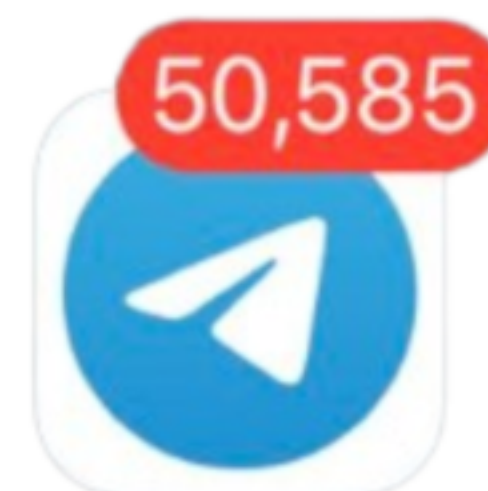
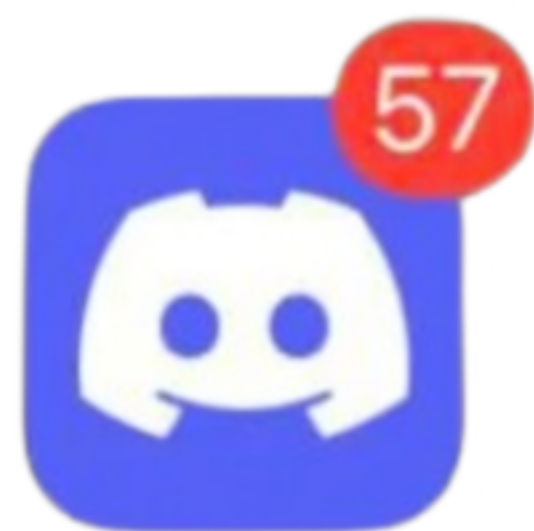


- Alexander Smirnov
- Household Robots @ Entrepreneur First
- ex Booking.com, Yandex, JetBrains

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|>John Smith
>>> damn, can't get around the 135 time limit
>>> trying to do everything super optimally, but no luck<|im_end|>
<|im_start|>Alexander Smirnov
>>> yeah same
>>> you still going with the same idea?<|im_end|>
<|im_start|>John Smith
>>> dunno, I think we're on the same page
>>> as you said
>>> going with the reversed string in a try and trying to find something there
>>> seems like real shit because z function ruins everything.....<|im_end|>
<|im_start|>Alexander Smirnov
>>> don't get where z comes into this<|im_end|>
<|im_start|>John Smith
>>> dunno seems like I'm doing everything iteratively anyway, but yeah gotta reverse some strings to build the z function
>>> and it's just a random solution
>>> from discussions<|im_end|>
<|im_start|>Alexander Smirnov
>>> got it<|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

