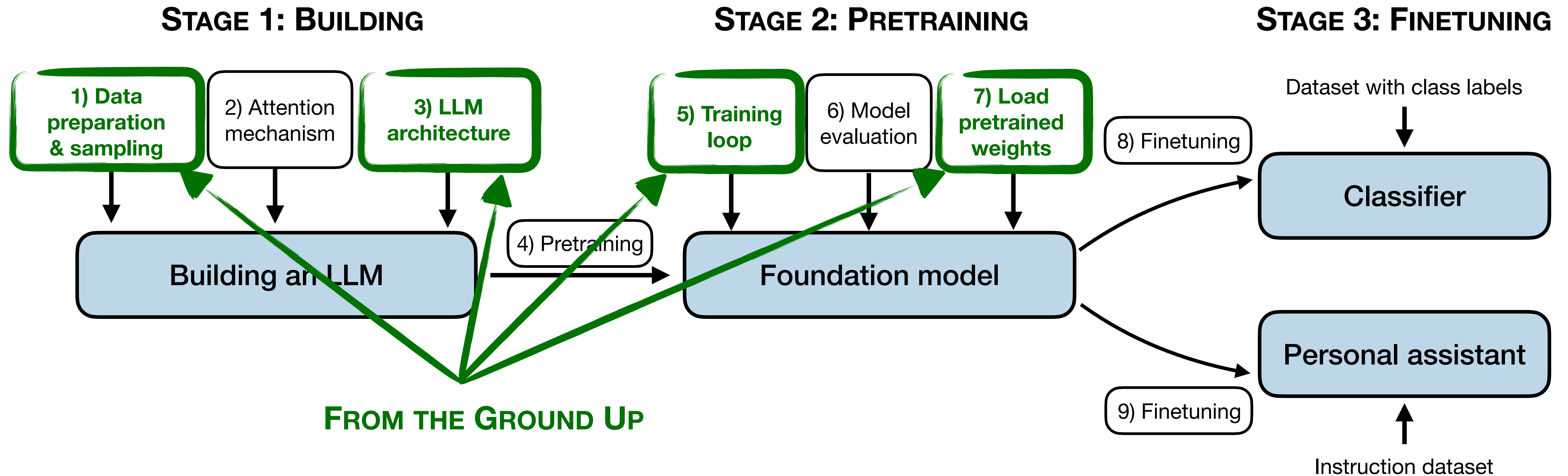


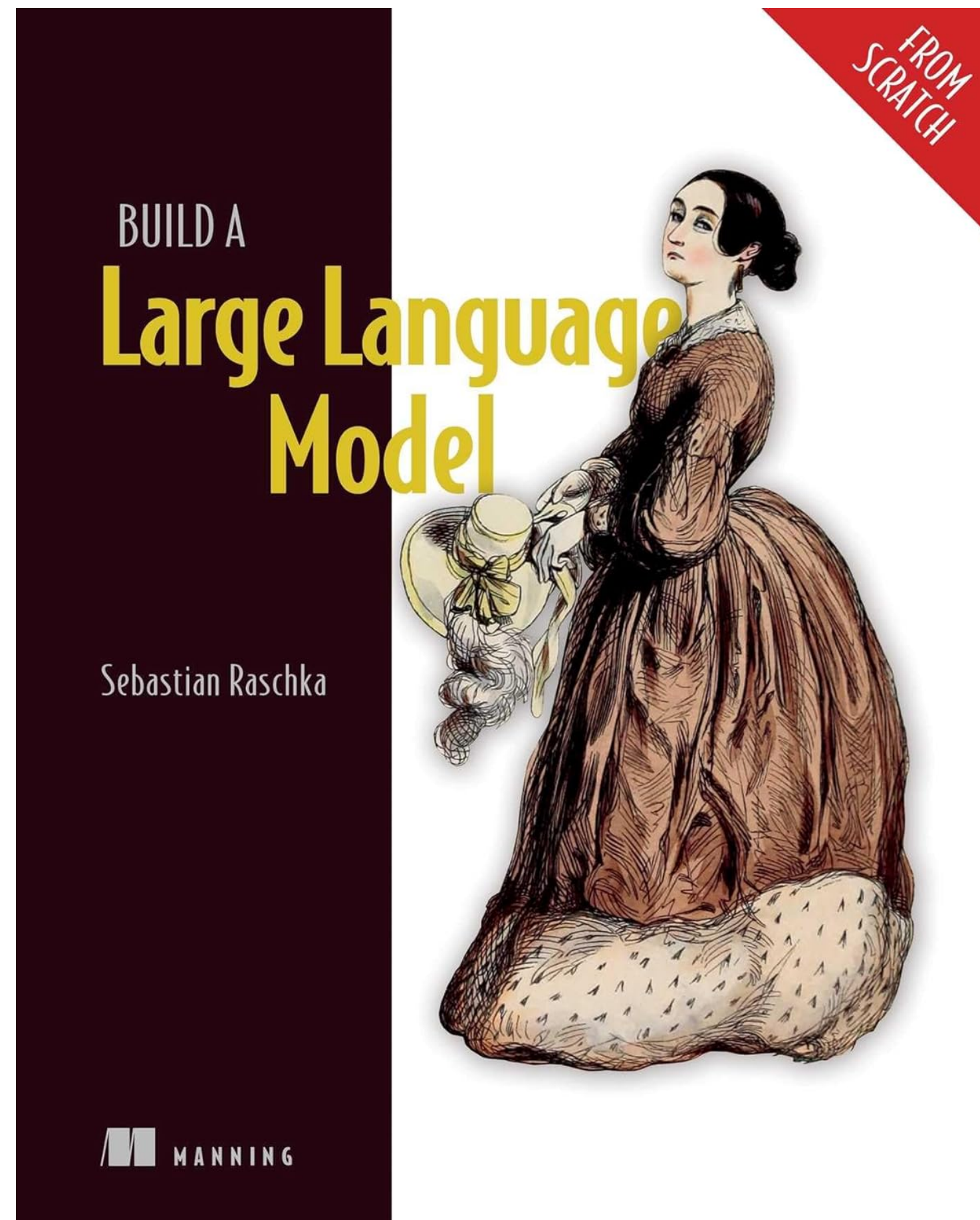
A nighttime photograph of the Seattle skyline. The Space Needle is the central focus, illuminated against the dark sky. Other skyscrapers are visible, their windows glowing with city lights. In the background, the dark silhouette of Mount Rainier is visible under a deep purple and blue twilight sky. The foreground shows some lower-rise buildings and trees, also partially illuminated.

