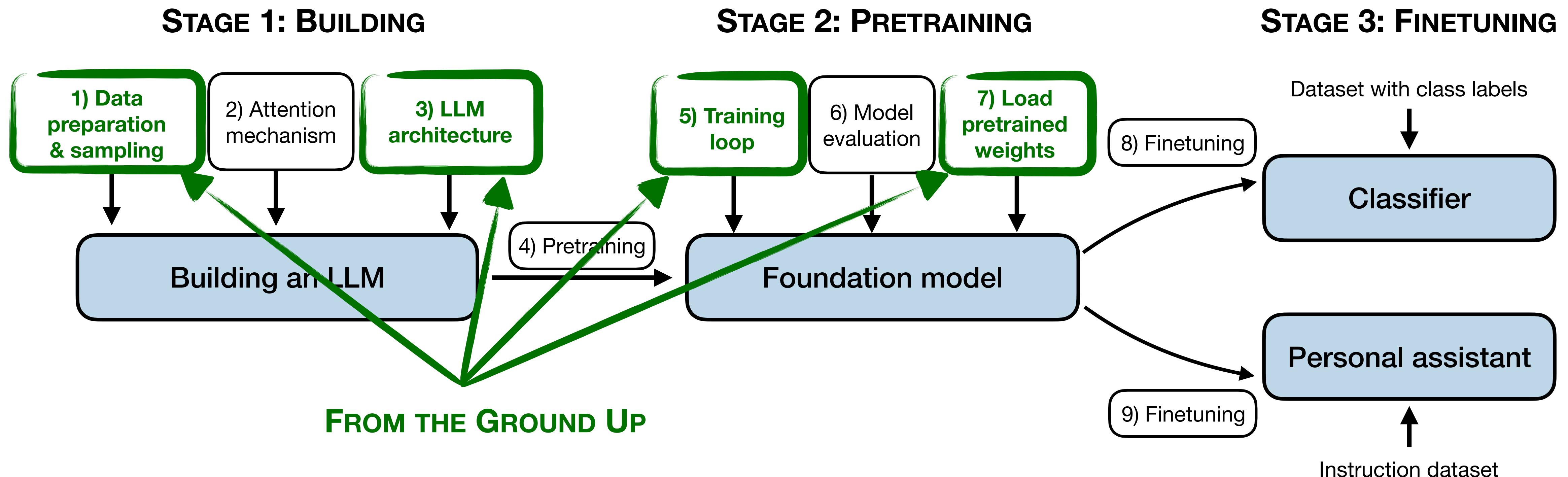


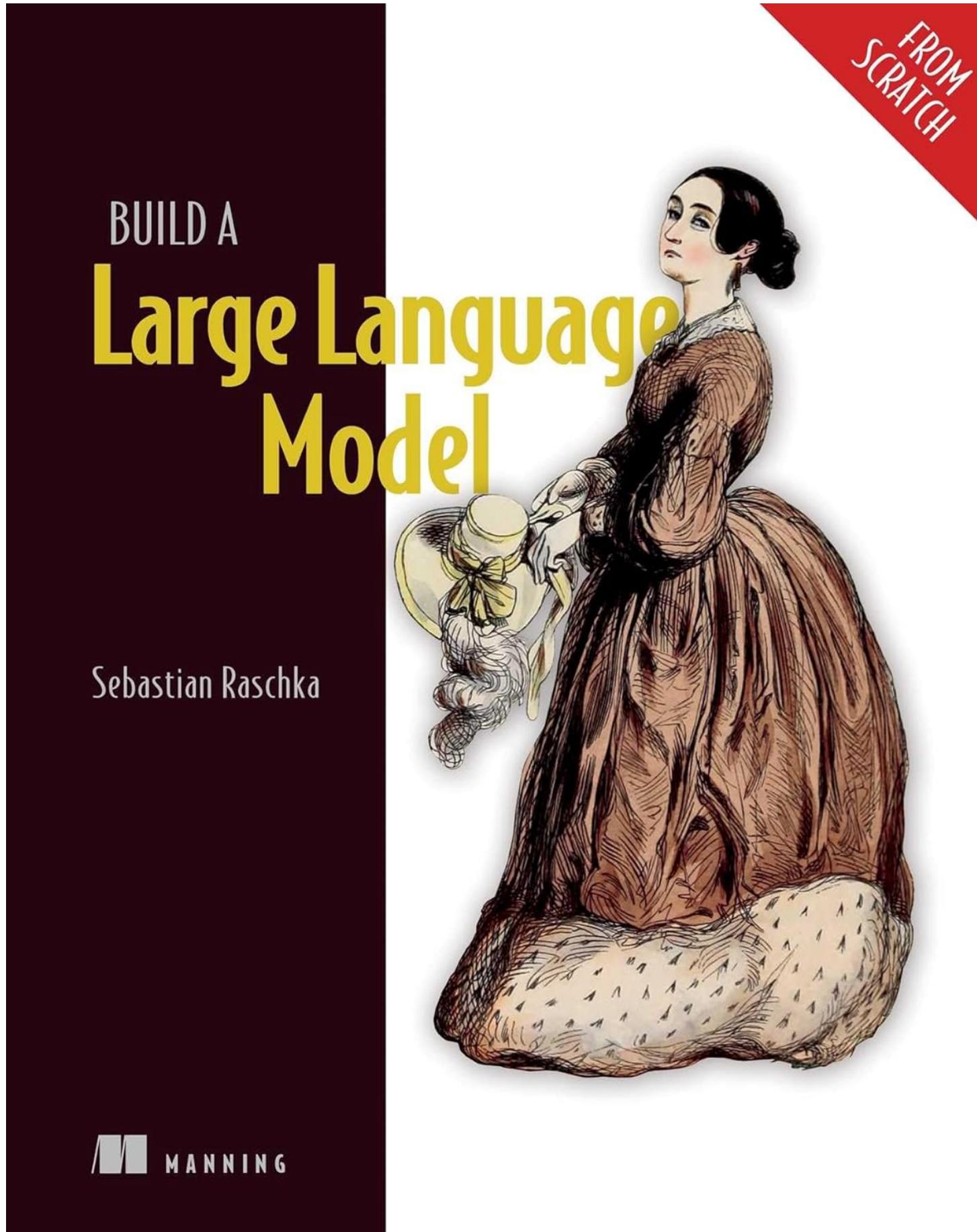
A nighttime photograph of the Seattle skyline, featuring the Space Needle and various skyscrapers. In the background, Mount Rainier is visible against a dark sky.

