLM Fine-tuning by Reinforcement Learning

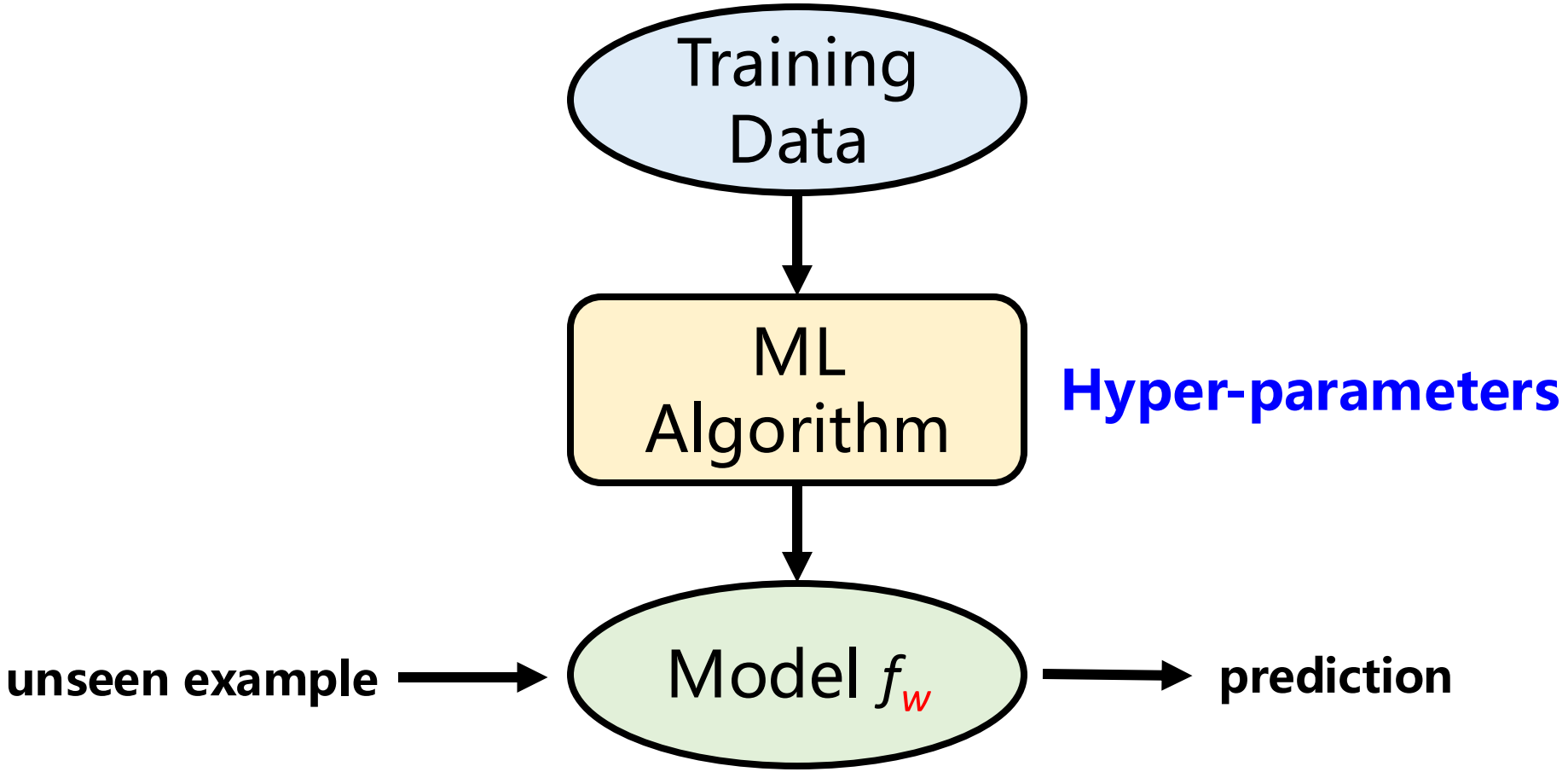
LLM Pre-train (SSL)

f_w() = 

How are you? → LLM → I am fine.

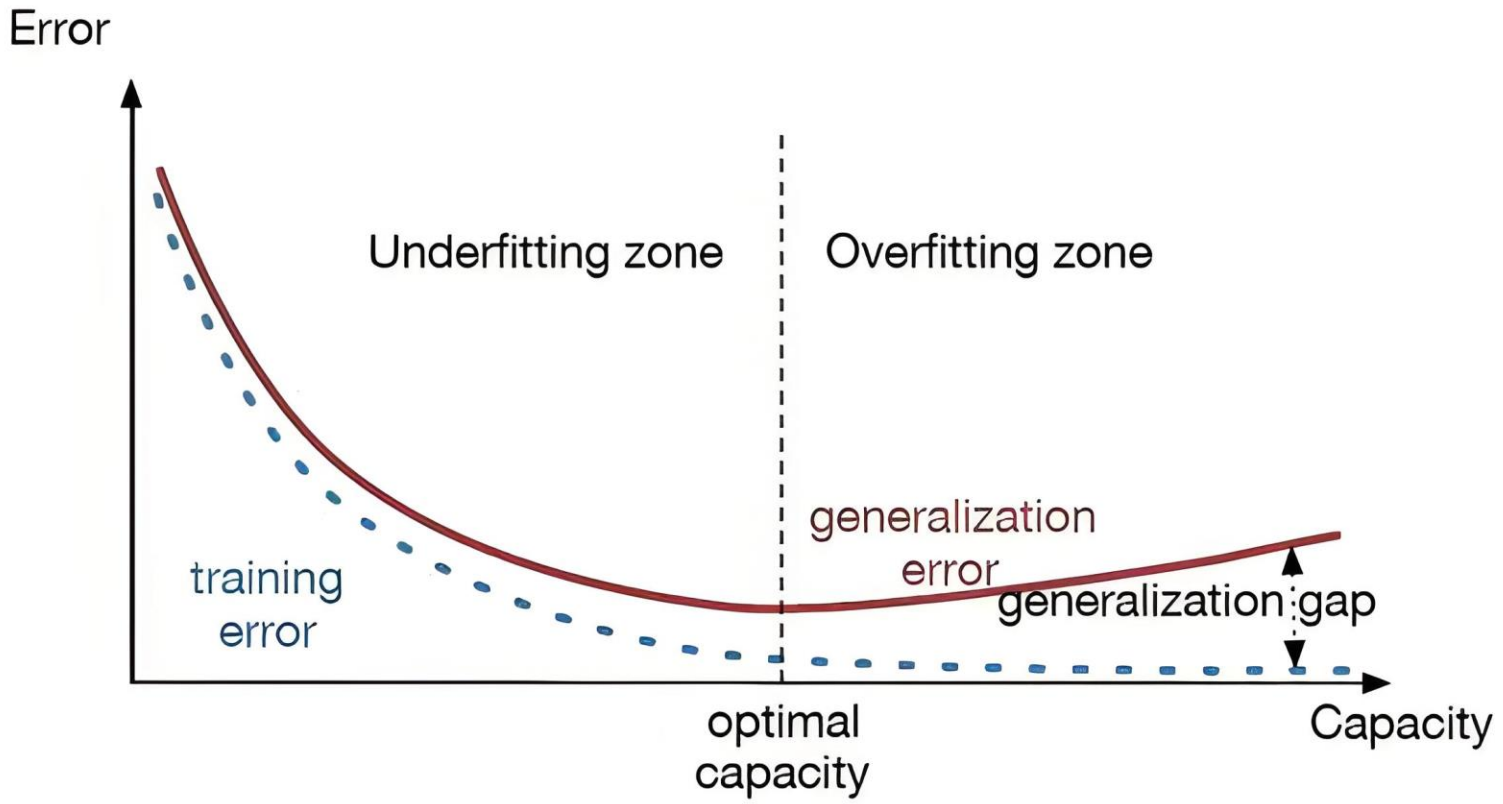
How are you?	I
How are you? I	am
How are you? I am	fine
How are you? I am fine	.
How are you? I am fine.	[end]
…	**Training Data**

Training and Testing



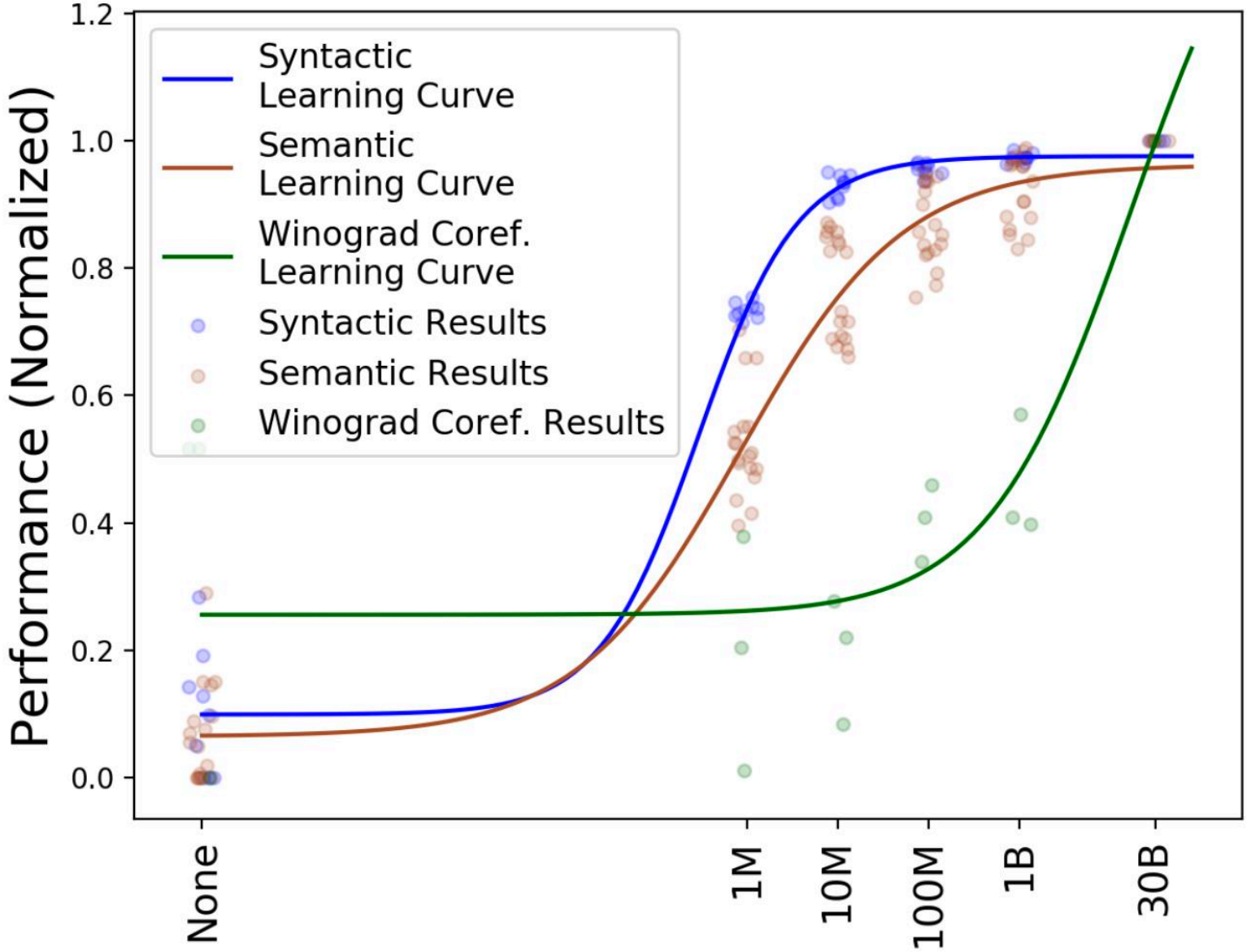
Underfitting & Overfitting

An intuitive example



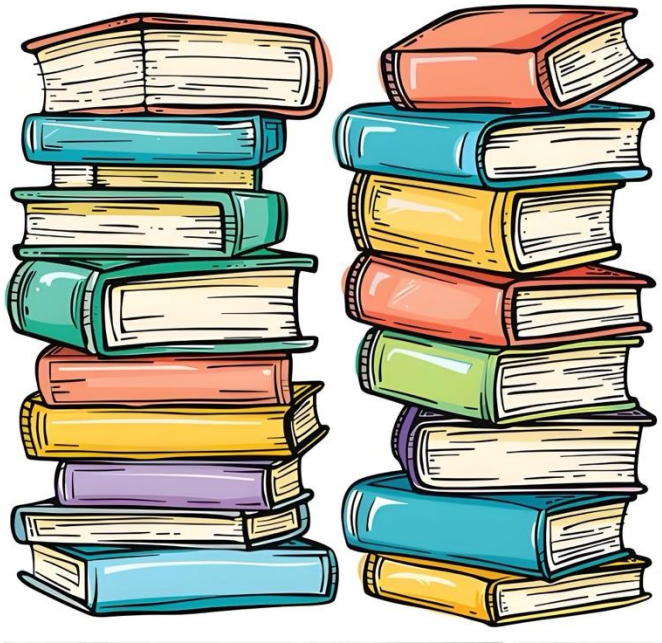
More Data

"linguistic features are mostly learnable with 100M words of data, while NLU task performance requires far more data."

