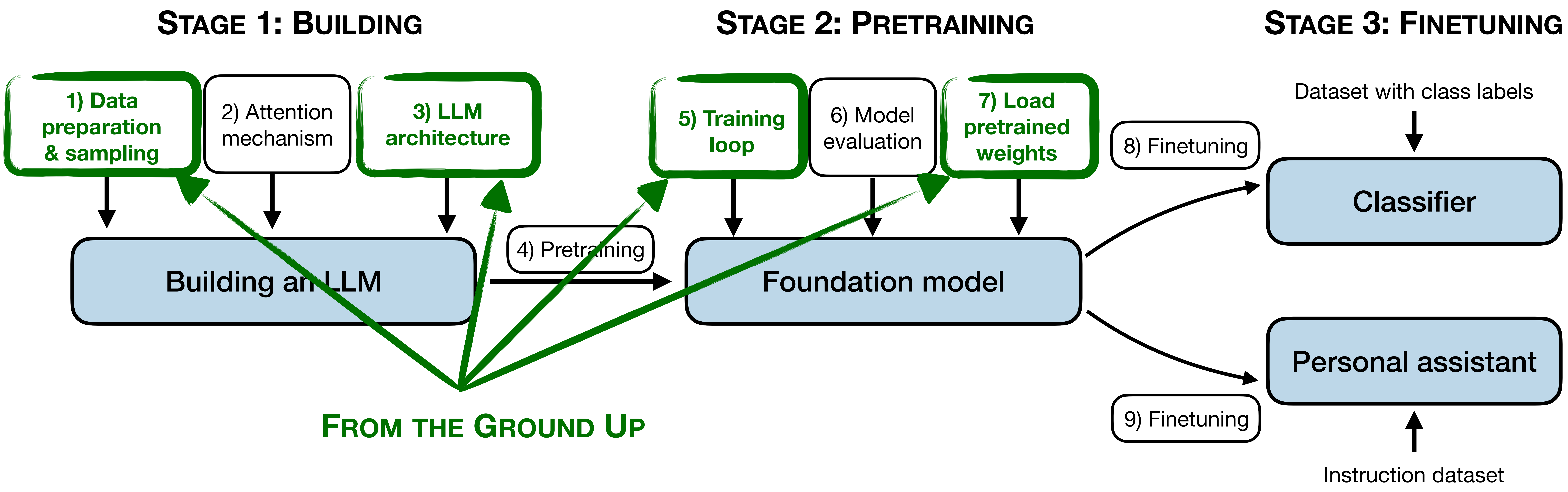
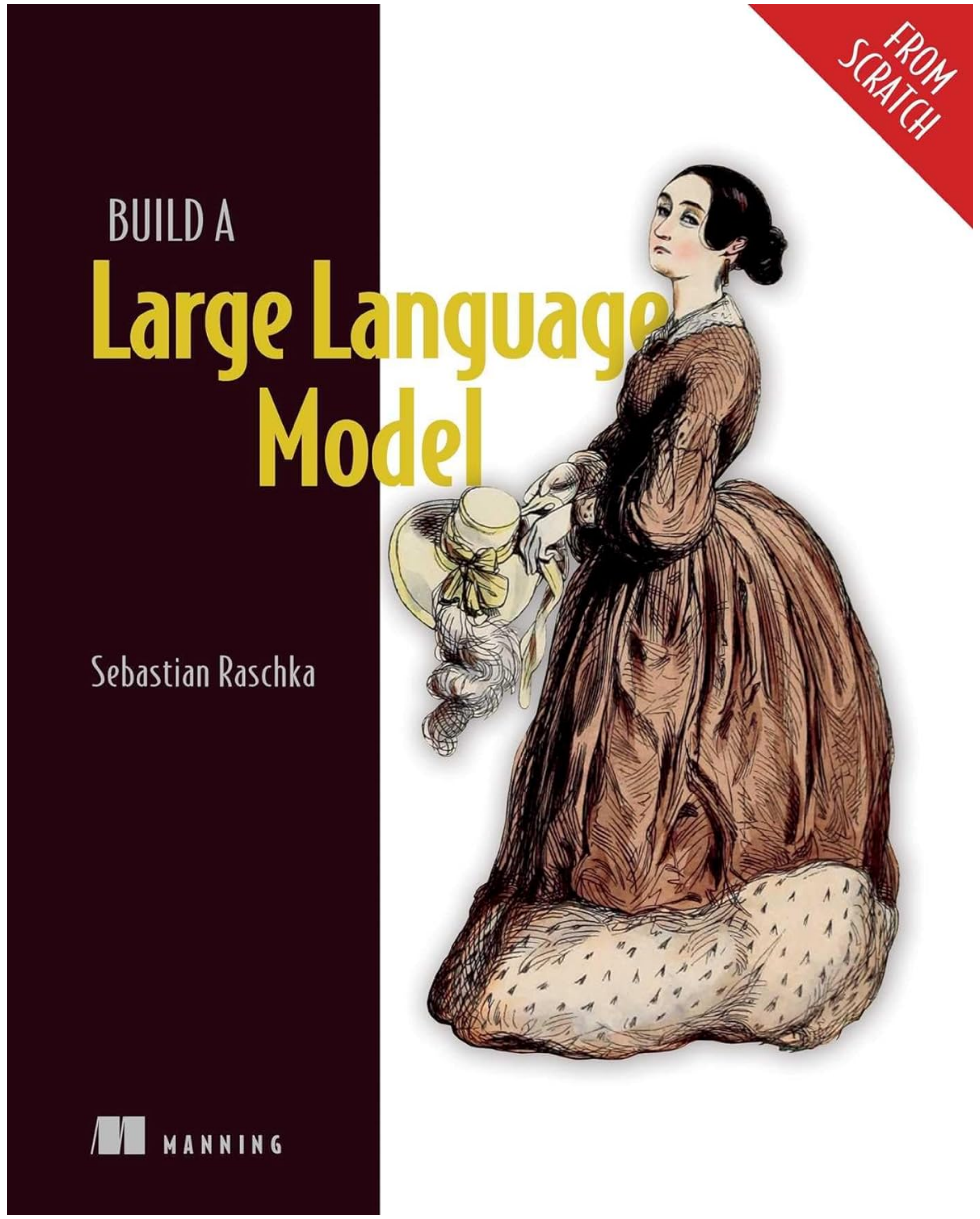
Pretraining and Finetuning LLMs

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



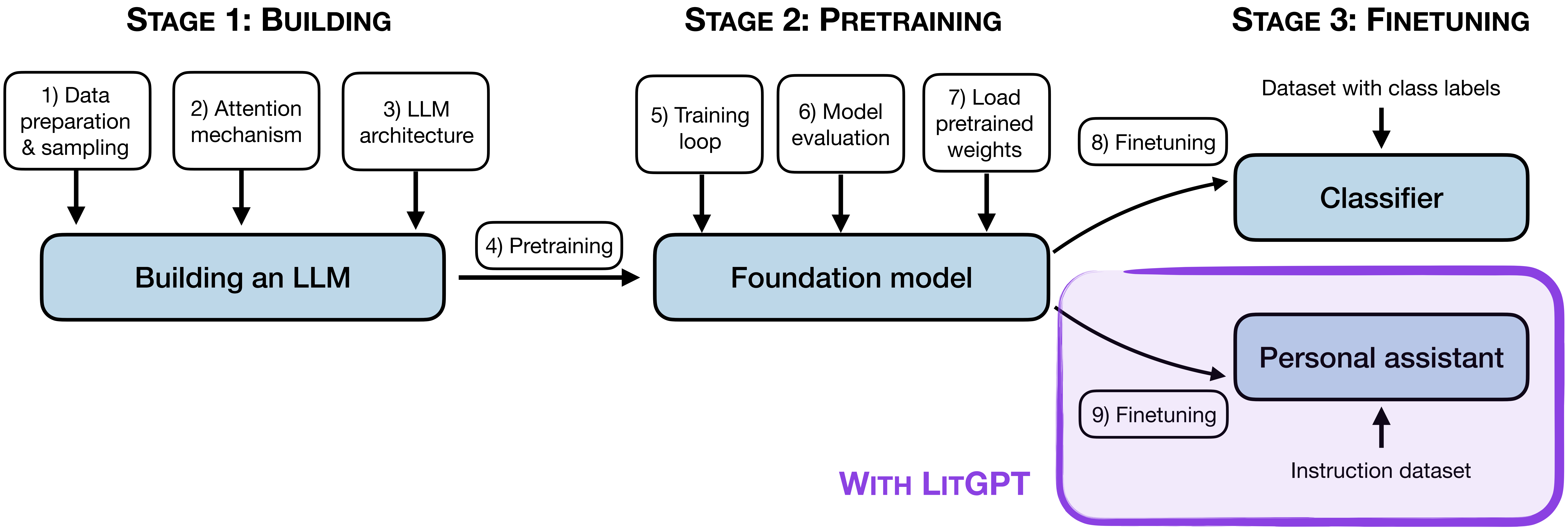


