Generative AI
Fine-Tuning Tutorial

Custom LLMs for fun and profit

charlesmartin14@gmail.com
+1 (415) 298-6783

Fine-Tuning LLMs : Custom 'ChatGPTs'

Specific (private) Knowledge

Gigantic web-scale dataset

pre-training

Base LLM

Falcon : open-source ChatGPT

Supervised fine-tuning

Fine-tuned LLM

Fine-Tuning LLMs : 'Which Model ?'

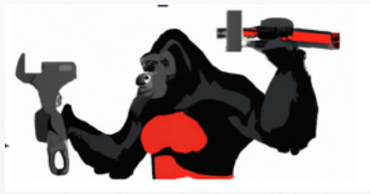
Falcon

FLAN-T5

Llama2

Meditron

Mistral

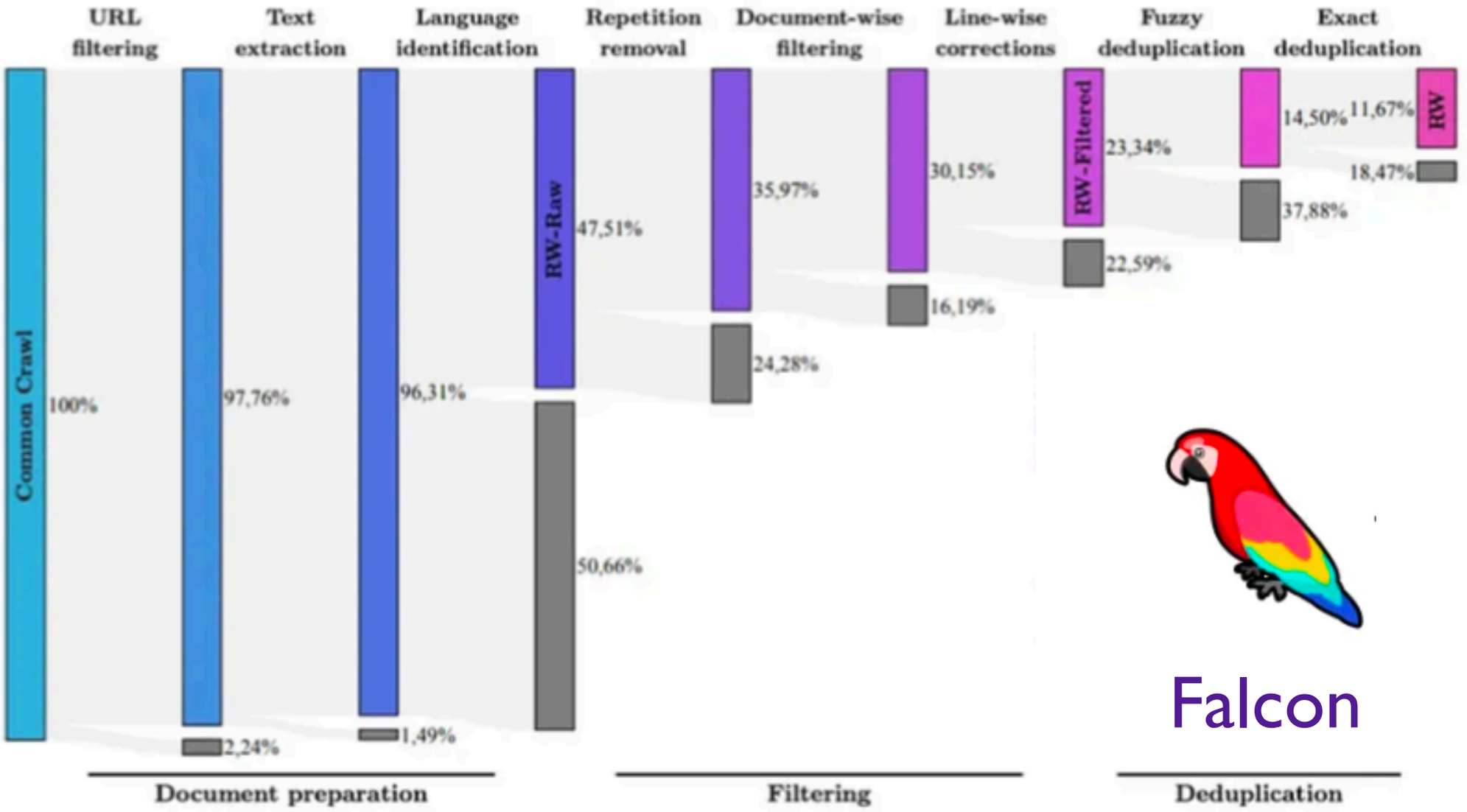


Gorilla

BERT

ALBERT

Fine-Tuning LLMs : Data Prep



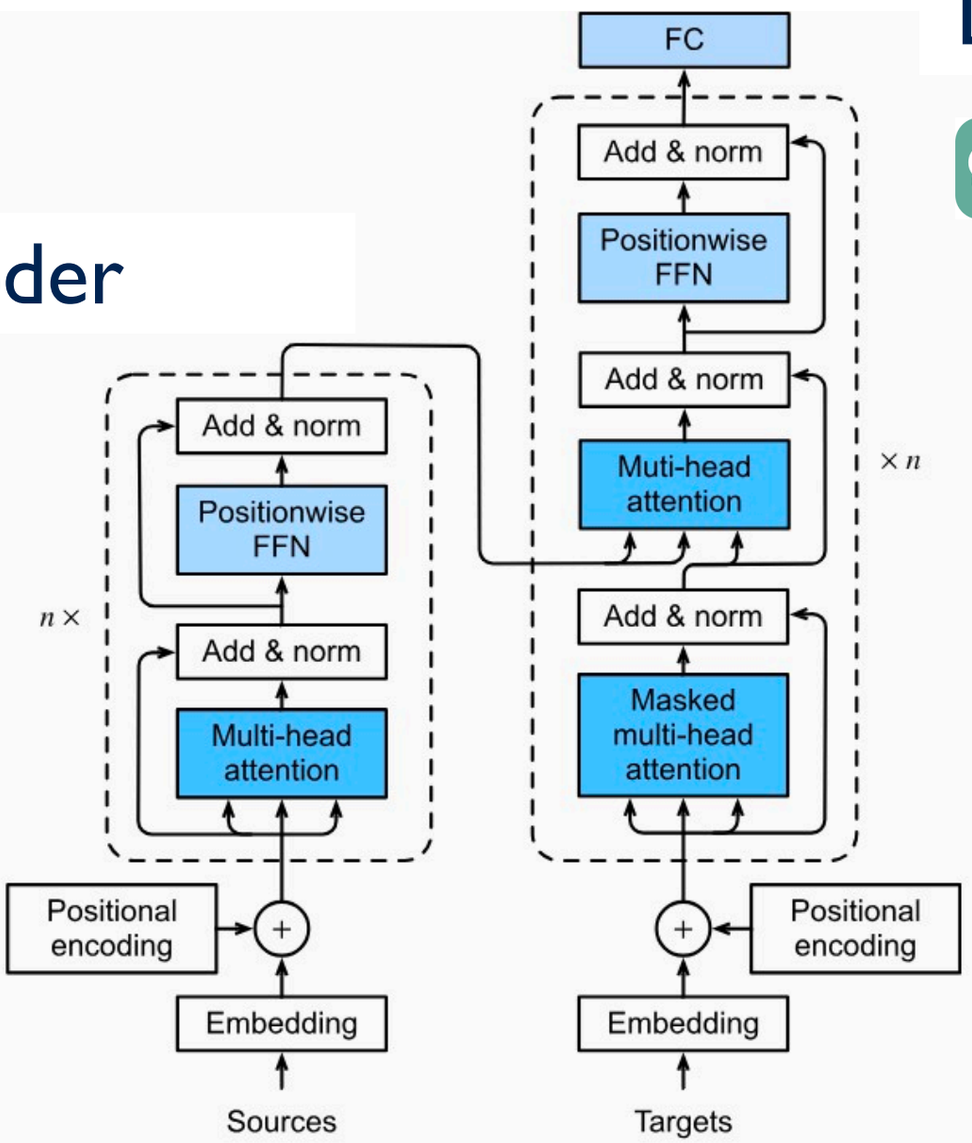
Falcon

LLM Internals : Transformer Architecture

Decoder

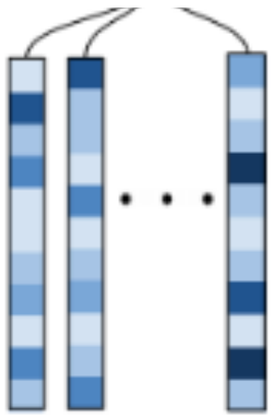
ChatGPT

Encoder

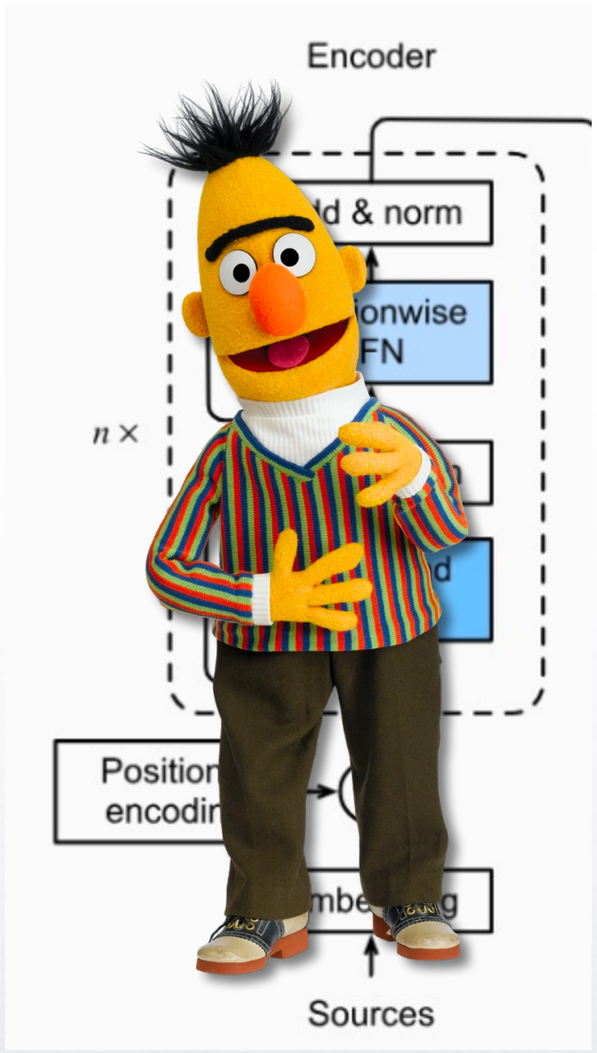


BERT : Encoder Only

Text Data



Vectors



Classify Text

Supervised: Predict Label

GPT : Decoder Only

Input Text