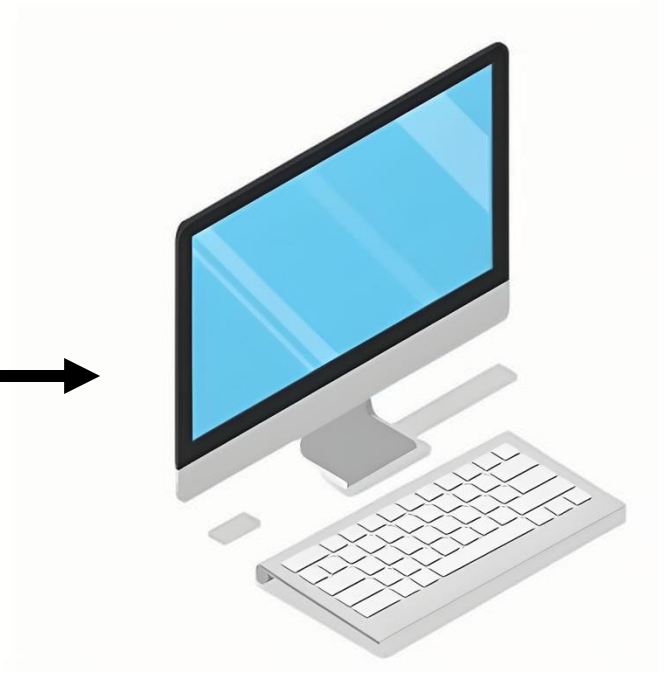


Data is everywhere

Internet



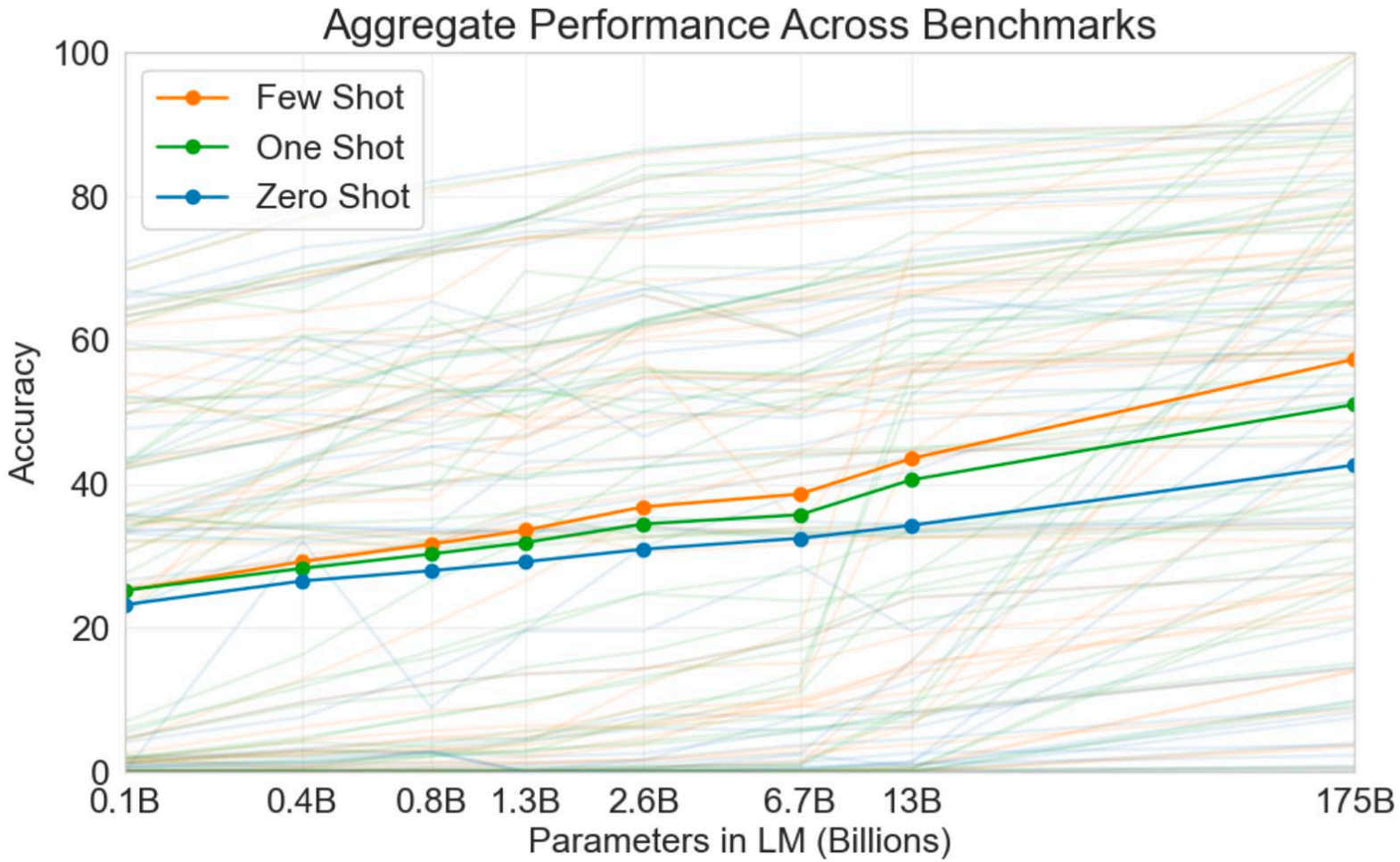
Big Data



Cleaning&
Processing

GPT-3

GPT	Model size	Data size
GPT-1 (2018)	117M	7000 Books
GPT-2 (2019)	1542M	40 GB
GPT-3 (2020)	175B	580 GB



PaLM

Model	# of Parameters (in billions)	Accelerator chips	Model FLOPS utilization
GPT-3	175B	V100	21.3%
Gopher	280B	4096 TPU v3	32.5%
Megatron-Turing NLG	530B	2240 A100	30.2%
PaLM	540B	6144 TPU v4	46.2%

Model input

The square root of x is the cube root of y. What is y to the power of 2, if x = 4?

PaLM 540B output

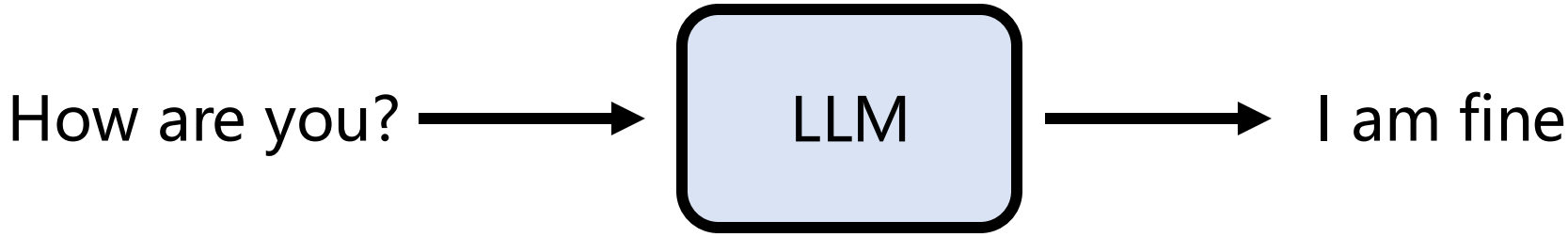
Q. The square root of x is the cube root of y. What is y to the power of 2, if x = 8?

Q. The square root of x is the cube root of y. What is y to the power of 2, if x = 12?

Q. The square root of x is the cube root of y. What is y to the power of 2, if x = 16?

✖ **(keeps asking more questions)**

LLM Fine-tuning (SL)



How are you?	I am fine
Which is the highest mountain in Shenzhen?	Wutong Mountain
Where is Sustech?	Shenzhen
...	**Training Data**

USER: Which is the highest mountain in Shenzhen? AI: Wutong Mountain

USER: Which is the highest mountain in Shenzhen? AI: Wutong

USER: Which is the highest mountain in Shenzhen? AI: Wutong Mountain

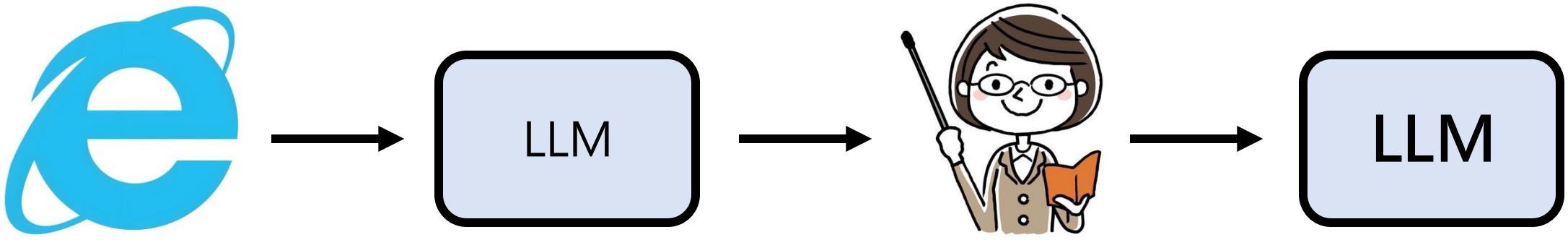
USER: Which is the highest mountain in Shenzhen? AI: Wutong Mountain [END]

Which is the highest mountain in Shenzhen? Wutong Mountain

USER: Which is the highest mountain in Shenzhen? AI: Wutong Mountain [END]

USER: Which is the highest mountain in Shenzhen? Wutong Mountain AI: Right

Pre-train vs Instruction Fine-tuning



Big Data

e.g., GPT-3

Small Data

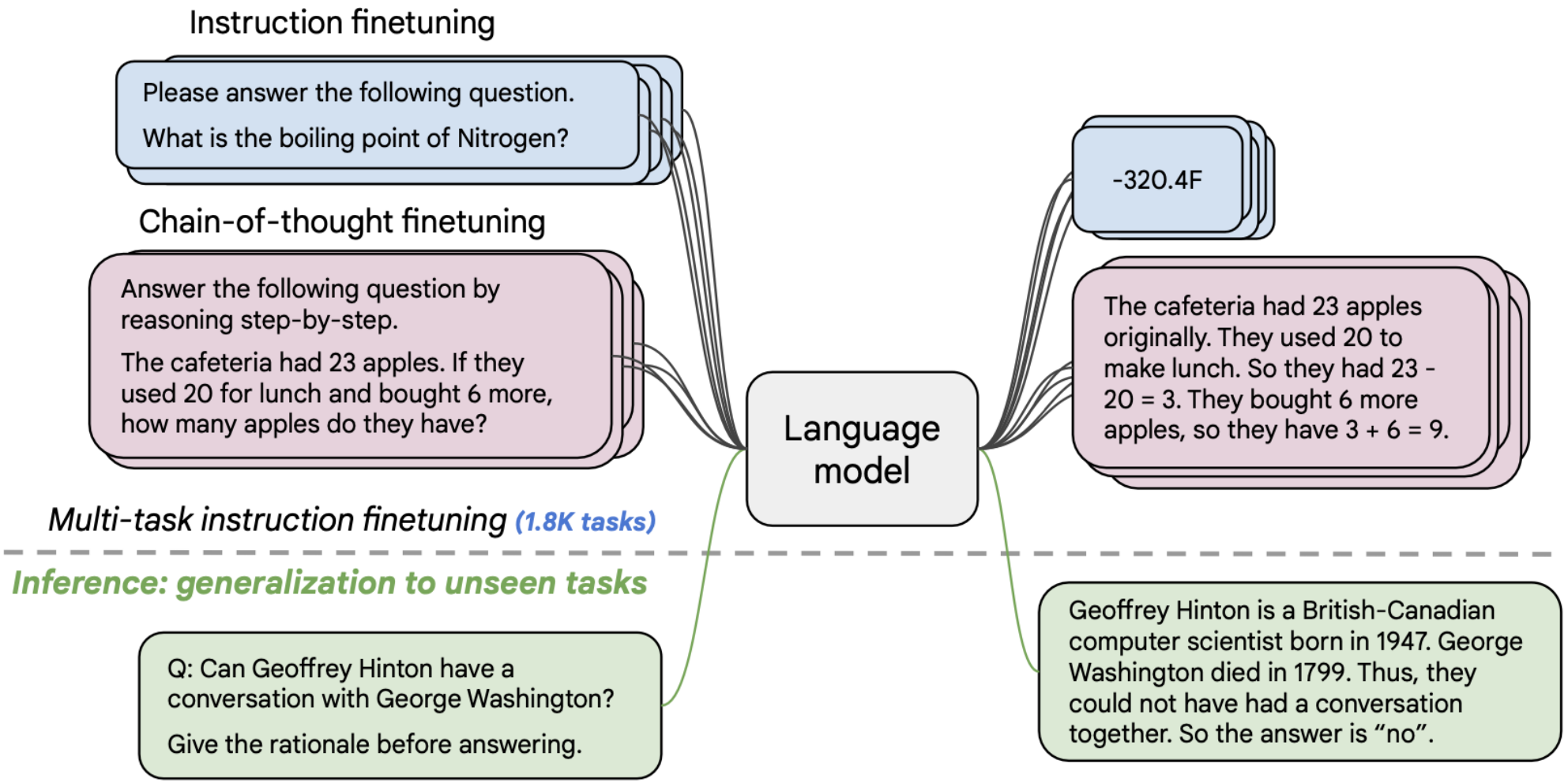
SFT Data			RM Data			PPO Data		
split	source	size	split	source	size	split	source	size
train	labeler	11,295	train	labeler	6,623	train	customer	31,144
train	customer	1,430	train	customer	26,584	valid	customer	16,185
valid	labeler	1,550	valid	labeler	3,488			
valid	customer	103	valid	customer	14,399			

