Introduction to Language Models

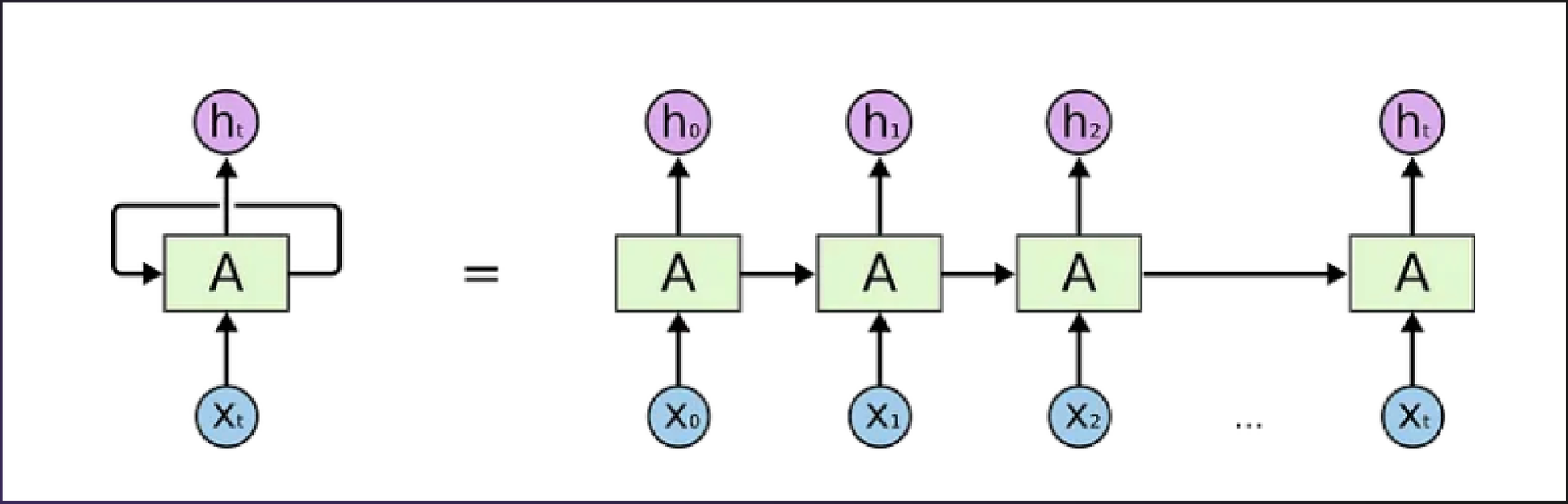
Pranav Reddy
Cofounder, Xylem AI
@pranavreddyg (Twitter/X)

xylem.ai

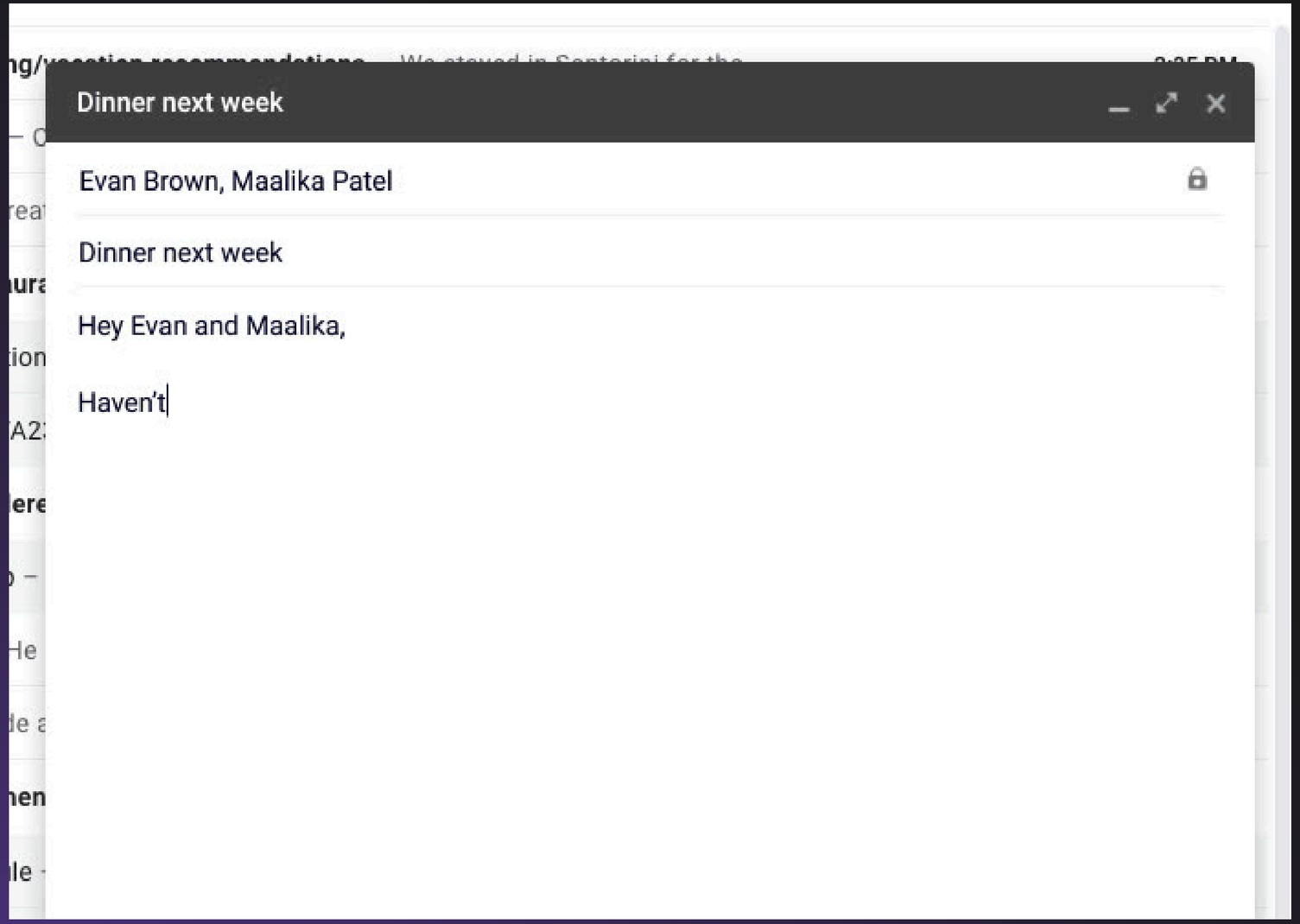

I DON'T KNOW WHAT
A LARGE LANGUAGE MODEL IS
AND AT THIS POINT
I AM TOO AFRAID TO ASK
imgflip.com

xylem.ai

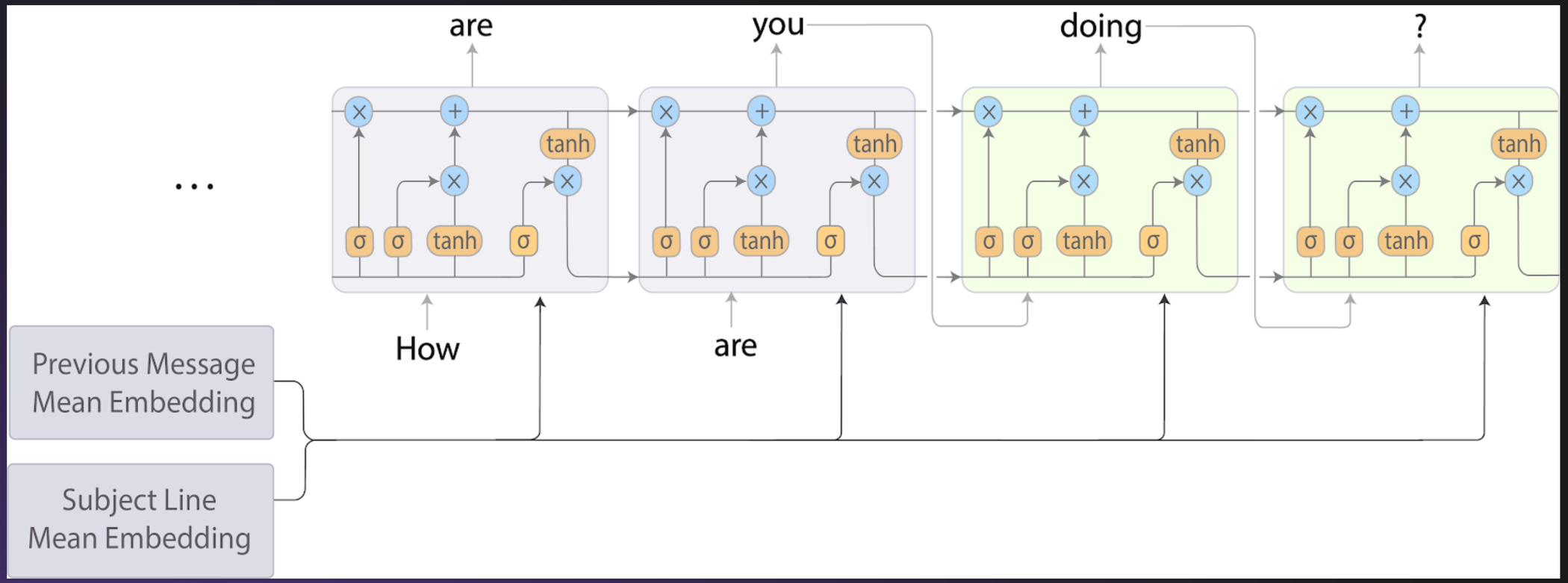
Recurrent Neural Networks



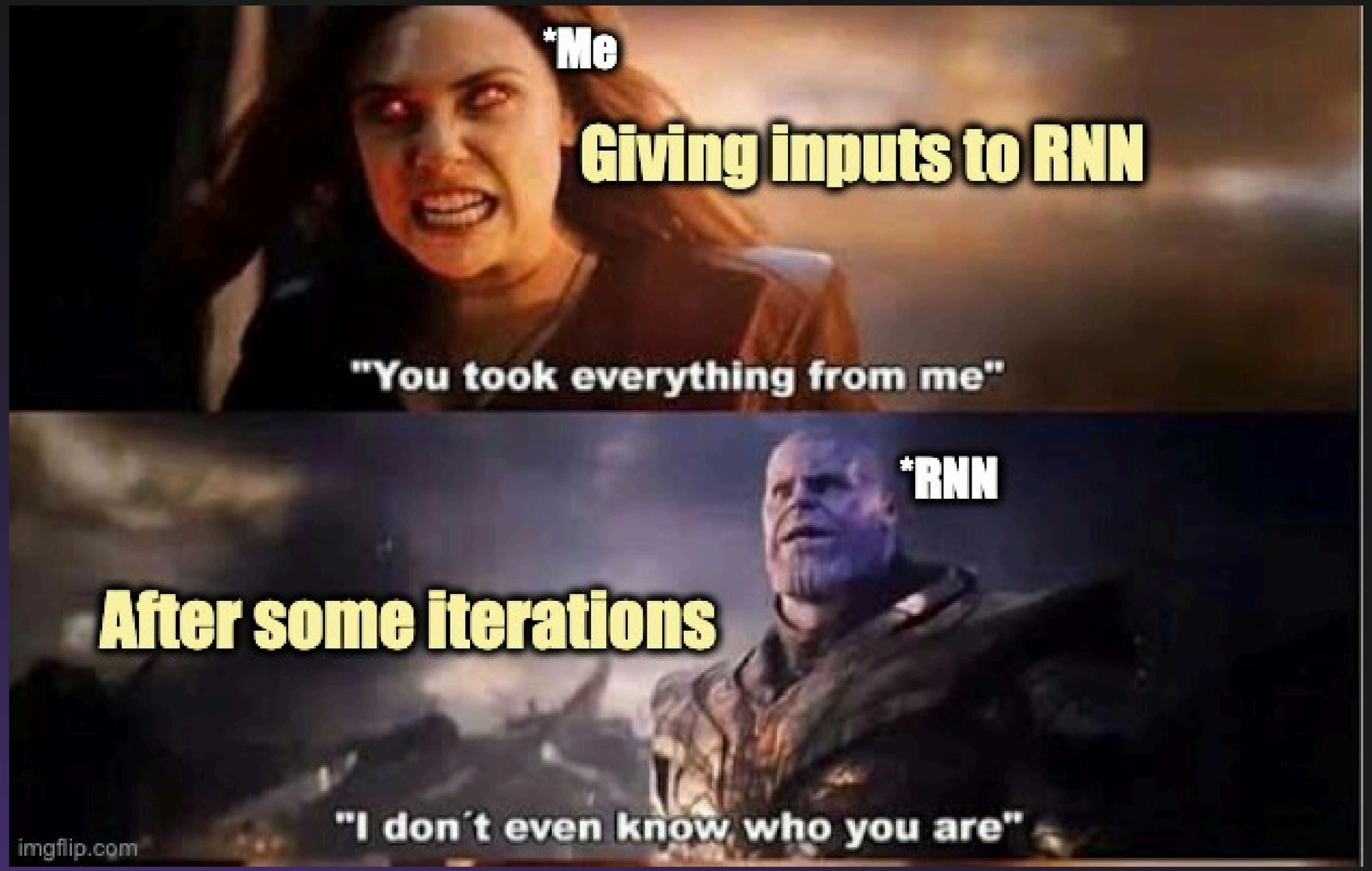
Smart Compose (Gmail)