Jack went to the university

Vectors

ChatGPT

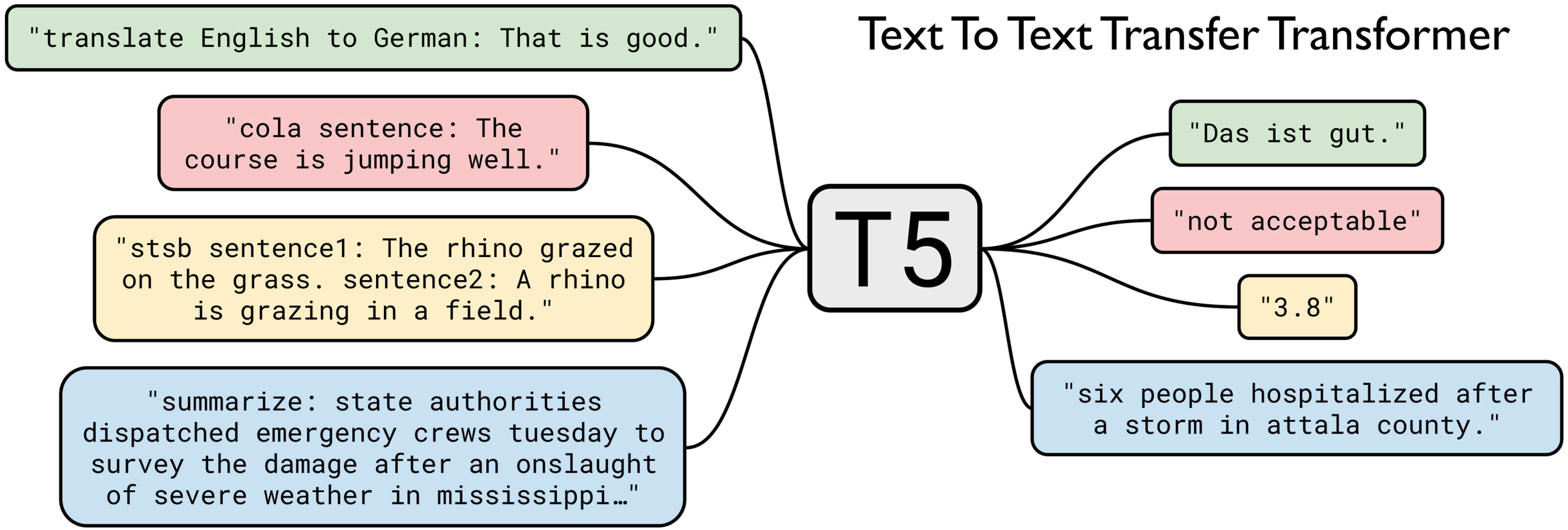
Generate Text

Jack went to the university in the early 1970s as a graduate student and started the company that would become McDonald's. Though he was a long time away from his family, he was an avid cook. "My mom is a huge foodie," he says. "She likes to know things about what I'm eating." Her recipes helped him develop an idea of how to go about the food business and eventually made him a fast-food millionaire in 1993. "I made sure that my family was always watching," he says. "And I always liked to tell them about everything I was doing on a regular basis."
That early obsession with recipes led to his own idea of what a McDonald's menu would look like: a menu that would make people say hello to their food, even though they've been away.

