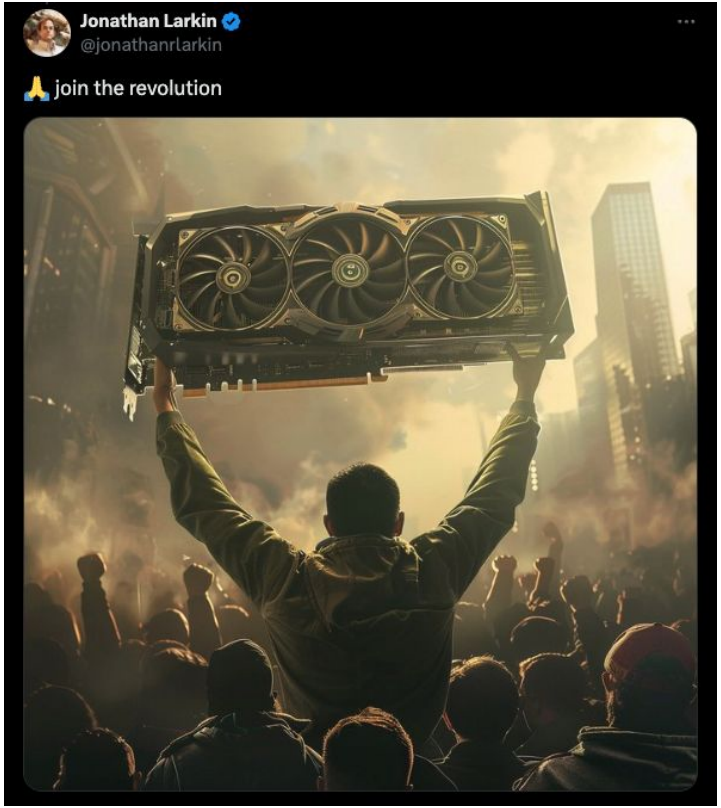
Mastering LLMs

A Conference For
Developers & Data Scientists

Growing Into Something Bigger

With world-class guest instructors:
- **Jeremy Howard:** Co-Founder Answer.AI & Fast.AI
- **Sophia Yang:** Head of Developer Relations, Mistral AI
- **Wing Lian:** Creator of Axolotl library for LLM fine-tuning
- **Simon Willison:** Creator of Datasette
- **Mark Saroufim**: PyTorch developer @ Meta
- **Paige Bailey:** DevRel Lead, GenAI, Google
- **Shreya Shankar:** LLMOps and LLM Evaluations researcher
- **Zach Mueller:** Lead maintainer of HuggingFace accelerate
- **Bryan Bischof:** Director of AI Engineering at Hex
- **Jason Liu:** Creator of Instructor
- **Abhishek Thakur** leads AutoTrain at HuggingFace.
- **Johno Whitaker:** R&D at AnswerAI
- **Charles Frye:** AI Engineer at Modal Labs
- **Eugene Yan:** Senior Applied Scientist @ Amazon
- **Harrison Chase**: CEO of LangChain
- **Travis Addair**: Co-Founder & CTO of Predibase
- **John Berryman:** Author of O'Reilly Book Prompt Engineering for LLMs
- **Joe Hoover**: Lead ML Engineer at Replicate
- **Ben Clavié**: R&D at AnswerAI



How To Be Successful With This Conference

- Tinker with the tools we show you -> Office Hours
- The importance of blogging. Tips.
- How to share your work (blogs, projects, etc)
  - Axolotl - @winglian, @axolotl_ai, @hamelhusain
  - Deepspeed/FSDP/Accelerate - @TheZachMueller
  - Modal - @charles_irl

Plan For Today

- What is Axolotl and how to use it to fine-tune model
  - Honeycomb example
  - Convo with Wing Lian
- Parallelism & HF Accelerate w/ Zach Mueller
- Fine-tuning on Modal
- Q&A