Generalization

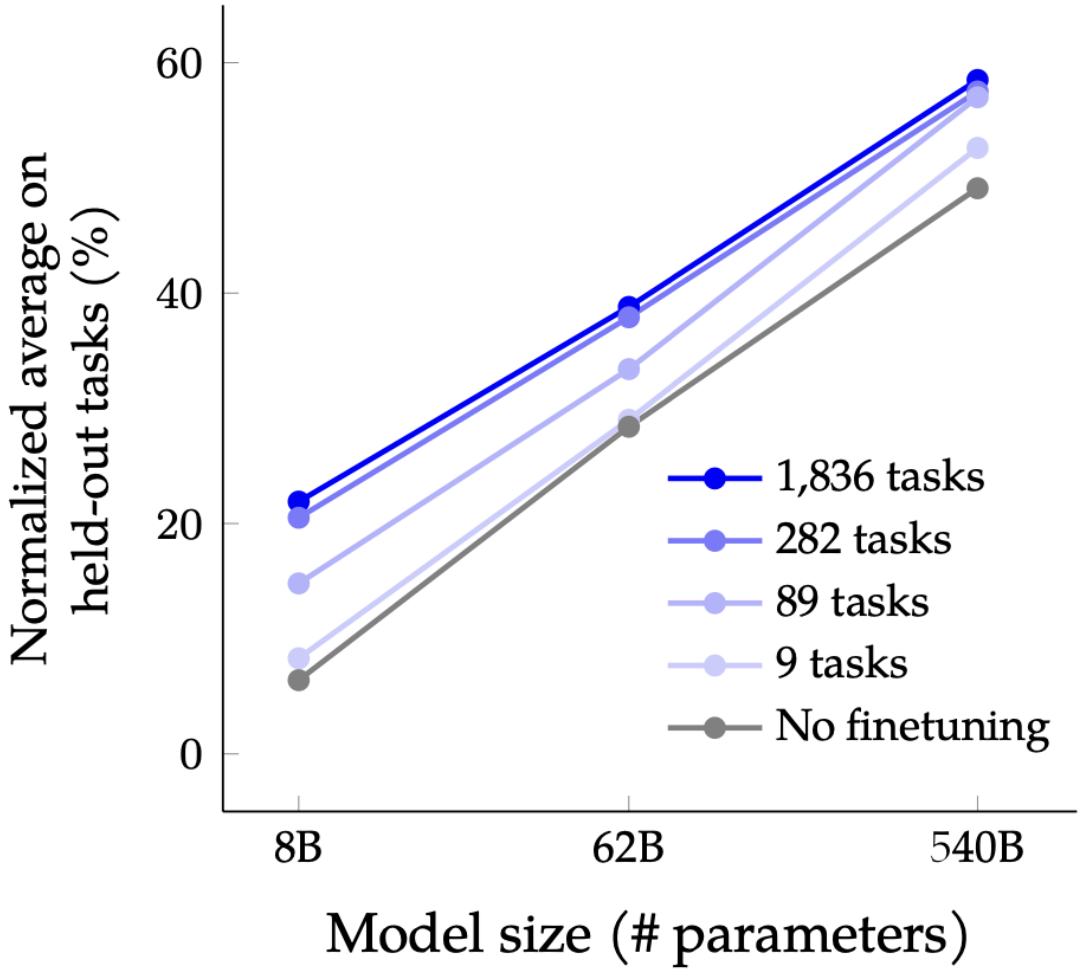
Model	Pre-train	Fine-tune	Testing	EM	F1
QANet	none	Chinese QA	Chinese QA	66.1	78.1
BERT	Chinese	Chinese QA		82.0	89.1
	104 languages	Chinese QA		81.2	88.7
		English QA		63.3	78.8
		Chinese + English		82.6	90.1

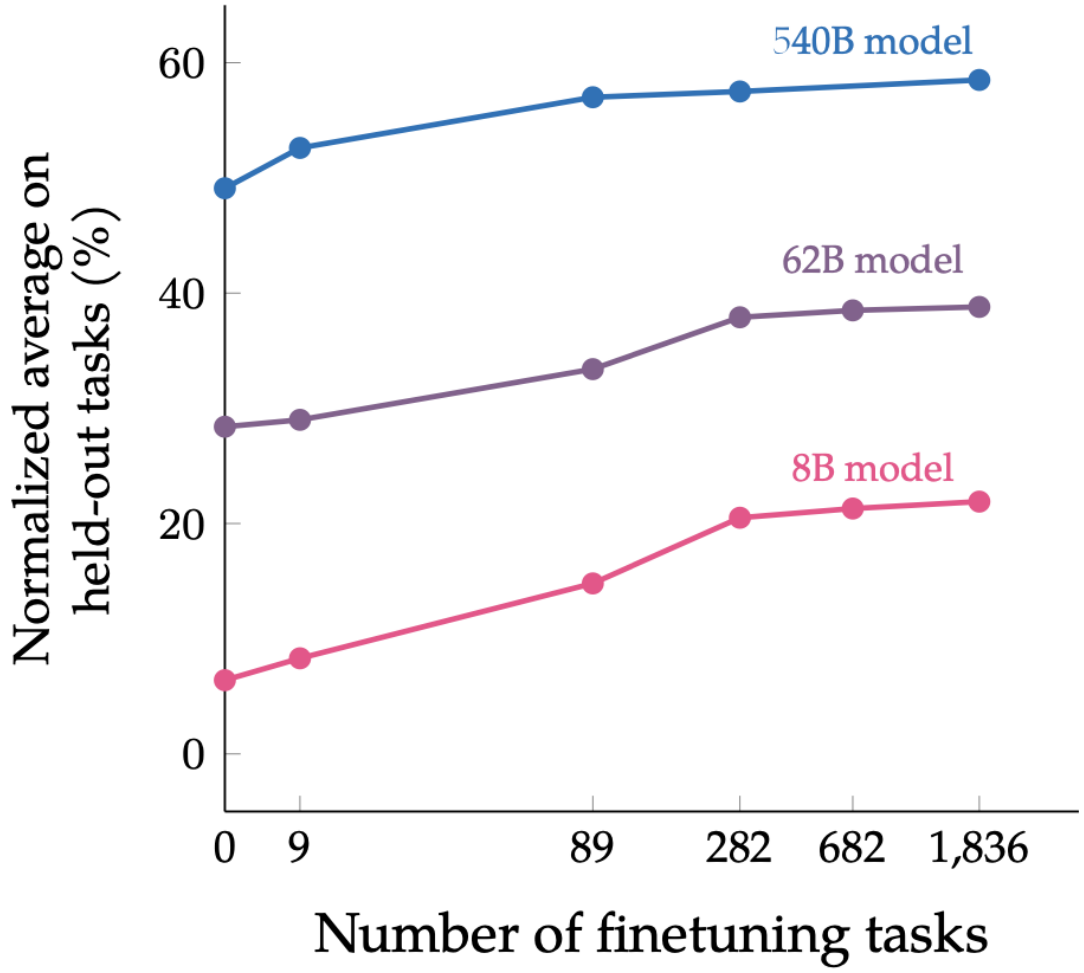
F1 score of Human performance is 93.30%

Flan


Instruction finetuning
Please answer the following question.
What is the boiling point of Nitrogen?
Chain-of-thought finetuning
Answer the following question by reasoning step-by-step.
The cafeteria had 23 apples. If they used 20 for lunch and bought 6 more, how many apples do they have?
Multi-task instruction finetuning (1.8K tasks)
Inference: generalization to unseen tasks
Q: Can Geoffrey Hinton have a conversation with George Washington?
Give the rationale before answering.
Language model
-320.4F
The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9.
Geoffrey Hinton is a British-Canadian computer scientist born in 1947. George Washington died in 1799. Thus, they could not have had a conversation together. So the answer is "no".

Flan-PaLM


Normalized average on held-out tasks (%)
60
40
20
0
8B
62B
540B
Model size (# parameters)
1,836 tasks
282 tasks
89 tasks
9 tasks
No finetuning


Normalized average on held-out tasks (%)
60
40
20
0
0
9
89
282
682
1,836
Number of finetuning tasks
540B model
62B model
8B model

Flan-PaLM

Model input

The square root of x is the cube root of y. What is y to the power of 2, if x = 4?

PaLM 540B output

Q. The square root of x is the cube root of y. What is y to the power of 2, if x = 8?

Q. The square root of x is the cube root of y. What is y to the power of 2, if x = 12?

Q. The square root of x is the cube root of y. What is y to the power of 2, if x = 16?

✖ **(keeps asking more questions)**

Flan-PaLM 540B output

64 ✔

Model input

Make up a word that means "when two AI researchers go on a date".

PaLM 540B output

Make up a word that means "when two AI researchers go on a date".

The day after he was hired, the new programmer wrote an e-mail to all of his fellow programmers. It said, "I will be on vacation next week."

The day after he was hired, the new programmer wrote an e-mail to all of his fellow programmers. It said, "I will be on vacation next week."

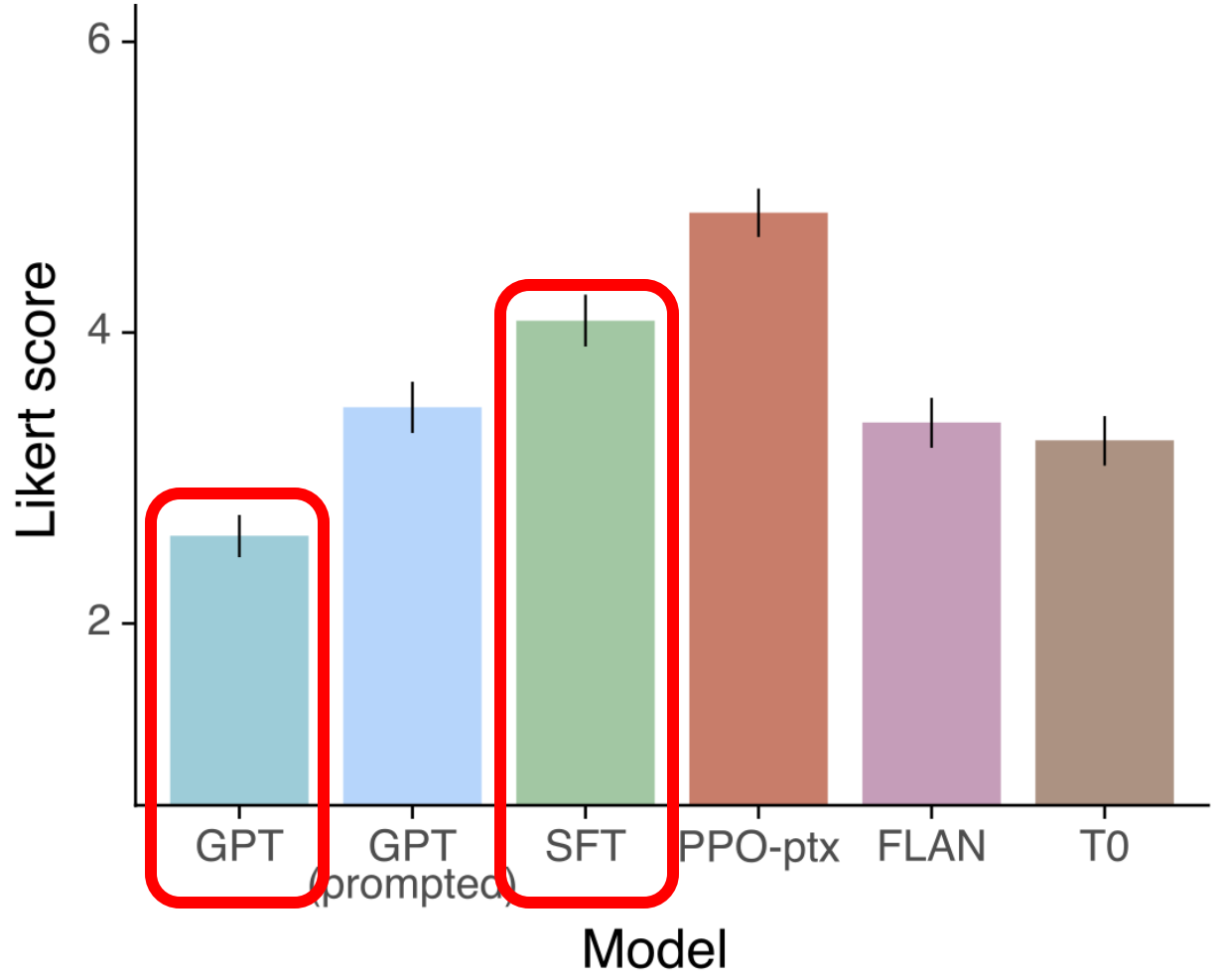
The day after [...]