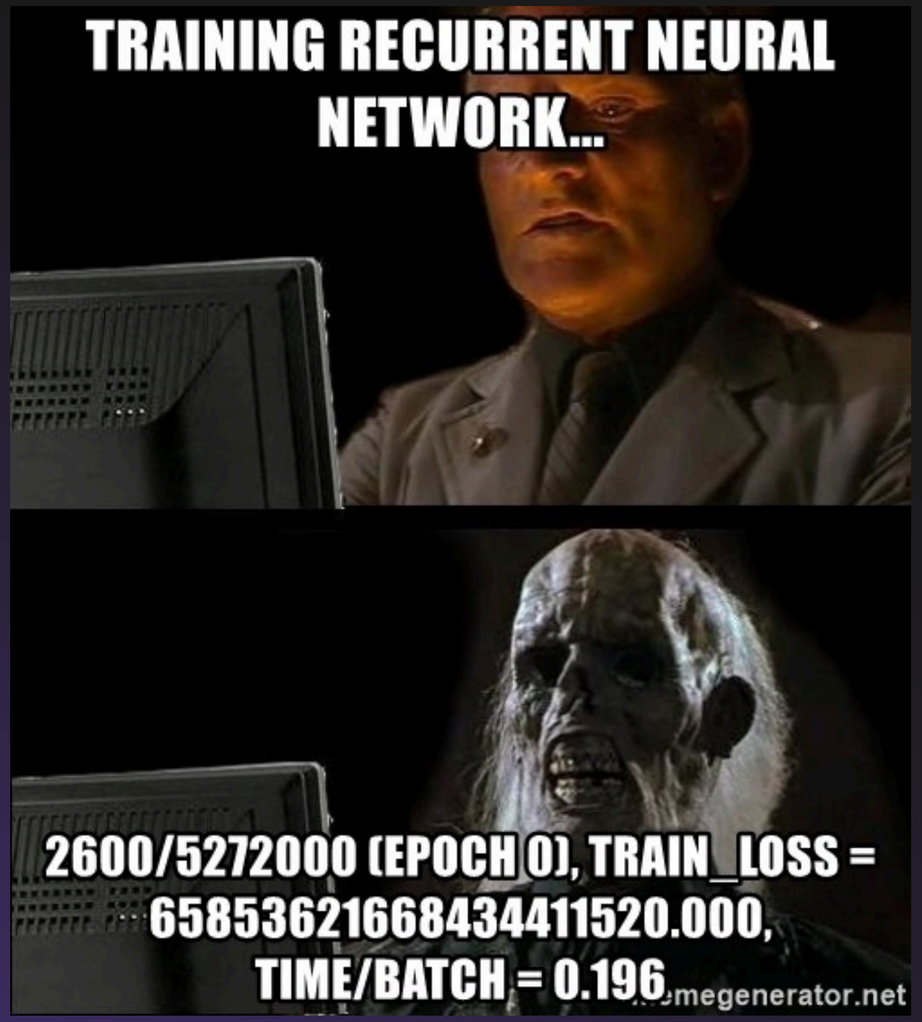


How it works


are
you
doing
?
...
tanh
σ
How
are
Previous Message
Mean Embedding
Subject Line
Mean Embedding


I CAN MAKE THE BAD GUYS
GOOD FOR A WEEKEND
I CAN MAKE THE BAD GUYS
TPGMP
HAVE A LIFE-CHANGING JOURNEY
AND BECOME GREAT PEOPLE
FOR THE REST OF THEIR LIVES
ifunny.co


*Me
Giving inputs to RNN
"You took everything from me"
*RNN
After some iterations
"I don´t even know who you are"
imgflip.com


TRAINING RECURRENT NEURAL NETWORK...
2600/5272000 (EPOCH 0), TRAIN_LOSS = 6585362166843441152O.000, TIME/BATCH = 0.196
megenerator.net

xylem.ai

Problems with RNNs

- Slow computation for long sequences
- Vanishing or exploding gradients
- Difficulty in retaining context in long sequences
- Very hard to train

Attention is all you need

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

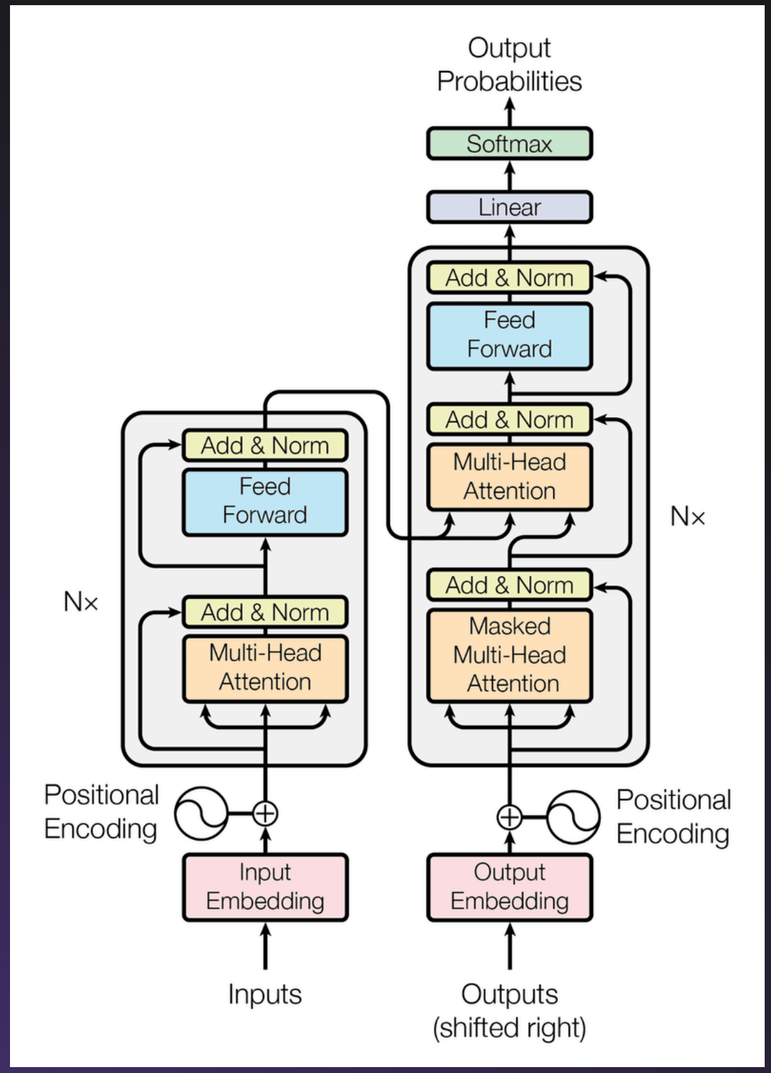
Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

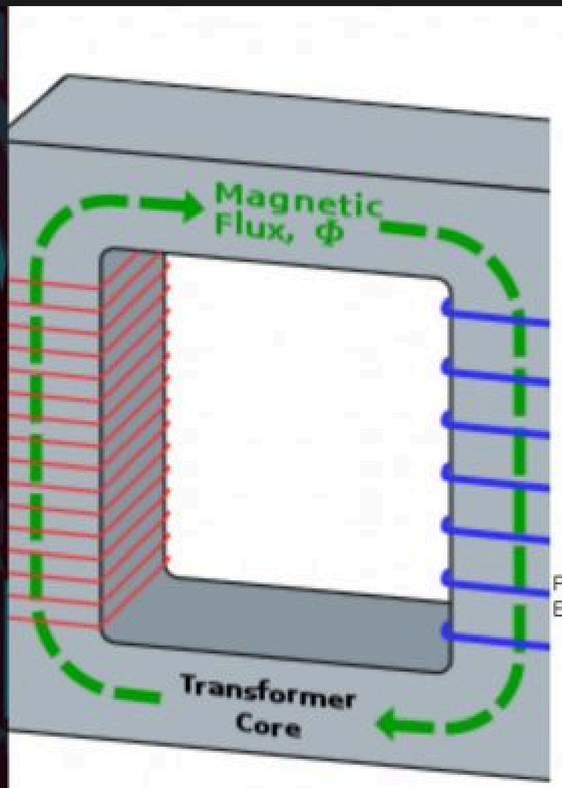
Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly

Transformer architecture



xylem.ai



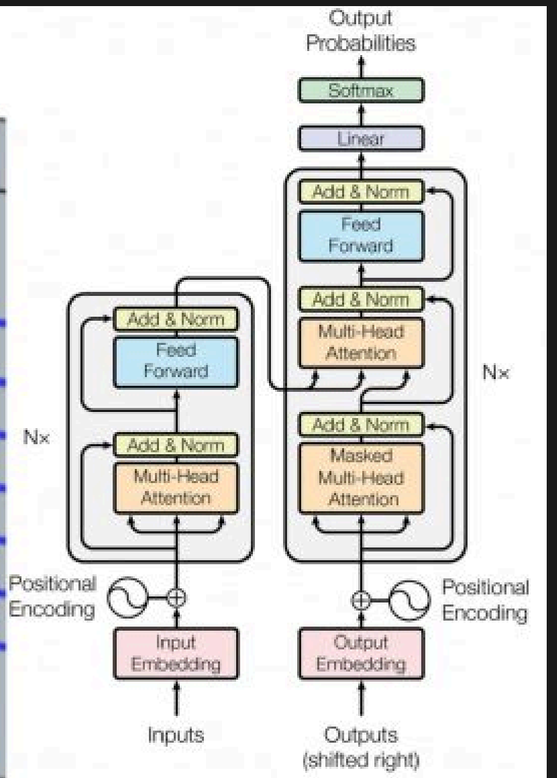


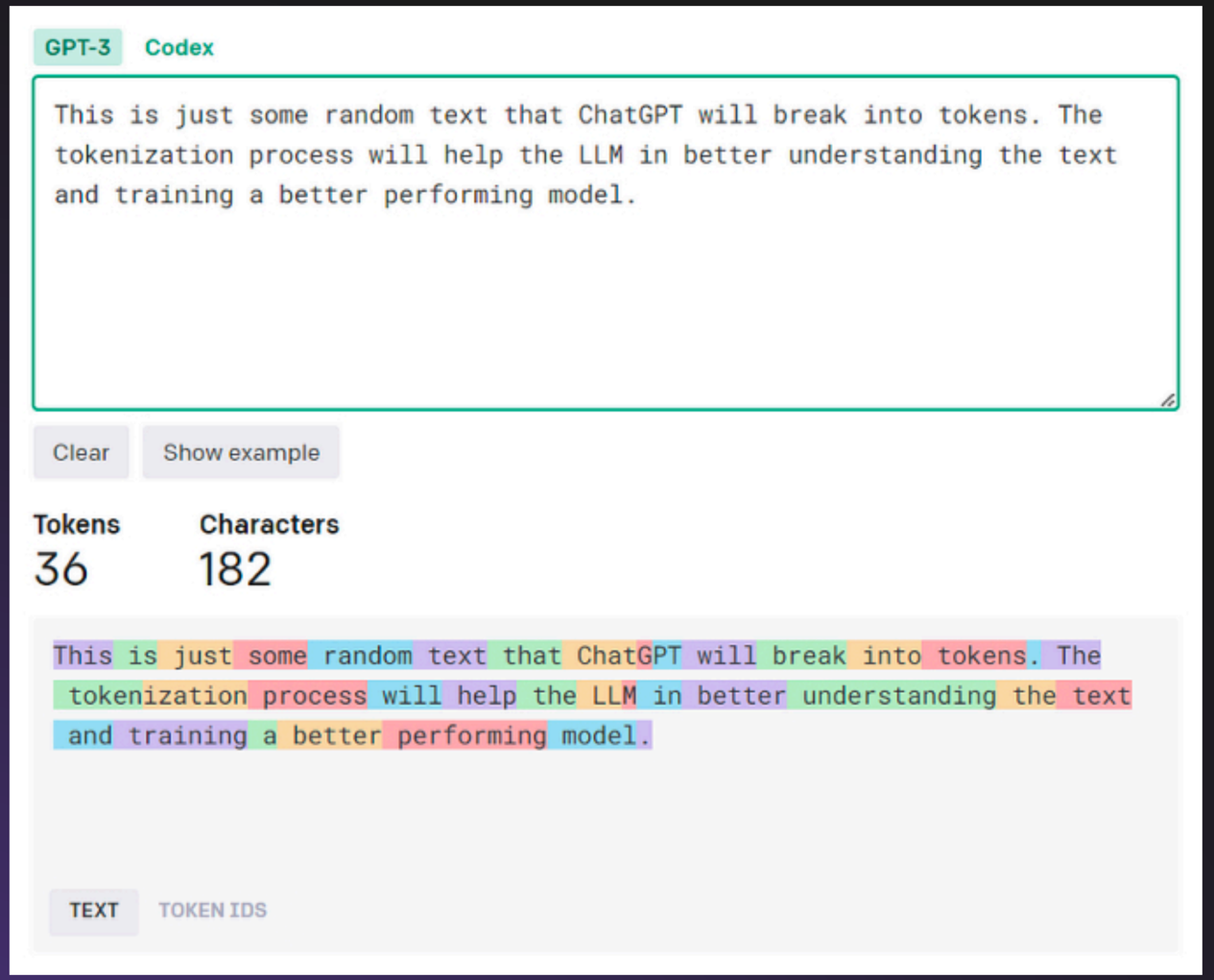
Figure 1: The Transformer - model architecture.