Pretraining and Finetuning LLMs from the Ground Up

<h2>Workshop topics</h2>	
1	Introduction to LLMs
2	Understanding LLM input data
3	Coding an LLM architecture
4	Pretraining LLMs
5	Loading pretrained weights
6	Finetuning LLMs

Developing an LLM





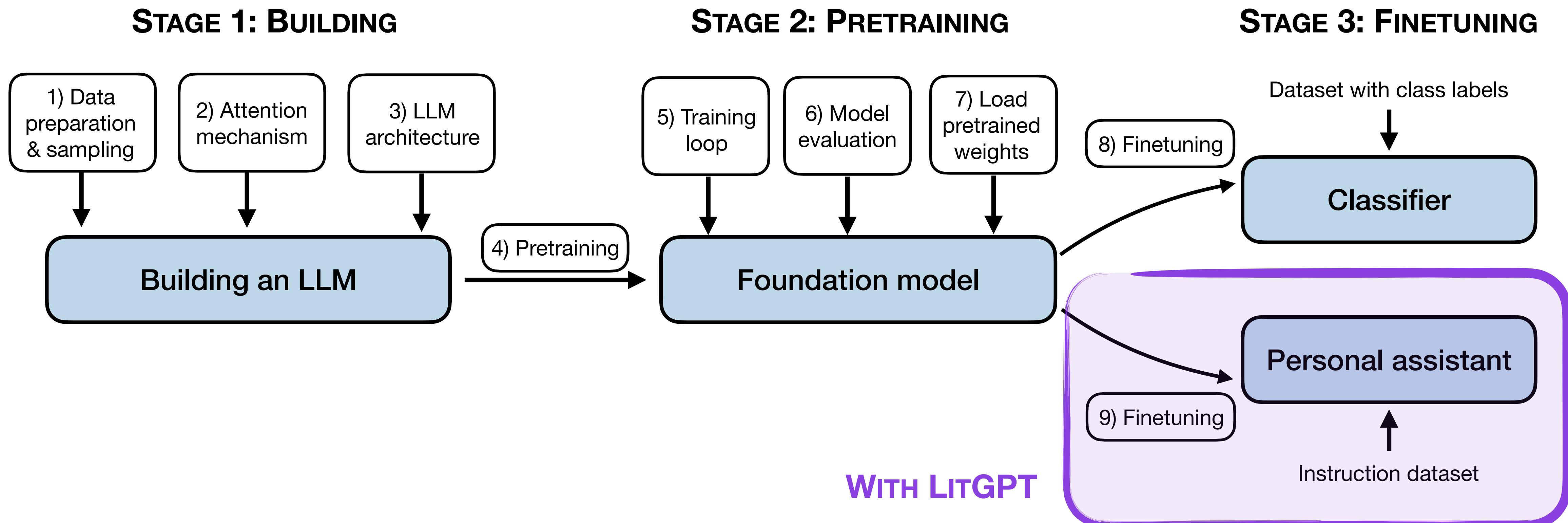
FROM
SCRATCH

<https://mng.bz/lrp2>

<https://github.com/rasbt/LLMs-from-scratch>

(Source for most figures and code)

Developing an LLM



LitGPT

20+ high-performance LLM implementations with recipes to pretrain, finetune, deploy at scale.