Model output

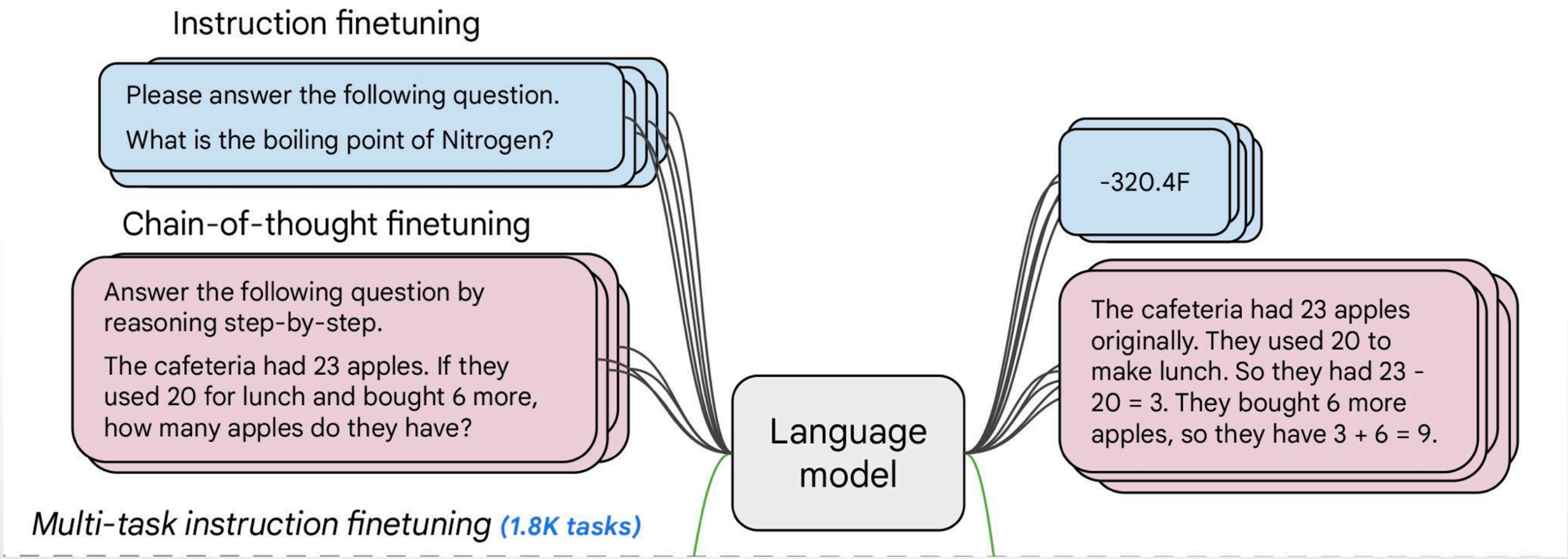
Self-Supervised: Predict Next Word

T5/Flan-T5 : Encoder-Decoder



Every problem is "Text-to-Text": translate, summarize, …

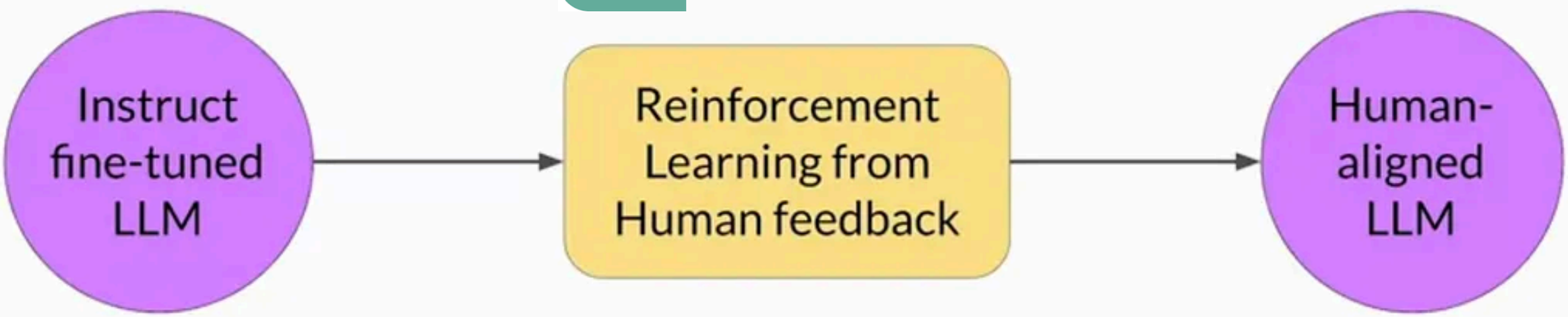
Flan-T5 : Fine-tuned Language Net - T5



Instruction Fine-tuning: predict token sequences

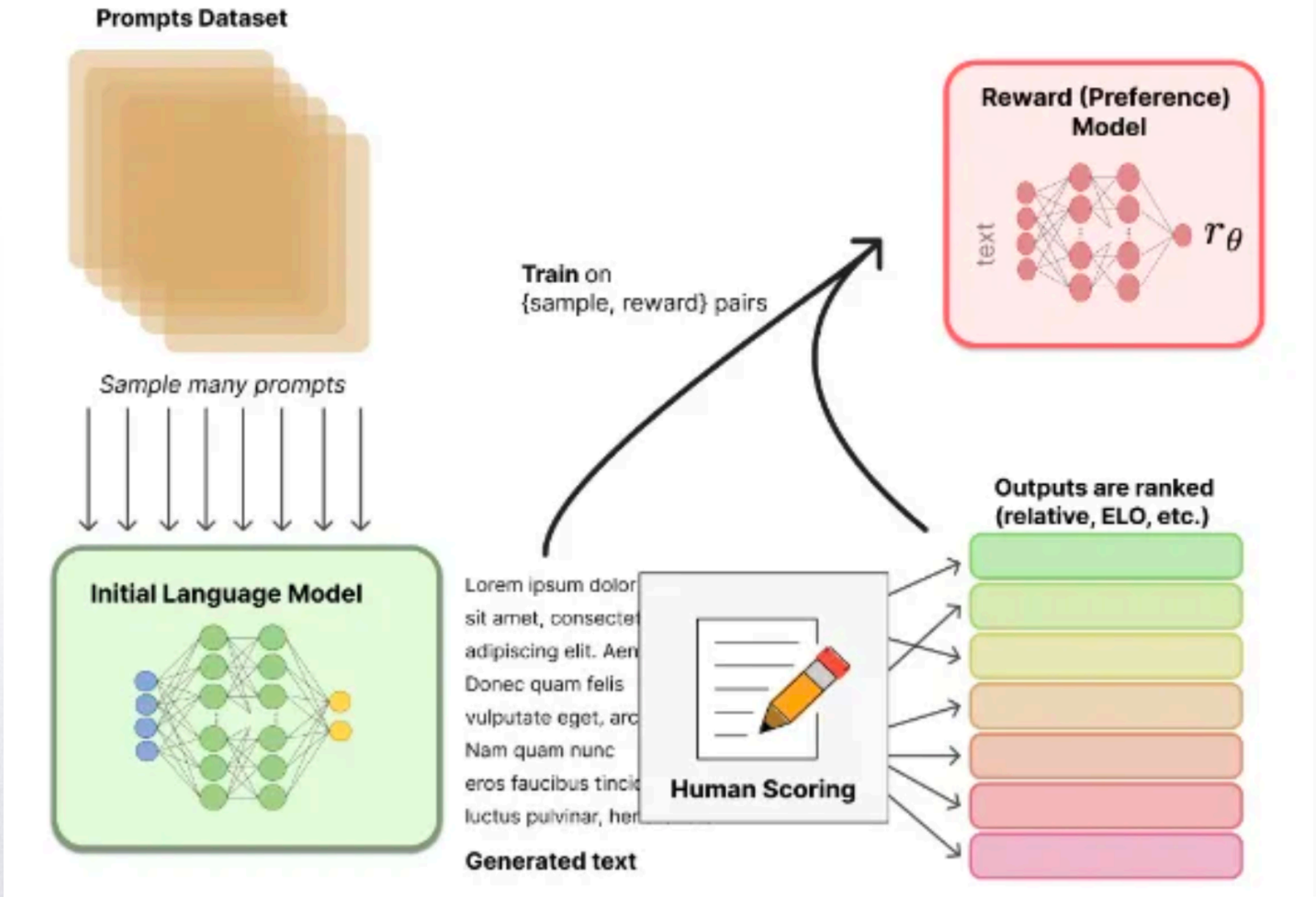
Reinforcement learning from human feedback (RLHF)