Transformers at school

Transformers at college

Transformers today

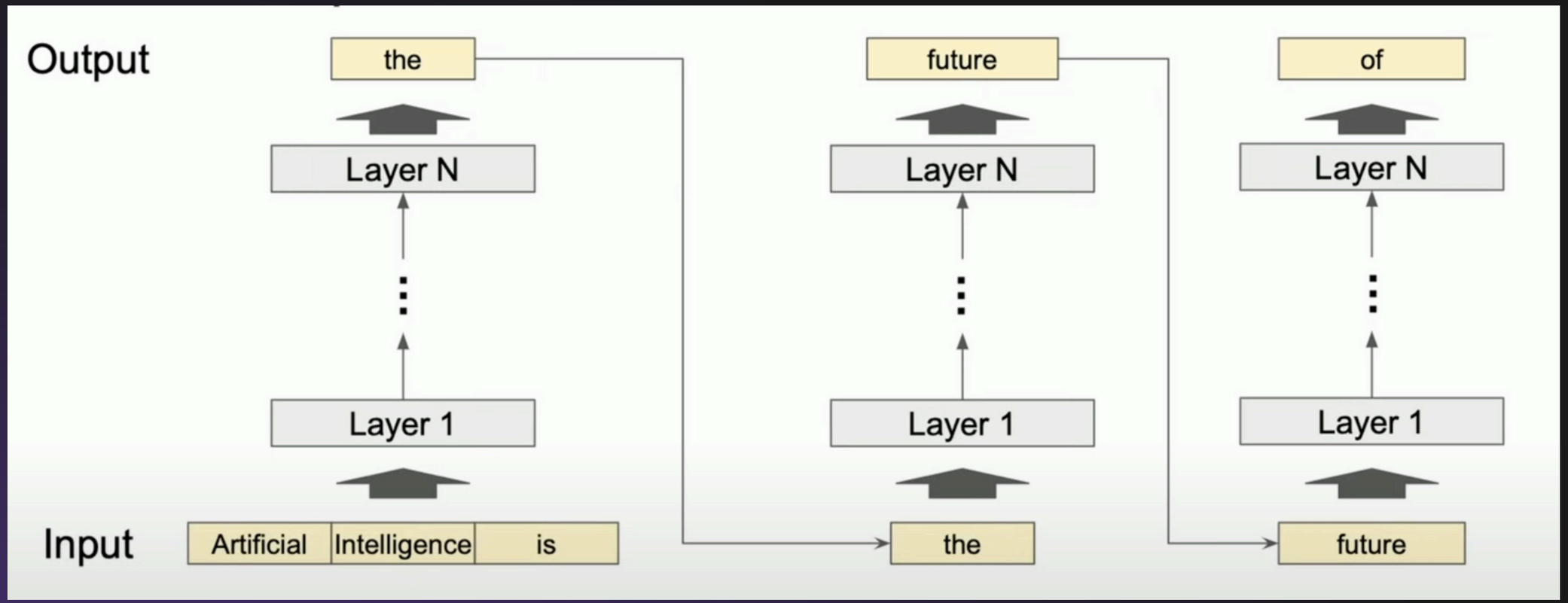
Tokenization



OpenAI Tokenizer

xylem.ai

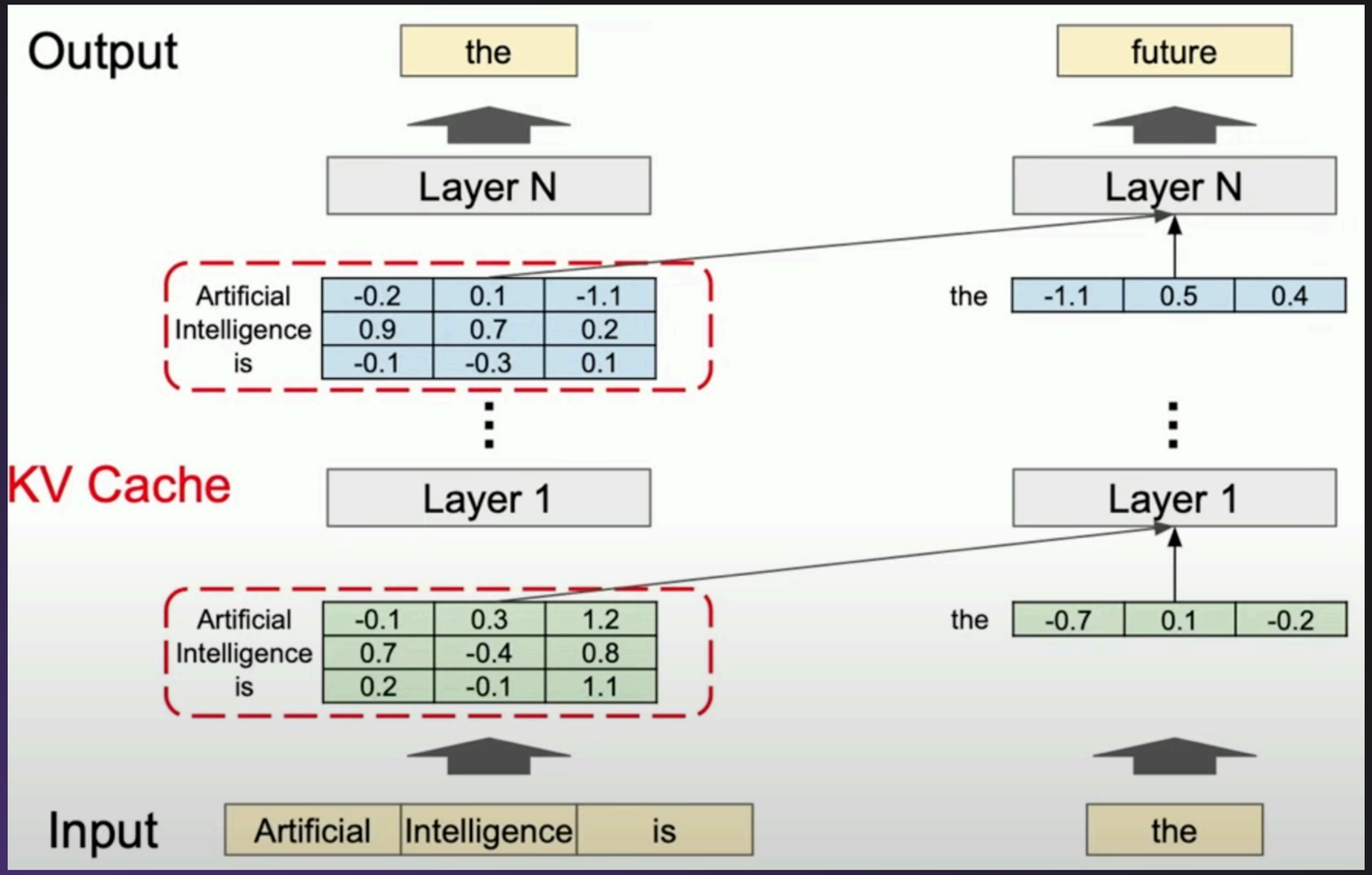
Inference process



Easy, Fast, and Cheap LLM Serving with PagedAttention - Anyscale

xylem.ai

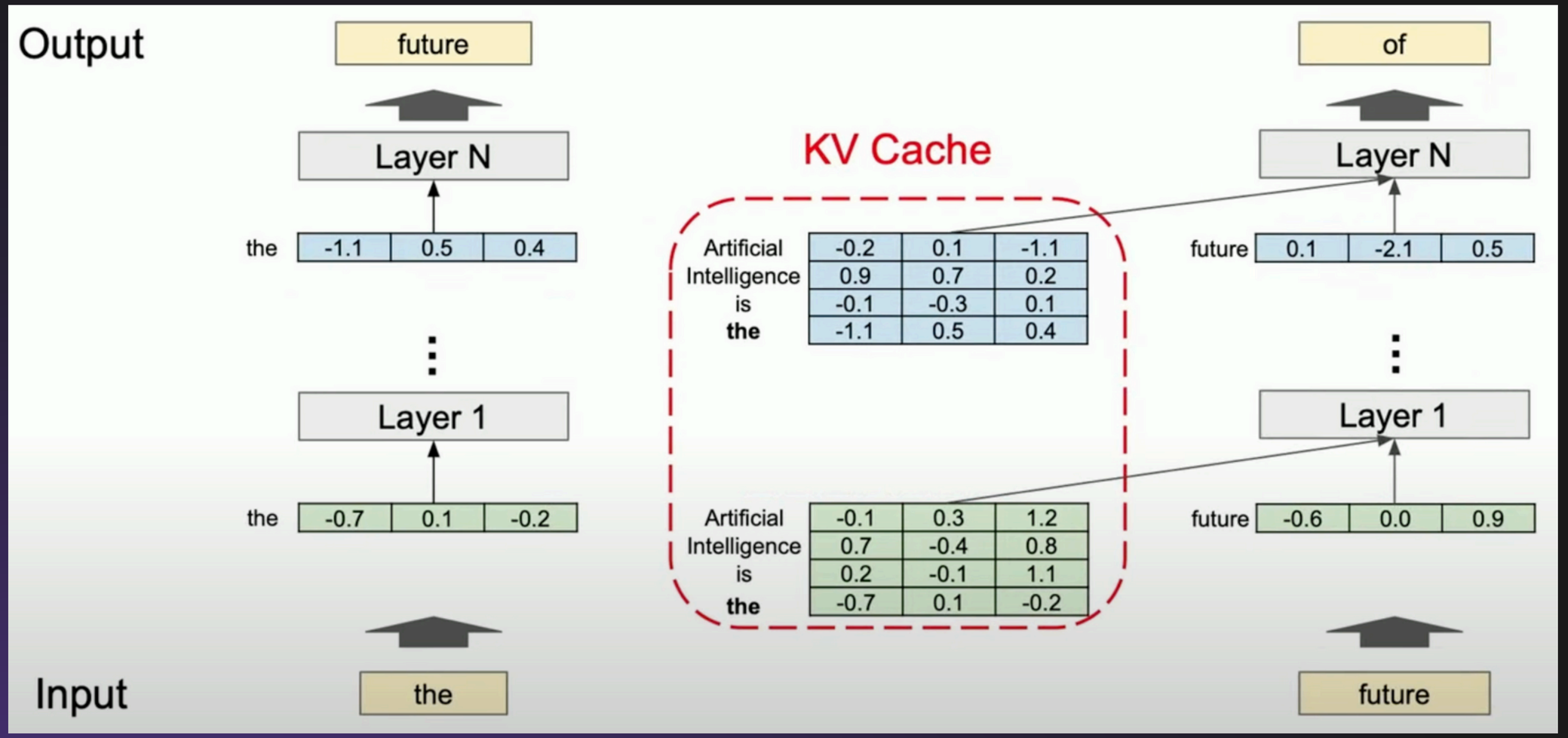
KV Cache



Easy, Fast, and Cheap LLM Serving with PagedAttention - Anyscale

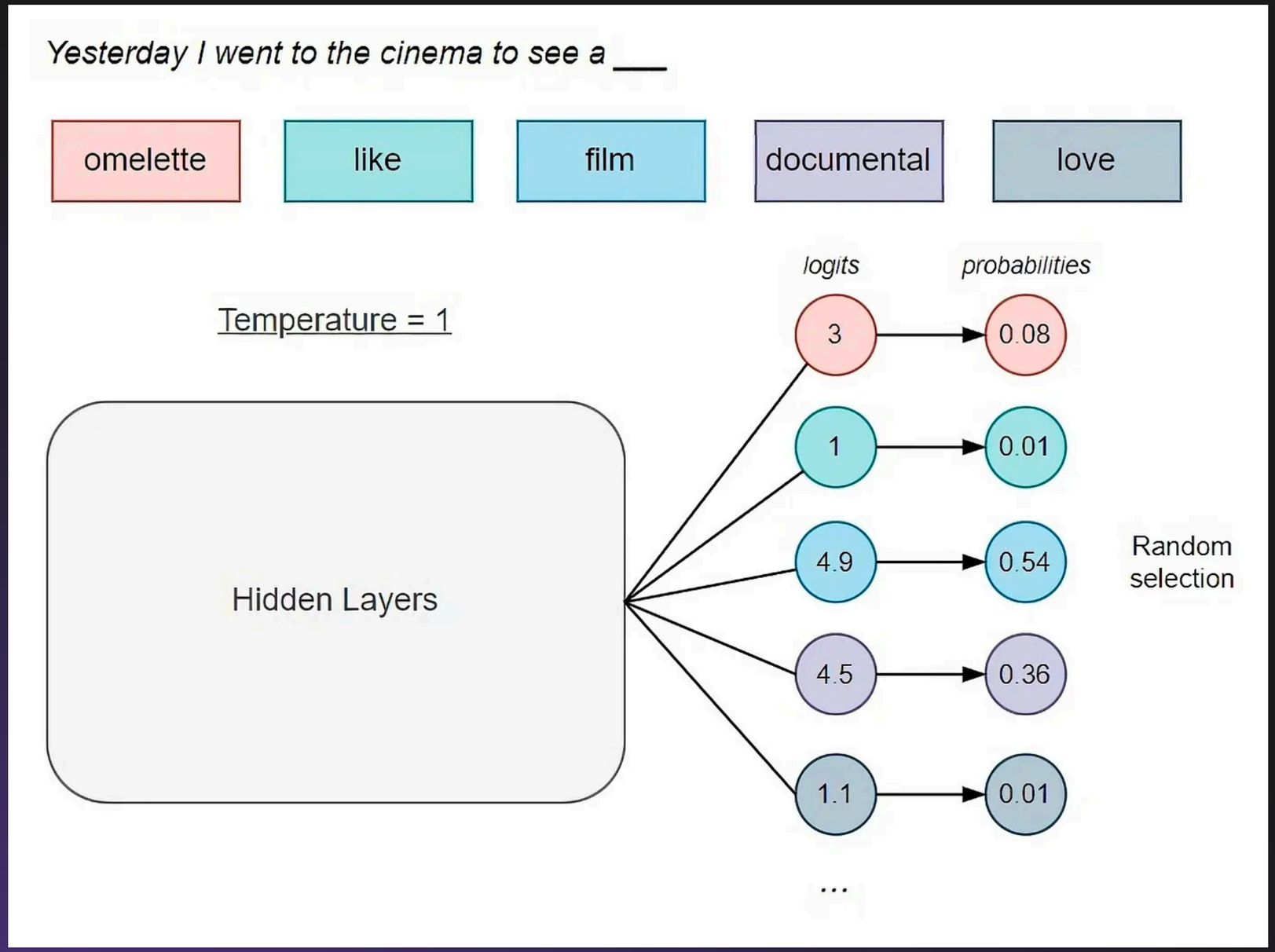