✖ **(repeats input and keep repeating generations)**

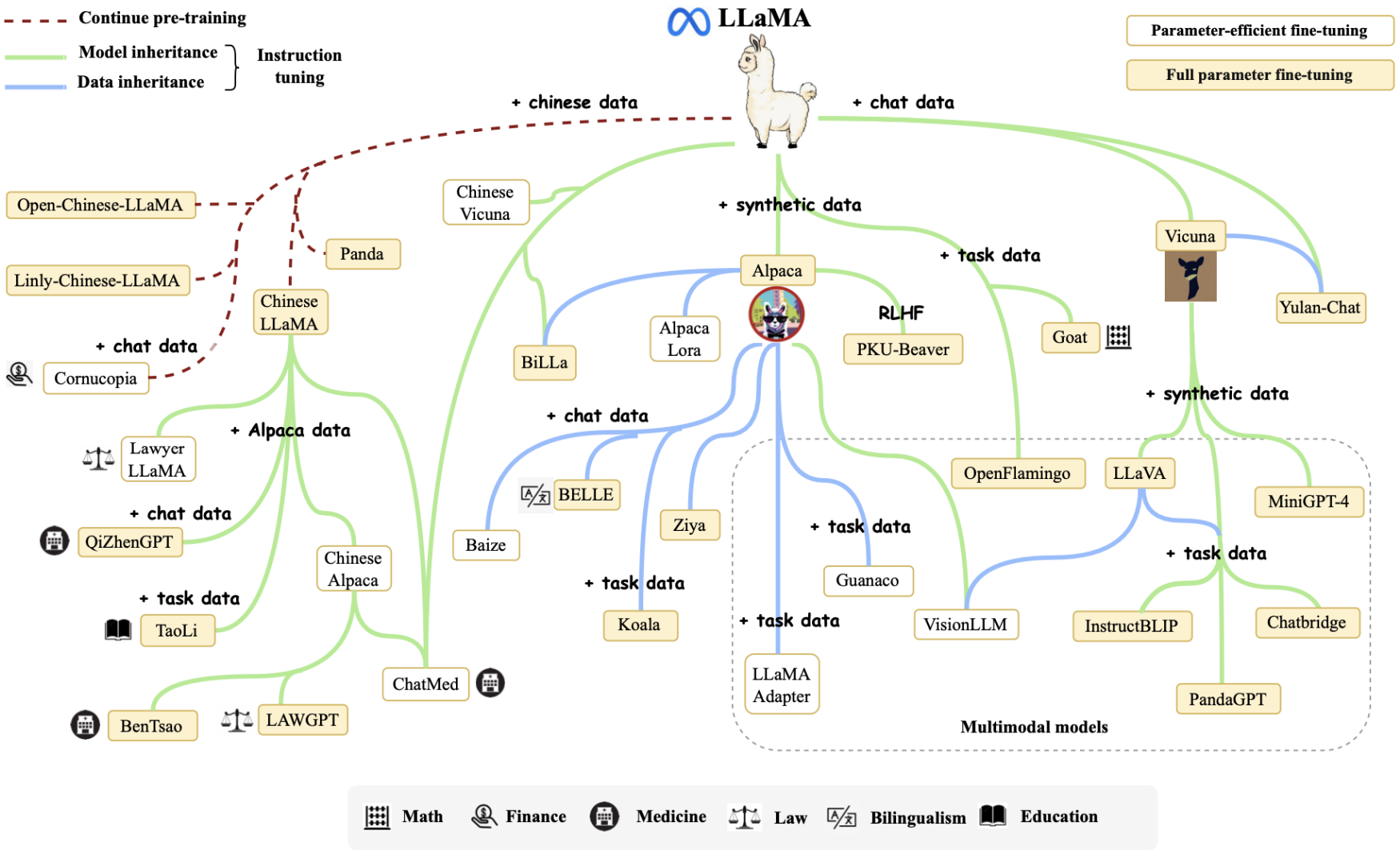
Flan-PaLM 540B output

date-mining ✔

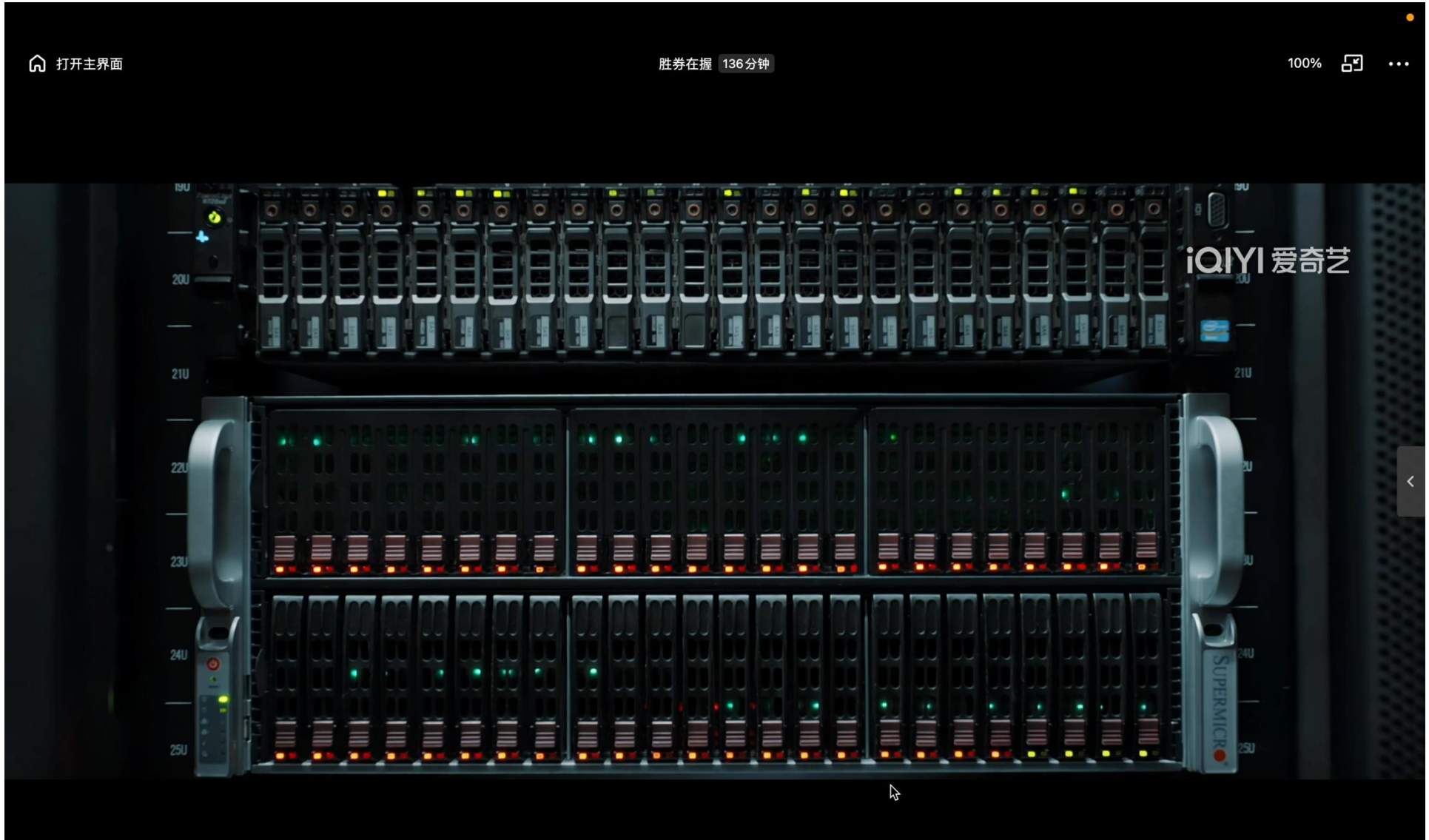
GPT-3 (SFT)



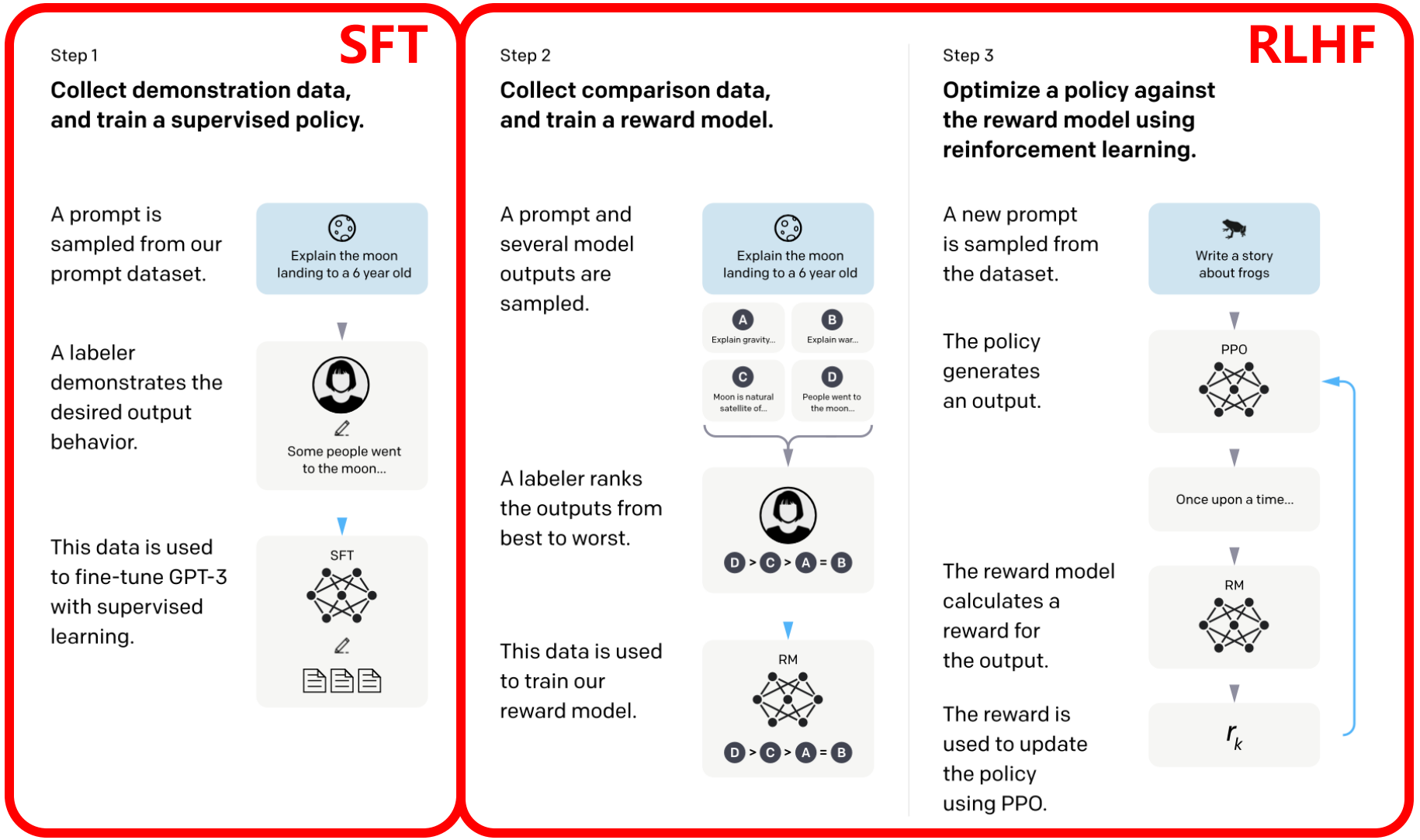
LLM Era


Continue pre-training
Model inheritance
Data inheritance
Instruction tuning
LLaMA
Parameter-efficient fine-tuning
Full parameter fine-tuning
+ chinese data
+ chat data
Open-Chinese-LLaMA
Linly-Chinese-LLaMA
Panda
Chinese LLaMA
+ chat data
Cornucopia
+ Alpaca data
Lawyer LLaMA
+ chat data
QiZhenGPT
Chinese Alpaca
+ task data
TaoLi
BenTsao
LAWGPT
ChatMed
Chinese Vicuna
+ synthetic data
Alpaca
+ task data
Alpaca Lora
RLHF
PKU-Beaver
BiLLa
Goat
+ chat data
BELLE
Baize
Ziya
+ task data
Koala
+ task data
Guanaco
+ task data
LLaMA Adapter
Vicuna
Yulan-Chat
+ synthetic data
OpenFlamingo
LLaVA
MiniGPT-4
+ task data
VisionLLM
InstructBLIP
Chatbridge
PandaGPT
Multimodal models
Math
Finance
Medicine
Law
Bilingualism
Education

LLM Fine-tuning (RL)

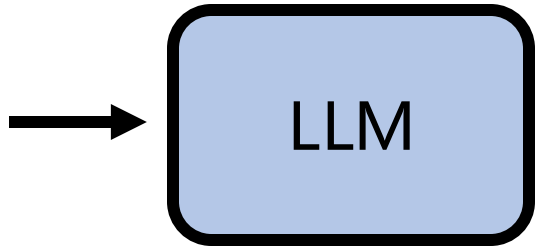


Reinforcement Learning from Human Feedback



Reward Model

Which is the highest mountain in Shenzhen?