Modeling Choices

- **Base model**

- LoRA vs Full Fine Tune

Choosing a Base Model

Model Size

meta-llama/Llama-2-7b-hf
Text Generation • Updated Apr 17 • 1.1M • 1.43k

Note The base 7B model in HF transformers format

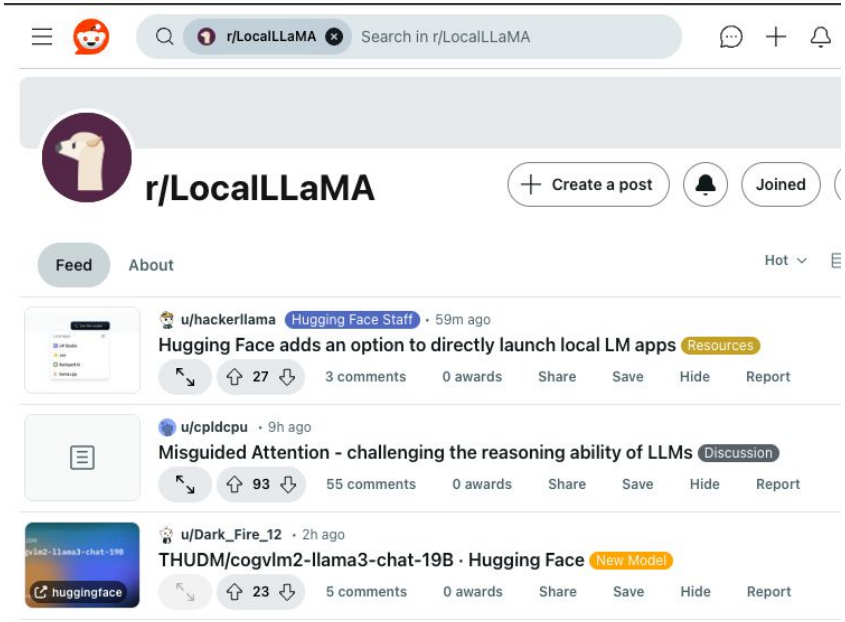
meta-llama/Llama-2-13b-hf
Text Generation • Updated Apr 17 • 319k • 542