xylem.ai

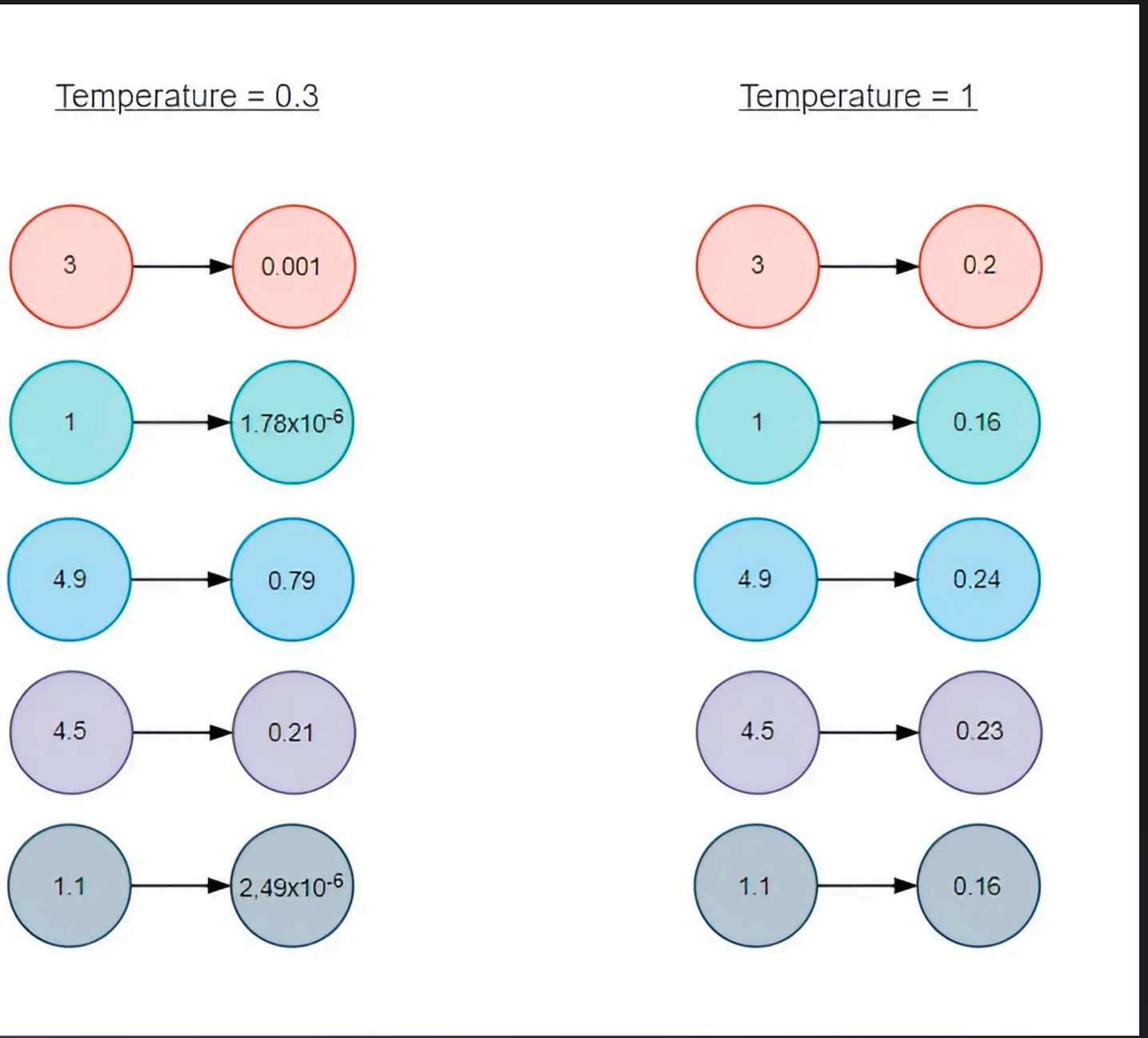
KV Cache

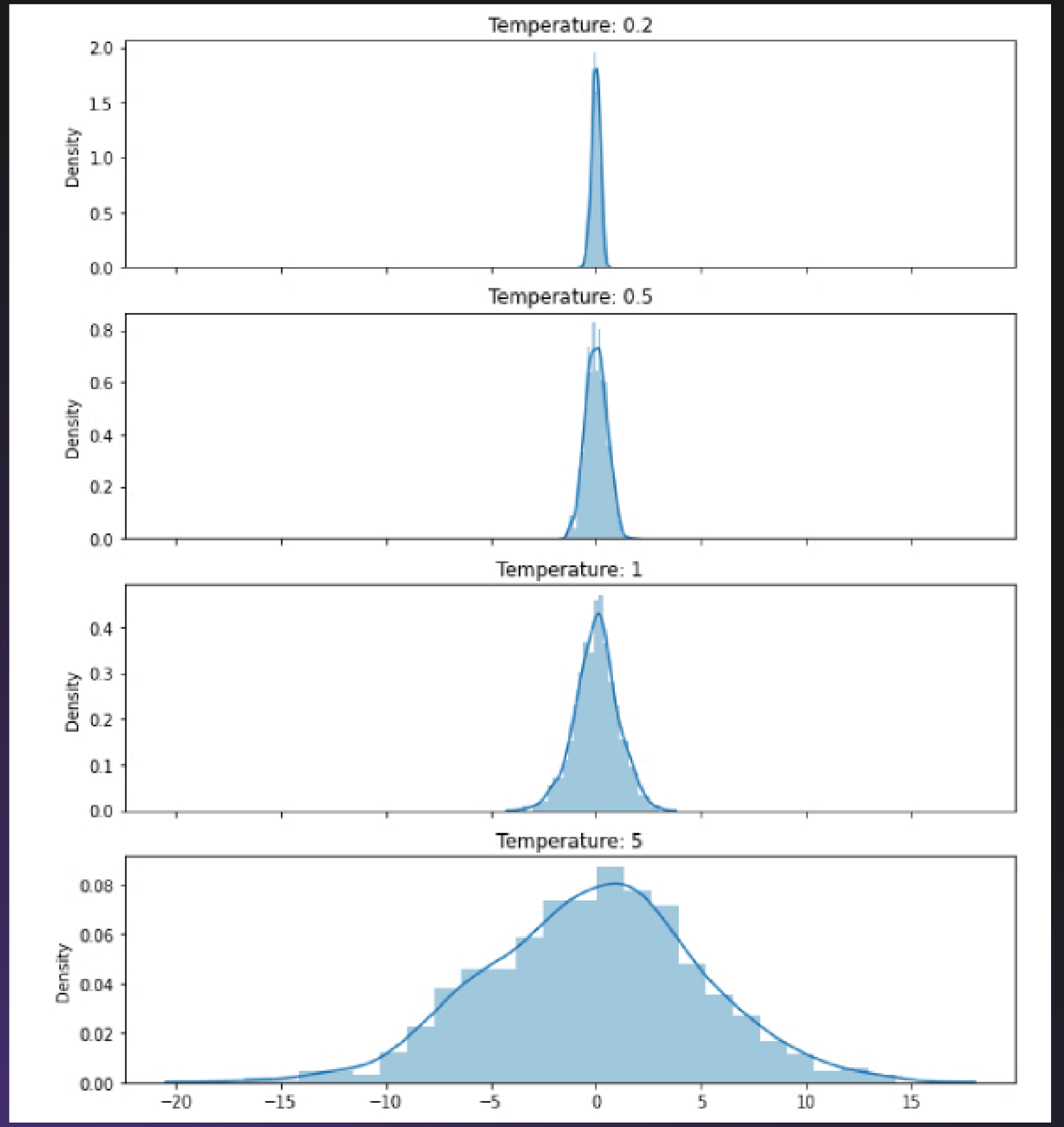


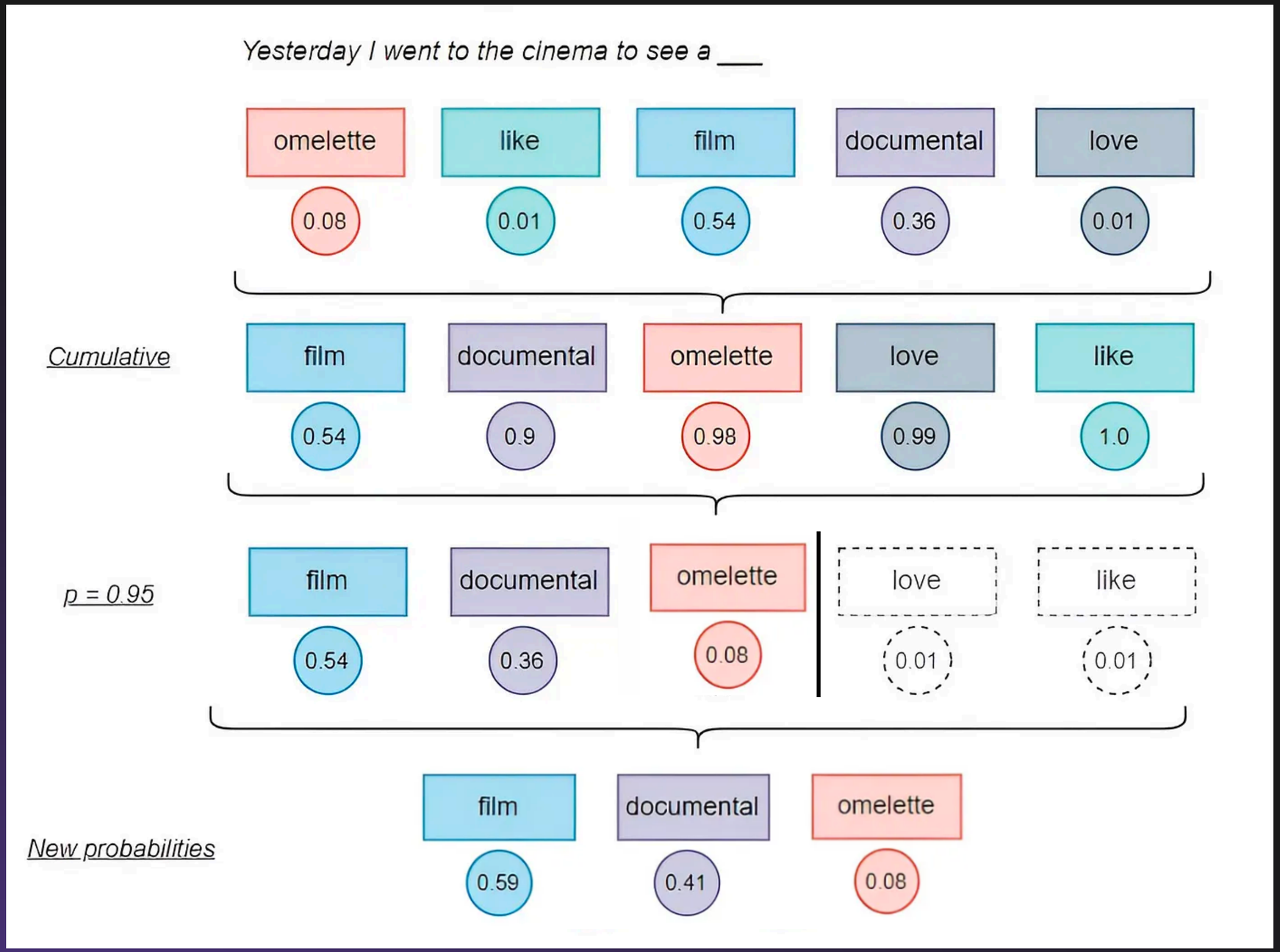
Easy, Fast, and Cheap LLM Serving with PagedAttention - Anyscale

xylem.ai

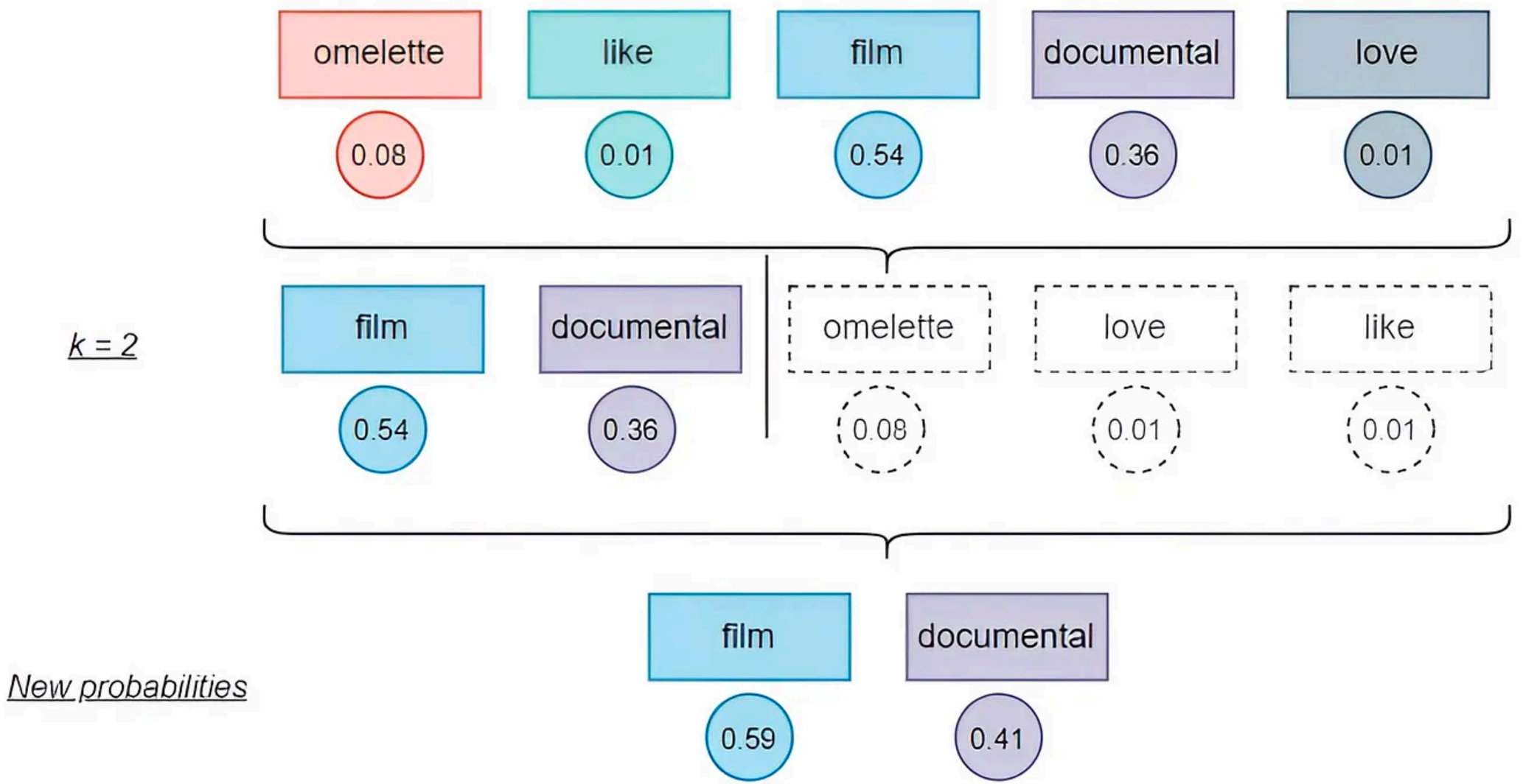

Yesterday I went to the cinema to see a ___
omelette
like
film
documental
love
logits
probabilities
Temperature = 1
3
0.08
1
0.01
4.9
0.54
4.5
0.36
1.1
0.01
Hidden Layers
Random selection
...