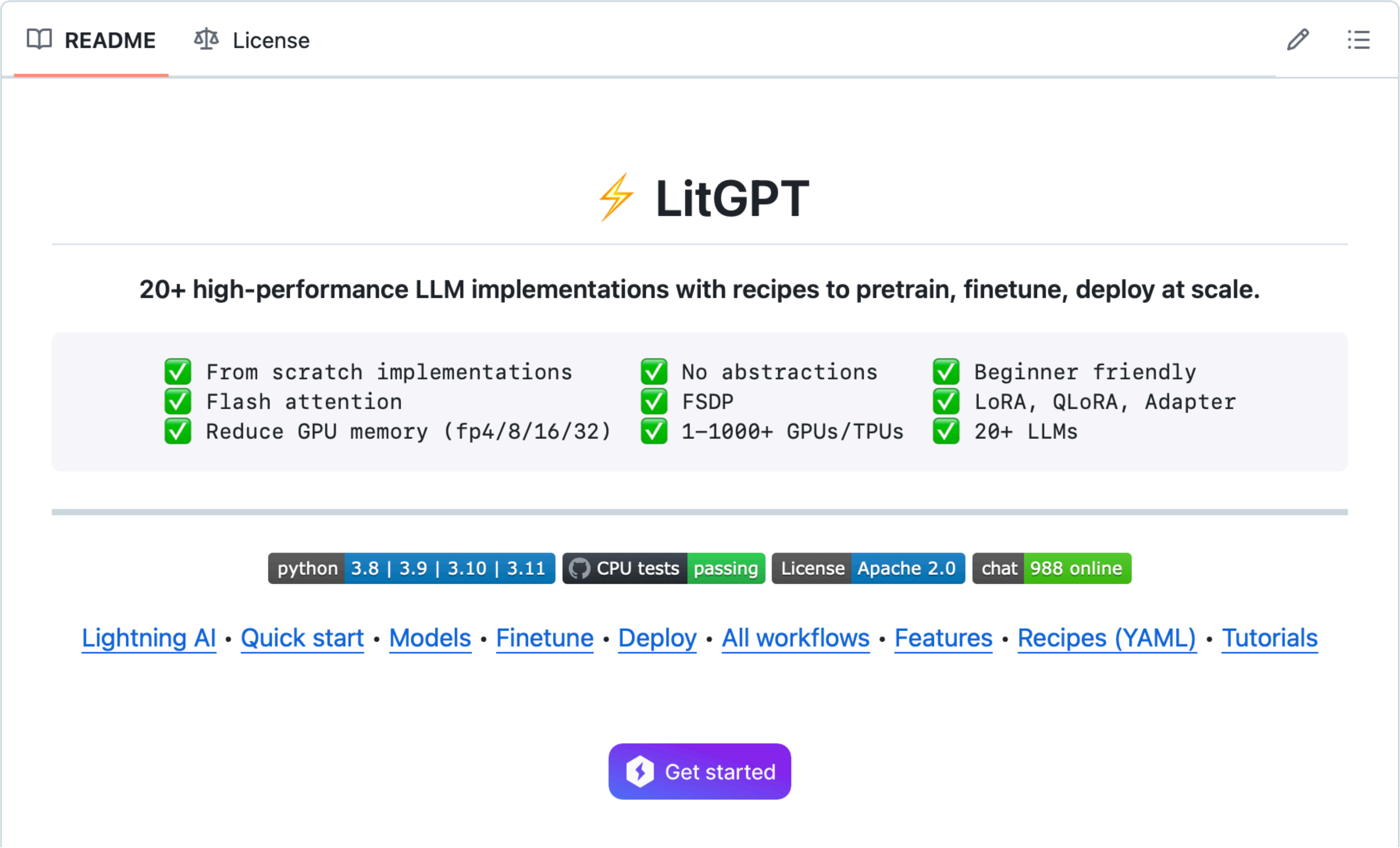
https://mng.bz/lrp2

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     ✅ No abstractions     ✅ Beginner friendly
✅ Flash attention                  ✅ FSDP                ✅ LoRA, QLoRA, Adapter