Note The base 13B model in HF transformers format

meta-llama/Llama-2-70b-hf
Text Generation • Updated Apr 17 • 421k • 805

Note The base 70B model in HF transformers format

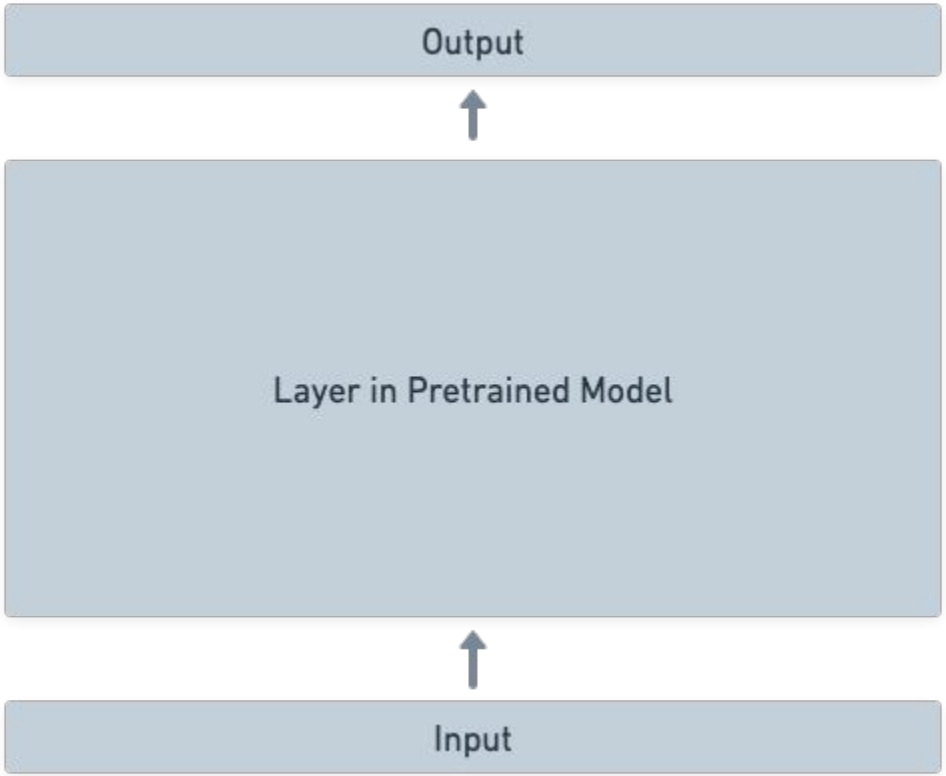
Model Family



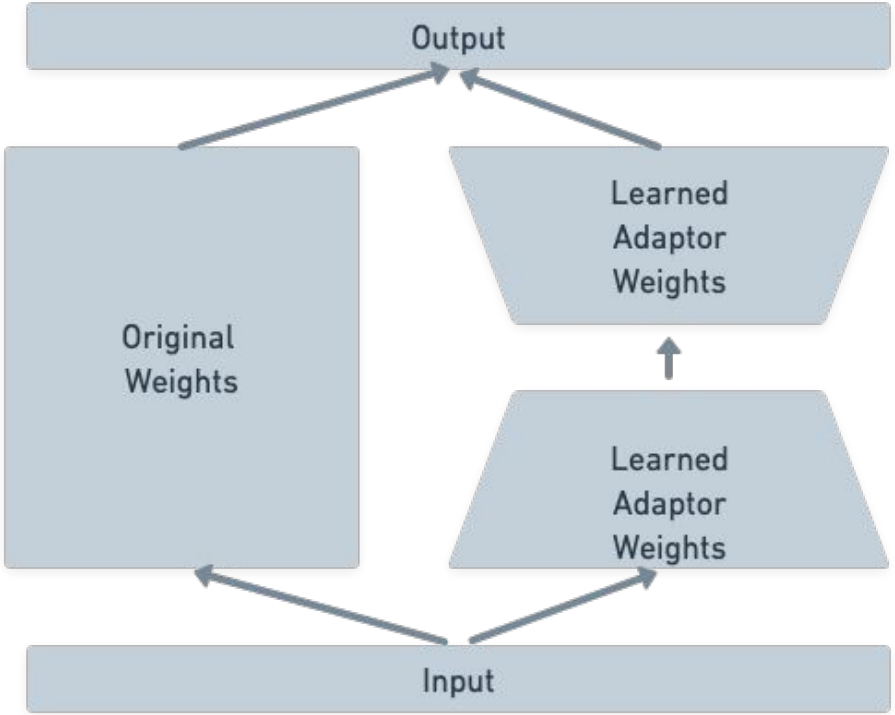
Modeling Choices

- Base model
- **LoRA vs Full Fine Tune**

LoRA In A Nutshell



LoRA In A Nutshell

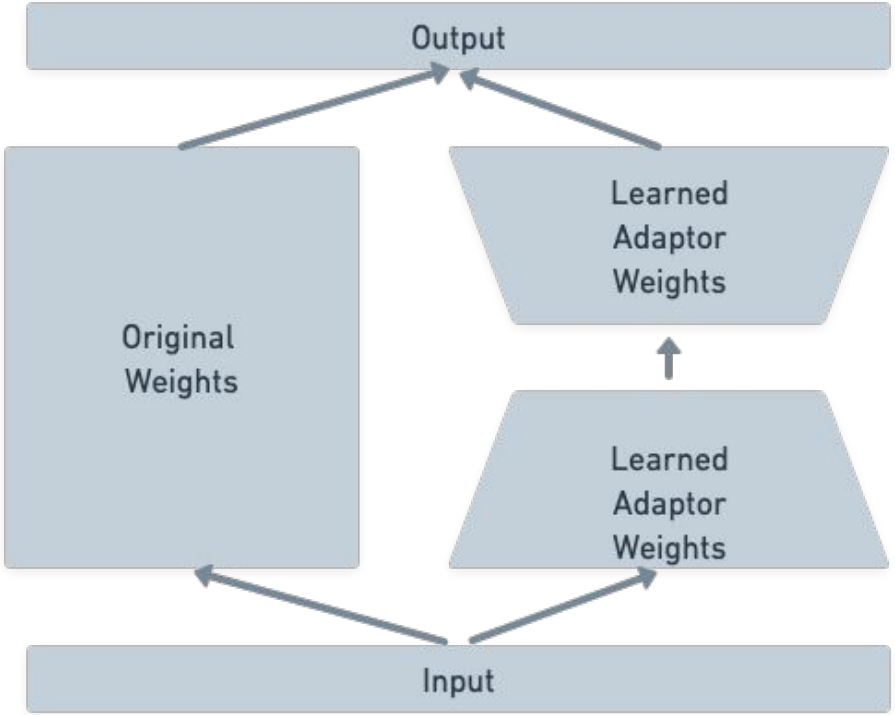


Input: 4000 dimensions

Output: 4000 dimensions

Original weights: 16M

