



To train or not to train your LLM

How to strike the right balance

Oren Razon, CO-Founder & CEO @ Superwise | oren.razon@superwise.ai | [linkedin/oren-razon](https://www.linkedin.com/in/oren-razon)

Gad Benram, Founder & CTO @ TensorOps | gad@tensorops.ai | [linkedin/gad-benram](https://www.linkedin.com/in/gad-benram)

About us



Gad Benram

Founder & CTO @ **TensorOps**



Oren Razon

CO-Founder & CEO @ **Superwise**



Jose Bastos

AI Engineer @ **TensorOps**



Model observability

built for scale

We empower data science, ML engineering, and operational teams with visibility and control to **scale AI activities**



The AI Experts

We simply help machines learn

We build end-to-end AI solutions for businesses; Specializing in time-series forecasting, search, OpenAI and Google Cloud.

Klarna.



monday.com

riskified

Fundbox

Panaya



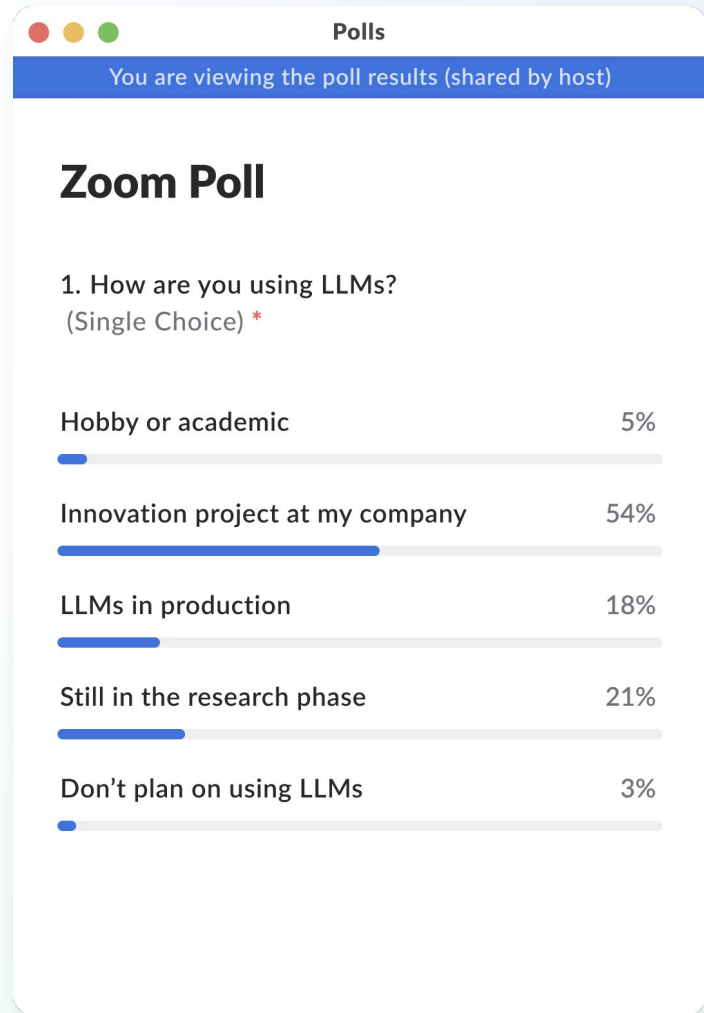
onebeat



Agenda

- Typical use cases
- LLM types
- To train or not to train?
- Pros & cons
- Types of training/tuning
- Monitoring LLMs
- Takeaways & going beyond tuning

How are you using LLMs?



