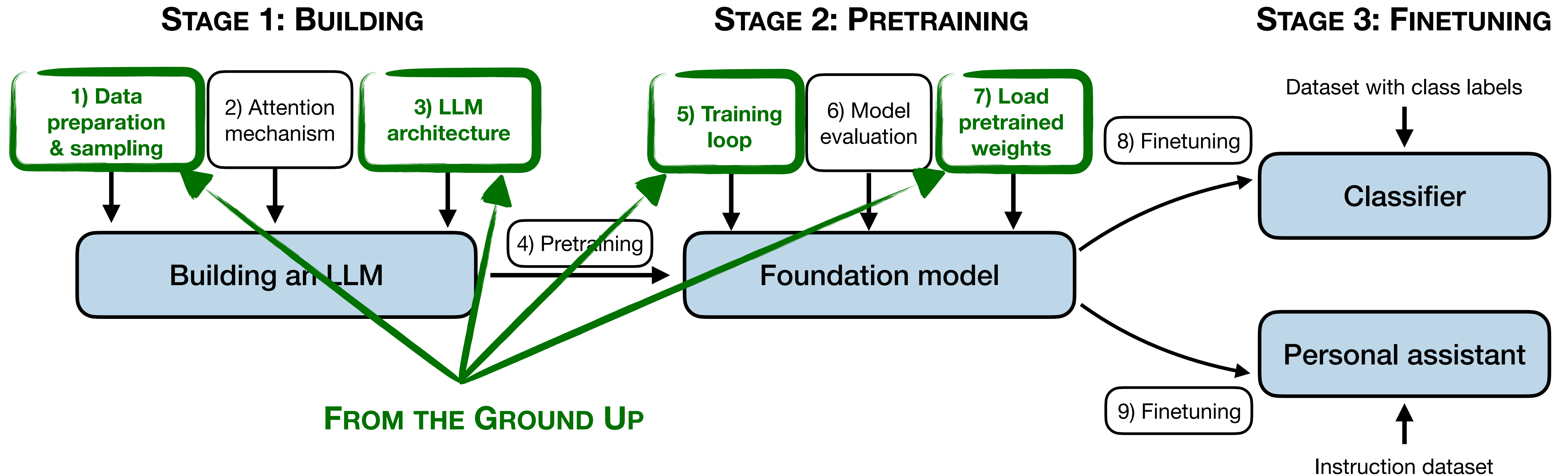
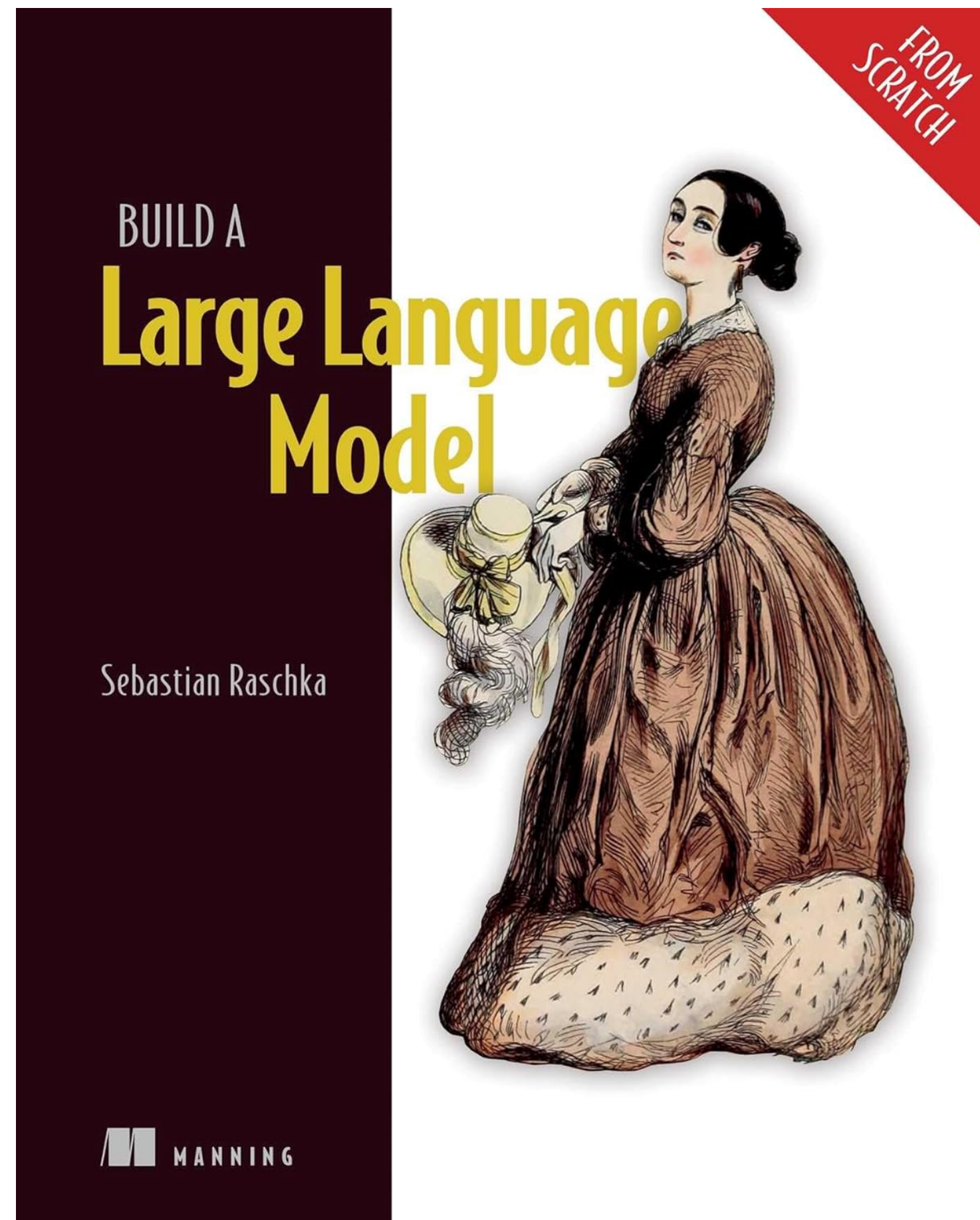