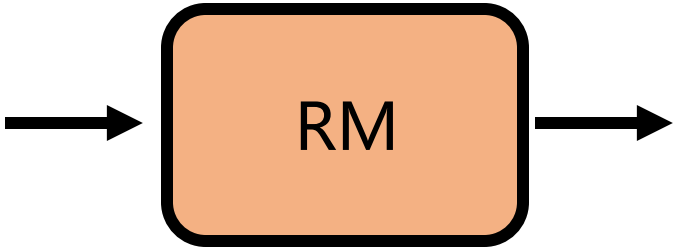


Wutong Mountain

v

Huangshan Mountain

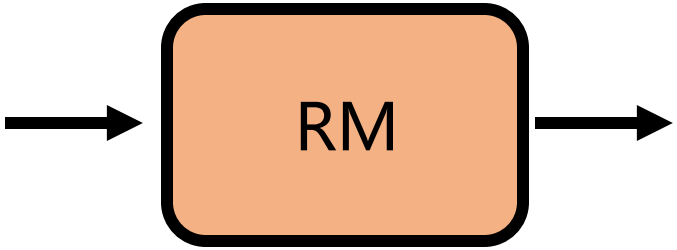
Which is the highest mountain in Shenzhen? Wutong Mountain



Score

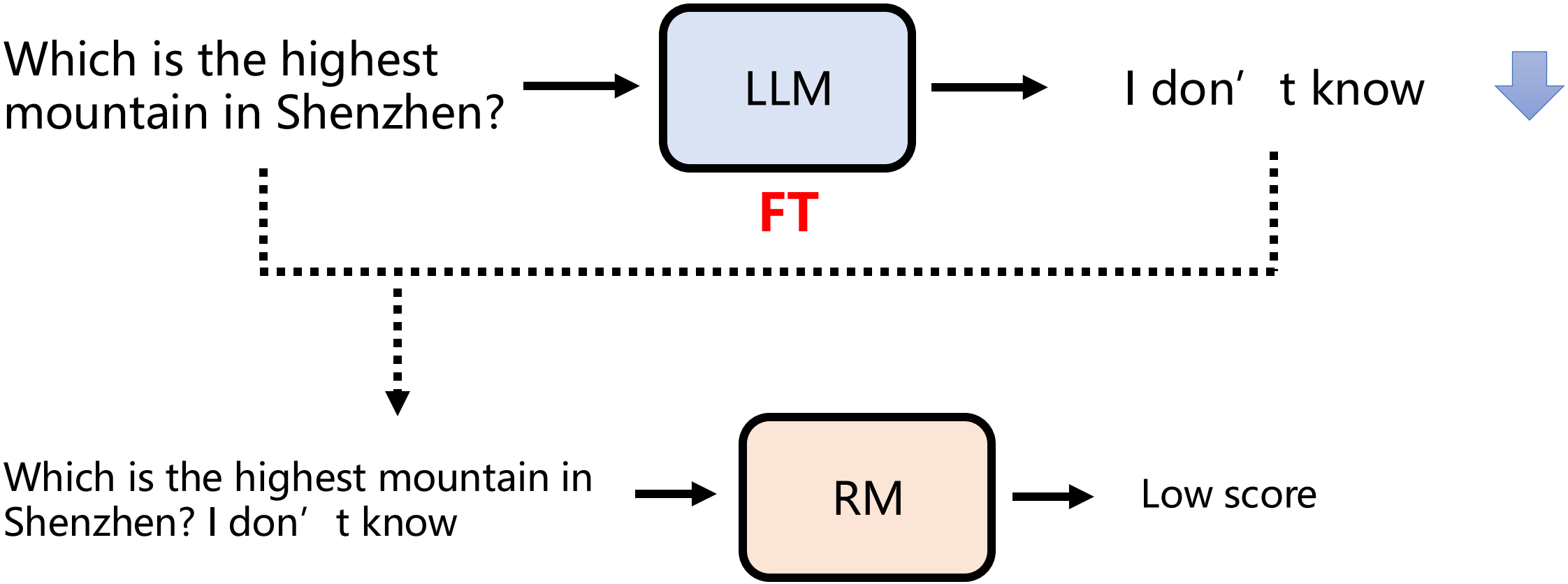
v

Which is the highest mountain in Shenzhen? Huangshan Mountain

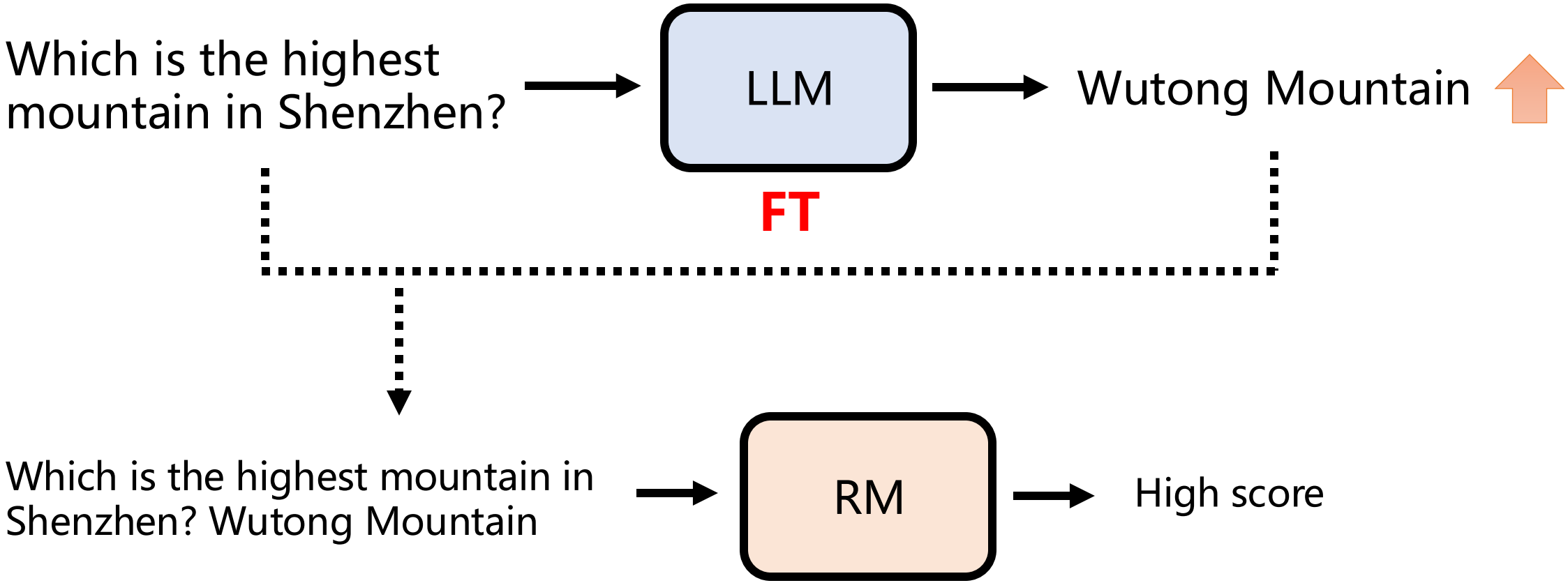


Score

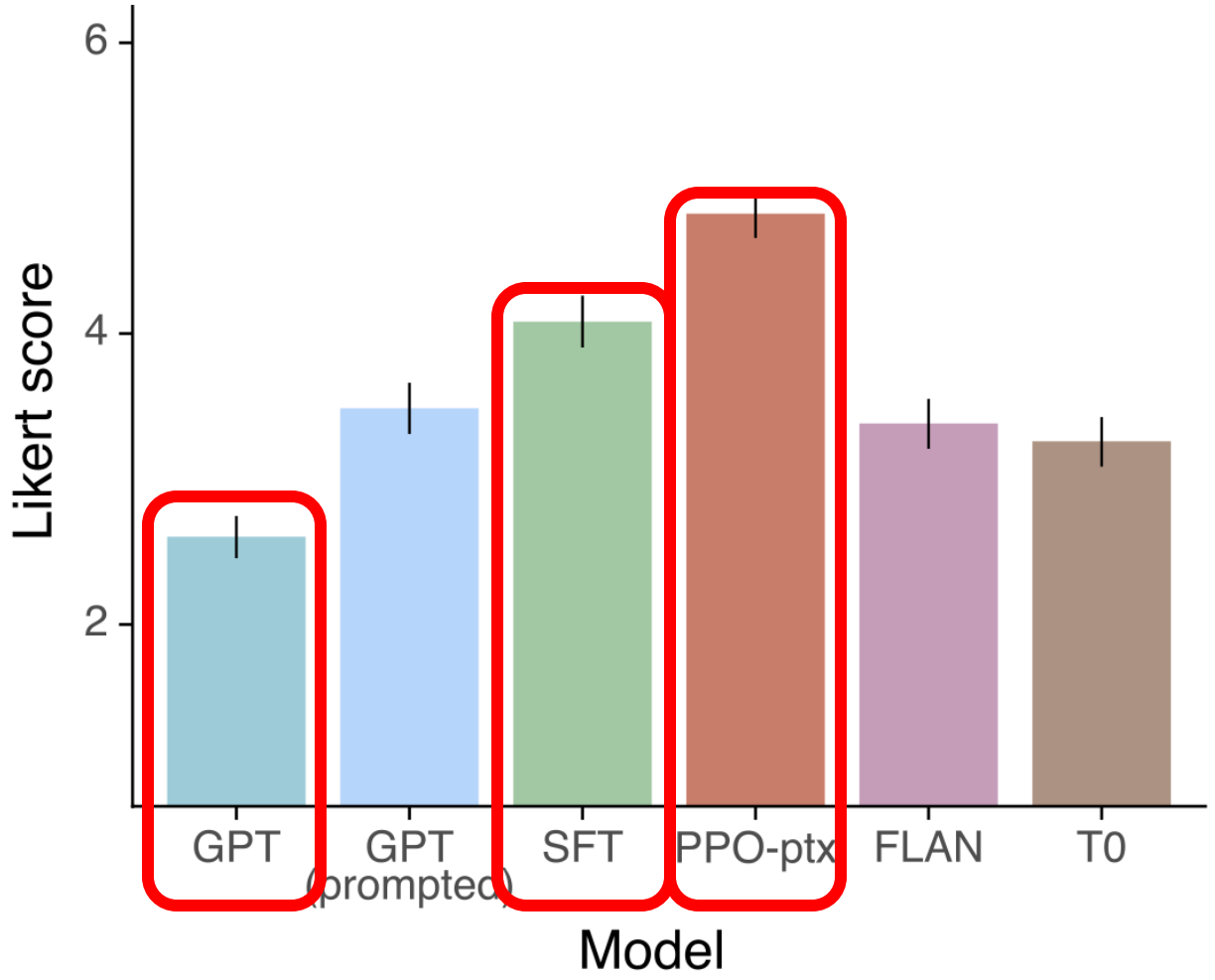
Reward Model



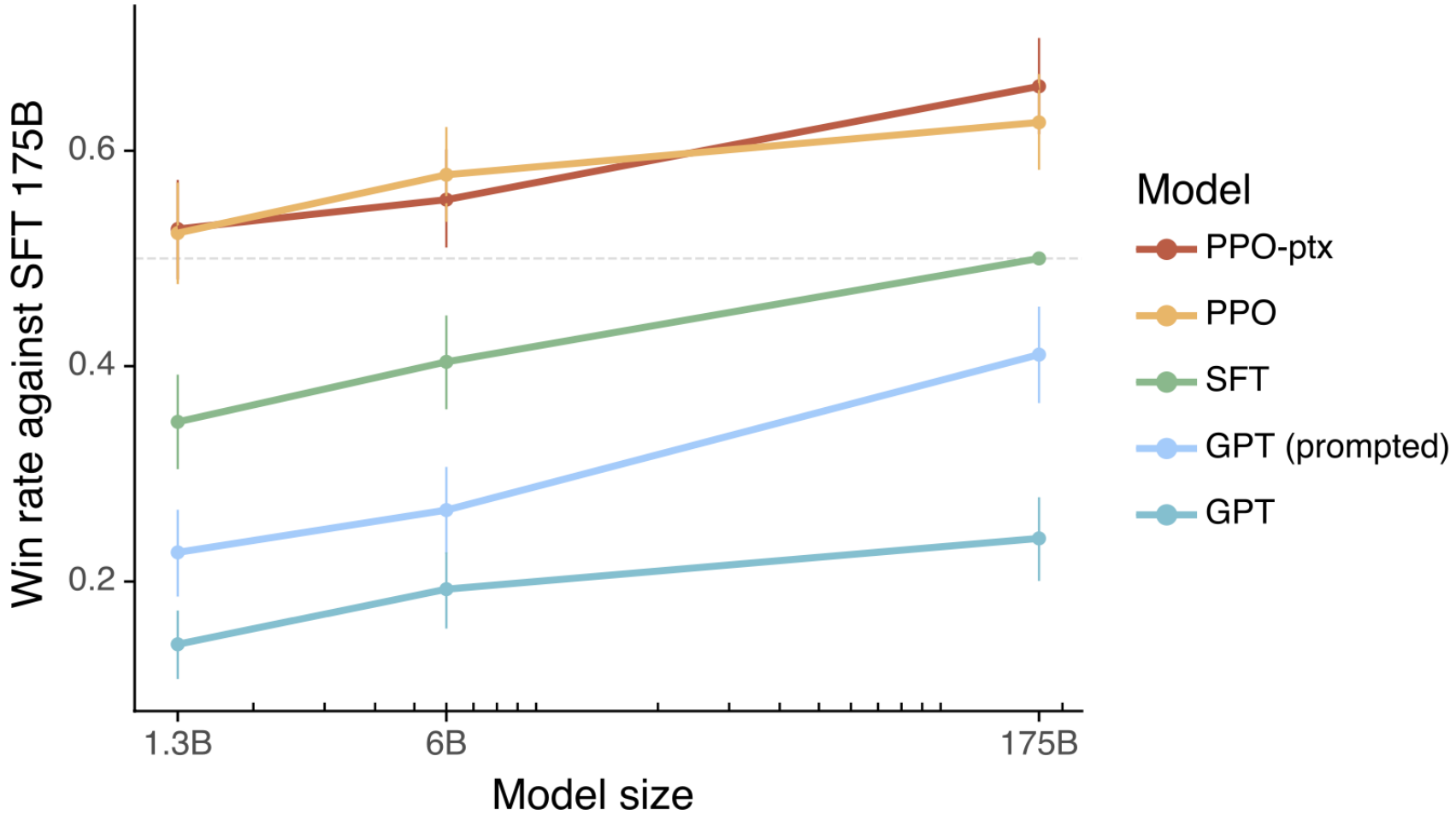
Reward Model



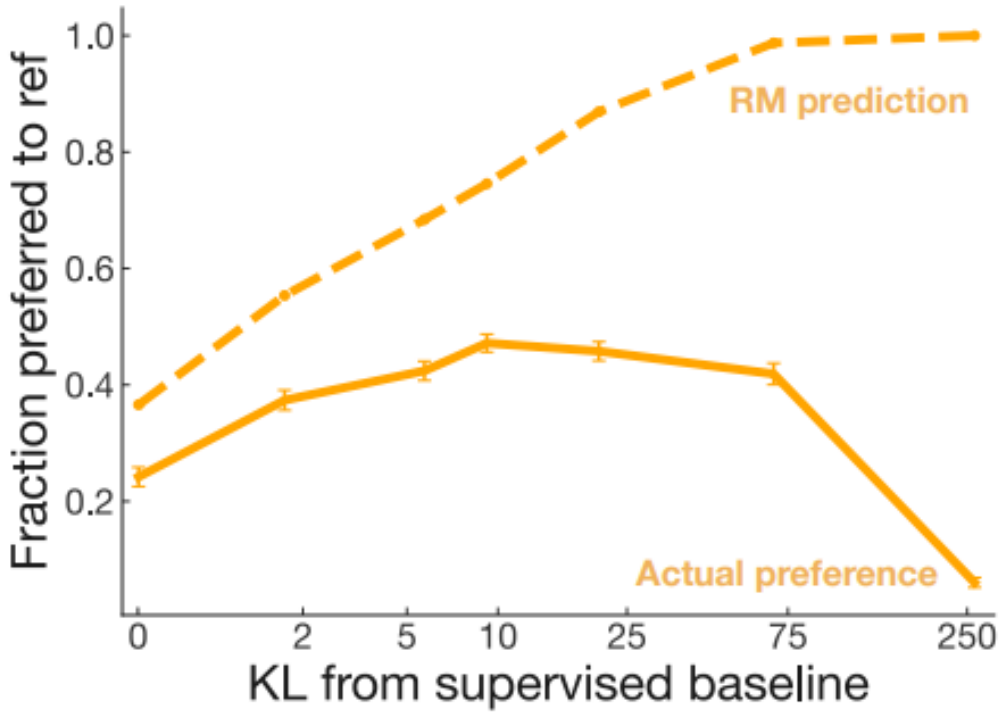
InstructGPT



InstructGPT



Overoptimization



Overoptimization in ChatGPT

- Some overoptimization symptoms we've seen:
  - Excessive verbosity (lists of lists of lists)
  - Excessive apologies, self-doubt
  - "As an AI language model"
  - Hedging language, "there's no one-size-fits-all-solution …."
  - Over-refusals

Summary: LLM Training

	Training Data
• Phase 1: Self-supervised Learning	How are you? I am fine.