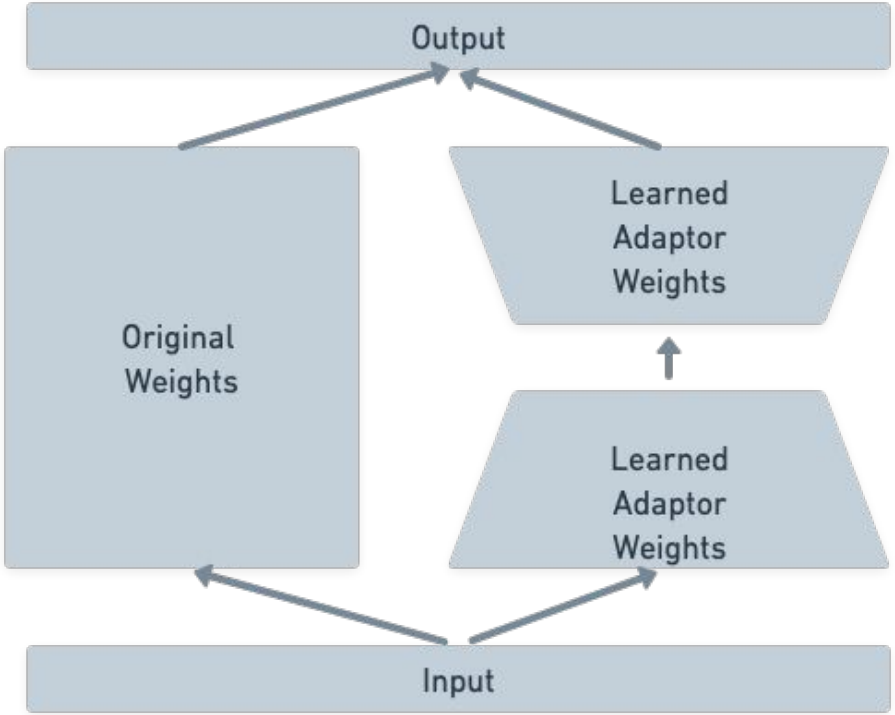
LoRA In A Nutshell



16 x 4000

4000 x 16

LoRA In A Nutshell



Input: 4000 dimensions

Output: 4000 dimensions

Original weights: 16M

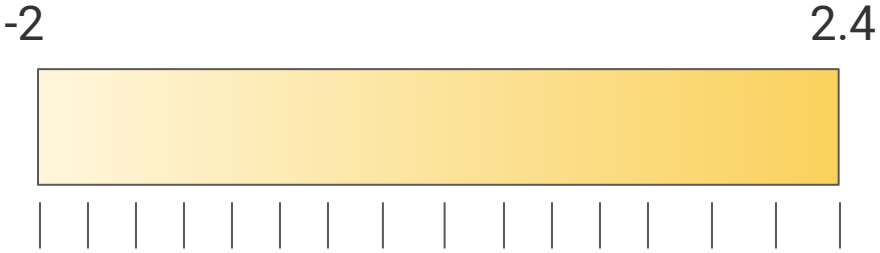
Adaptor rank: 16

LoRA weights: 2 * 16 * 4000

= 128,000

QLoRA

- LoRA at lower precision
- Memory savings with possible loss in quality




FIDDLE WITH HYPERPARAMETERS
IMPROVE YOUR DATA

What Is Axolotl

- Wrapper for Hugging Face tools
- Easy to use. So you can focus on your data
- Best practices built-in