Typical LLM use cases

Tasks

- Chatbots
- Embedding
- Organizational knowledge
- Productivity
- Code completion
- Marketing
- Raising funding in 2023

Real-world

- Jasper
- Semantic search
- Google gen app builder
- Notion
- Co-Pilot
- SlidesAI.io
- Mistral AI

Types of LLMs

API

No or limited visibility into code and data, but easier to deploy out of the box



Open source

Offer the most transparency and flexibility but require time and expertise to set up



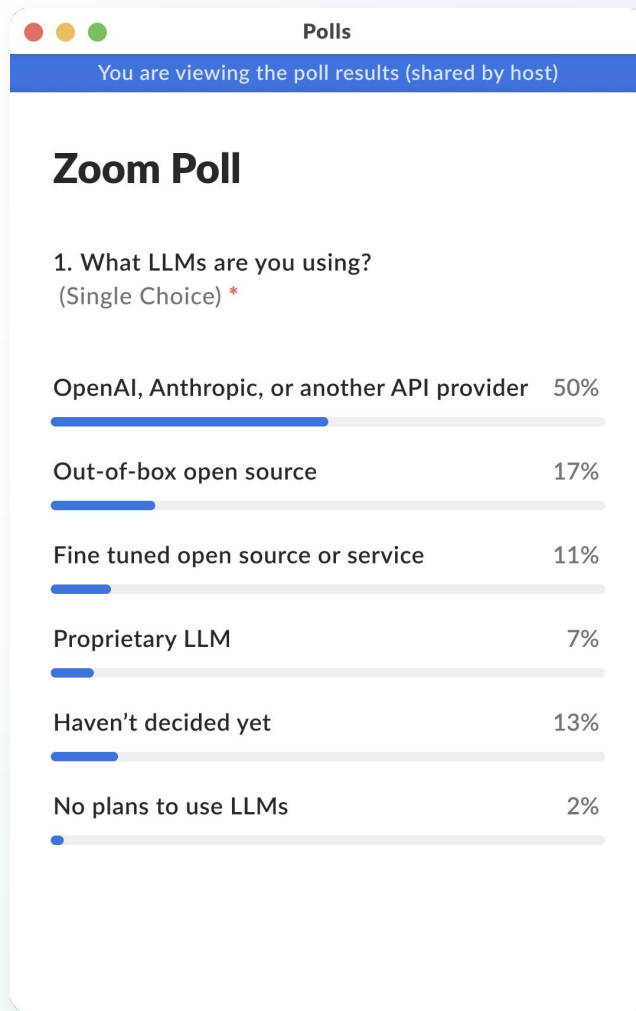
LLaMA

Proprietary

Built in-house from scratch for the organization

Bloomberg

What LLMs are you using?



**Out of
the box**

vs

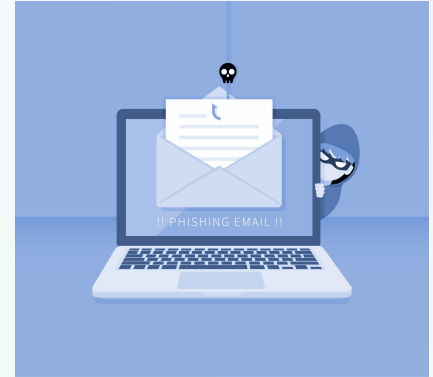
Tuned

LLMs in the wild

Real world use-case

Phishing email detector

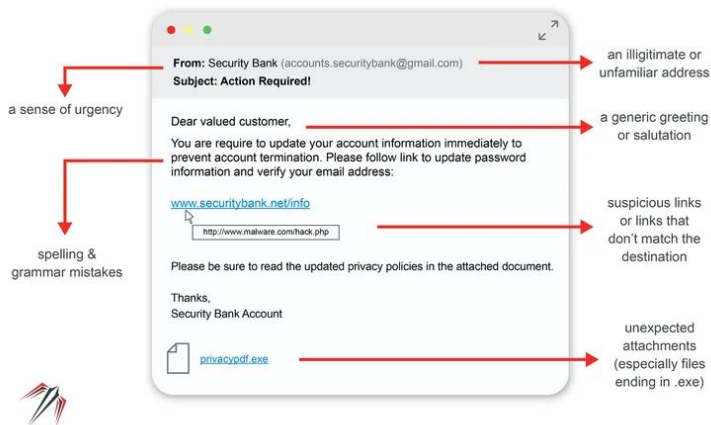
- MailProtect is a (made up) company that offers email protection service for large organizations
- They have a solution for detecting and blocking phishing emails



LLMs in the wild

Cyber LLM use-cases

WATCH OUT FOR...



Classical ML

Deals with anomaly detection and tabular data.