ChatGPT

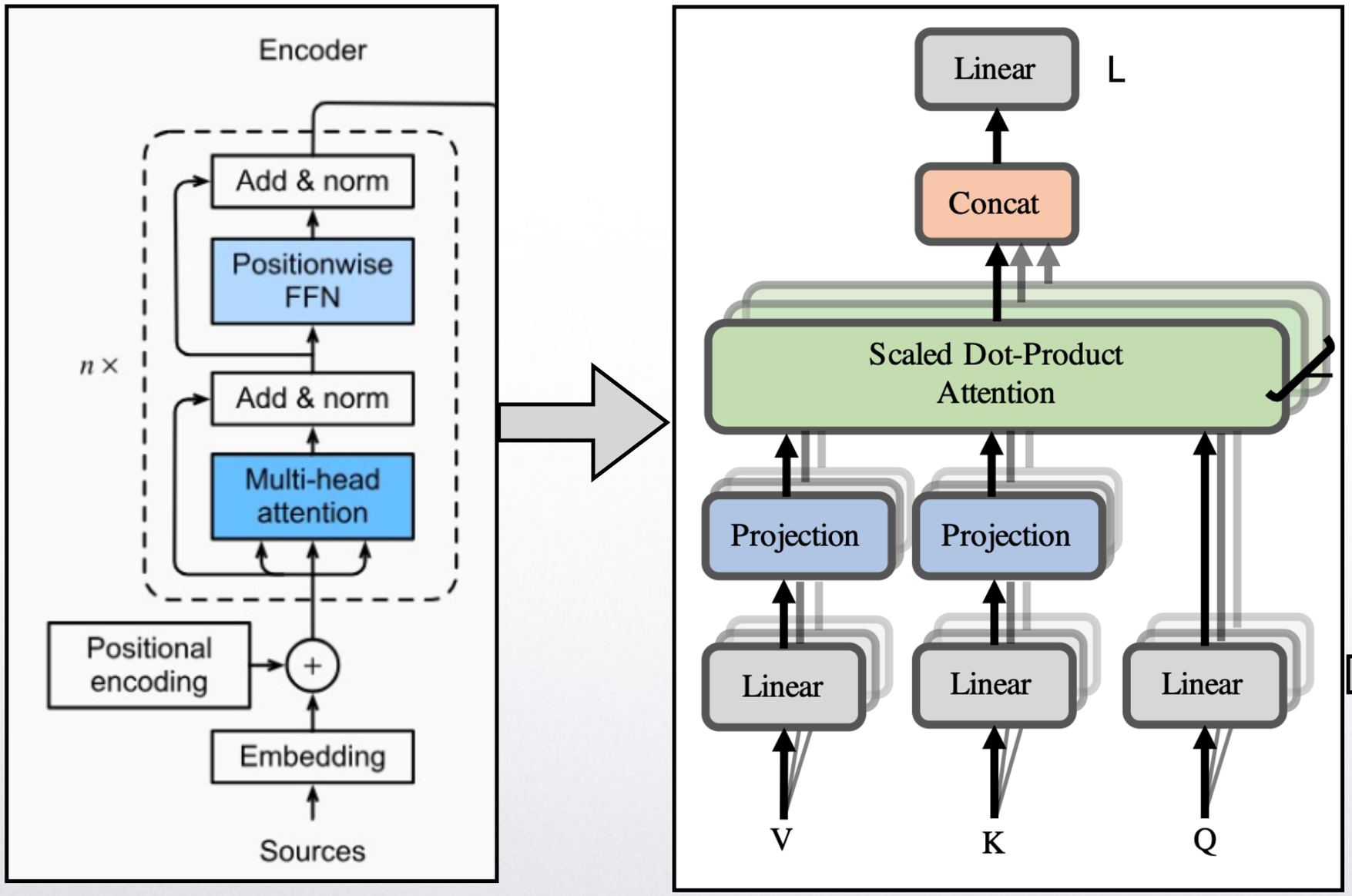


- Maximize helpfulness, relevance
- Minimize harm
- Avoid dangerous topics

RLHF : Reinforcement Learning from Human Feedback



Fine Tuning : Weight Matrices (W)



L

1	2	-1	-4	-20
-8	-3	4	2	1
3	8	10	1	3
-4	-1	1	7	-6

V

1	2	-1	-4	-20
-8	-3	4	2	1
3	8	10	1	3
-4	-1	1	7	-6

K

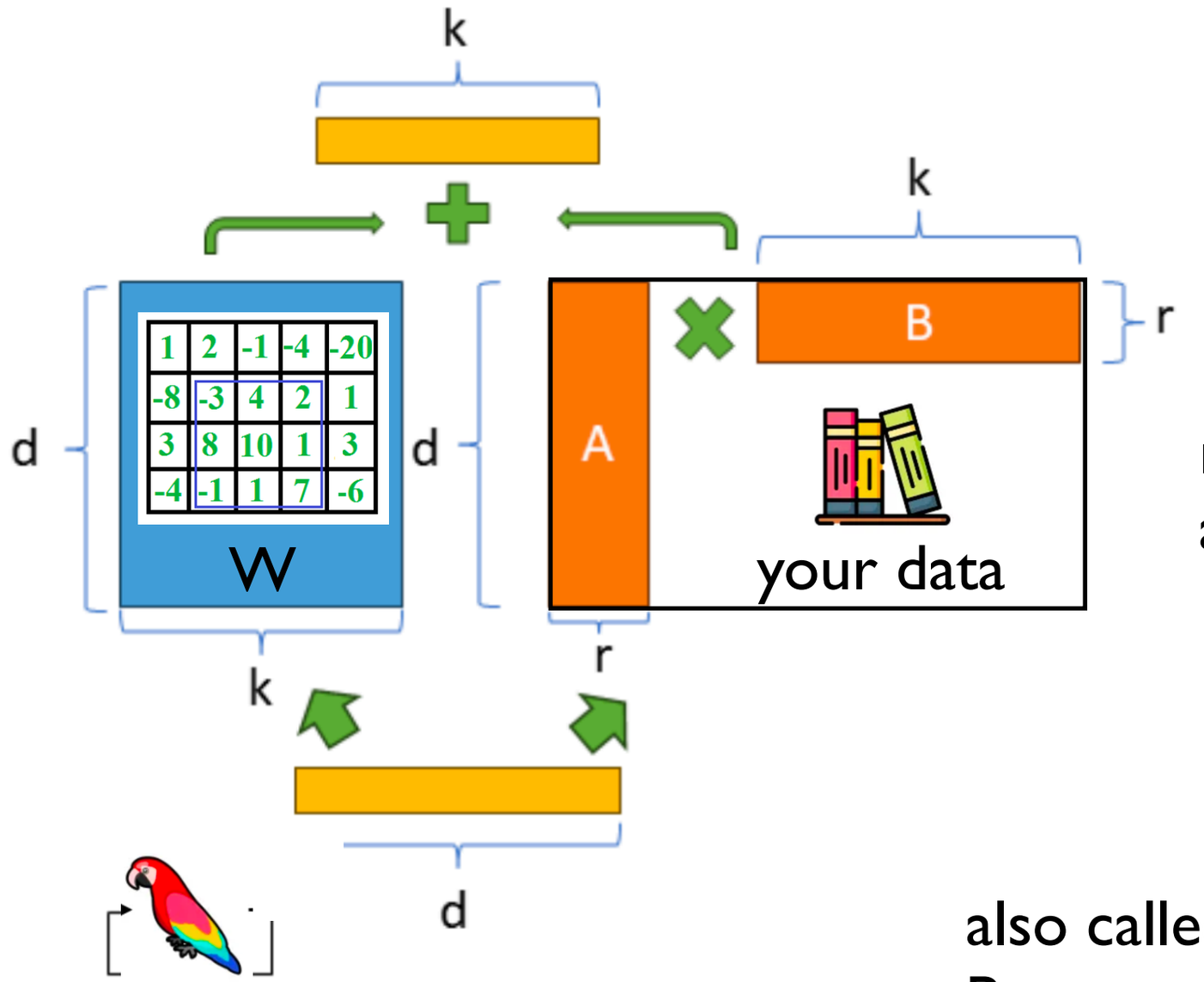
1	2	-1	-4	-20
-8	-3	4	2	1
3	8	10	1	3
-4	-1	1	7	-6