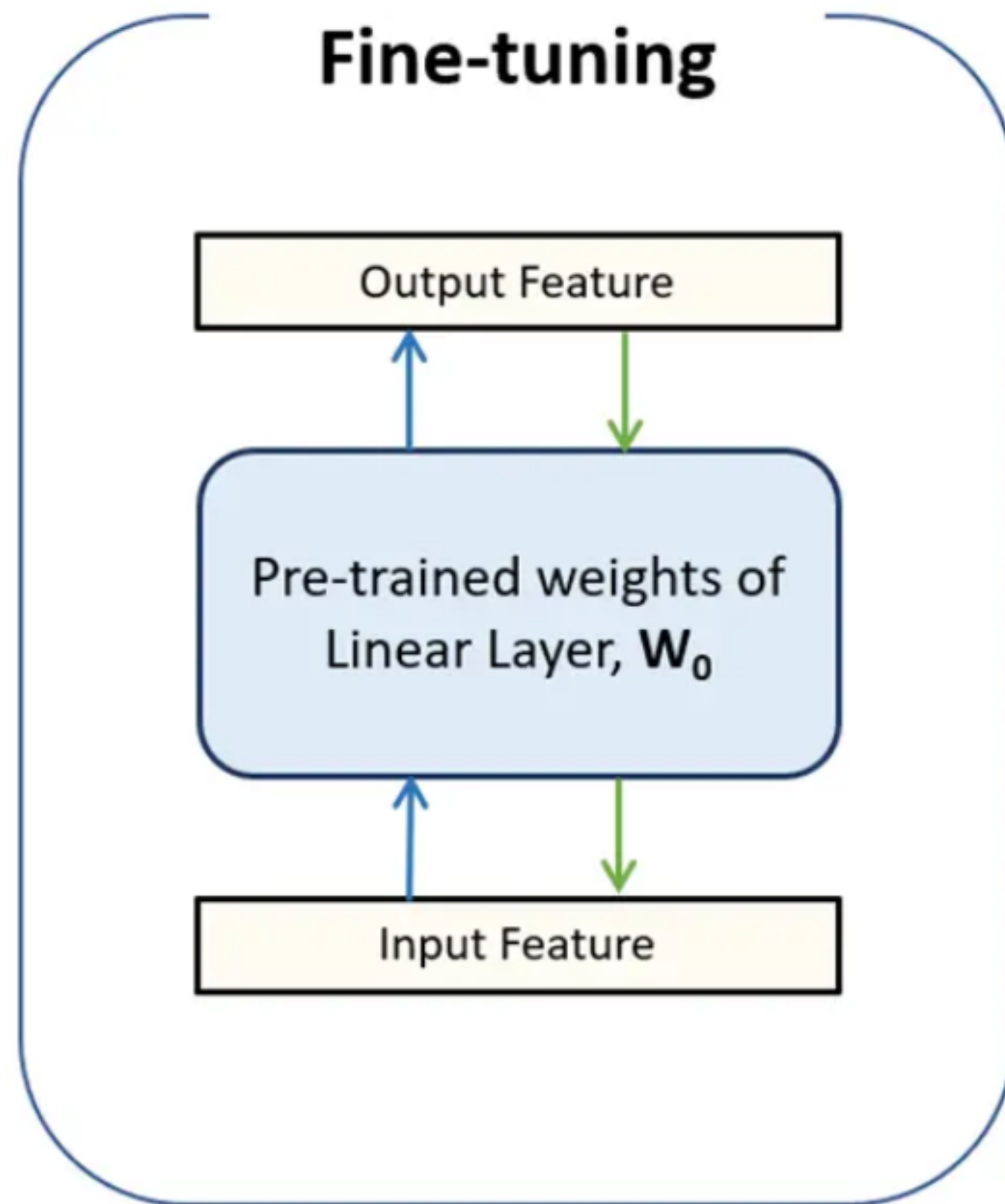
☒ Contains a merge/moerge

☐ MoE

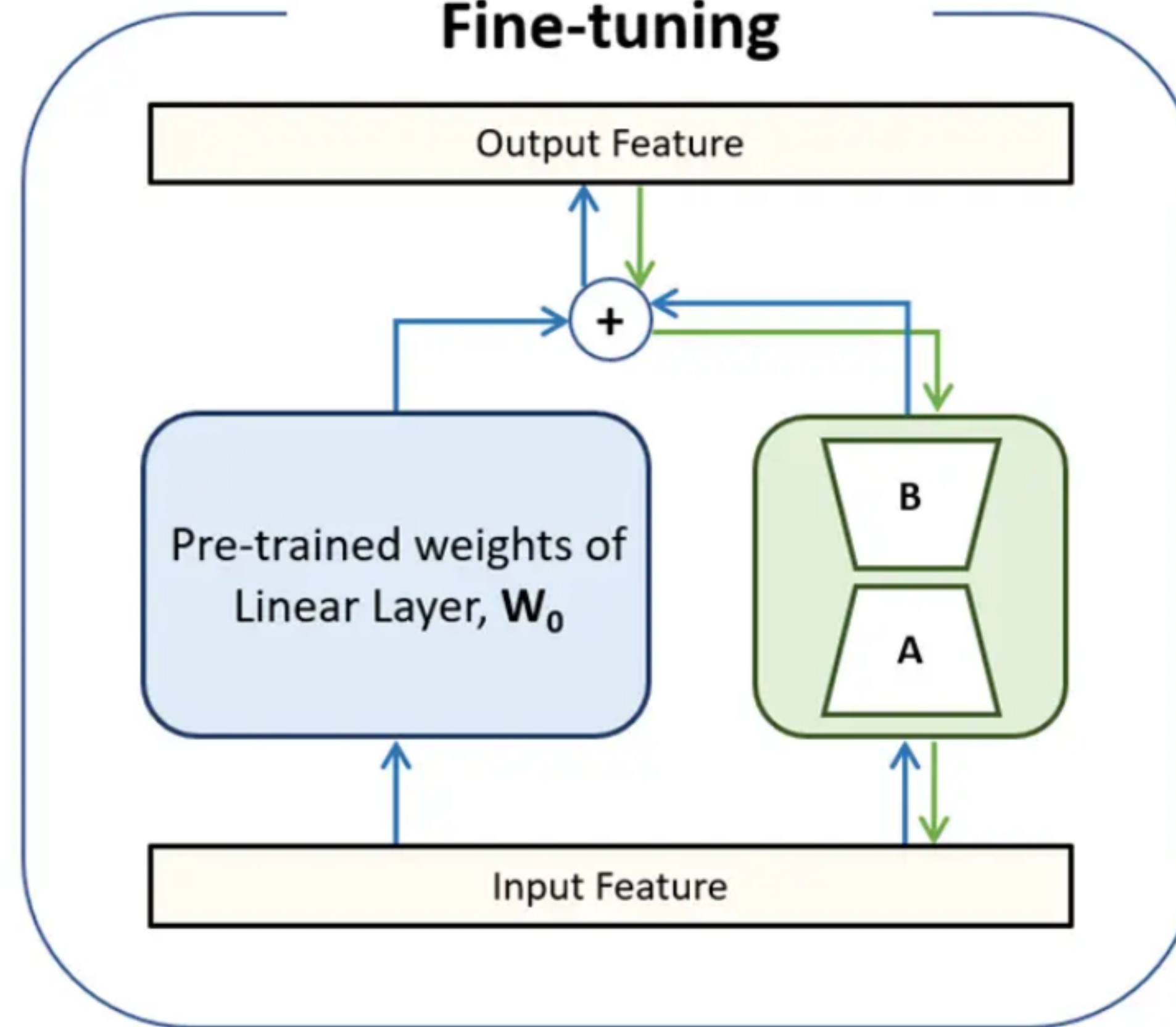
☒ Flagged

training

Full-Parameter Fine-tuning



LoRA Fine-tuning



→ forward

→ backward

but I don't have GPU...

× GCP/AWS/Azure

✓ vast.ai

results


```
>>> hey
>>> what's up?
>>> I'm good, what about you?
>>> nothing much
>>> what did you do today?
>>> I worked.
>>> then I did a few tasks in Python and studied data processing
libraries.
>>> in the evening I'll do wrestling training.
>>> have you ever been into sports or do you have any hobbies?
>>> I see
>>> want to watch a movie today?
>>> maybe, if I find the time.
>>> what's your favorite music? Electronic, rock, or other genres?
```


LORA

```
>>> hey
>>> what's up?
>>> ok, we show yesterday's product to searchcom, today we write
endpoints for them so that someone else can request our api
>>> cool
>>> what problems do we have now?
>>> no problems
>>> everything is working
>>> I added an option to choose index
>>> so now we can attach several indexes
>>> and separate collections within single index
>>> what we have to do?
>>> test
>>> yes ok
>>> are they always will be sending post-script?
```

- 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


```
>>> hey
>>> what's up?
>>> hi, good
>>> let's meet in half an hour?
>>> yep
>>> what did you want?
>>> damn I haven't finished the script for recalculating models yet
>>> and I need to finish it to send it to them
>>> this recalculation takes a lot of time, so I just want to turn
around and start doing it and not waste time on preparation
>>> got it
>>> what is the plan for today?
>>> meet in half an hour to discuss what we are doing
>>> if decided, then I send an application to them and wait for a
response
>>> if we haven't decided, then we continue to work
>>> who do you mean by them?
>>> investors
>>> https://www.ycombinator.com/howtoapply
>>> there are two tests)
```

- 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

asmirnov.xyz