LLMs improve textual pipelines

Improve textual pipelines that rely on classification, segmentation, generation, interpretation.

LLMs in the wild

**ChatGPT
works!**

J

Classify this email as phishing or not. Return only yes or no.

From: fms@cinci.rr.com

Sent: Wednesday, June 21, 2023 11:40 AM

To: Me

Subject: Invoice input

Good Morning,

Called you a few times without success. Decided to reach by email. I need to know the status of this invoice bellow, it's way past due. <https://apgproperty.ca/invoice-number-747853>

Warmest Regards,

merc@merconstructioninc.com

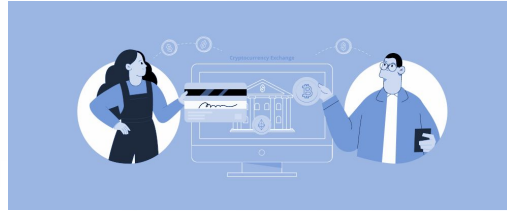


Yes.



LLMs in the wild

ChatGPT works?



Classify this email as phishing or not .
Return only a json with the following structure {"is_phishing": BOOL}

Dear Finance team,
We recently submitted a transaction from account
1CGpXZ9LLYwi1baonweGfZDMsyA35sZXCW and it seems like there was an issue
processing this request.
Is it possible to cancel the transaction and return the money to my wallet at
1LNL2n9RGf3oKctR1kWrrLojxxDbRdujS2?



```
{"is_phishing": true}
```

Tuning in the wild

Let's look
at an example from
a financial use case

J

Classify this email as phishing or not. Return only yes or no.

From: fms@cinci.rr.com

Sent: Wednesday, June 21, 2023 11:40 AM

To: Me

Subject: Invoice input

Good Morning,

Called you a few times without success. Decided to reach by email. I need to know the status of this invoice bellow, it's way past due. <https://apgproperty.ca/invoice-number-747853>

Warmest Regards,

merc@merconstructioninc.com



Yes.

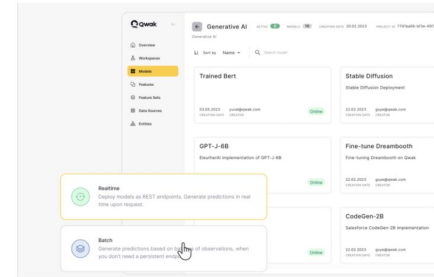


Demo:

Does tuning really work?

Qwak

Deploy Your Generative AI Models with Qwak



Using Qwak Model

FLAN T5

Ask Qwak:

What is an index fund?

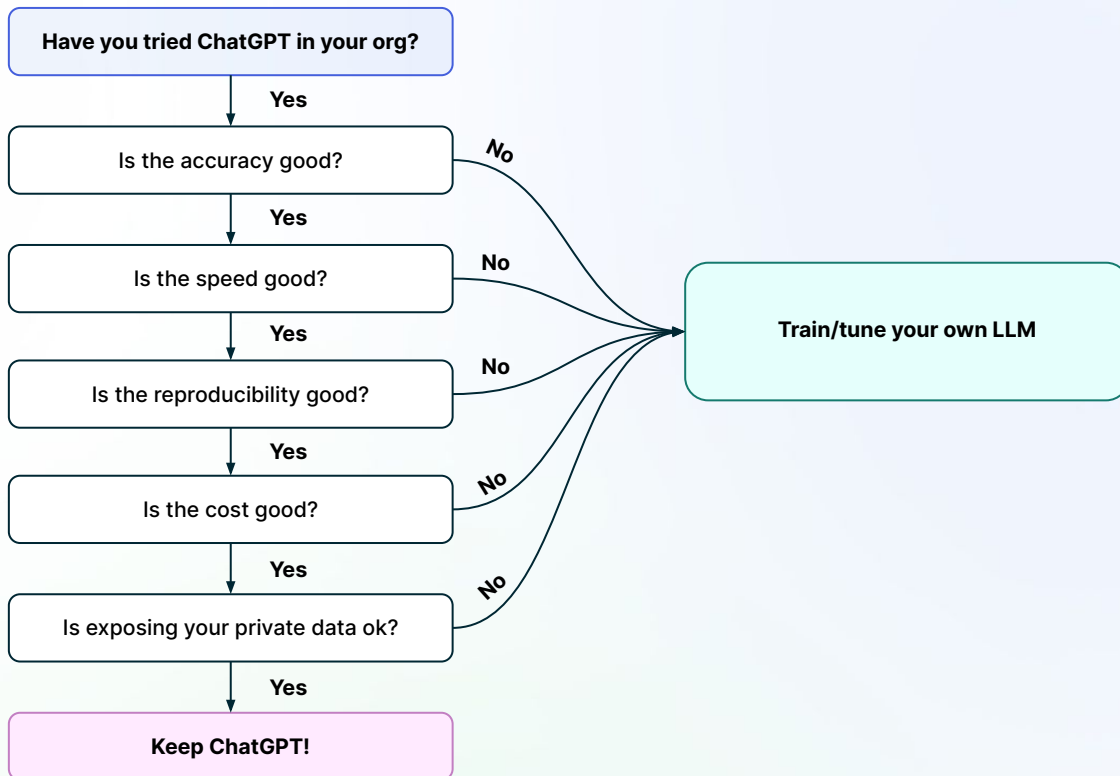
What is an index fund?