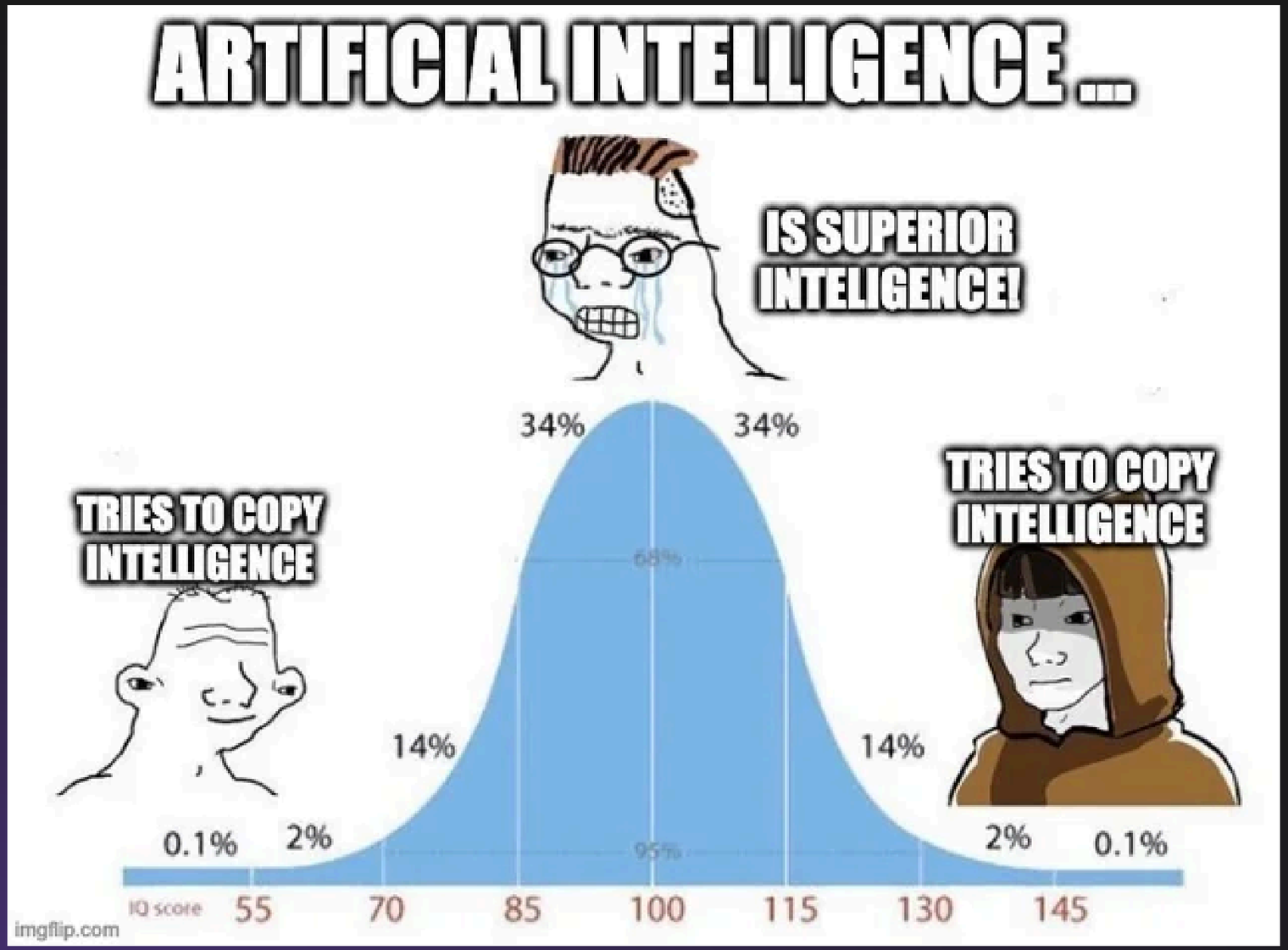

Temperature = 0.3
3
0.001
1
1.78x10-6
4.9
0.79
4.5
0.21
1.1
2,49x10-6
Temperature = 1
3
0.2
1
0.16
4.9
0.24
4.5
0.23
1.1
0.16


Temperature: 0.2
Density
2.0
1.5
1.0
0.5
0.0
Temperature: 0.5
Density
0.8
0.6
0.4
0.2
0.0
Temperature: 1
Density
0.4
0.3
0.2
0.1
0.0
Temperature: 5
Density
0.08
0.06
0.04
0.02
0.00
−20
−15
−10
−5
0
5
10
15


Yesterday I went to the cinema to see a ___
omelette
like
film
documental
love
0.08
0.01
0.54
0.36
0.01
Cumulative
film
documental
omelette
love
like
0.54
0.9
0.98
0.99
1.0
p = 0.95
film
documental
omelette
love
like
0.54
0.36
0.08
0.01
0.01
New probabilities
film
documental
omelette
0.59
0.41
0.08

Yesterday I went to the cinema to see a ___




ARTIFICIAL INTELLIGENCE ...
IS SUPERIOR INTELIGENCE!
TRIES TO COPY INTELLIGENCE
TRIES TO COPY INTELLIGENCE
34%
34%
14%
14%
0.1%
2%
2%
0.1%
68%
95%
IQ score
55
70
85
100
115
130
145
imgflip.com

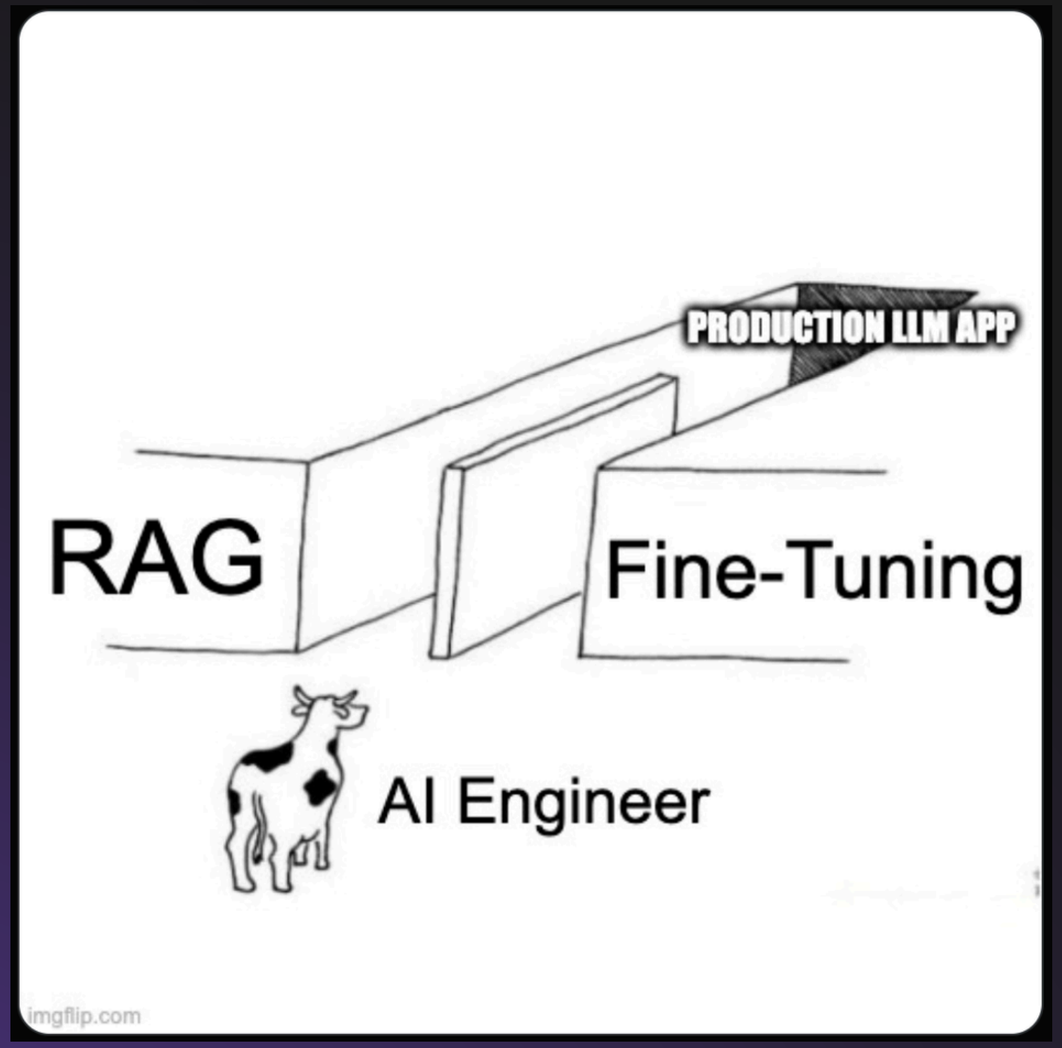

GENERATIVE AI
"Let's see who you really are!"
TRANSFORMER

Finetuning LLMs

Pranav Reddy
Cofounder, Xylem AI
@pranavreddyg (Twitter/X)

xylem.ai

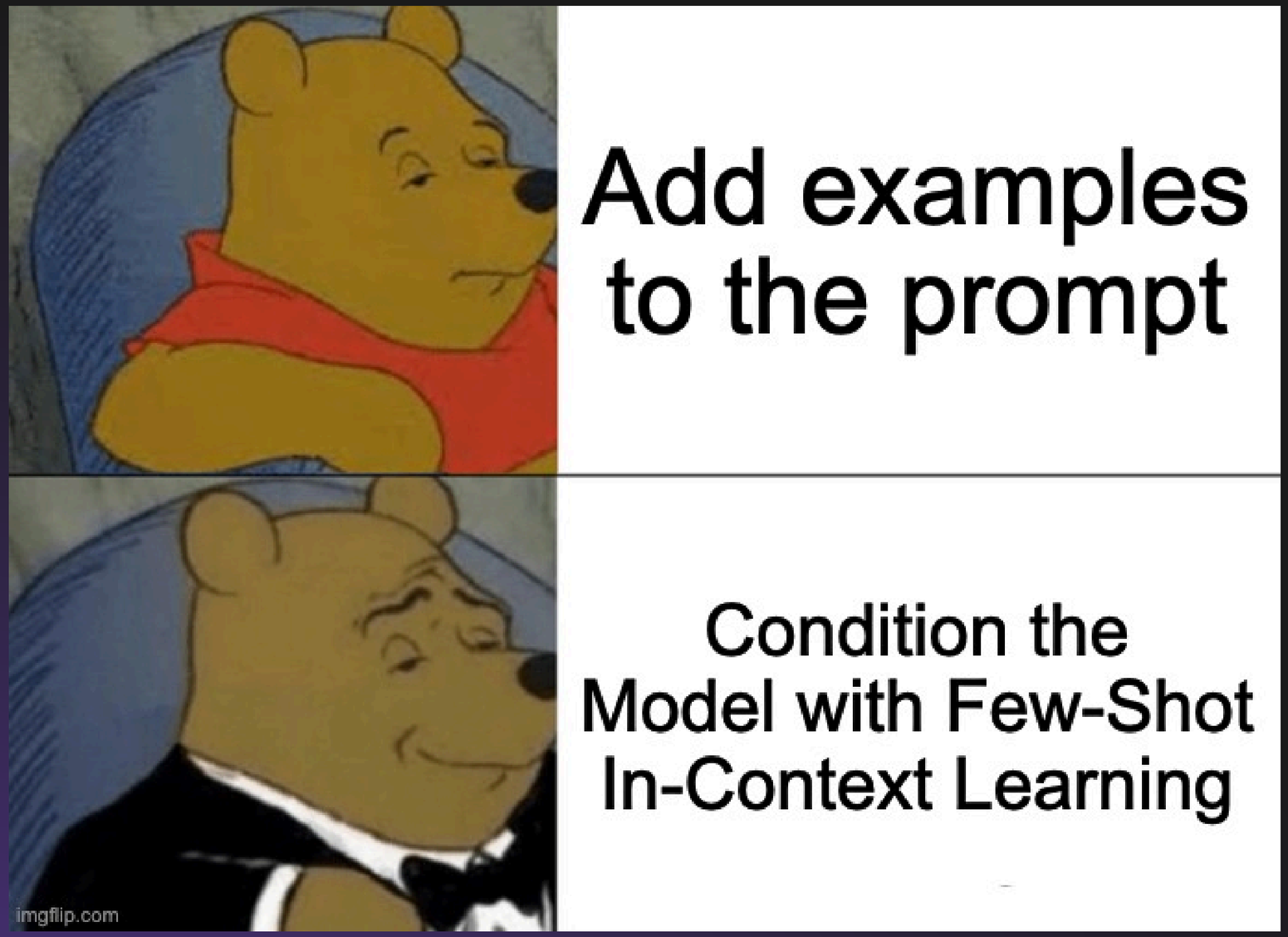
Finetuning/RAG?