Using Axolotl

github.com/OpenAccess-AI-Collective/axolotl

README | Code of conduct | Apache-2.0 license | Security

Phorm Ask AI

Table of Contents

Axolotl provides a unified repository for fine-tuning
a variety of AI models with ease

Go ahead and Axolotl questions!!

jquesnelle add save_only_model option (#1634) ✓ 702a669 · 2 days ago 1,448 Commits

.github	cloud image w/o tmux (#1628)	3 days ago
.vscode	feat: enable trl's autounwrap (#1060)	4 months ago
cicd	make sure to install causal_conv1d in docker (#1459)	2 months ago
deepspeed_configs	DBRX Model Support (#1462)	last month
devtools	Bootstrap Hosted Axolotl Docs w/Quarto (#1429)	2 months ago
docker	fix ray install (#1630)	3 days ago
docs	docs(config.qmd): add loraplus example (#1577)	2 weeks ago
examples	update outputs path so that we can mount workspace to /...	3 days ago
image	badge (#739)	7 months ago
scripts	more fixes to work with runpod + skypilot (#1629)	3 days ago

axolotl / examples / mistral / qlora.yml

winglian update outputs path so that we can mount workspace to /workspace/data (...

Code Blame 77 lines (66 loc) · 1.23 KB

```yaml
base_model: mistralai/Mistral-7B-v0.1
model_type: MistralForCausalLM
tokenizer_type: LlamaTokenizer

load_in_8bit: false
load_in_4bit: true
strict: false

datasets:
  - path: mhenrichsen/alpaca_2k_test
    type: alpaca
dataset_prepared_path: last_run_prepared
val_set_size: 0.1
output_dir: ./outputs/qlora-out

adapter: qlora
lora_model_dir:

sequence_len: 8192
sample_packing: true
pad_to_sequence_len: true

lora_r: 32
lora_alpha: 16
```

Datasets: mhenrichsen/**alpaca_2k_test** like 10

Tags: Croissant License: apache-2.0

Dataset card | Viewer | Files and versions | Community

Dataset Viewer

Auto-converted to Parquet | API | View in Dataset Viewer

Split (1)
train · 2k rows

Search this dataset

instruction string · *lengths* 17 193	input string · *lengths* 0 624	output string · *lengths* 1 2.88k	text string · *lengths* 184 3.1k
text.	Waterloo	central Belgium that is known…	describes a task, paired with a…
Who is the world's most famous painter?		The title of the world's most famous painter is subjective an…	Below is an instruction that describes a task. Write a…
Explain the concept of cogging torque.		Cogging torque is a torque that's produced by the…	Below is an instruction that describes a task. Write a…
Look up the boiling point of water.		The boiling point of water is 100 degrees Celsius or 212…	Below is an instruction that describes a task. Write a…
Describe the life and reign of King Charles II.		King Charles II was the monarch of England, Scotland, and…	Below is an instruction that describes a task. Write a…

```yaml
wandb_project:
wandb_entity:
wandb_watch:
wandb_name:
wandb_log_model:

gradient_accumulation_steps: 4
micro_batch_size: 2
num_epochs: 1
optimizer: adamw_bnb_8bit
lr_scheduler: cosine
learning_rate: 0.0002

train_on_inputs: false
group_by_length: false
bf16: auto
fp16:
tf32: false

gradient_checkpointing: true
early_stopping_patience:
resume_from_checkpoint:
local_rank:
logging_steps: 1
xformers_attention:
flash_attention: true
```

github.com/OpenAccess-AI-Collective/axolotl

README | Code of conduct | Apache-2.0 license | Security

Phorm Ask AI

Table of Contents

Axolotl provides a unified repository for fine-tuning
a variety of AI models with ease