Pretraining and Finetuning LLMs from the Ground Up

	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



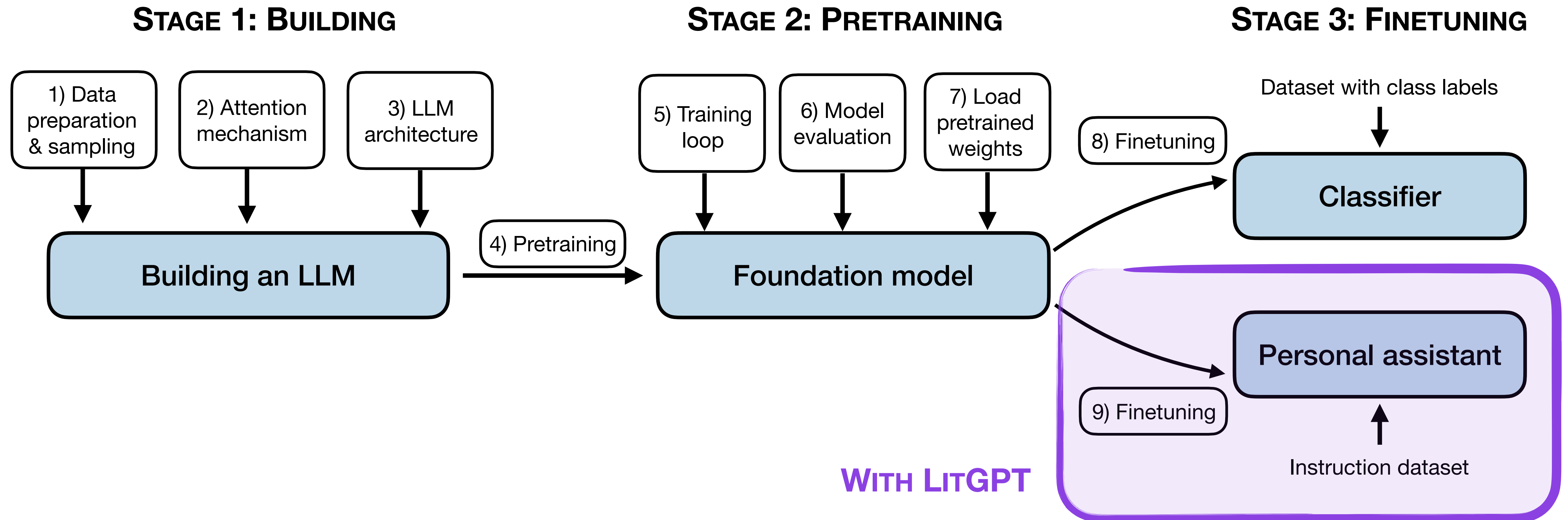


<https://mng.bz/lrp2>

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

(Source for most figures and code)