xylem.ai

Finetuning/RAG? Asking the right questions

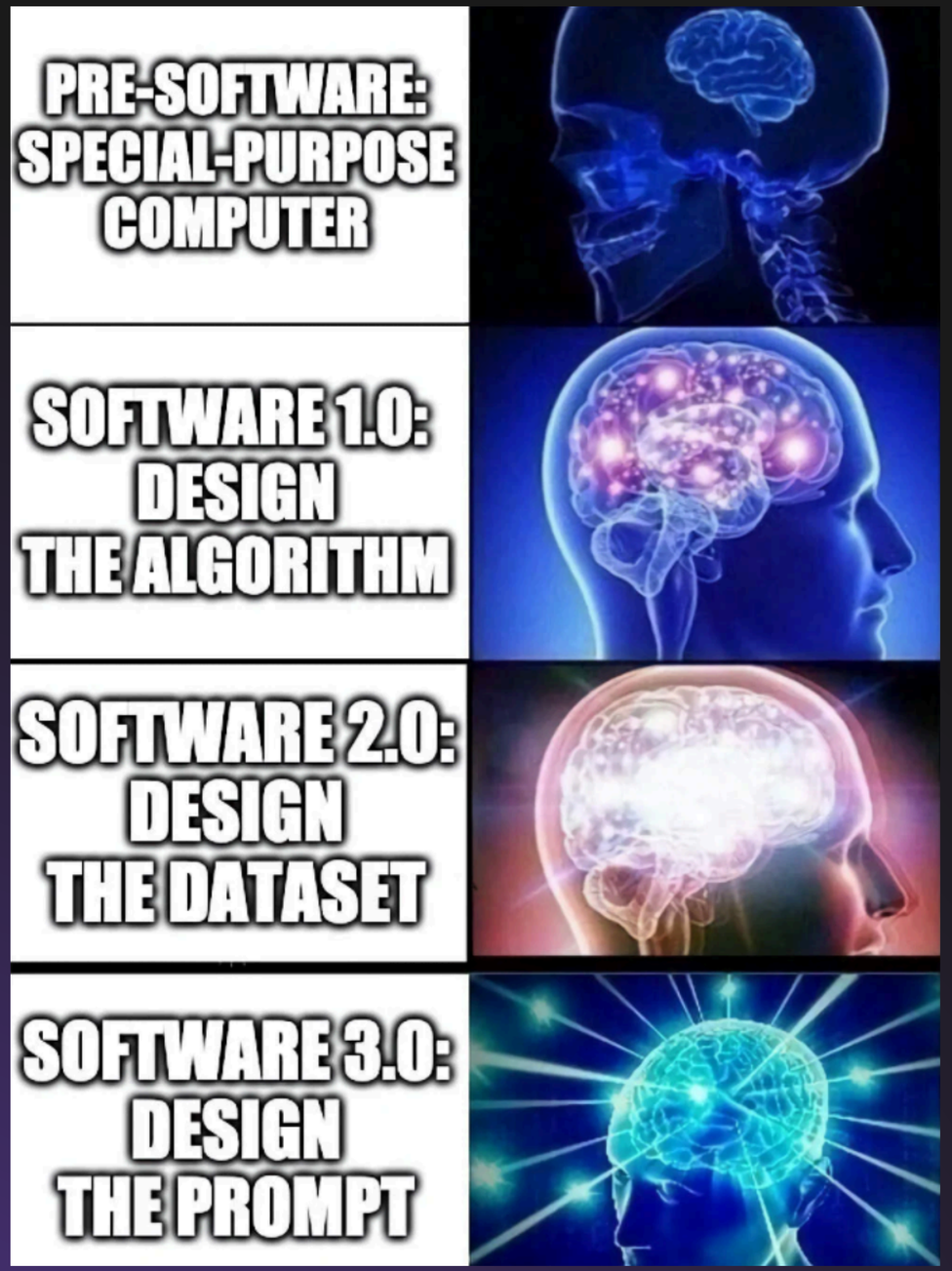
- Does the model lack knowledge of a particular domain?
- Does the context keep changing?
- Do the outputs need to follow any particular format?
- Does prompting solve the issue?
- Does the model not respond with the desired output?
- How often do I need to train?


Add examples to the prompt
Condition the Model with Few-Shot In-Context Learning
imgflip.com

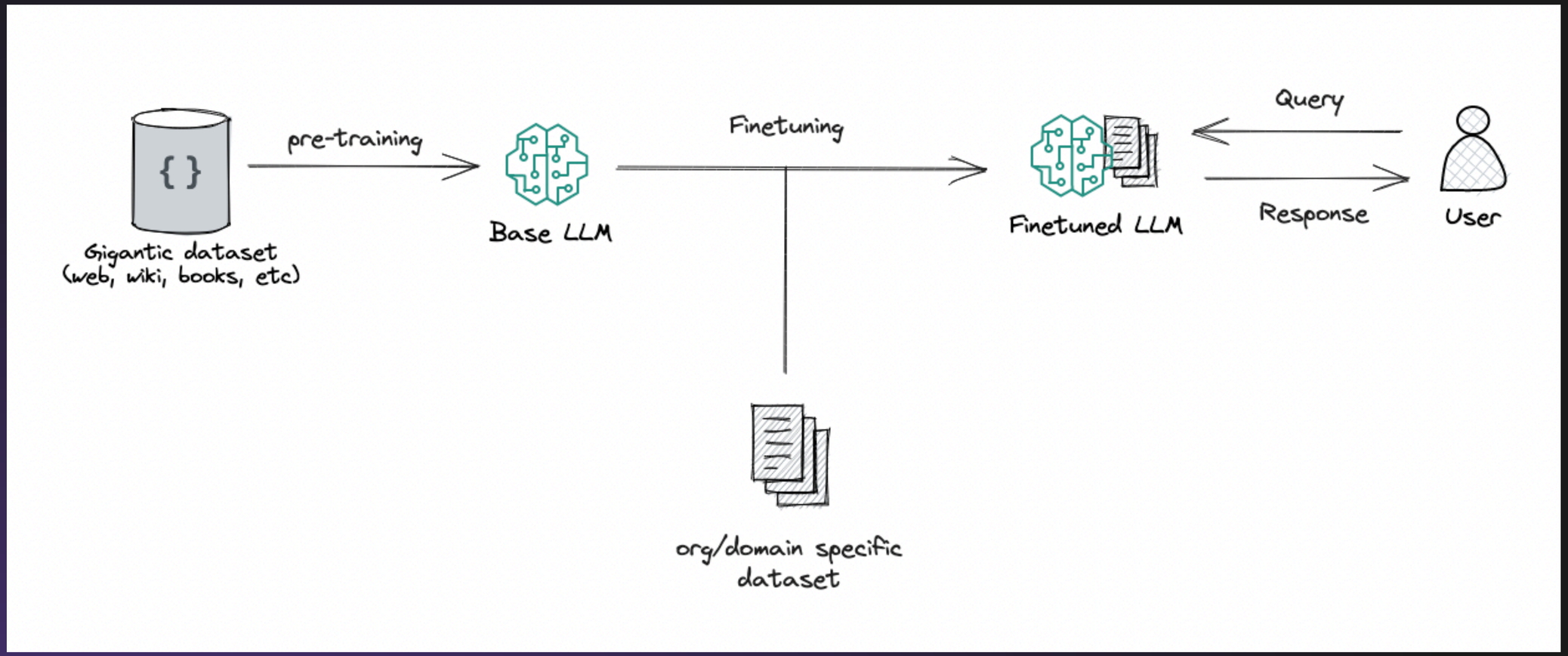
xylem.ai


Asking GPT to be truthful
Fine tuning
Prompt chaining
Pre/post processing user input
All of the above
imgflip.com

xylem.ai


PRE-SOFTWARE:
SPECIAL-PURPOSE
COMPUTER
SOFTWARE 1.0:
DESIGN
THE ALGORITHM
SOFTWARE 2.0:
DESIGN
THE DATASET
SOFTWARE 3.0:
DESIGN
THE PROMPT

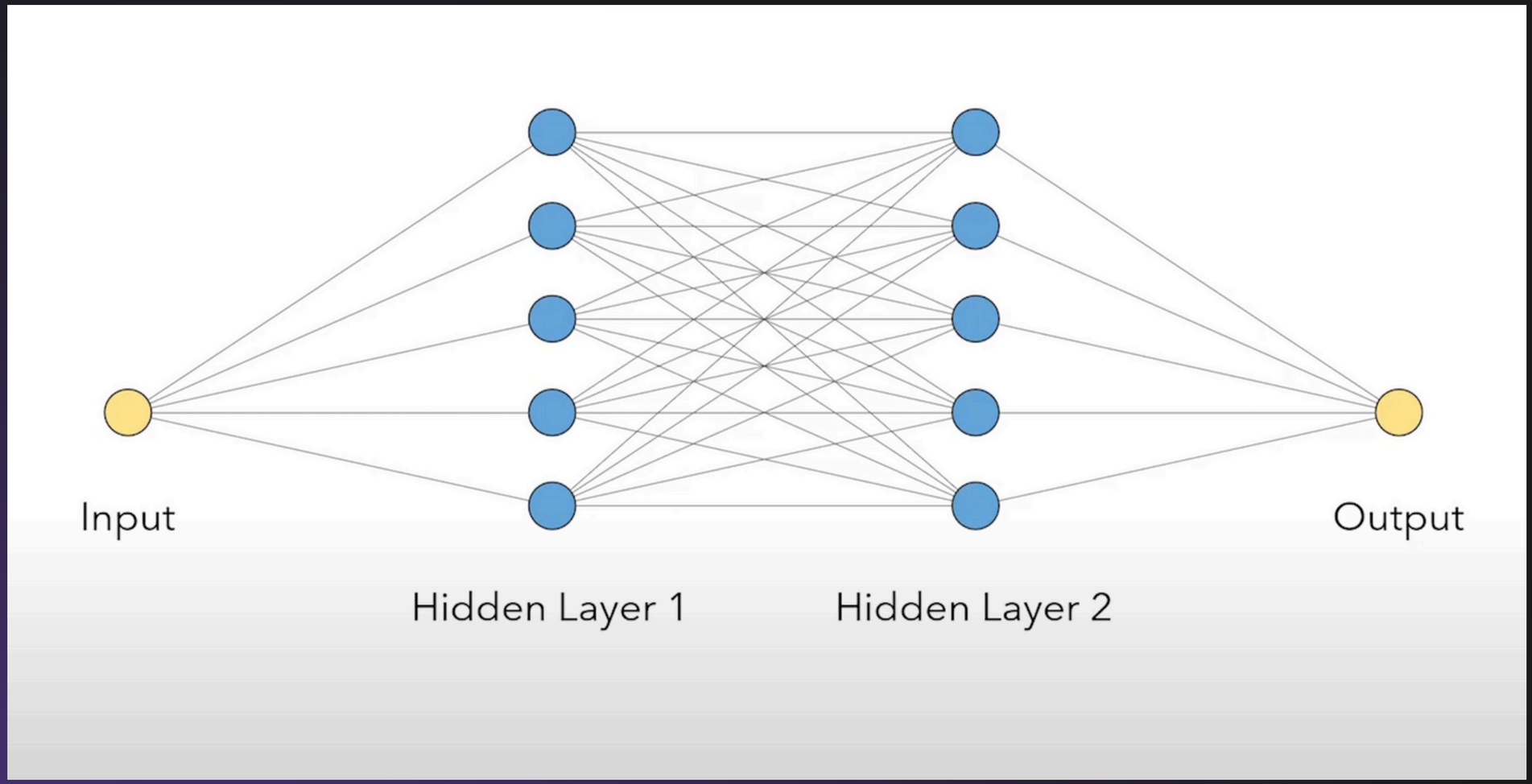
Fine-tuning


{ }
pre-training
Gigantic dataset
(web, wiki, books, etc)
Base LLM
Finetuning
org/domain specific
dataset
Finetuned LLM
Query
Response
User


Prompt engineering
Finetuning LLM
Pretraining LLM
Building foundational LLM from scratch
imgflip.com