✅ Reduce GPU memory (fp4/8/16/32)  ✅ 1-1000+ GPUs/TPUs   ✅ 20+ LLMs
```

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

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

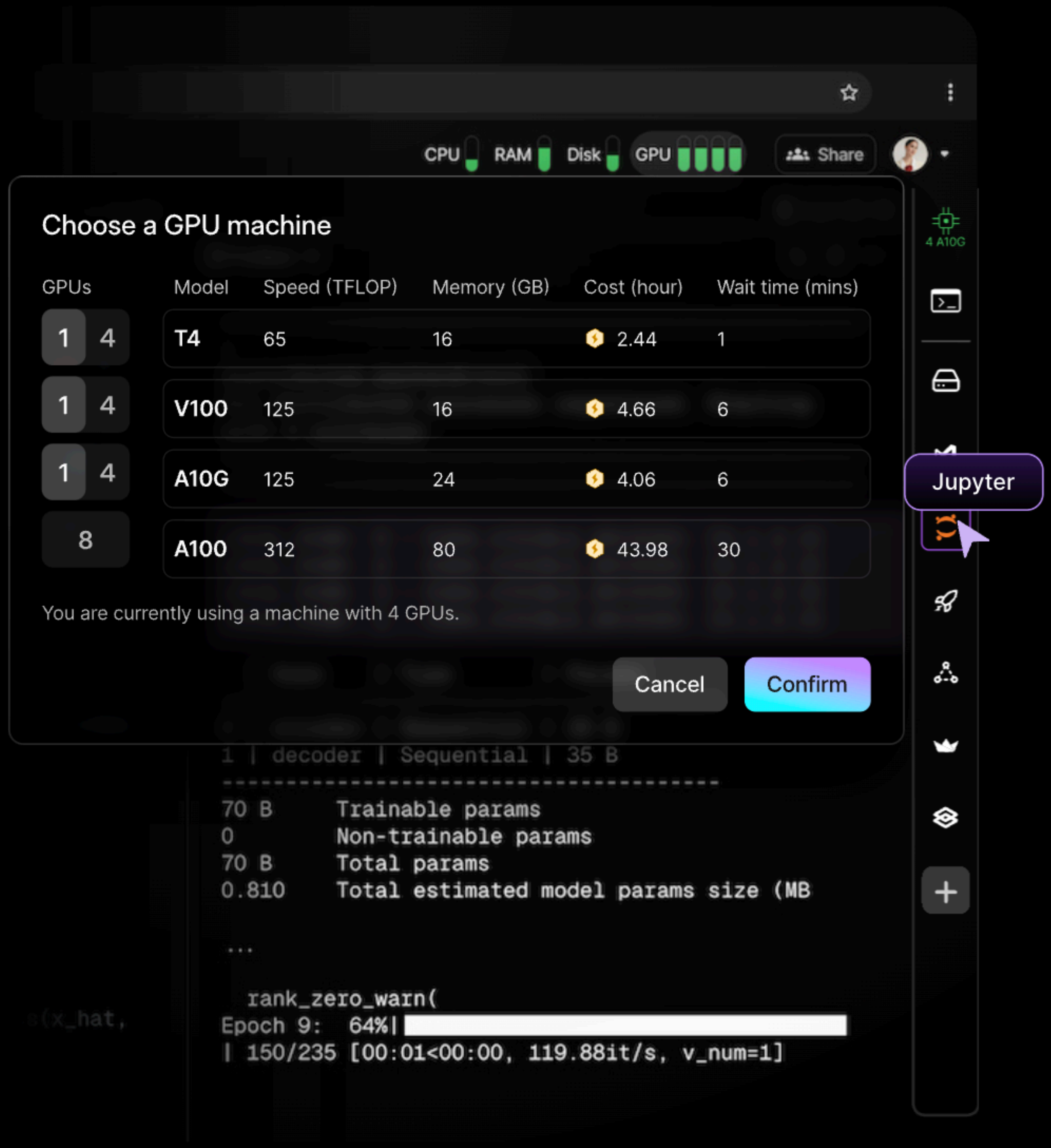
Lightning AI

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.



Choose a GPU machine

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

You are currently using a machine with 4 GPUs.