Developing an LLM



⚡ 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>

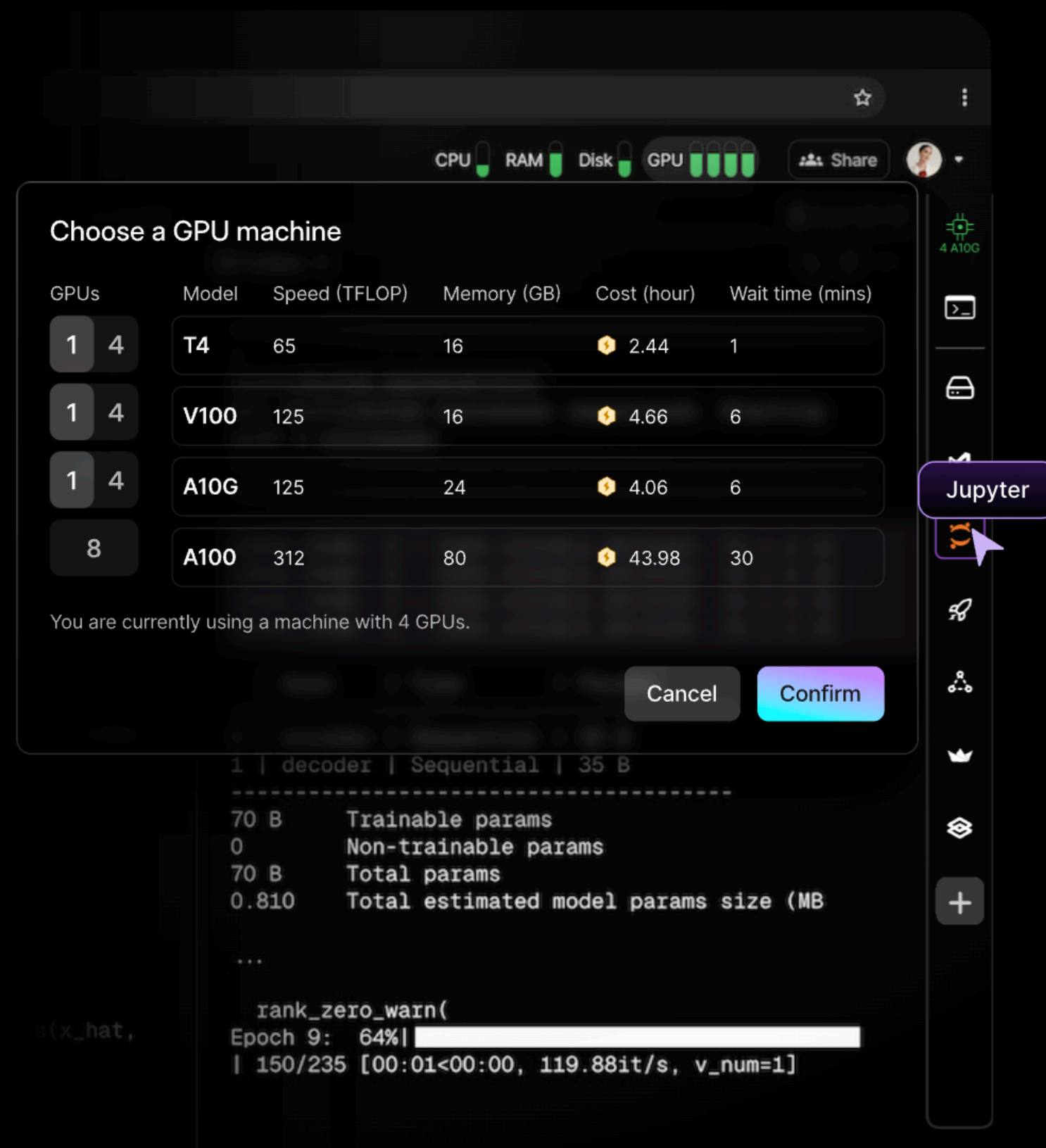


Creators of PyTorch Lightning

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

```

graph LR
    TargetLLM[Target LLM  
e.g., Llama2 70B] --> Sum((+))
    SmallBaseLLM[Small Base LLM  
e.g., Llama2 7B] --> Sum
    Sum --> ProxyTunedLLM[Proxy-tuned LLM  
e.g., now as good as Llama2 70B Chat]
    SmallBaseLLM --> Diff(( ))
    Diff --> ComputeDiff[Compute difference in outputs]
    ComputeDiff --> ProxyTunedLLM
  
```

Embed Wikipedia English under 5 dollars

English
6,776,000+ articles

Español
1,927,000+ artículos

日本語
1,401,000+ 記事

Italiano
1,845,000+ voci

العربية
١,٩٢٧,٠٠٠+ مقالة



Français
2,587,000+ articles

Русский
1,960,000+ статей

Deutsch
2,876,000+ Artikel

中文
1,401,000+ 条目 / 條目

فارسی
١,٩٢١,٠٠٠+ مقاله

Embed English Wikipedia under 5 dollars



thomasgridai





26







2.73 K

A screenshot of a YouTube video player. The video title is "Finetune Hugging Face BERT with PyTorch Lightning" in white text on a dark background. The channel name "JG justin" is visible in the bottom left corner. The video thumbnail shows a person wearing headphones and glasses, looking at a laptop screen. The video player interface includes a play button, a progress bar, and a timestamp of 97. The video description is partially visible in the bottom right corner, showing "Finetune Hugging Face BERT with PyTorch Lig...".

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	Low
WebDataset	Medium
PyTorch Lightning Data	High

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

24.66 K

Optimized Inference API for Mistral 7B with vLLM

Contact

 @rasbt  in/sebastianraschka

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

 <https://lightning.ai>