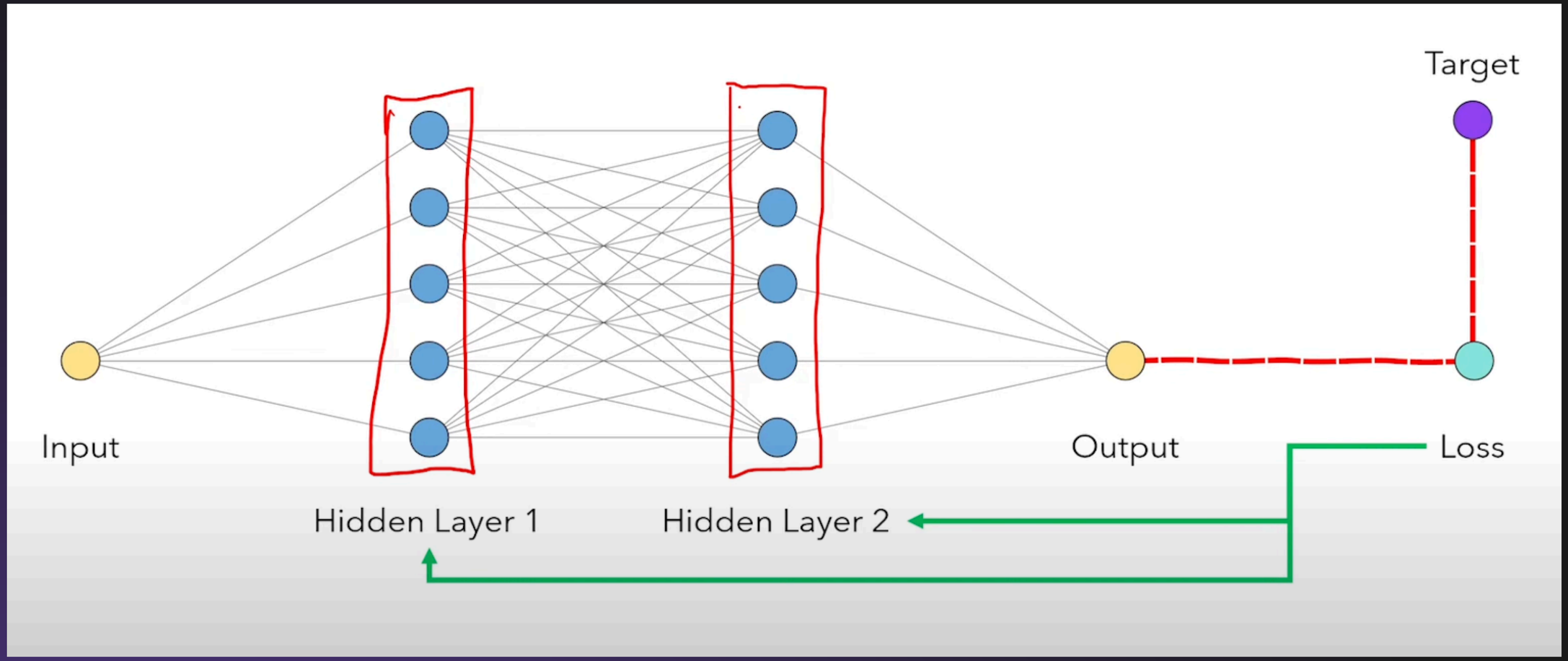
xylem.ai

Neural Networks

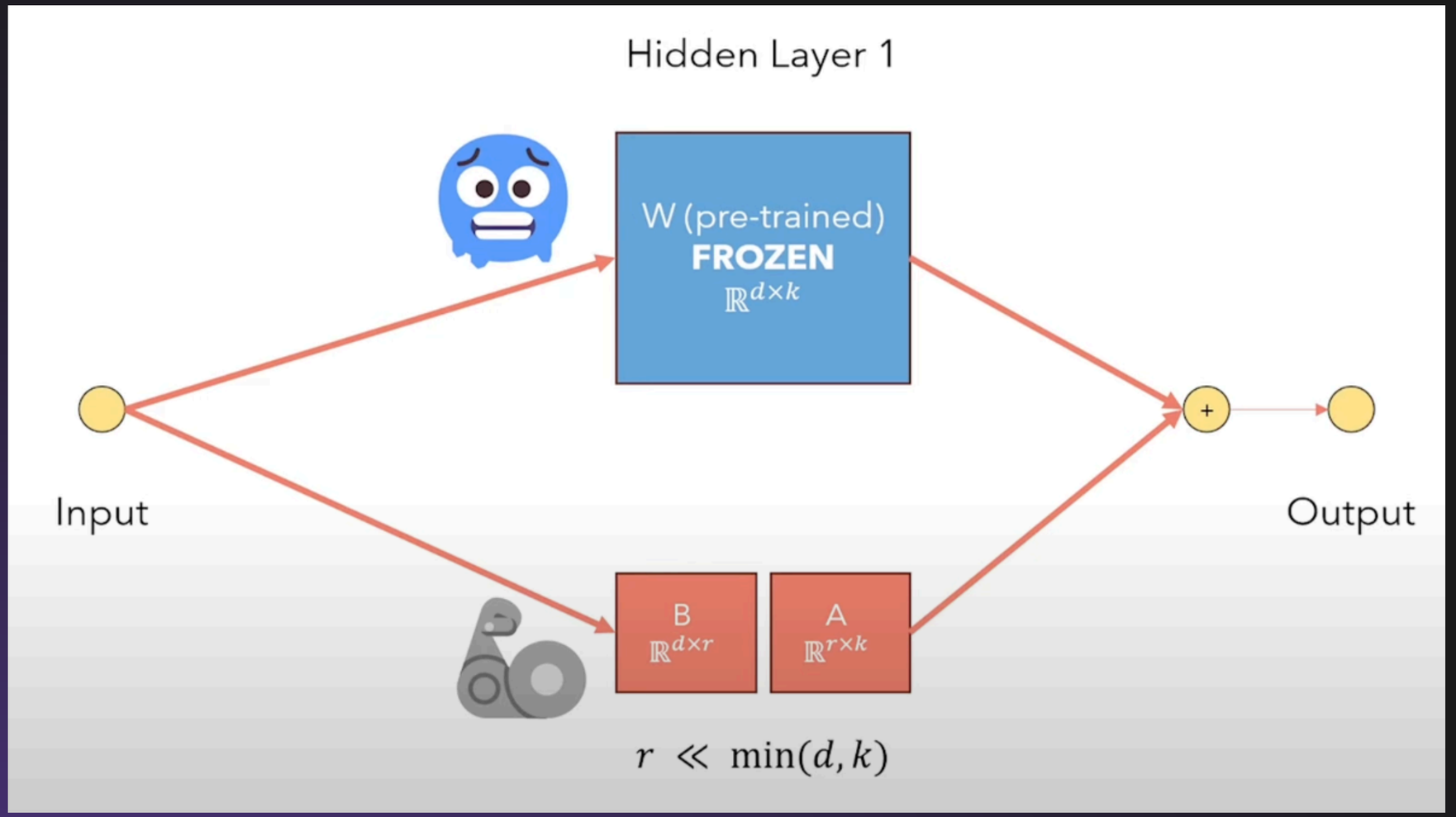


xylem.ai

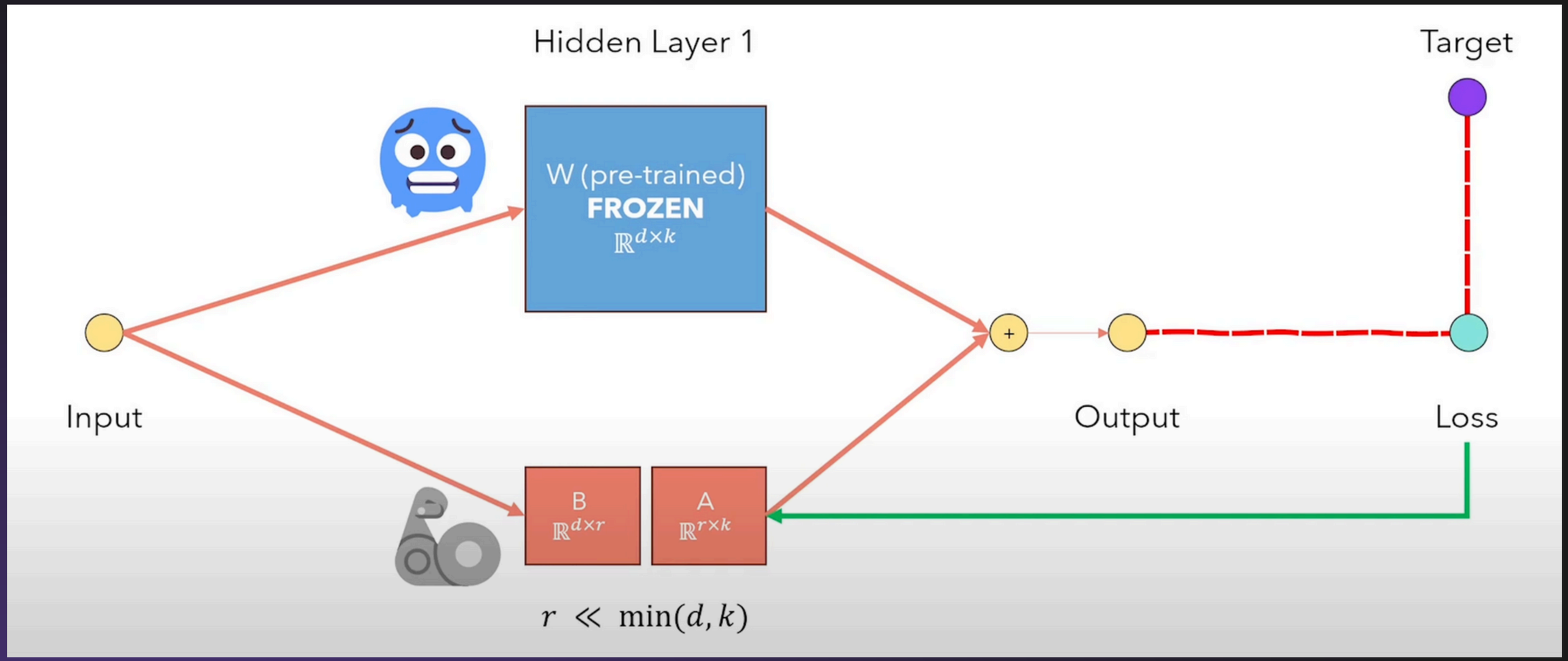
Training Neural Networks

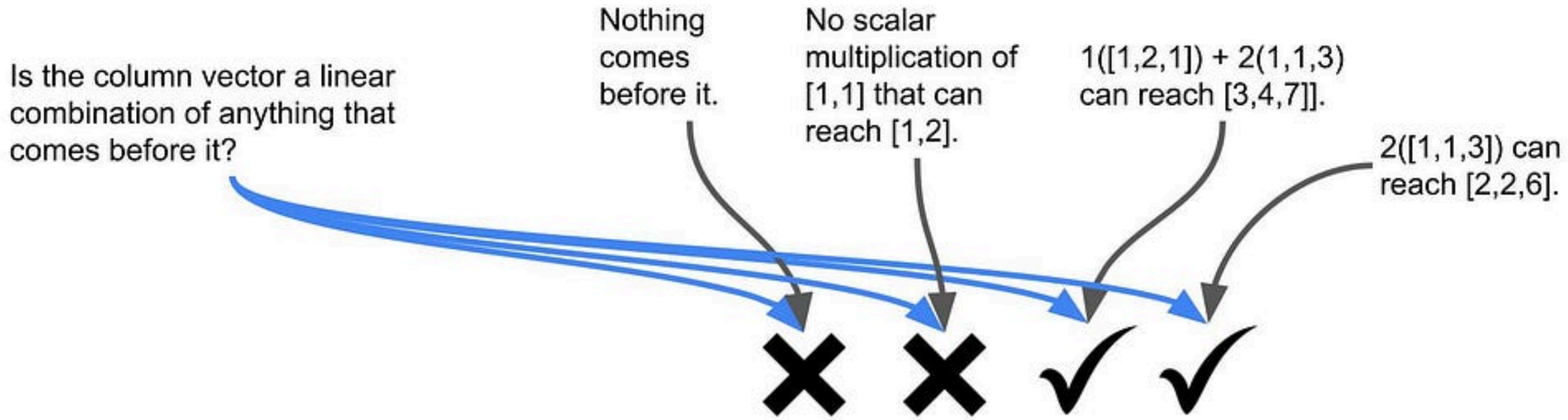


Low Rank Adaptation of LLMs (LoRA)



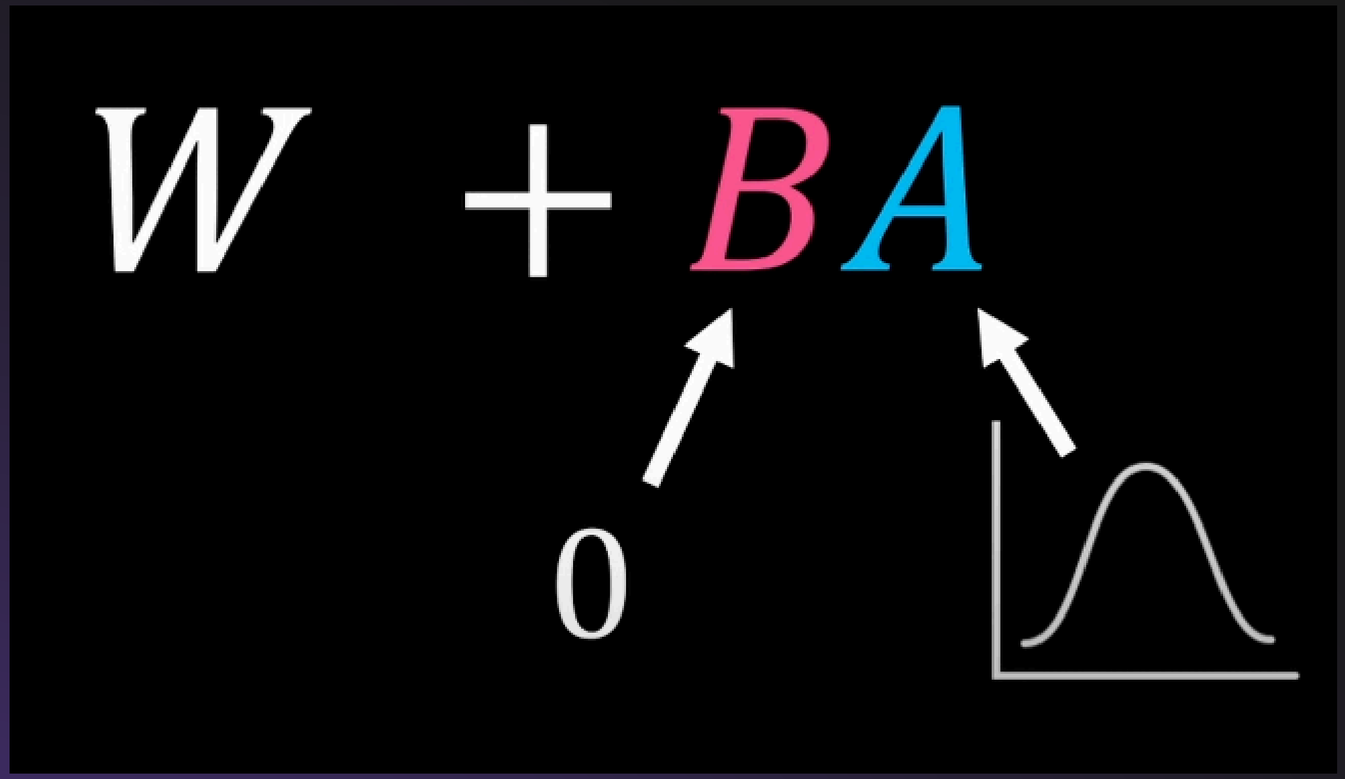
Low Rank Adaptation of LLMs (LoRA)