Q

1	2	-1	-4	-20
-8	-3	4	2	1
3	8	10	1	3
-4	-1	1	7	-6

W = V | Q | K | L

Fine-Tuning w/LoRA : Low Rank Adaptation

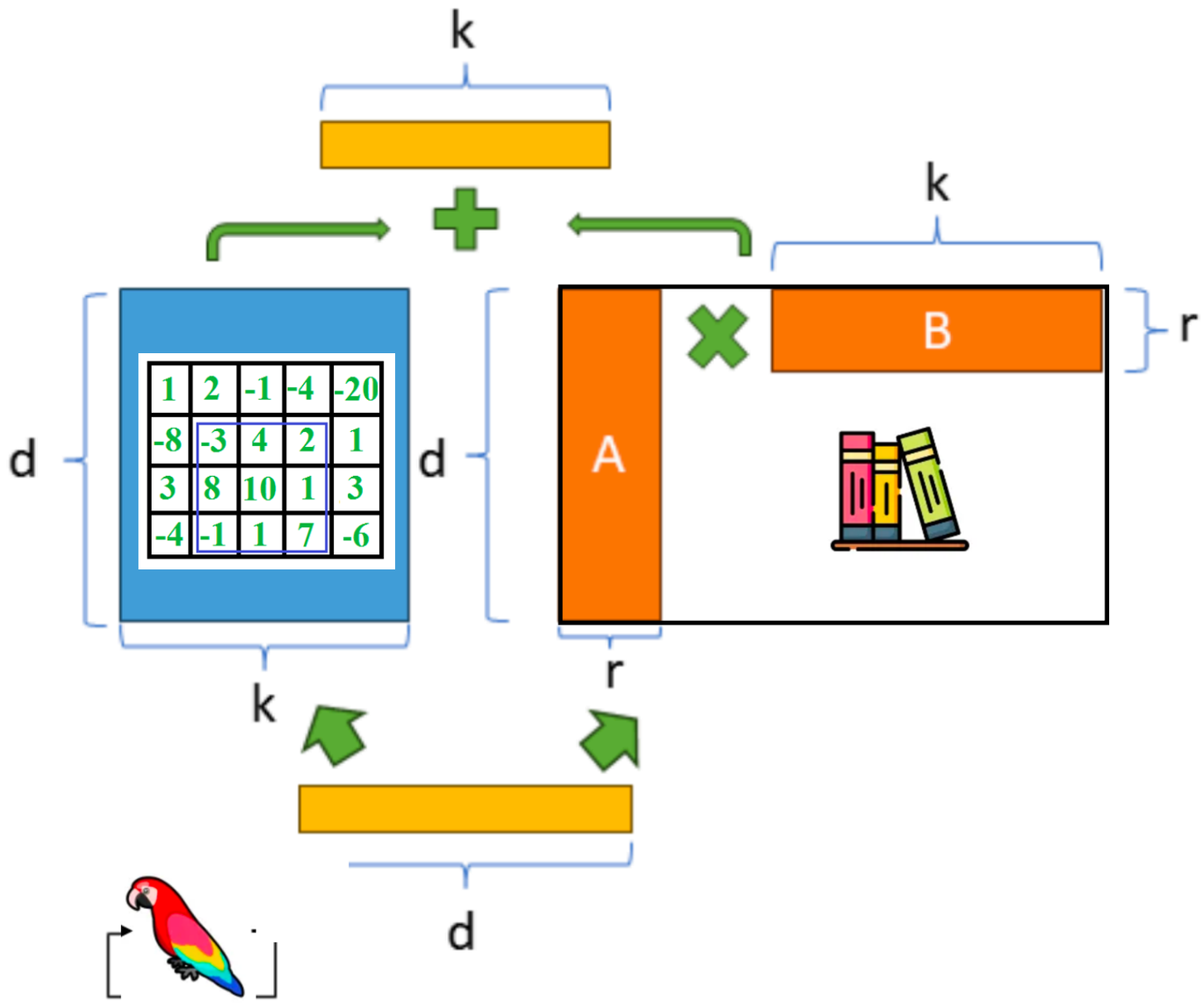


rank r matrix
added to W

open source LLM

also called PEFT:
Parameter Efficient Fine-Tuning

Fine-Tuning w/LoRA : Low Rank Adaptation



<u>easy fine-tuning</u>

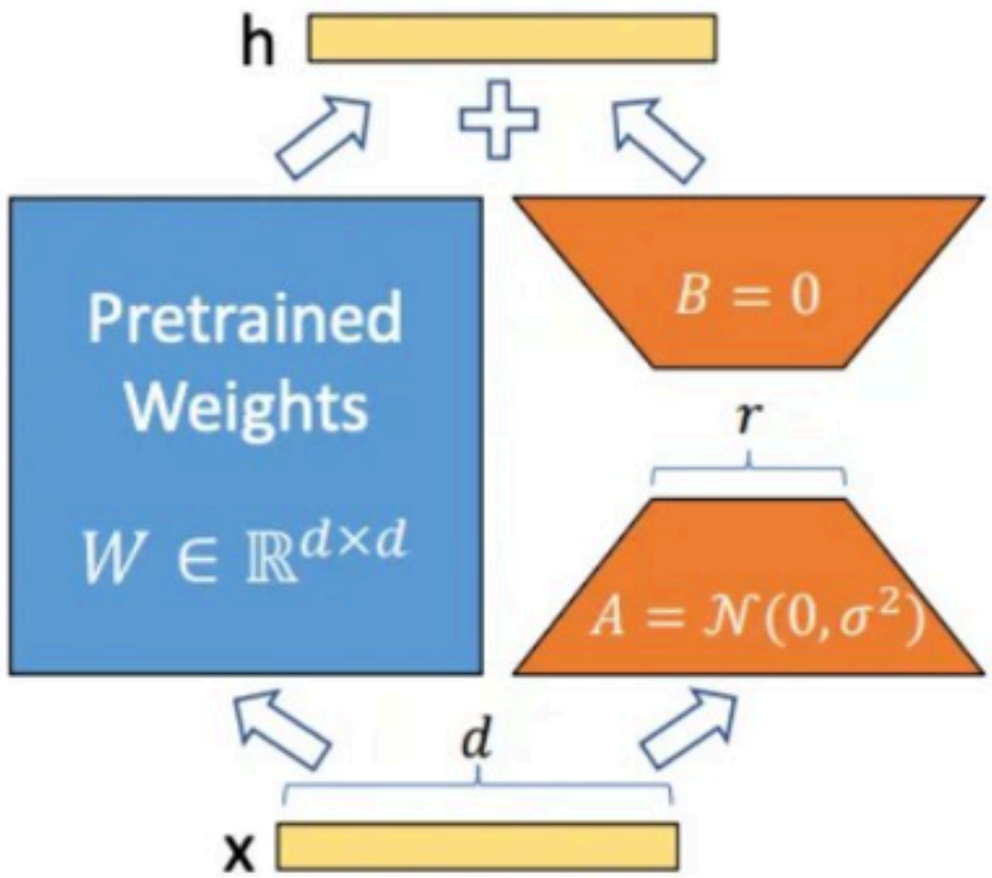
- very fast
- low memory

<u>choices:</u>

- rank (r=64|128)
- # epochs (3)
- learning rate ?