-  From scratch implementations
-  Flash attention
-  Reduce GPU memory (fp4/8/16/32)

-  No abstractions
-  FSDP
-  1–1000+ GPUs/TPUs

-  Beginner friendly
-  LoRA, QLoRA, Adapter
-  20+ LLMs

python 3.8 | 3.9 | 3.10 | 3.11

 CPU tests passing

License Apache 2.0

chat 988 online

[Lightning AI](#) • [Quick start](#) • [Models](#) • [Finetune](#) • [Deploy](#) • [All workflows](#) • [Features](#) • [Recipes \(YAML\)](#) • [Tutorials](#)

 Get started

<https://github.com/Lightning-AI/litgpt>



Lightning AI

Creators of PyTorch Lightning

Simple. Powerful.

Zero setup. Persistent. Always ready.

Studio marries the simplicity of a **local development experience** with the power of **1,000s of cloud GPUs**, unlimited storage and multiplayer collaboration.

The screenshot shows the Lightning Studio interface. On the left, a modal dialog titled "Choose a GPU machine" lists four GPU options: T4, V100, A10G, and A100. The A100 option is selected, showing 312 TFLOPs, 80 GB memory, a cost of \$43.98 per hour, and a wait time of 30 minutes. On the right, a Jupyter notebook cell displays code related to a decoder model, including trainable and non-trainable parameters, and a rank_zero_warn message indicating an epoch progress of 64%.

GPUs	Model	Speed (TFLOP)	Memory (GB)	Cost (hour)	Wait time (mins)
1 4	T4	65	16	\$ 2.44	1
1 4	V100	125	16	\$ 4.66	6
1 4	A10G	125	24	\$ 4.06	6
8	A100	312	80	\$ 43.98	30

No environment setup.

Code in the browser or connect your local IDE.

Switch from CPU to GPU with zero environment changes.

Host and share AI apps. Streamlit. Gradio. React JS.

Code together.

Infinite storage. Upload, share files and connect S3 buckets.

<https://lightning.ai/>

Lightning AI

Home Studio templates Agents Teamspaces Community Docs

Source

Lightning AI Public

Explore

Featured

Trending

Recent

All studios

My studios

Educational

Blogs

Papers

Tutorials

Workflows

Data processing

Endpoints

Training

Serving

Other

Model types

Audio

Image

Multimodal

Text

Tabular

RAG 102 Chat with Documents

★ Featured

Document Chat Assistant using RAG

aniket 269 6.54 K

Improve LLMs via Proxy-Tuning

★ Featured

Proxy-tuned LLM (e.g., Llama 2 70B)

Target LLM (e.g., Llama 2 70B)

Small Base LLM (e.g., Llama 2 7B)

Small Tuned LLM (e.g., Llama 2 7B Chat)

Add output

Composite difference in outputs

Improve LLMs With Proxy-Tuning

sebastian 47 7.88 K

Embed Wikipedia English under 5 dollars

★ Featured

WIKIPEDIA The Free Encyclopedia

English 6,776,000+ articles

French 2,587,000+ articles

Español 1,927,000+ artículos

日本語 1,401,000+記事

Italiano 1,845,000+ voci

Deutsch 2,876,000+ Artikel

中文 1,401,000+ 条目

عربی ٢,٣٧٦,٠٠٠+ مقالات

فارسی ۱,۷۷۷,۰۰۰+ مقاله

Embed English Wikipedia under 5 dollars

thomasgridai 26 2.73 K

Finetune Hugging Face BERT

★ Featured

with PyTorch Lightning

Finetune Hugging Face BERT with PyTorch Lig...

JG justin 97 1.98 K

Ingest documents (text, pdf, markdown, docx) in a vector database for Retrieval Augmented Generation (RAG)

Document Search and Retrieval using RAG

aniket 676 7.10 K

Data streaming benchmarks for ImageNet

★ Featured

Mosaic ML

WebDataset

PyTorch Lightning Data

Benchmark cloud data-loading libraries

thomasgridai 23 1.05 K

SlimPajama & Starcoder

★ Featured

1 trillion tokens

Prepare the TinyLlama 1T token dataset

thomasgridai 38 1.64 K

LoRA from Scratch

★ Featured

Forward pass with updated model weights

Embedding h

Pretrained weights W

A

B

Inputs x

class LoRALayer(torch.nn.Module):

```
def __init__(self, in_dim, out_dim, rank, alpha):
    super().__init__()
    std_dev = 1 / torch.sqrt(torch.tensor(rank).float())
    self.W_a = torch.nn.Parameter(torch.randn(in_dim, rank) * std_dev)
    self.W_b = torch.nn.Parameter(torch.zeros(rank, out_dim))
    self.alpha = alpha

def forward(self, x):
    x = self.alpha * (x @ self.W_a @ self.W_b)
    return x
```

Code LoRA from Scratch

sebastian 229 24.66 K

Optimized Inference API for Mistral 7B with vLLM

★ Featured

Mistral-7B

Mistral-7B x AWQ

vLLM 7B LLM benchmark

Optimized LLM inference API for Mistral 7B usi...

aniket 50 7.63 K

Contact

 @rasbt  in/sebastianraschka

 <https://sebastianraschka.com/contact/>

 <https://lightning.ai>