Go ahead and Axolotl questions!!

```
# preprocess datasets - optional but recommended
CUDA_VISIBLE_DEVICES="" python -m axolotl.cli.preprocess examples/openllama-3b/lora.yml

# finetune lora
accelerate launch -m axolotl.cli.train examples/openllama-3b/lora.yml

# inference
accelerate launch -m axolotl.cli.inference examples/openllama-3b/lora.yml \
    --lora_model_dir="./lora-out"

# gradio
accelerate launch -m axolotl.cli.inference examples/openllama-3b/lora.yml \
    --lora_model_dir="./lora-out" --gradio
```

axolotl/examples/mistral/qlo mhenrichsen/alpaca_2k_tes env@d7492ffaa770

39x8o4f9ceor7i-19123-vxraclvcstcm50gh1y64.proxy.runpod.net

Finish update

```
root@d7492ffaa770:/workspace/demo#
```

[ssh_tmux]0:bash* "d7492ffaa770" 18:28 18-May-24

```
<start>
Below is an instruction that describes a task, paired with an input that provi
des further context. Write a response that appropriately completes the request
.

### Instruction:
{instruction}

### Input:
{input}

### Response:
{output}
<end>
```

Datasets: mhenrichsen/**alpaca_2k_test** like 10

Tags: Croissant License: apache-2.0

Dataset card | Viewer | Files and versions | Community

Dataset Viewer

Auto-converted to Parquet | API | View in Dataset Viewer

Split (1)
train · 2k rows

Search this dataset

instruction string · *lengths* 17 193	input string · *lengths* 0 624	output string · *lengths* 1 2.88k	text string · *lengths* 184 3.1k
text.	Waterloo	central Belgium that is known…	describes a task, paired with a…
Who is the world's most famous painter?		The title of the world's most famous painter is subjective an…	Below is an instruction that describes a task. Write a…
Explain the concept of cogging torque.		Cogging torque is a torque that's produced by the…	Below is an instruction that describes a task. Write a…
Look up the boiling point of water.		The boiling point of water is 100 degrees Celsius or 212…	Below is an instruction that describes a task. Write a…
Describe the life and reign of King Charles II.		King Charles II was the monarch of England, Scotland, and…	Below is an instruction that describes a task. Write a…

```
<start>
Below is an instruction that describes a task, paired with an input that provi
des further context. Write a response that appropriately completes the request
.

### Instruction:
{instruction}

### Input:
{input}

### Response:
{output}
<end>
```

(-100, 2899) that(-100, 369) appropri(-100, 6582) ately(-100, 1999) complet(-1
00, 2691) es(-100, 274) the(-100, 272) request(-100, 2159) .(-100, 28723) <0x0
A>(-100, 13) <0x0A>(-100, 13) ###(-100, 27332) Inst(-100, 3133) ruction(-100,
3112) :(-100, 28747) <0x0A>(-100, 13) rew(-100, 2516) rite(-100, 1967) this(-1
00, 456) sentence(-100, 12271) as(-100, 390) a(-100, 264) question(-100, 2996)
 <0x0A>(-100, 13) <0x0A>(-100, 13) ###(-100, 27332) Input(-100, 11232) :(-100,
 28747) <0x0A>(-100, 13) My(-100, 5183) mom(-100, 1948) made(-100, 1269) me(-1
00, 528) a(-100, 264) delicious(-100, 15992) dinner(-100, 7854) .(-100, 28723)
 <0x0A>(-100, 13) <0x0A>(-100, 13) ###(-100, 27332) Response(-100, 12107) :(-1
00, 28747) <0x0A>(-100, 13) Did(7164, 7164) your(574, 574) mom(1948, 1948) mak
e(1038, 1038) you(368, 368) a(264, 264) delicious(15992, 15992) dinner(7854, 7
854) ?(28804, 28804) </s>(2, 2)

Case Study

Honeycomb - NL to Query

User: “latency distribution by status code”

Input 1: NLQ

Candidate Fields:

['trace.trace_id', 'trace.span_id', 'trace.parent_id', 'duration_ms', 'name', 'faas.instance', 'faas.id', 'filter', 'telemetry.instrumentation_library', 'library.name', 'faas.name' ...