a fund

Do you even need tuning?

Have you taken
ChatGPT to
production?



Pros and cons of training LLMs

Pros

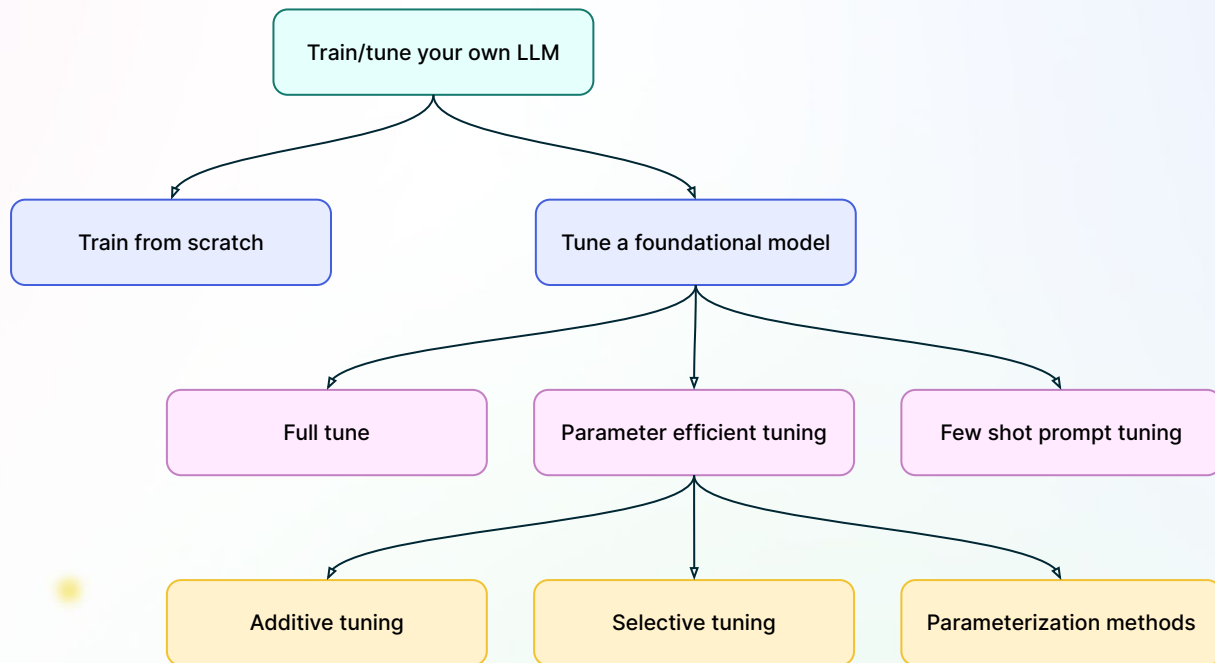
- Better accuracy
- Potential to reduce inference cost
- Control data
- Control infrastructure