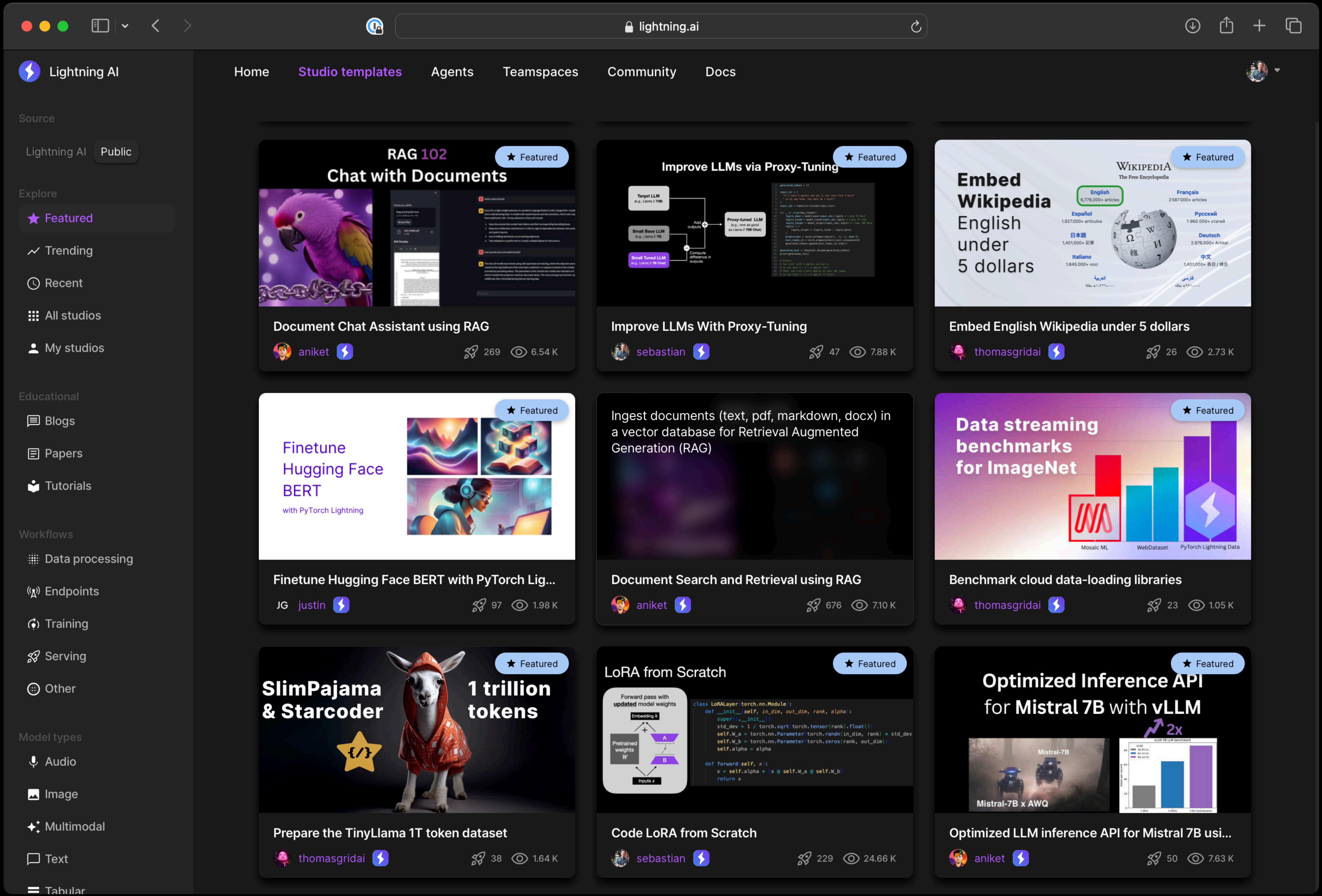
```
1 | decoder | Sequential | 35 B
-----------------------------------------
70 B      Trainable params
0         Non-trainable params
70 B      Total params
0.810     Total estimated model params size (MB

...

  rank_zero_warn(
Epoch 9:  64%|
| 150/235 [00:01<00:00, 119.88it/s, v_num=1]
```

- 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

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

Title	Author	Runs	Views
Document Chat Assistant using RAG	aniket	269	6.54 K
Improve LLMs With Proxy-Tuning	sebastian	47	7.88 K
Embed English Wikipedia under 5 dollars	thomasgridai	26	2.73 K
Finetune Hugging Face BERT with PyTorch Lig...	justin	97	1.98 K
Document Search and Retrieval using RAG	aniket	676	7.10 K
Benchmark cloud data-loading libraries	thomasgridai	23	1.05 K
Prepare the TinyLlama 1T token dataset	thomasgridai	38	1.64 K
Code LoRA from Scratch	sebastian	229	24.66 K
Optimized LLM inference API for Mistral 7B usi...	aniket	50	7.63 K

Featured
RAG 102
Chat with Documents
Improve LLMs via Proxy-Tuning
Embed Wikipedia English under 5 dollars
Wikipedia
The Free Encyclopedia
Finetune Hugging Face BERT
with PyTorch Lightning
Ingest documents (text, pdf, markdown, docx) in a vector database for Retrieval Augmented Generation (RAG)
Data streaming benchmarks for ImageNet
Mosaic ML
WebDataset
PyTorch Lightning Data
SlimPajama & Starcoder
1 trillion tokens
LoRA from Scratch
Optimized Inference API for Mistral 7B with vLLM
2x
Mistral-7B x AWQ

Contact

@rasbt in/sebastianraschka

https://sebastianraschka.com/contact/

https://lightning.ai