Fine-Tuning w/LoRA : Low Rank Adaptation

LoRA leaves the pretrained layers of the LLM fixed and injects a trainable rank decomposition matrix in each layer of the transformer.



Finetuned Weights — Weight Update

$$W_{\mathtt{ft}} = W_{\mathtt{pt}} + \Delta W$$

Pretrained Weights

Rank Decomposition Matrix

$$W_{\mathtt{ft}} = W_{\mathtt{pt}} + \Delta W = W_{\mathtt{pt}} + AB$$

Approximation

where $W_{\mathtt{ft}}, W_{\mathtt{pt}}, \Delta W, AB \in \mathbb{R}^{d \times d}$

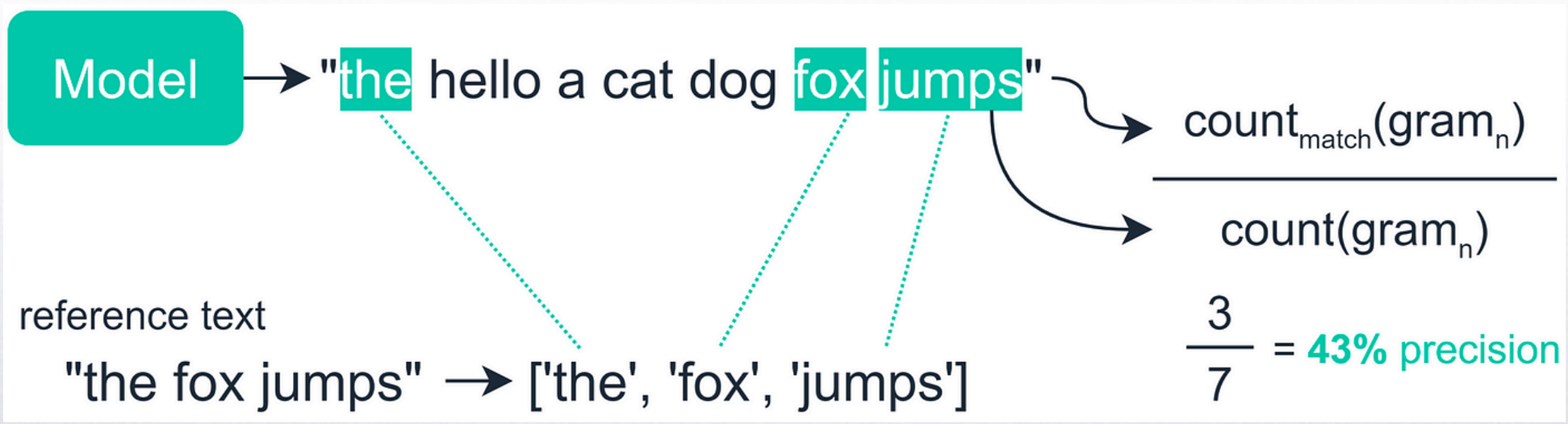
and $A \in \mathbb{R}^{d \times r}, B \in \mathbb{R}^{r \times d}$

Low Rank

Evaluations : Text Generation is Tricky

BLEU: ~ precision (# outputs in inputs)
ROUGE: ~ recall (# inputs in output)

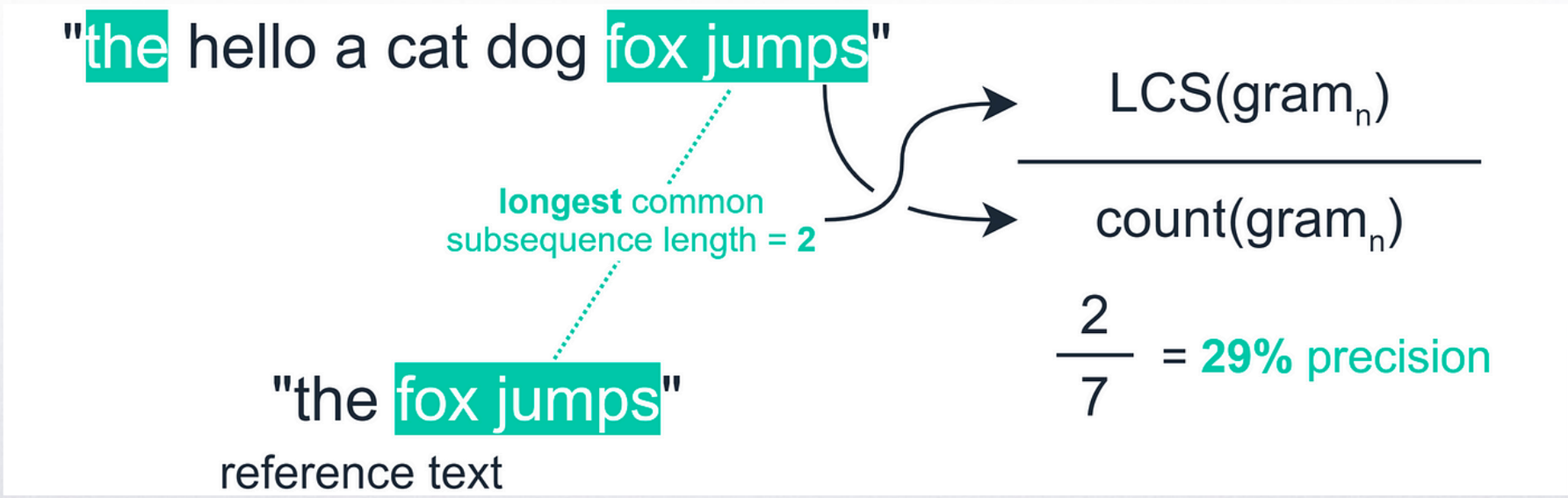
i.e ROUGE-N: # ngrams from ref text in gen text



Evaluations : Text Generation is Tricky

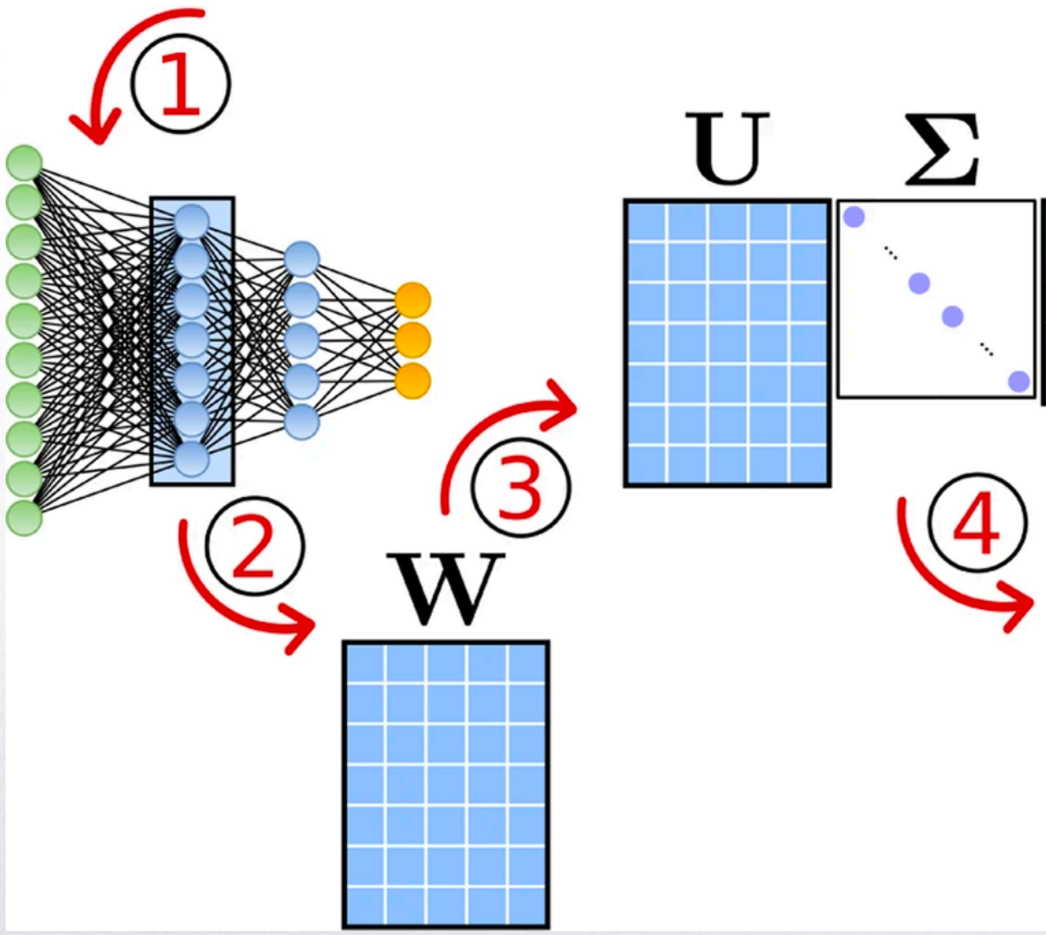
BLEU: ~ precision (# outputs in inputs)
ROUGE: ~ recall (# inputs in output)