Input 2: Schema

GPT 3.5 w/prompt

Honeycomb Query

```
{'calculations': [{'column': 'duration_ms',
                   'op': 'MAX'}],
 'filters': [{'column': 'trace.parent_id',
              'join_column': '',
              'op': 'does-not-exist'}],
 'limit': 1,
 'orders': [{'column': 'duration_ms',
             'op': 'MAX',
             'order': 'descending'}],
 'time_range': 7200}
```

Honeycomb Case Study

https://github.com/parlance-labs/ftcourse

Debugging Axolotl

How-To Guides > Debugging

Debugging

How to debug Axolotl

This document provides some tips and tricks for debugging Axolotl. It also provides an example configuration for debugging with VSCode. A good debugging setup is essential to understanding how Axolotl code works behind the scenes.

Table of Contents

https://openaccess-ai-collective.github.io/axolotl/docs/debugging.html

Debugging Axolotl

General Tips

While debugging it's helpful to simplify your test scenario as much as possible. Here are some tips for doing so:

> [!Important] All of these tips are incorporated into the example configuration for debugging with VSCode below.

1. **Make sure you are using the latest version of axolotl**: This project changes often and bugs get fixed fast. Check your git branch and make sure you have pulled the latest changes from `main`.
2. **Eliminate concurrency**: Restrict the number of processes to 1 for both training and data preprocessing:
   - Set `CUDA_VISIBLE_DEVICES` to a single GPU, ex: `export CUDA_VISIBLE_DEVICES=0`.
   - Set `dataset_processes: 1` in your axolotl config or run the training command with `--dataset_processes=1`.
3. **Use a small dataset**: Construct or use a small dataset from HF Hub. When using a small dataset, you will often have to make sure `sample_packing: False` and `eval_sample_packing: False` to avoid errors. If you are in a pinch and don't have time to construct a small dataset but want to use from the HF Hub, you can shard the data (this will still tokenize the entire dataset, but will only use a fraction of the data for training. For example, to shard the dataset into 20 pieces, add the following to your axolotl config): `yaml dataset: ... shards: 20`
4. **Use a small model**: A good example of a small model is TinyLlama/TinyLlama-1.1B-Chat-v1.0.
5. **Minimize iteration time**: Make sure the training loop finishes as fast as possible, with these settings.
   - `micro_batch_size: 1`
   - `max_steps: 1`
   - `val_set_size: 0`
6. **Clear Caches:** Axolotl caches certain steps and so does the underlying HuggingFace trainer. You may want to clear some of these caches when debugging.
   - Data preprocessing: When debugging data preprocessing, which includes prompt template formation, you may want to delete the directory set in `dataset_prepared_path:` in your axolotl config. If you didn't set this value, the default is `last_run_prepared`.
   - HF Hub: If you are debugging data preprocessing, you should clear the relevant HF cache HuggingFace cache, by deleting the appropriate `~/.cache/huggingface/datasets/...` folder(s).
   - **The recommended approach is to redirect all outputs and caches to a temporary folder and delete selected subfolders before each run. This is demonstrated in the example configuration below.**

https://openaccess-ai-collective.github.io/axolotl/docs/debugging.html#general-tips

Questions For Wing

Zach Mueller
Accelerate / FSDP

Training On Modal

Why Modal

- Feels local, but its remote (“code in production”)
- Massively parallel
- Python native
- Docs: https://modal.com/

