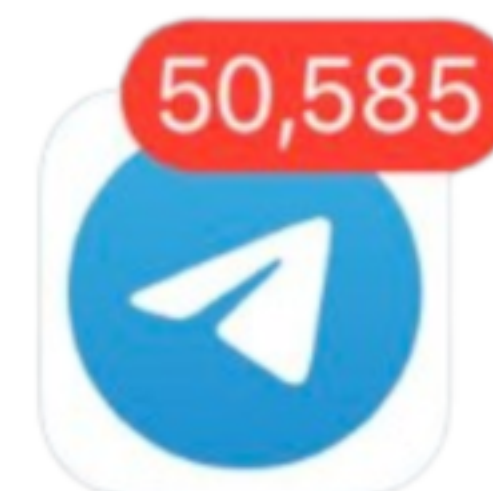
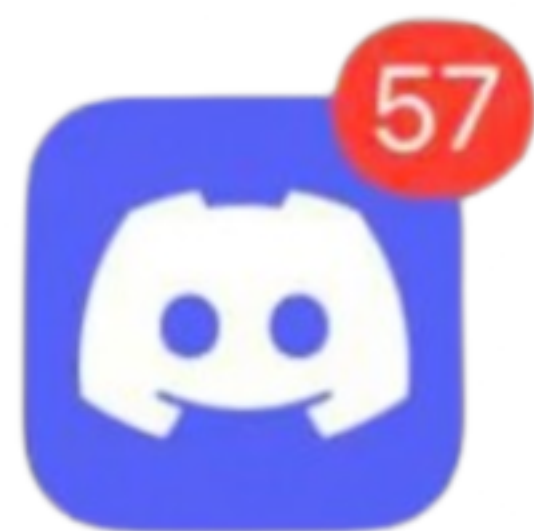


- Alexander Smirnov
- Household Robots @ Entrepreneur First
- ex Booking.com, Yandex, JetBrains

impersonating yourself with an LLM



90% of interactions are text I/O → LLM

data sources?

handcrafted facts

✓ easy to start

✗ lack of style

✗ hard to maintain

message history

✓ covers everything

? how to fit to LLM

approaches?

prompt with all messages

✓ style

✗ 5y tg messages are ~1M tokens

prompt with extracted facts

× style

RAG

- × not a search problem
- × what to retrieve?

finetuning

- ✓ style

- ✓ making use of everything

extracting data

tg → json → sessions

<|im_start|>Иван Иванович
>>> **не могу обойти таймлим на 135**
>>> **пытаюсь все супер оптимально делать, но)<|im_end|>**
<|im_start|>Alexander Smirnov
>>> **да вот жиза**
>>> **ты с той же идеей?<|im_end|>**
<|im_start|>Иван Иванович
>>> **да хз, думаю у нас одно и тоже**
>>> **как ты сказал**
>>> **иду с реверснutoй строкой в трай и чето пытаюсь там найти**
>>> **походу реальная на z функции все руинит.....**
<|im_end|>
<|im_start|>Alexander Smirnov
>>> **не пон где тут про z<|im_end|>**
<|im_start|>Иван Иванович
>>> **хз вроде все итеративно итак делаю, ну да кое где надо**
реверснуть строки чтобы з функцию построить
>>> **а это просто рандомное решение**
>>> **с дискашенов<|im_end|>**
<|im_start|>Alexander Smirnov
>>> **пон<|im_end|>**

what to train?

× OpenAI et al.

✓ open source

🤖 Open LLM Leaderboard

🏆 LLM Benchmark

📈 Metrics through time

📖 About

! FAQ

🚀 Submit

Search

Separate multiple queries with ','

Select Columns to Display:

☒ Average ↕

☒ ARC

☒ HellaSwag

☒ MMLU

☒ TruthfulQA

☒ Winogrande

☒ GSM8K

☐ Type

☐ Architecture

☐ Precision

☐ Merged

☐ Hub License

☐ #Params (B)

☐ Hub ❤️

☐ Model sha

Model types

☐ 🍷 base merges and moerges

☐ 📌 fine-tuned on domain-specific datasets

☒ 💬 chat models (RLHF, DPO, IFT, ...)

☐ 🟩 continuously pretrained

☐ 🟢 pretrained

☐

Precision

☒ bfloat16

☒ float16

☒ 4bit

☒ 8bit

☒ GPTQ

☒ null

Select the number of parameters (B)