Cons

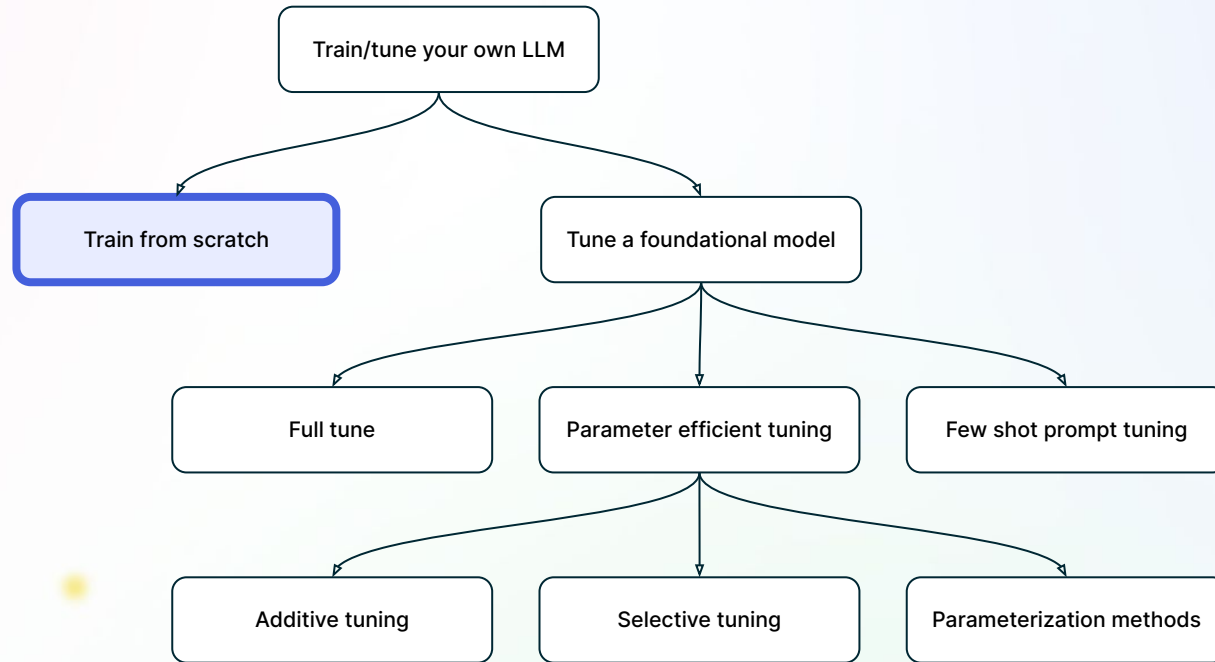
- Data leakage
- Engineering complexity
- Reduced performance
- Cost

What does
“training an LLM”
even mean?

To train or not to train **your LLM**?



Train from **scratch**



Train from scratch

Need huge amounts of data and compute power and takes some months to do it

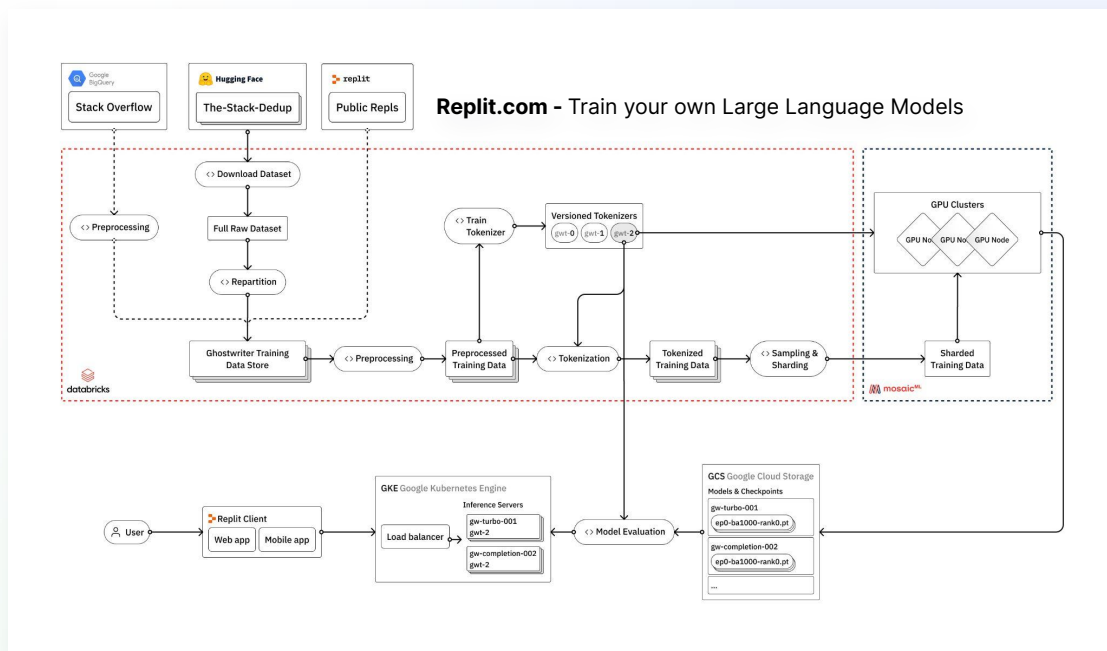
Pros:

- Best fit (?)
- Full ownership (model + data)

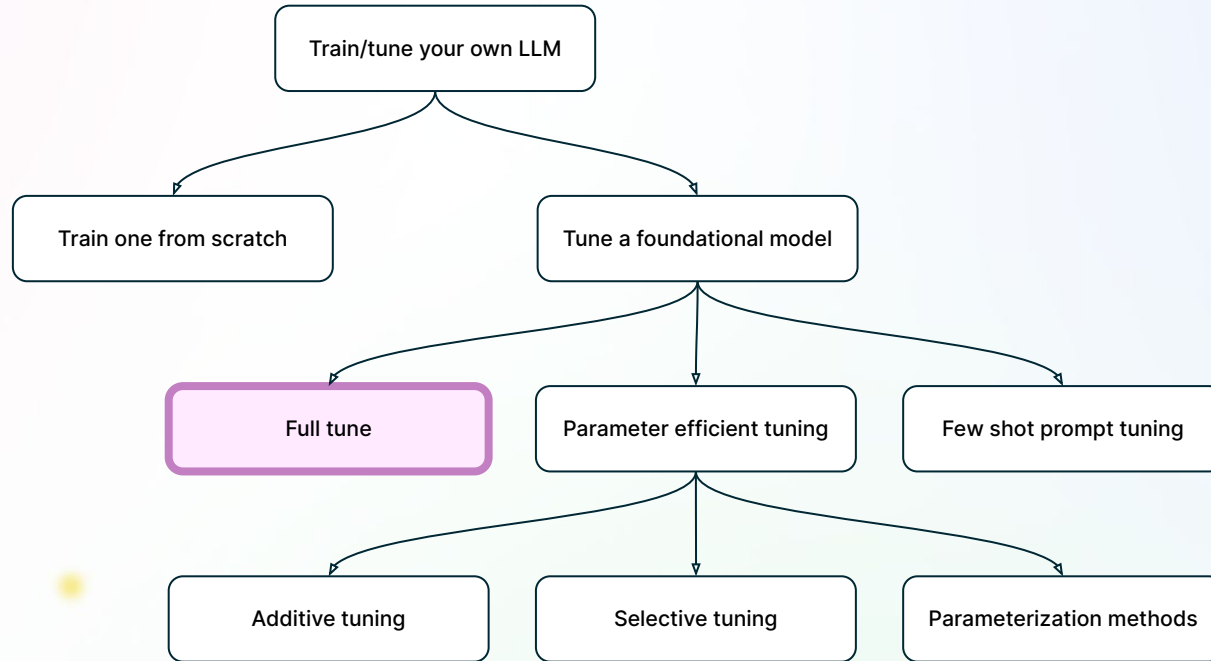
Cons:

- **Huge** amounts of data
- **Huge** amounts of compute power
- Takes **months**