Things I’ve built with modal

- Transcript Summarizer
- W&B Webhook

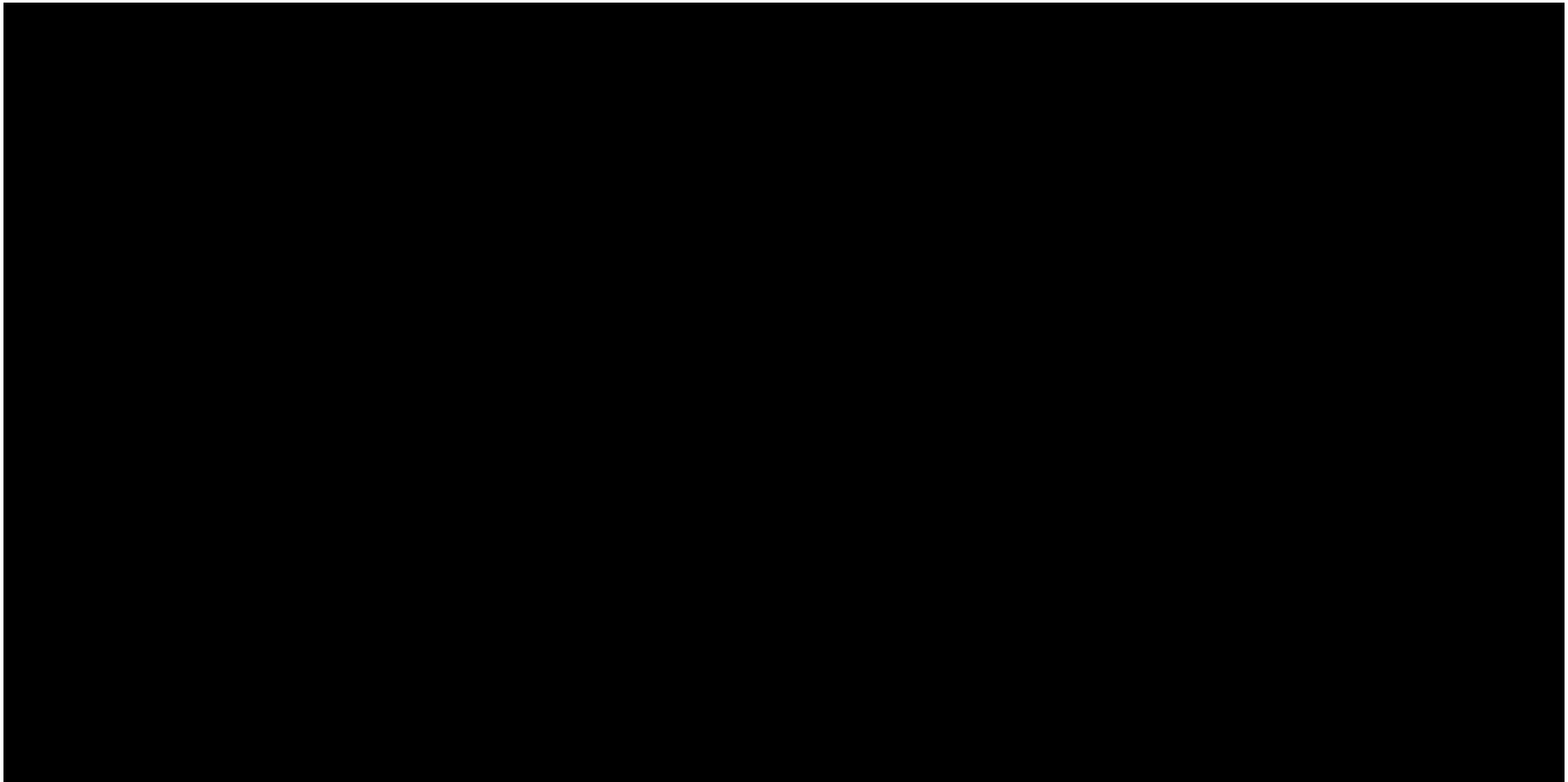
Modal - Axolotl Fine Tuning

https://github.com/modal-labs/llm-finetuning

Has additional defaults / some differences

- Merges LoRA back into the base model
- Use a –data flag instead of relying on the config
- Deepspeed config comes from the axolot repo that is cloned

Modal - Axolotl Fine Tuning



Modal - Debug Data

https://github.com/modal-labs/llm-finetuning/blob/main/nbs/inspect_data.ipynb

Tip: replace github.com with nbsanity.com to view notebooks

Q & A