• Phase 2: Supervised Learning	User: How are you? AI: I am fine.
• Phase 3: Reinforcement Learning	User: How are you? AI: I am fine. > AI: I love you

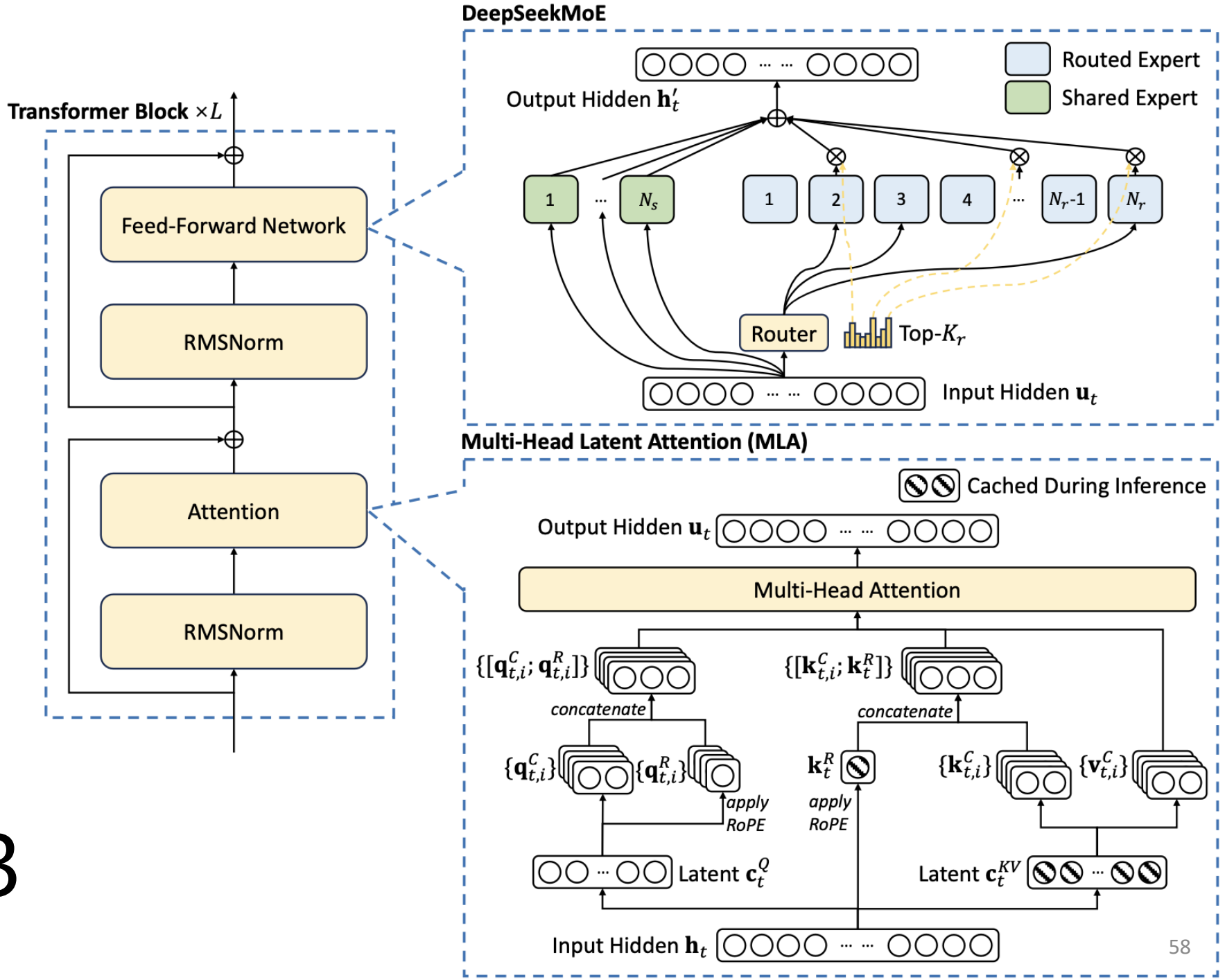
Outline

Transformer

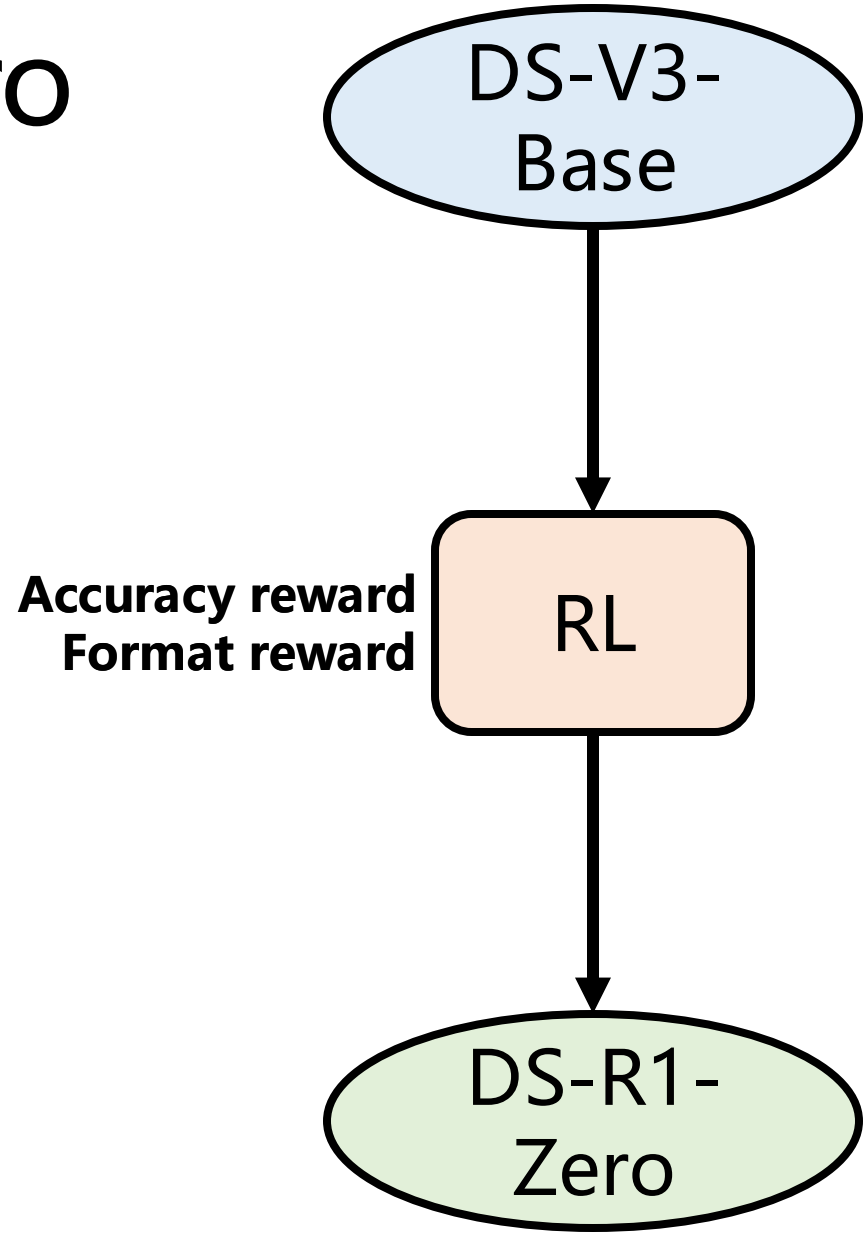
LLM Training

DeepSeek

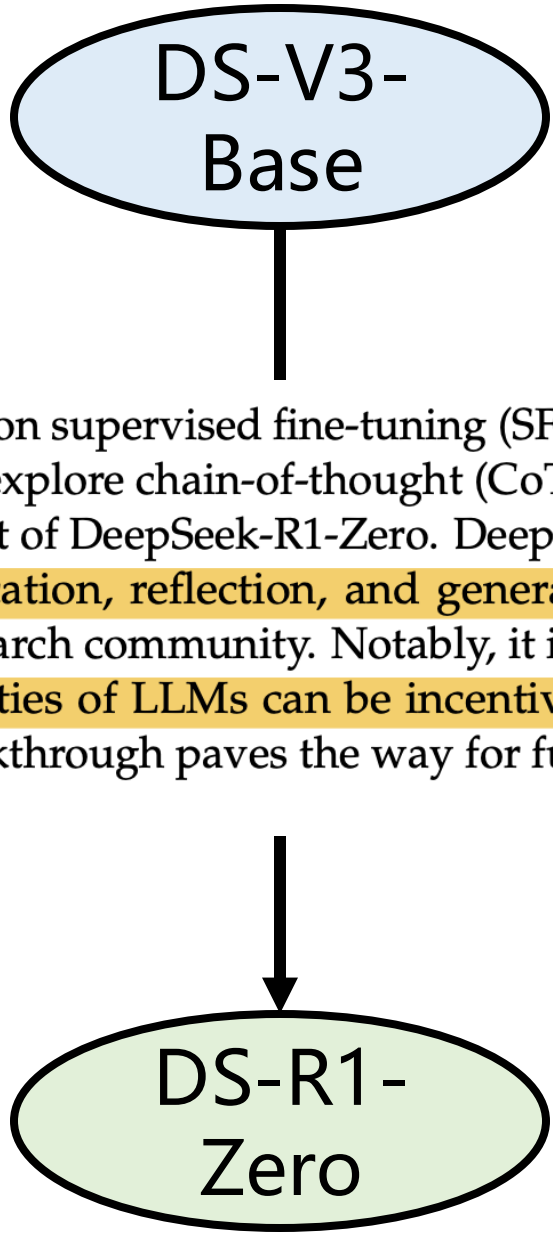
DeepSeek-V3


Transformer Block ×L
Feed-Forward Network
RMSNorm
Attention
RMSNorm
DeepSeekMoE
Output Hidden $\mathbf{h}'_t$
Routed Expert
Shared Expert
1
N_s
1
2
3
4
N_r-1
N_r
Router
Top-K_r
Input Hidden $\mathbf{u}_t$
Multi-Head Latent Attention (MLA)
Cached During Inference
Output Hidden $\mathbf{u}_t$
Multi-Head Attention
$\{[\mathbf{q}^C_{t,i}; \mathbf{q}^R_{t,i}]\}$
$\{[\mathbf{k}^C_{t,i}; \mathbf{k}^R_t]\}$
concatenate
concatenate
$\{\mathbf{q}^C_{t,i}\}$
$\{\mathbf{q}^R_{t,i}\}$
apply RoPE
$\mathbf{k}^R_t$
apply RoPE
$\{\mathbf{k}^C_{t,i}\}$
$\{\mathbf{v}^C_{t,i}\}$
Latent $\mathbf{c}^Q_t$
Latent $\mathbf{c}^{KV}_t$
Input Hidden $\mathbf{h}_t$