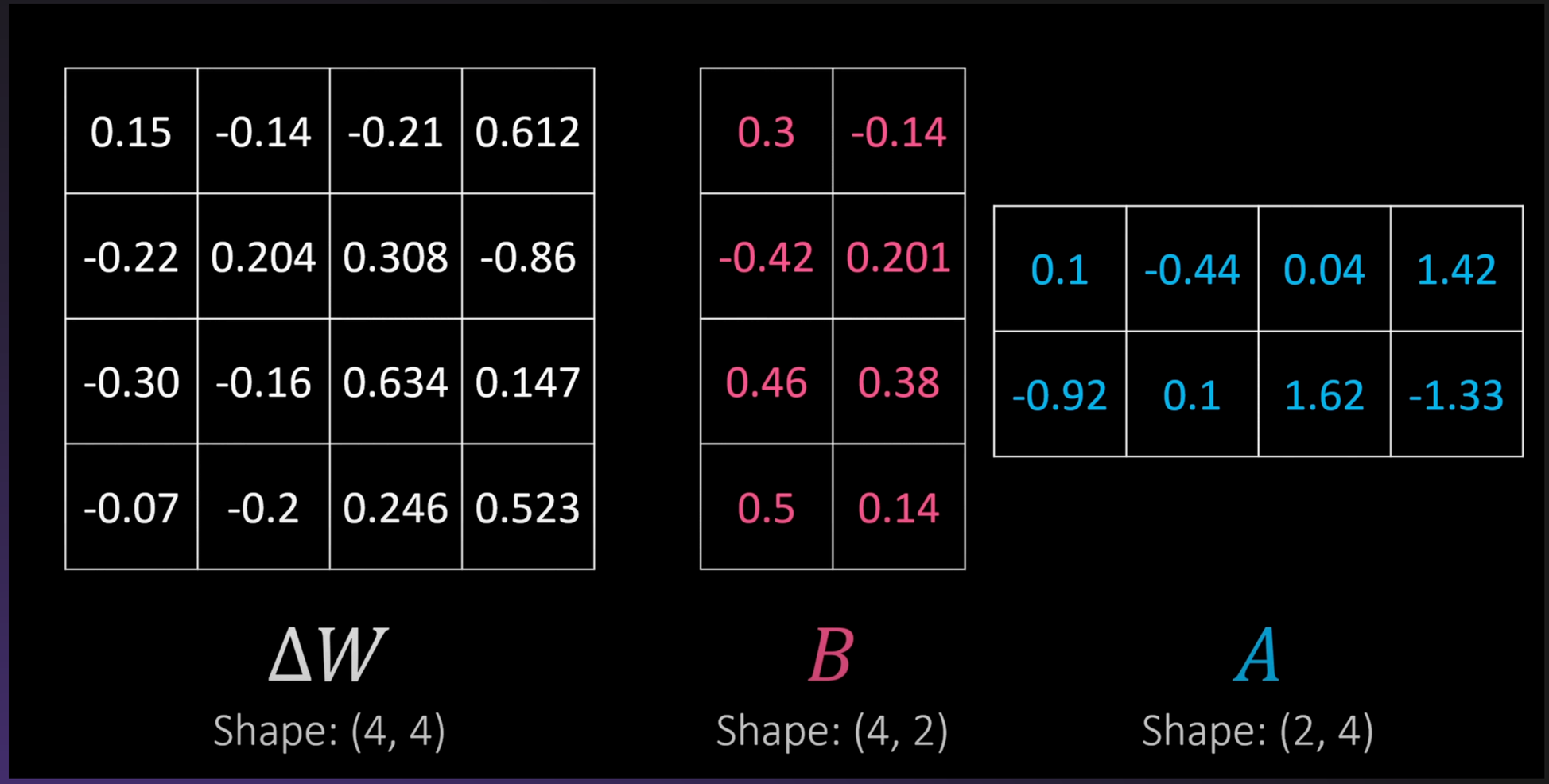
$$\begin{bmatrix} 1 & 1 & 3 & 2 \\ 2 & 1 & 4 & 2 \\ 1 & 3 & 7 & 6 \end{bmatrix} = rank\ 2$$

Rank decomposition

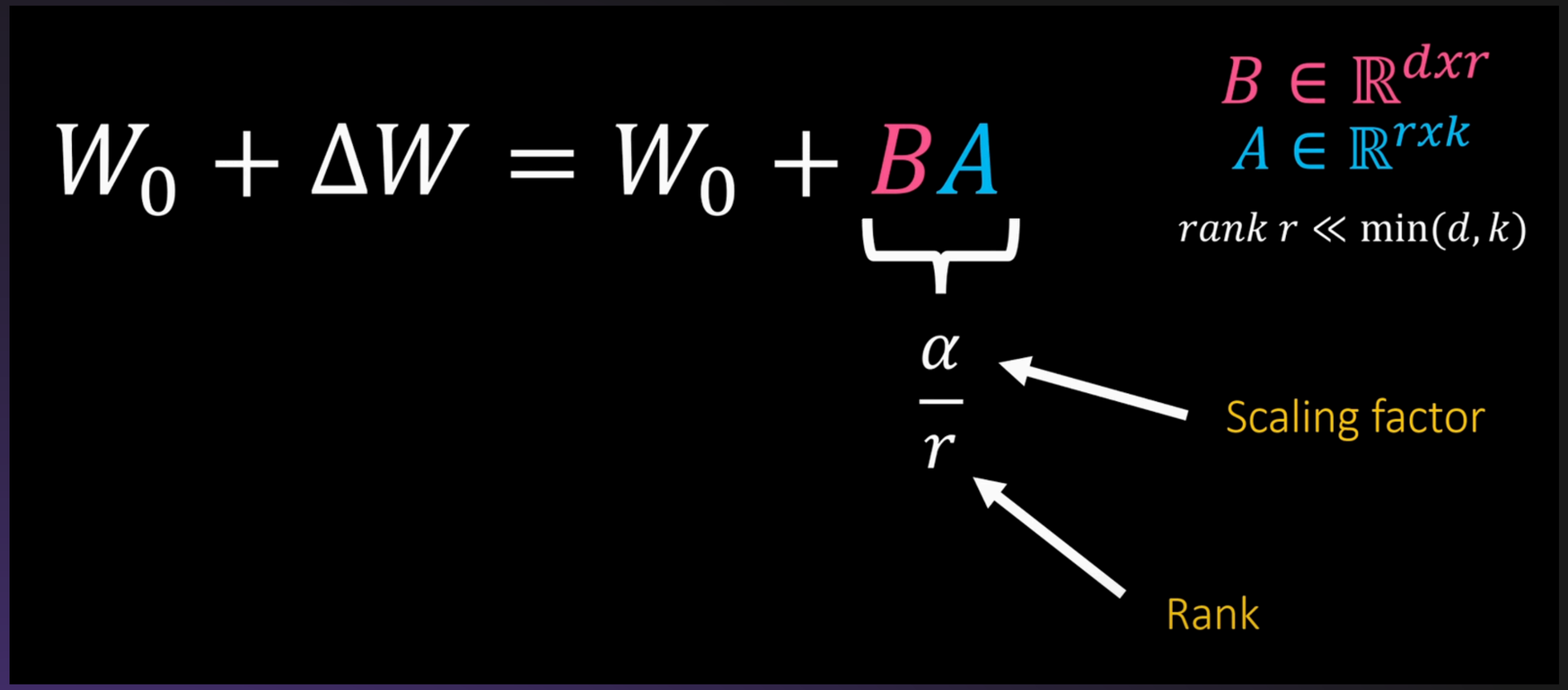


xylem.ai

Rank decomposition



Scaling factor



Other parameters

- **bnb_4bit_quant_type -** This sets the quantization data type in the bnb.nn.Linear4Bit layers. Options are FP4 and NF4 data types which are specified by fp4 or nf4.
- **bnb_4bit_compute_dtype -** This sets the computational type which might be different than the input time. For example, inputs might be fp32, but computation can be set to bf16 for speedups.
- **lora_alpha -** scaling factor for the weight matrices. alpha is a scaling factor that adjusts the magnitude of the combined result (base model output + low-rank adaptation). We have set it to 16.

Other parameters

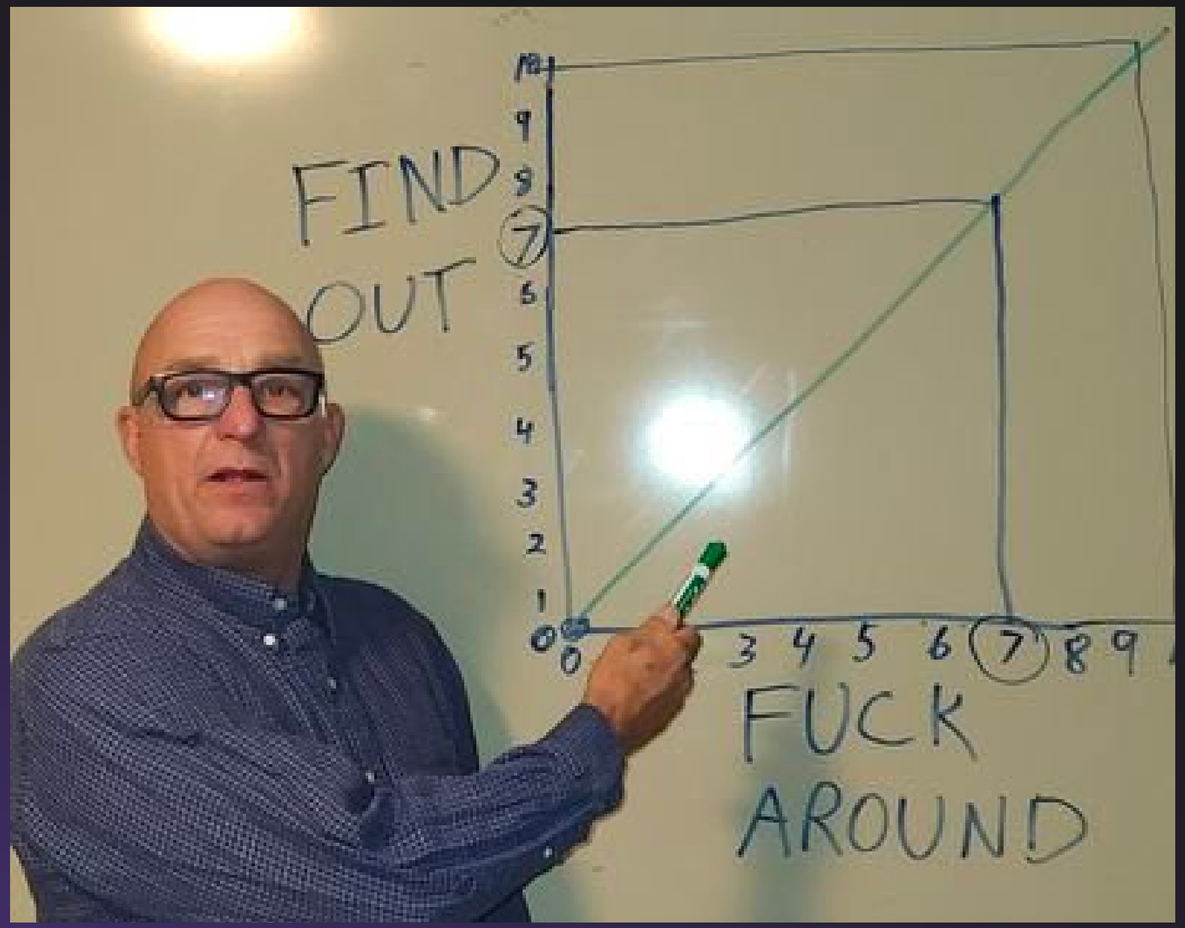
- **lora_dropout -** dropout probability of the LoRA layers. This parameter is used to avoid overfitting. This technique drop-outs some of the neurons during both forward and backward propagation, this will help in removing dependency on a single unit of neurons. We are setting this to 0.1 (which is 10%), which means each neuron has a dropout chance of 10%.
- **r -** This is the dimension of the low-rank matrix
- **bias -** We will not be training the bias in this example, so we are setting that to “none”. If we have to train the biases, we can set this to “all”, or if we want to train only the LORA biases then we can use “lora_only”


I just spent our entire AI budget on fine-tuning a model
But at least we have a working model now, right?
We have a working model, right?

The world if choosing the right hyperparameters was straightforward

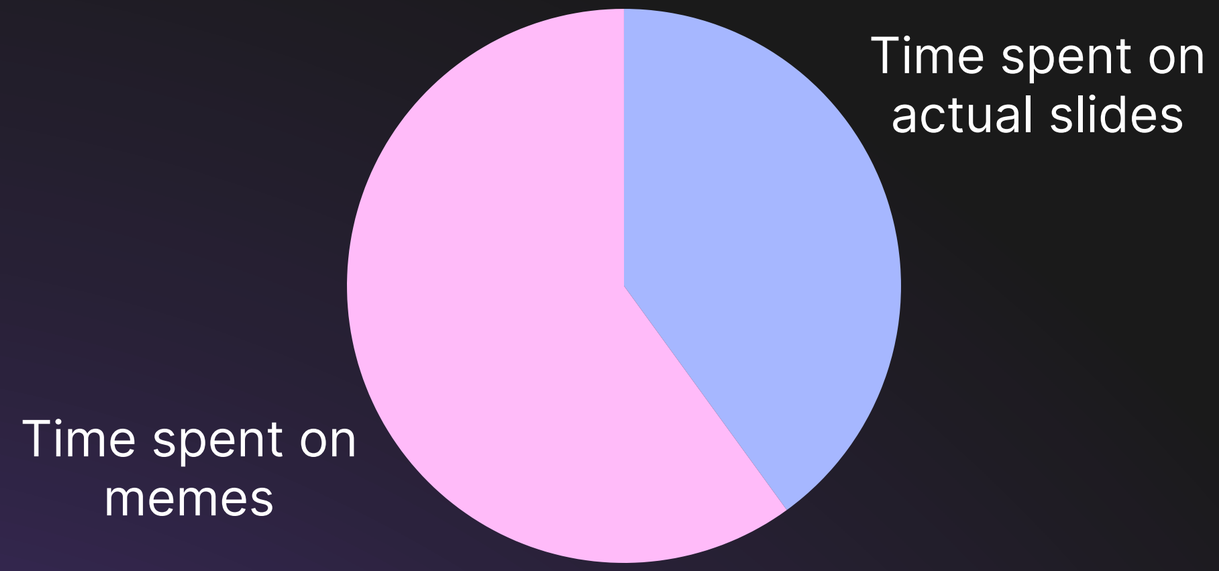
xylem.ai

How to find the best hyperparams?



xylem.ai

Thank you



Pranav Reddy
Cofounder, Xylem AI
@pranavreddyg (Twitter/X)

xylem.ai