	Workshop topics
1	Introduction to LLMs
2	Understanding LLM input data
3	Coding an LLM architecture
4	Pretraining LLMs
5	Loading pretrained weights
6	Finetuning LLMs

Developing an LLM



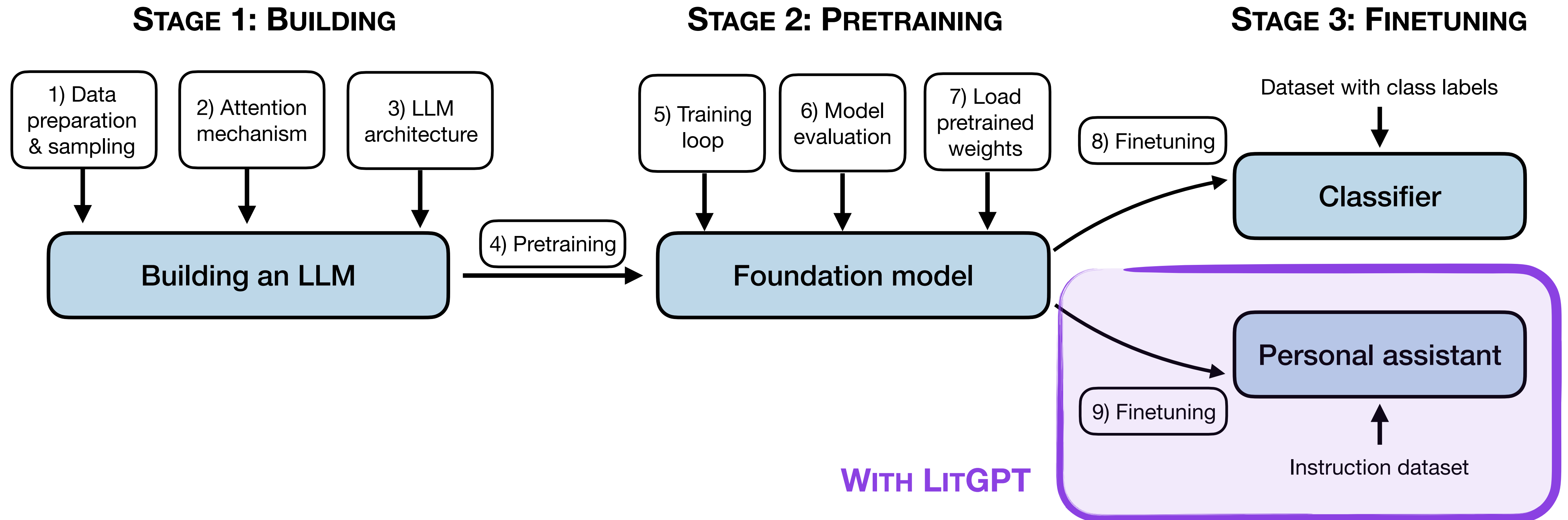


<https://mng.bz/lrp2>

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

(Source for most figures and code)

Developing an LLM





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

- | | | |
|-----------------------------------|---------------------|------------------------|
| ✓ From scratch implementations | ✓ No abstractions | ✓ Beginner friendly |
| ✓ Flash attention | ✓ FSDP | ✓ LoRA, QLoRA, Adapter |
| ✓ Reduce GPU memory (fp4/8/16/32) | ✓ 1–1000+ GPUs/TPUs | ✓ 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>