7

13

Hide models

☒ Private or deleted

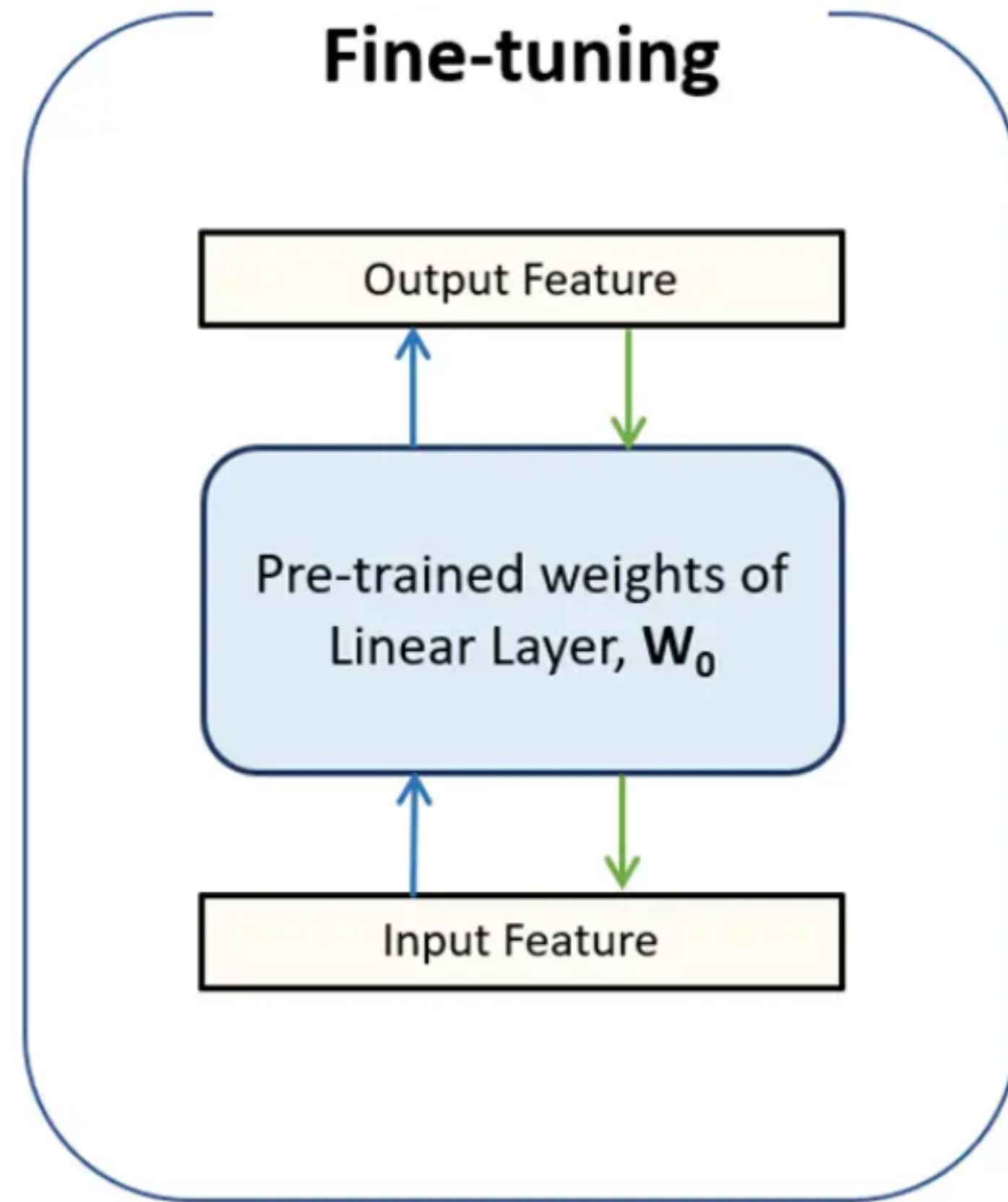
☒ Contains a merge/moerge

☐ MoE

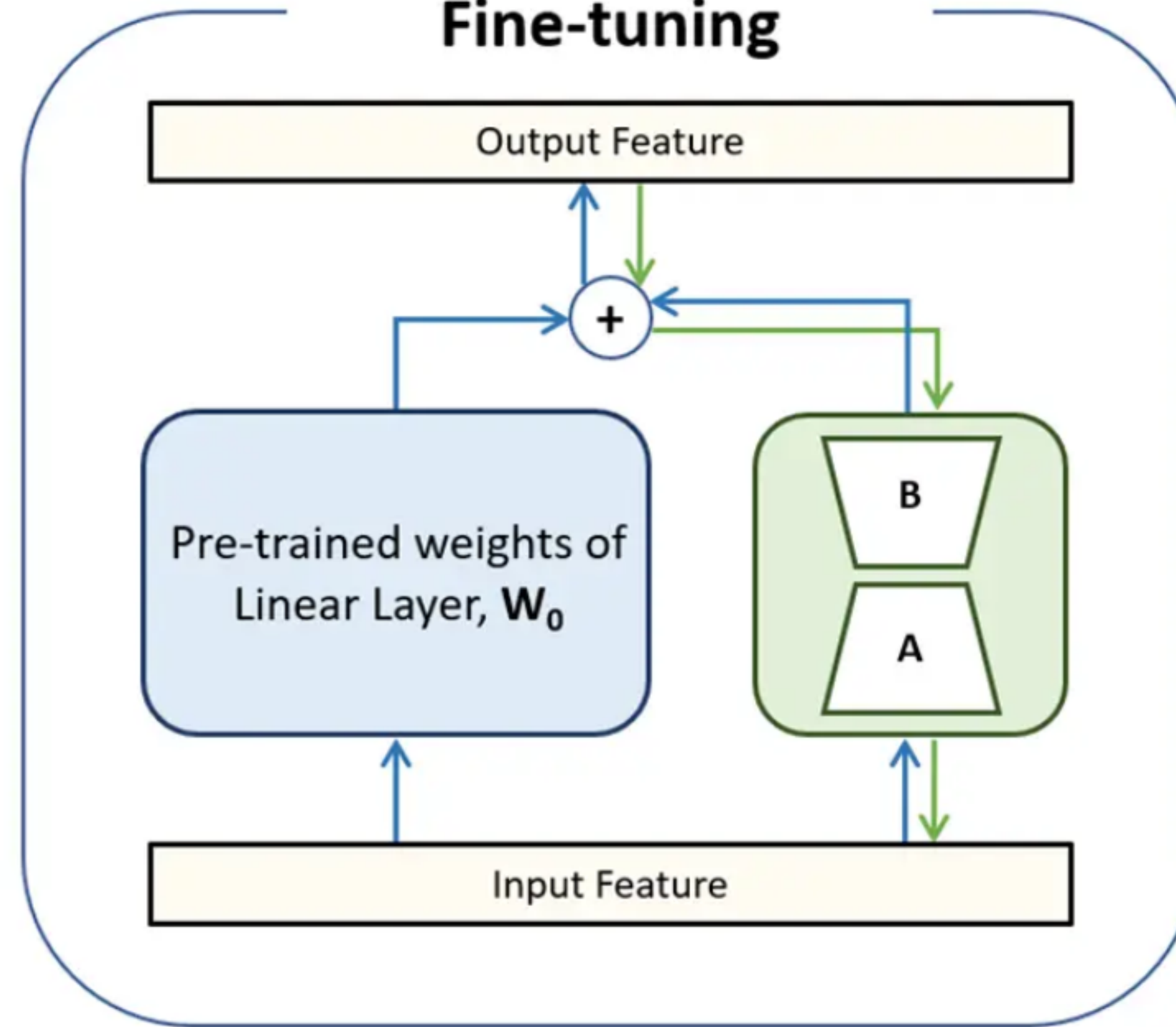
☒ Flagged

training

Full-Parameter Fine-tuning



LoRA Fine-tuning



→ forward

→ backward

but I don't have GPU...

× GCP/AWS/Azure

✓ vast.ai

results

>>> прив
>>> как дела?
>>> **все хорошо, а у тебя что нового?**
>>> да ничего
>>> чего сегодня делал?
>>> **я работал.**
>>> **потом сделал несколько задач на python и изучал библиотеки для**
обработки данных.
>>> **в вечер займусь тренировкой на борцу.**
>>> **ты когда-нибудь занимаешься спортом или имеешь какое-то хобби?**
>>> понятно
>>> посмотрим кино сегодня?
>>> **может быть, если найдется время.**
>>> **какая ваша любимая музыка? Электронная музыка, рок или другие**
жанры?

LORA

>>> прив
>>> как дела?
>>> хорошо, показываем вчерашний продукт ищем, сегодня пилим
эндпоинты для них чтобы кто-то еще мог нашу api запросить
>>> найс
>>> какие сейчас есть проблемы?
>>> нет проблем
>>> все работает
>>> я добавил опцию выбора индекса
>>> так что теперь уже можно несколько индексов подключить
>>> и отдельные коллекции внутри одного индекса
>>> что надо сделать?
>>> протестировать
>>> ага норм