i.e ROUGE-L: # least common sequence

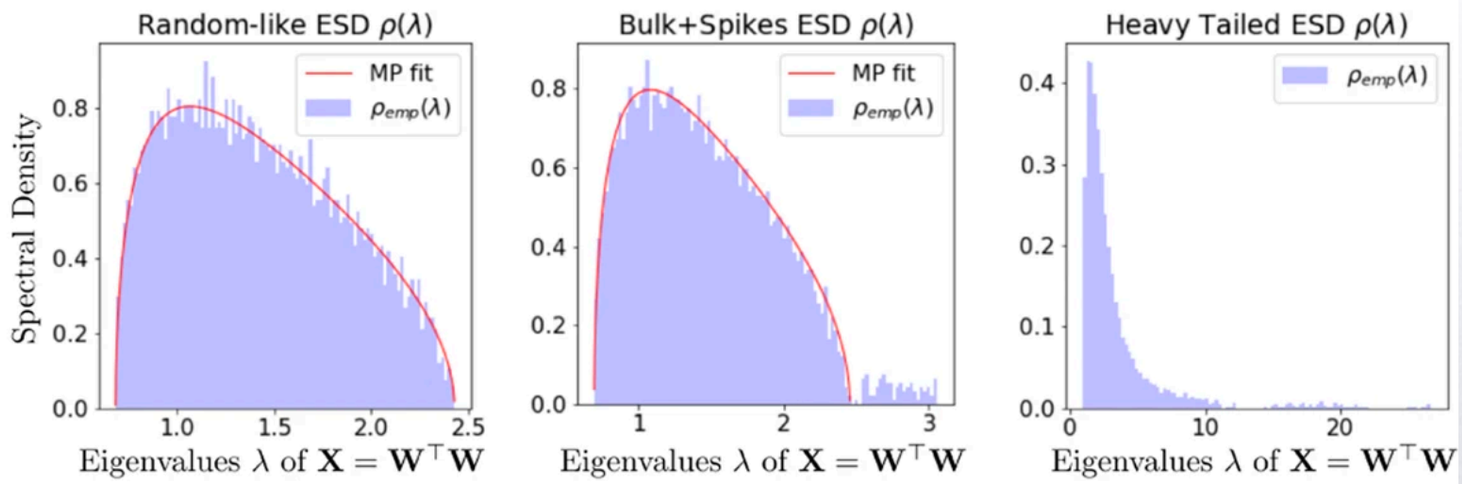


WeightWatcher : Data-Free Quality Metrics

https://weightwatcher.ai

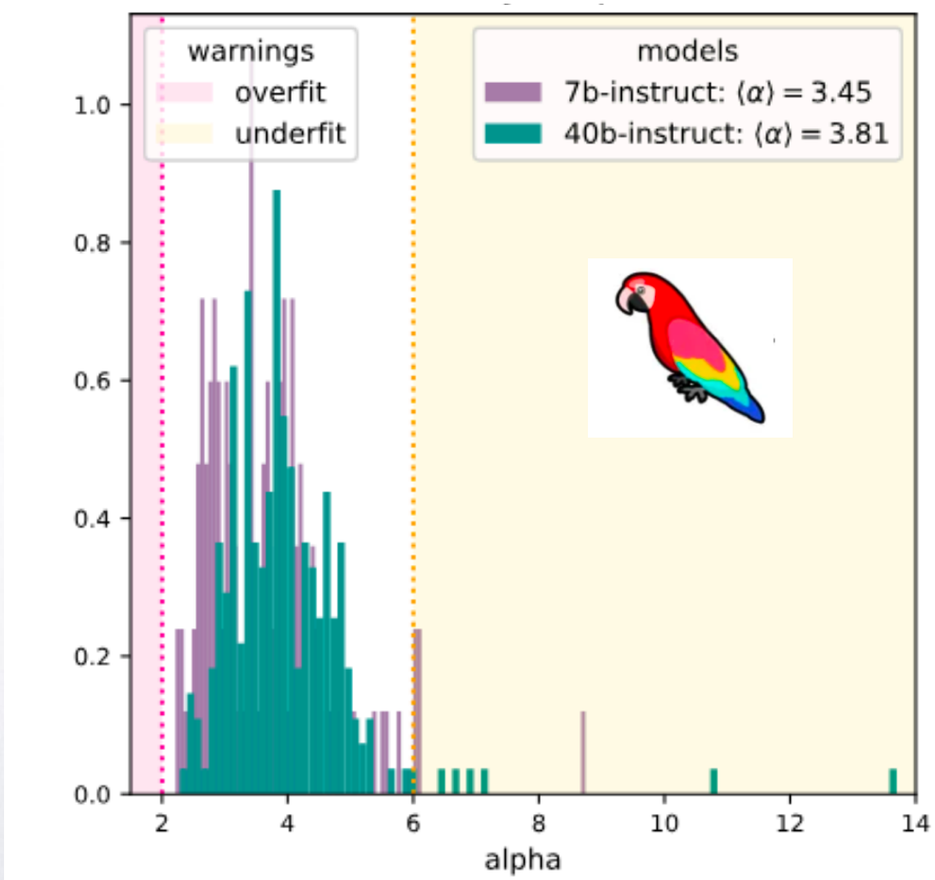


1. Take a model
   - 2. Take a weight matrix
     - 3. Do Spectral analysis
       - 4. Histogram of eigenvalues

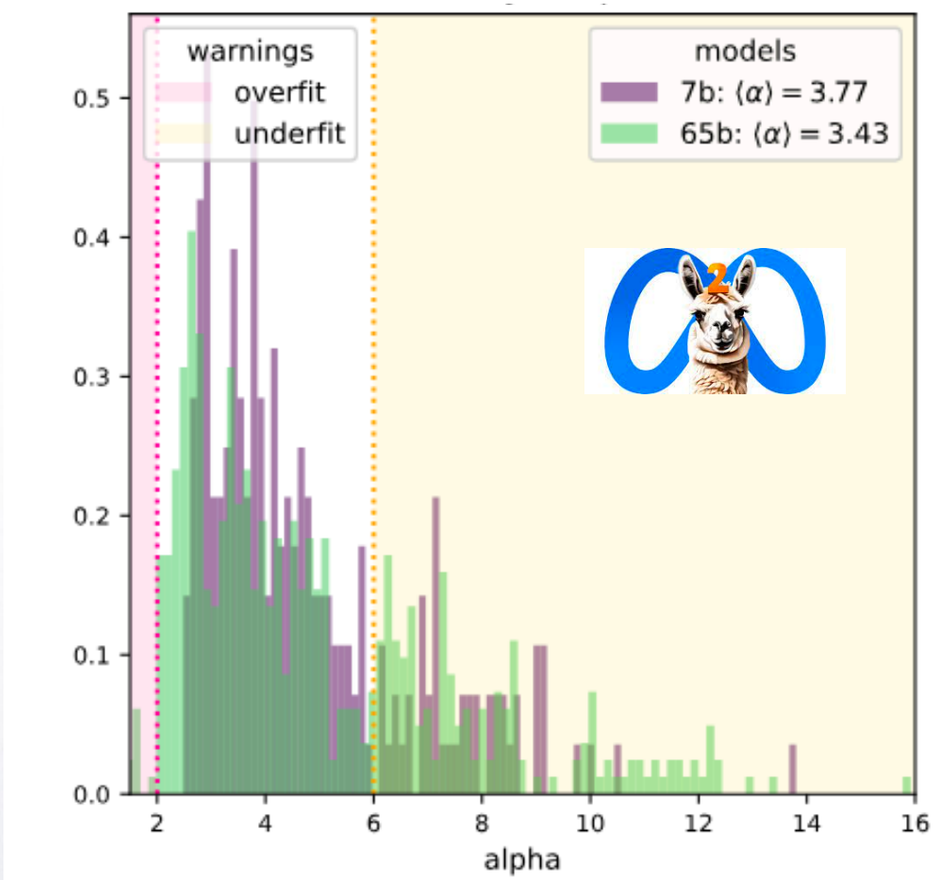


WeightWatcher : Data-Free Quality Metrics

Falcon



Llama

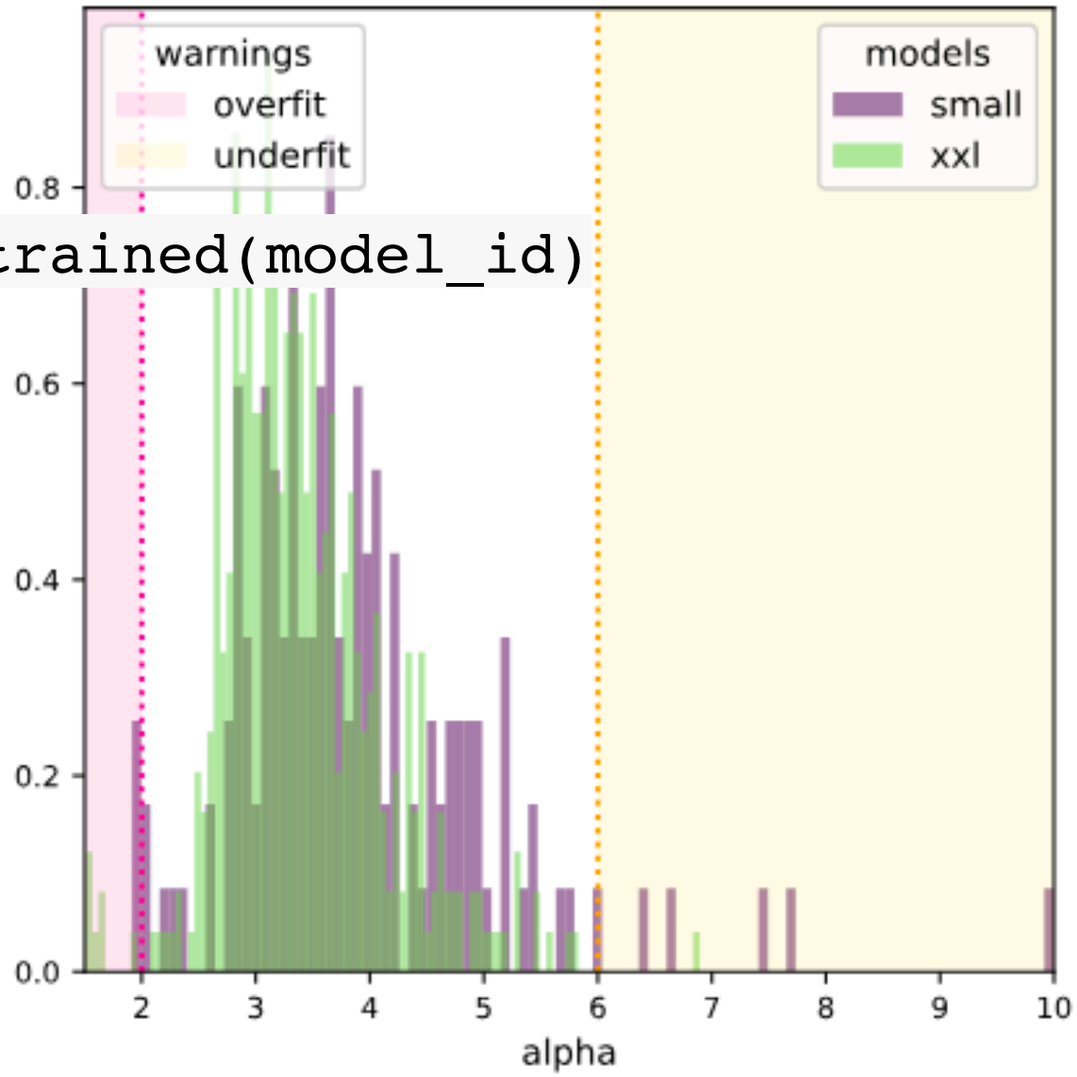


https://weightwatcher.ai

Fine-Tuning Examples : LoRA on FlanT5-XXL

```
from transformers import AutoTokenizer, AutoModelForSeq2SeqLM

model_id="google/flan-t5-xxl"
# Load tokenizer of FLAN-t5-XL
tokenizer = AutoTokenizer.from_pretrained(model_id)
```



Fine-Tuning Examples : SAMSum Dataset

```
from datasets import load_dataset
dataset = load_dataset("samsum")
```

13818513	Amanda: I baked cookies. Do you want some? Jerry: Sure! Amanda: I'll…	Amanda baked cookies and will bring Jerry some…
13728867	Olivia: Who are you voting for in this election? Oliver: Liberals as…	Olivia and Olivier are voting for liberals in this…
13681000	Tim: Hi, what's up? Kim: Bad mood tbh, I was going to do lots of stuf…	Kim may try the pomodoro technique recommended by Ti…
13730747	Edward: Rachel, I think I'm in ove with Bella.. rachel: Dont say…	Edward thinks he is in love with Bella. Rachel wants…
13728094	Sam: hey overheard rick say something Sam: i don't know what to…	Sam is confused, because he overheard Rick complaining…
13716343	Neville: Hi there, does anyone remember what date I got married on…	Wyatt reminds Neville his wedding anniversary is on…

< Previous 1 2 3 ... 148 Next >



Fine-Tuning Examples : PEFT Package

```
from peft import LoraConfig, get_peft_model,
                 prepare_model_for_int8_training, TaskType

# Define LoRA Config
lora_config = LoraConfig(
 r=16,
 lora_alpha=32,
 target_modules=["q", "v"],
 lora_dropout=0.05,
 bias="none",
 task_type=TaskType.SEQ_2_SEQ_LM
)
# prepare int-8 model for training
model = prepare_model_for_int8_training(model)
# add LoRA adaptor
model = get_peft_model(model, lora_config)
```