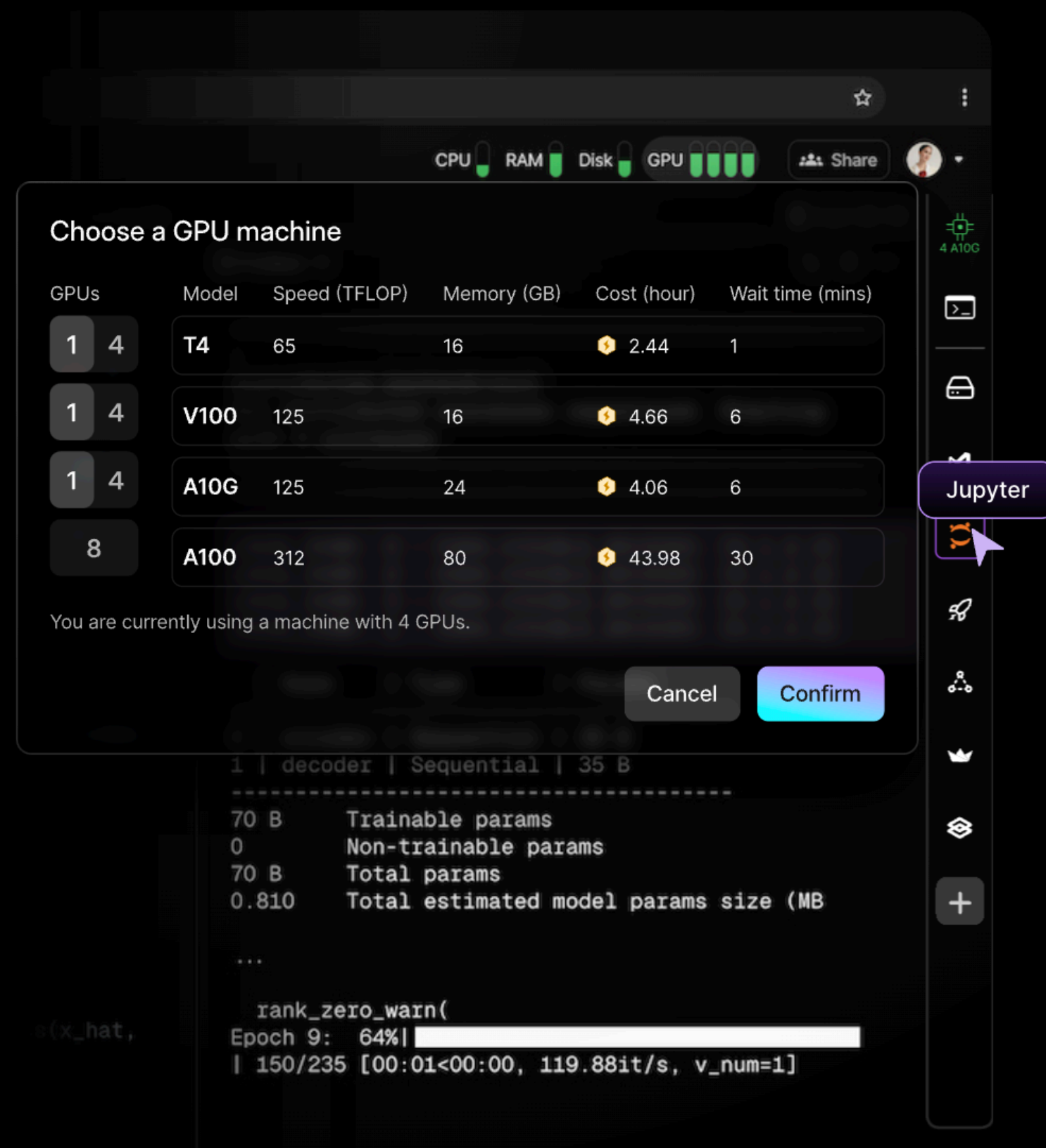


Creators of PyTorch Lightning

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.



- ⚡ 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 **Public**

★ Featured

↗ Trending


🕒 Recent

All studios

 My studios

 Blogs


Papers


 Tutorials

 Data processing

Endpoints

 Training



 Other

 Audio

 Image

✦ Multimodal

 Text

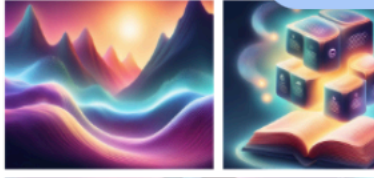
= Tabular

[illegible][illegible]

★ Featured



Finetune Hugging Face BERT



with PyTorch Lightning



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

Library	Performance (Relative)
Mosaic ML	Lowest
WebDataset	Medium
PyTorch Lightning Data	Highest

LoRA from Scratch

★ Featured

Forward pass with
updated model weights

```

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
 24.6K

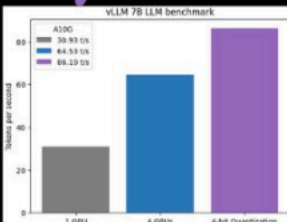
★ Featured

Optimized Inference API for Mistral 7B with vLLM



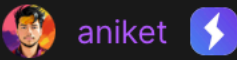
Mistral-7B

Mistral-7B x AWQ



vLLM 7B LLM Inference

Quantization	Tokens per second
8-bit	~25
4-bit	~50
4-bit Quantization	~100



aniket

⚡

🚀


50

👁️

7.63 K

Optimized LLM inference API for Mistral 7B using vLLM

Contact

 @rasbt  in/sebastianraschka

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

 <https://lightning.ai>