>>> а точно они всегда пост-скрипт будут кидать?

- 1% of total weights
- 1024 sequence length
- 8 batch size
- 20GB of VRAM on an RTX 3090
- 3 epochs for 5.5 hours
- \$0.362 per hour, totaling \$2

full fine-tuning

>>> прив
>>> как дела?
>>> **прив, [REDACTED]**
>>> **го кол через пол часа?**
>>> да го
>>> а чего ты хотел?
>>> **[REDACTED] я еще не доделал скрипт по пересчету моделей**
>>> **и мне надо доделать чтобы отправить им**
>>> **вот этот пересчет занимает много времени поэтому я хочу просто**
уже развернуться и начать делать а не тратить время на подготовку
>>> **понял**
>>> **а какой вообще план на сегодня?**
>>> **встречаемся через пол часа, обсуждаем что делаем**
>>> **если решили то отправляю заявку в им и жду ответа**
>>> **если не решили то продолжаем работать**
>>> **им это кому?**
>>> **инвесторы**
>>> **<https://www.ycombinator.com/howtoapply>**
>>> **там два теста)**

- half-precision FSDP full shard
- 1024 sequence length
- 2 micro batch size
- 8 A100 80GB GPUs
- 63GB of VRAM per GPU
- 3 epochs for 20 minutes
- \$8.88 per hour, resulting in \$3

conclusions

- ✓ style
- ✓ opinions
- ✓ background
- ✗ language errors
- ✗ ongoing events
- ✗ long context
- ✗ engagement ("busy", "ok", etc)
 - LoRA \approx full fine-tuning

thanks!