DeepSeek-R1-Zero

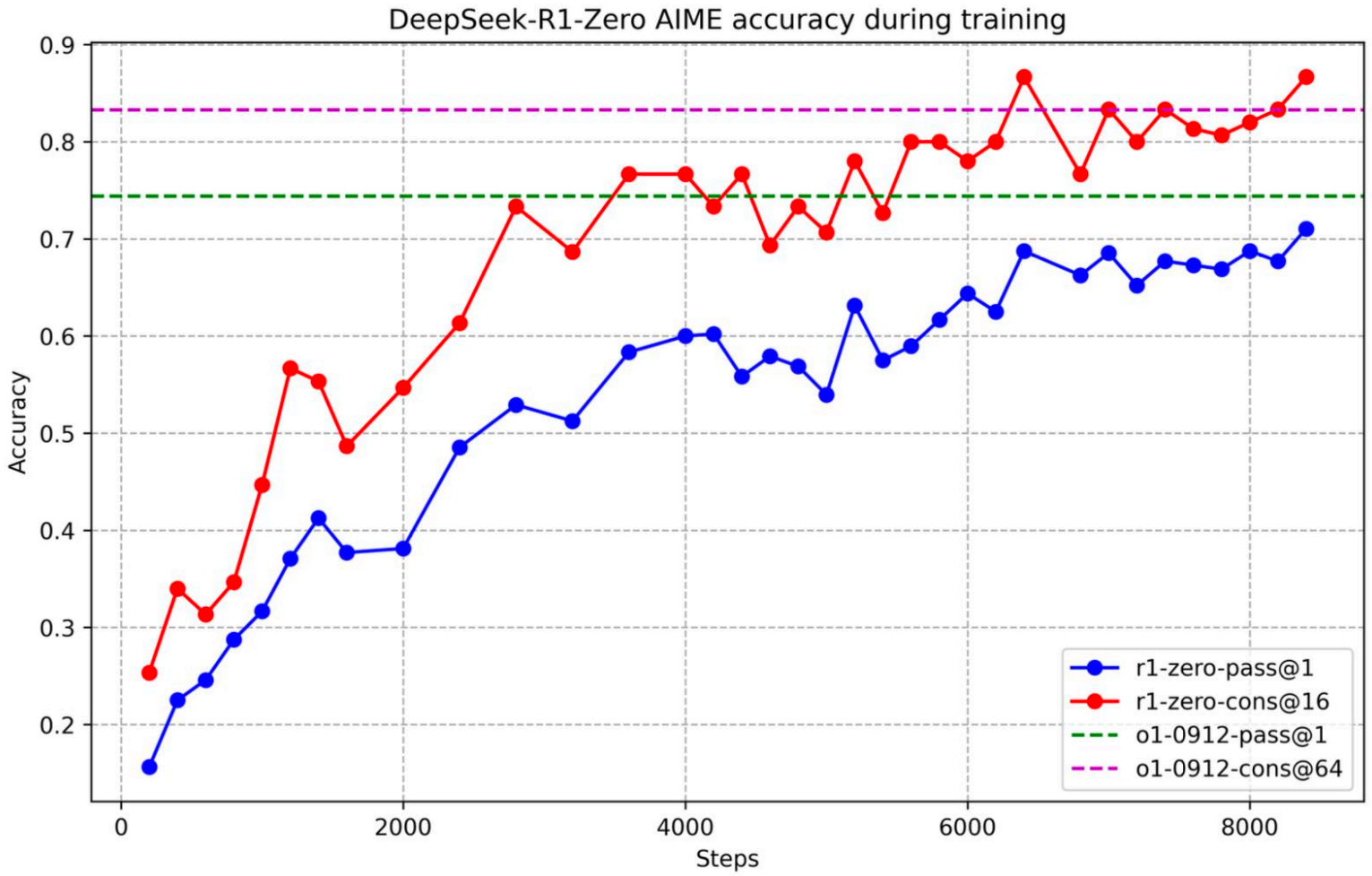


DeepSeek-R1-Zero

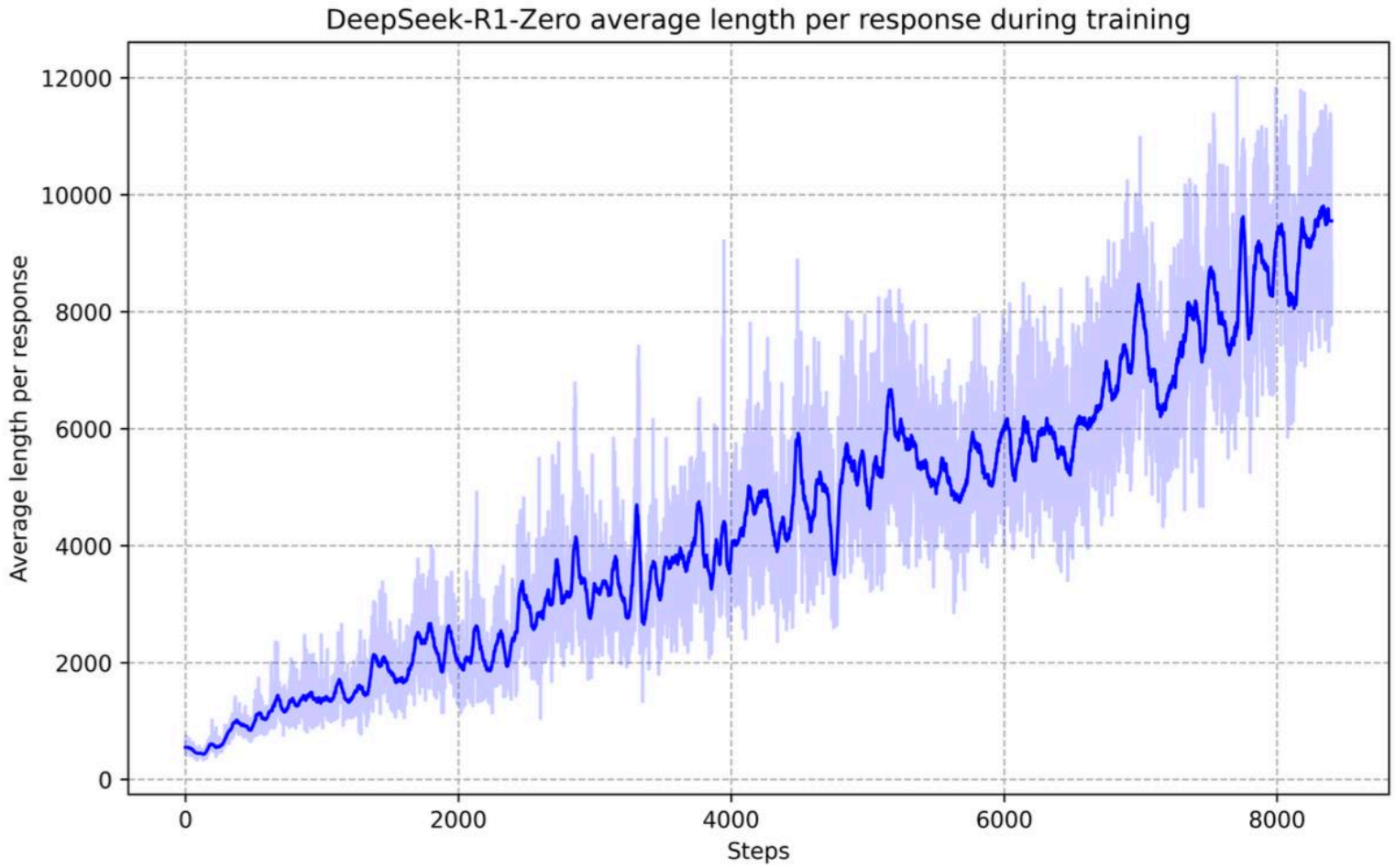


We directly apply RL to the base model without relying on supervised fine-tuning (SFT) as a preliminary step. This approach allows the model to explore chain-of-thought (CoT) for solving complex problems, resulting in the development of DeepSeek-R1-Zero. DeepSeek-R1-Zero demonstrates capabilities such as self-verification, reflection, and generating long CoTs, marking a significant milestone for the research community. Notably, it is the first open research to validate that reasoning capabilities of LLMs can be incentivized purely through RL, without the need for SFT. This breakthrough paves the way for future advancements in this area.

DeepSeek-R1-Zero



DeepSeek-R1-Zero



Aha Moment

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

Response: <think>

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

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

Rearrange to isolate the inner square root term:

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

$\cdots$

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

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

We started with the equation:

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

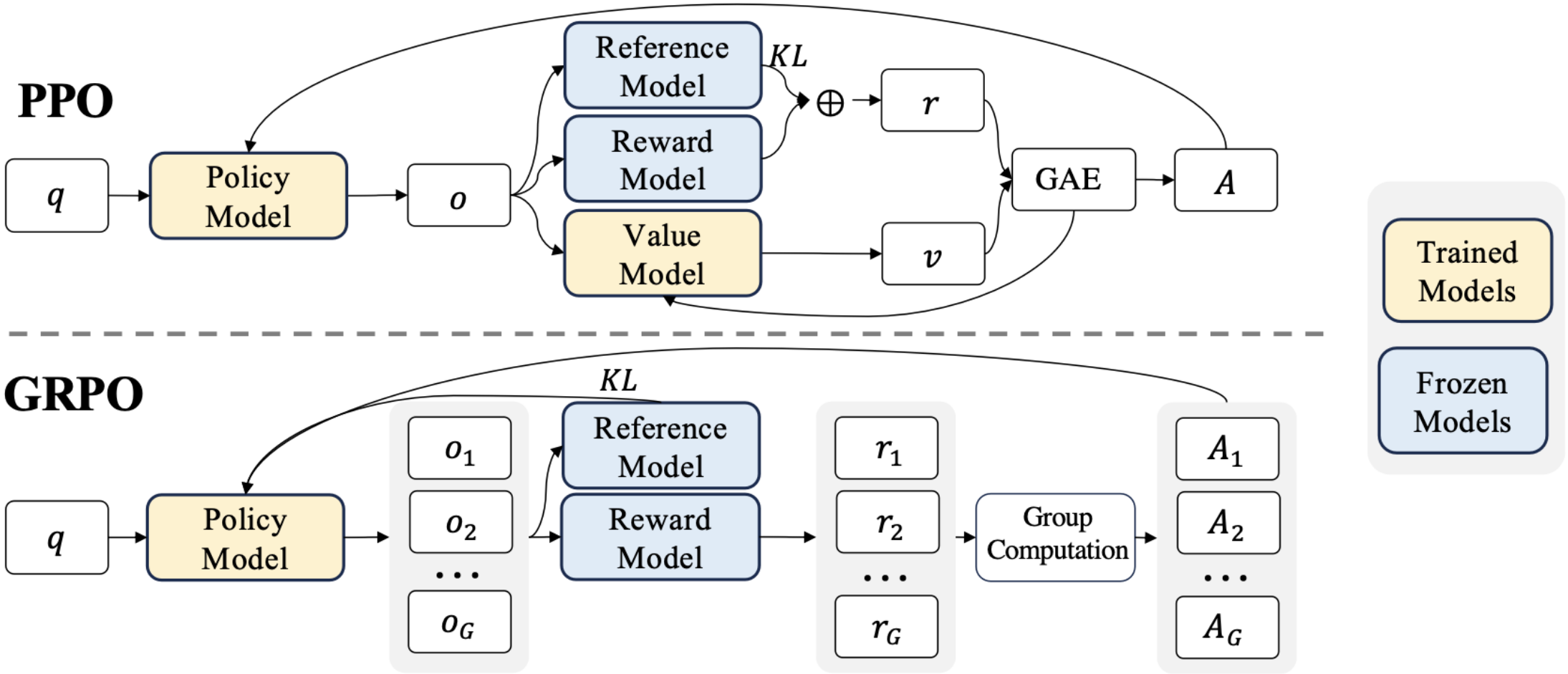
First, let's square both sides:

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

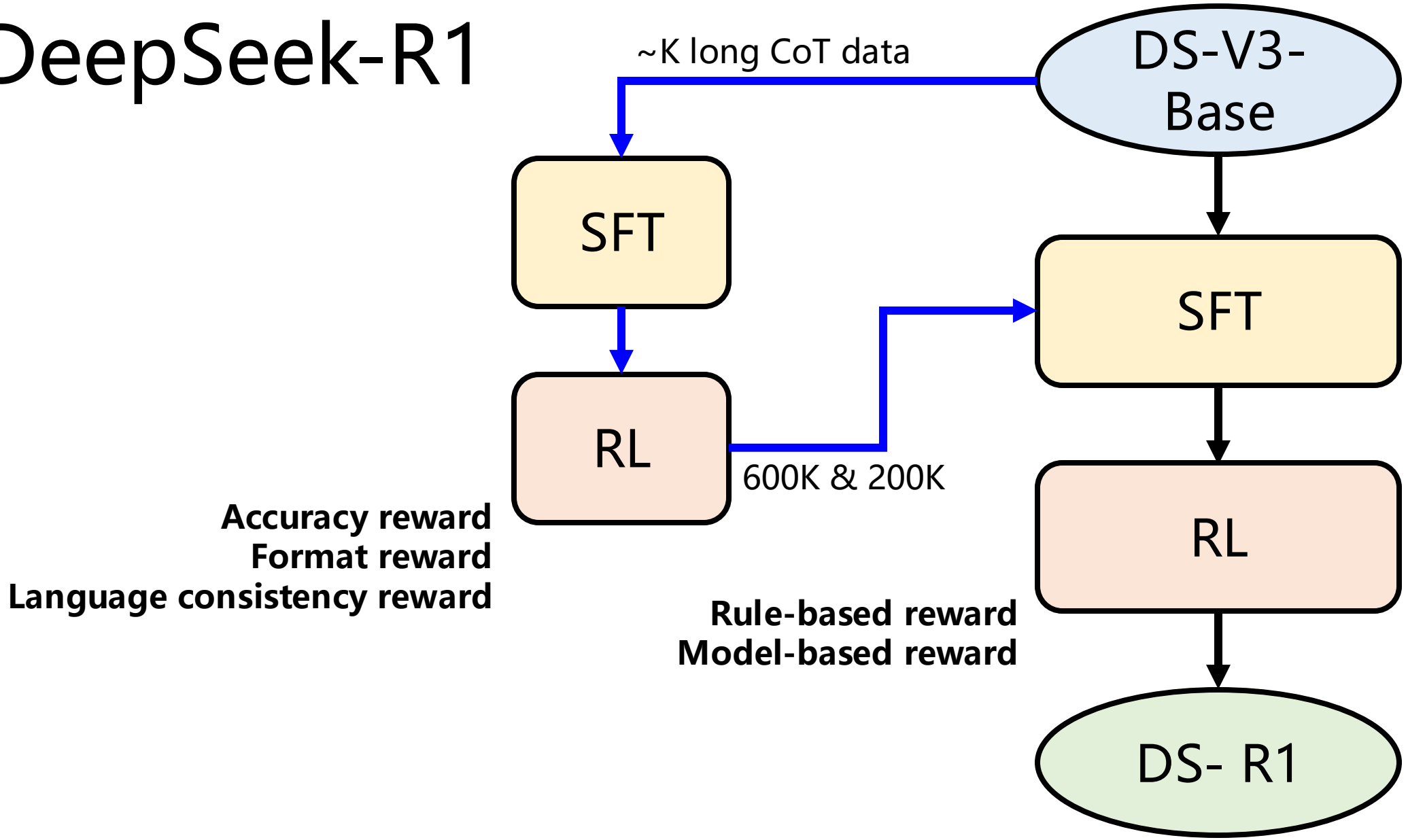
Next, I could square both sides again, treating the equation: $\cdots$

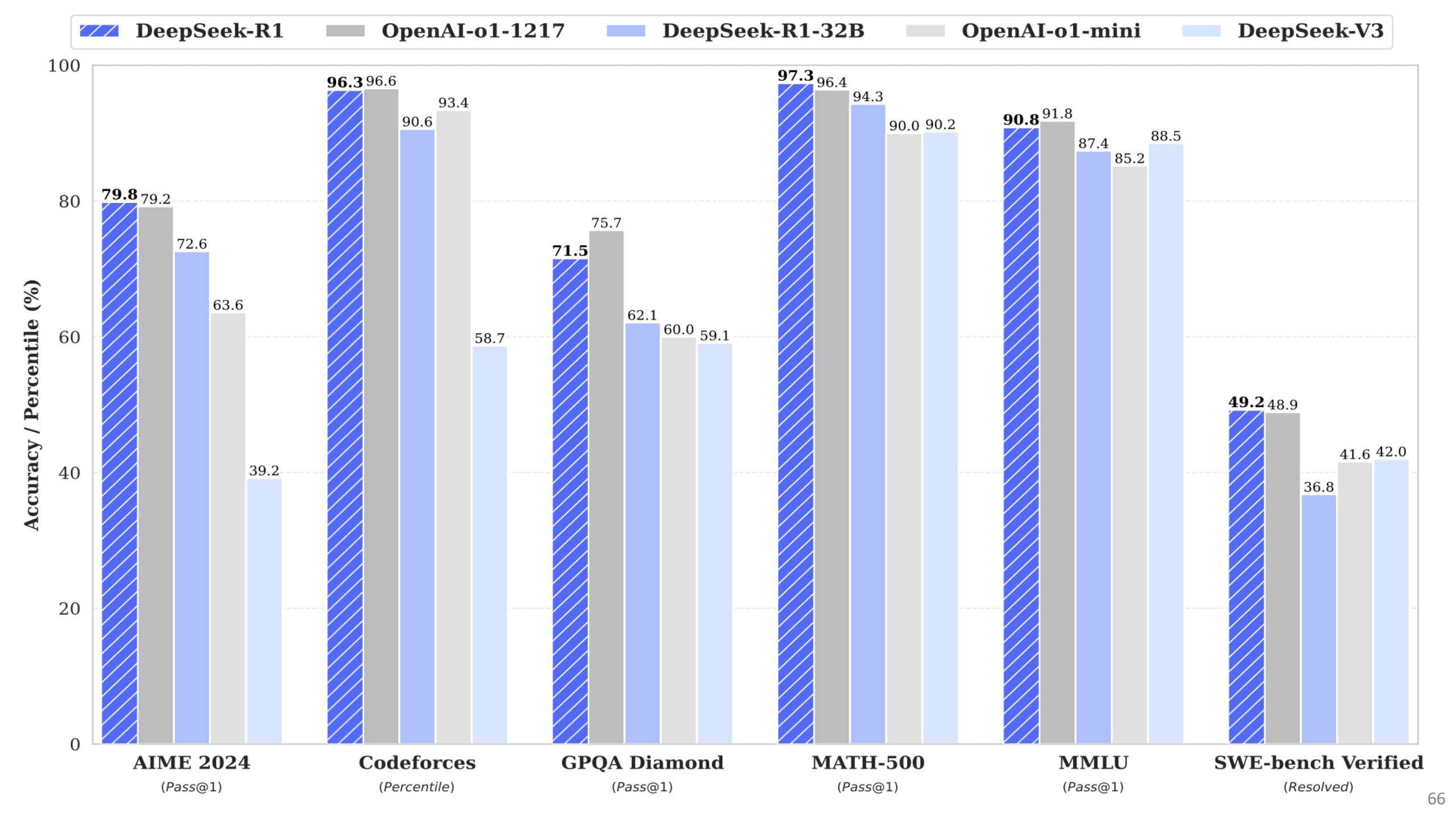
$\cdots$

Group Relative Policy Optimization



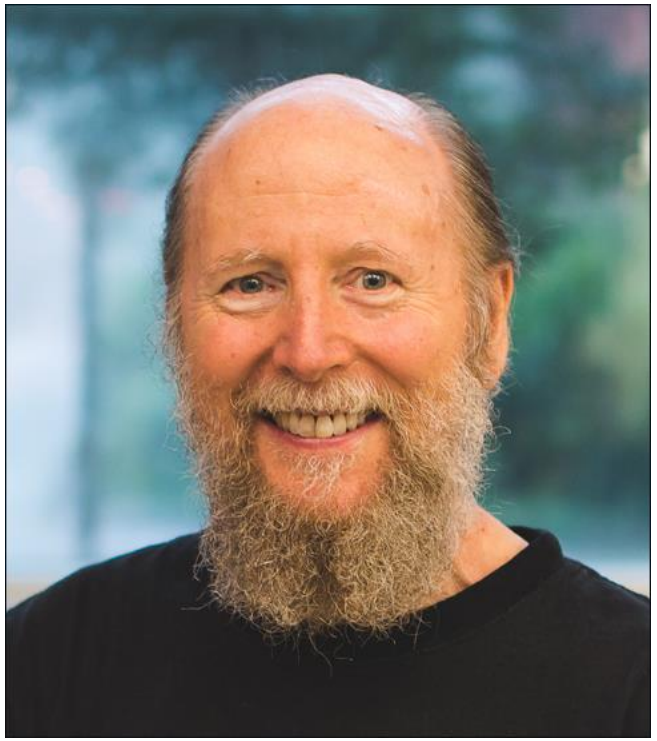
DeepSeek-R1


~K long CoT data
DS-V3-Base
SFT
SFT
RL
600K & 200K
RL
Accuracy reward
Format reward
Language consistency reward
Rule-based reward
Model-based reward
DS- R1


DeepSeek-R1
OpenAI-o1-1217
DeepSeek-R1-32B
OpenAI-o1-mini
DeepSeek-V3
Accuracy / Percentile (%)
100
80
60
40
20
0
79.8
79.2
72.6
63.6
39.2
96.3
96.6
90.6
93.4
58.7
71.5
75.7
62.1
60.0
59.1
97.3
96.4
94.3
90.0
90.2
90.8
91.8
87.4
85.2
88.5
49.2
48.9
36.8
41.6
42.0
AIME 2024
(Pass@1)
Codeforces
(Percentile)
GPQA Diamond
(Pass@1)
MATH-500
(Pass@1)
MMLU
(Pass@1)
SWE-bench Verified
(Resolved)

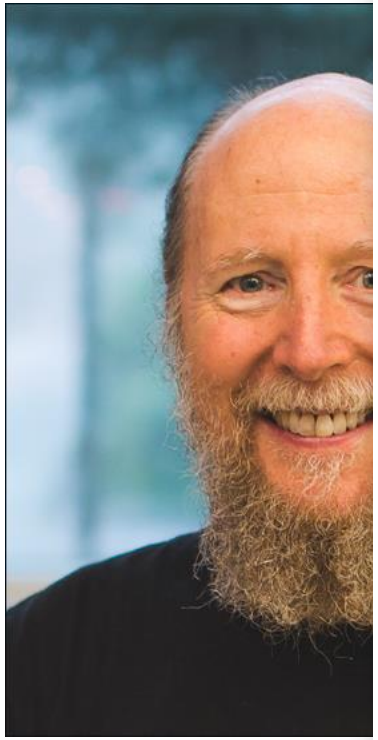
2024 Turing Award

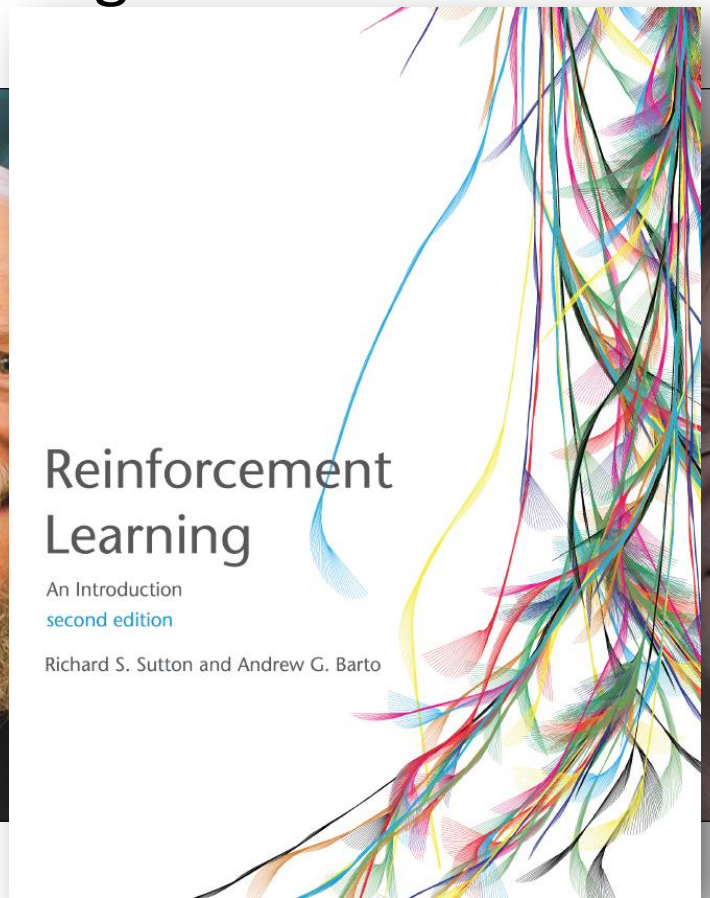
- Andrew Barto and Richard Sutton Recognized as Pioneers of Reinforcement Learning



2024 Turing Award

- Andrew Barto and Richard Sutton Recognized as Pioneers of Reinforcement Learning





THANK YOU