➔ **Huge costs!**



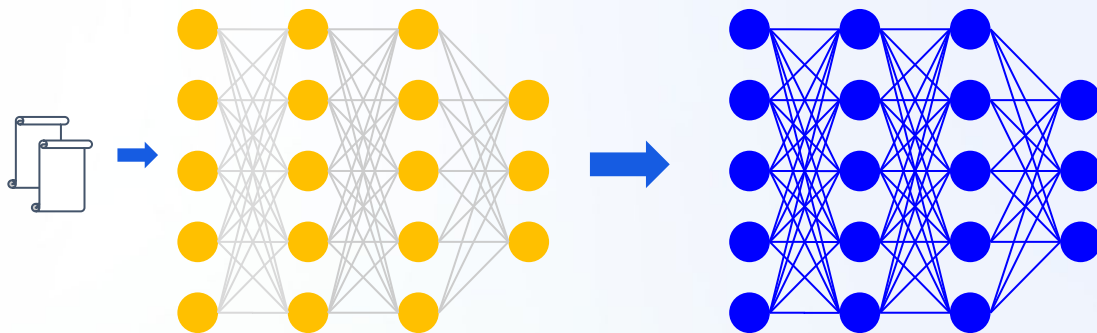
Full tune



Full tune

Fully tune an existing LLM

Customize an existing model with your specific use cases. Gradually change the model by feeding it examples of the expected behavior



Pros:

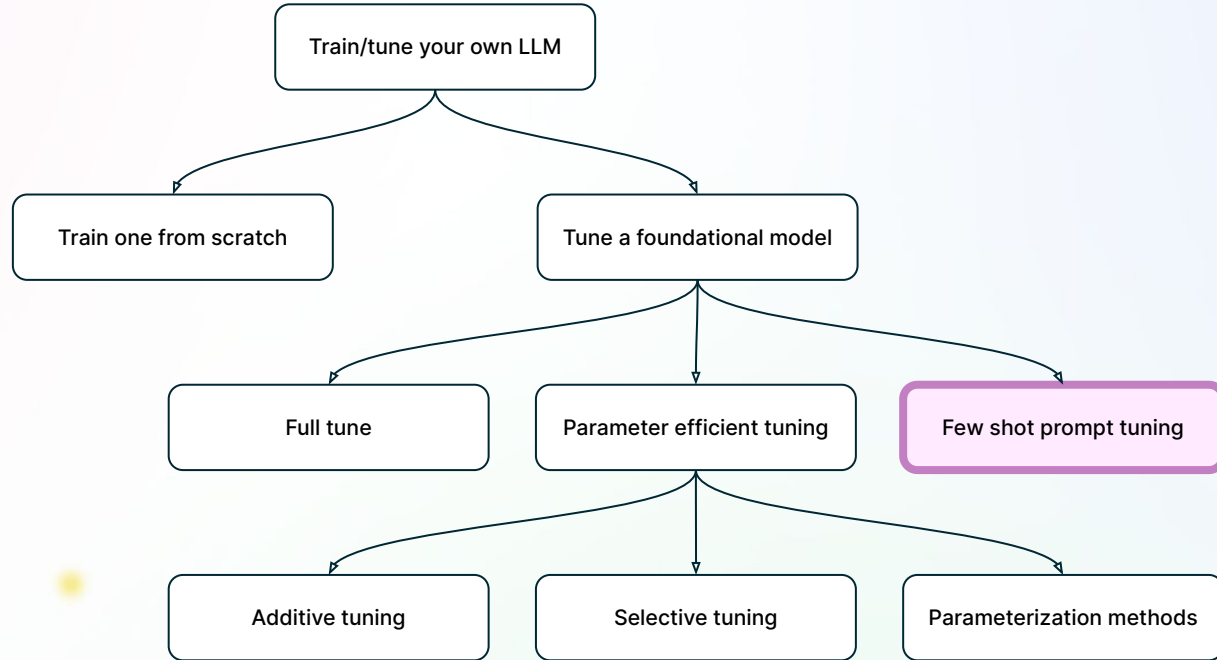
- Small downgrade in accuracy*
- Almost full ownership (model + tuning data)

Cons:

- **Moderate** amounts of data
- **Moderate** amounts of compute power
- Takes **days/weeks**

* For a considerable amount of tuning data

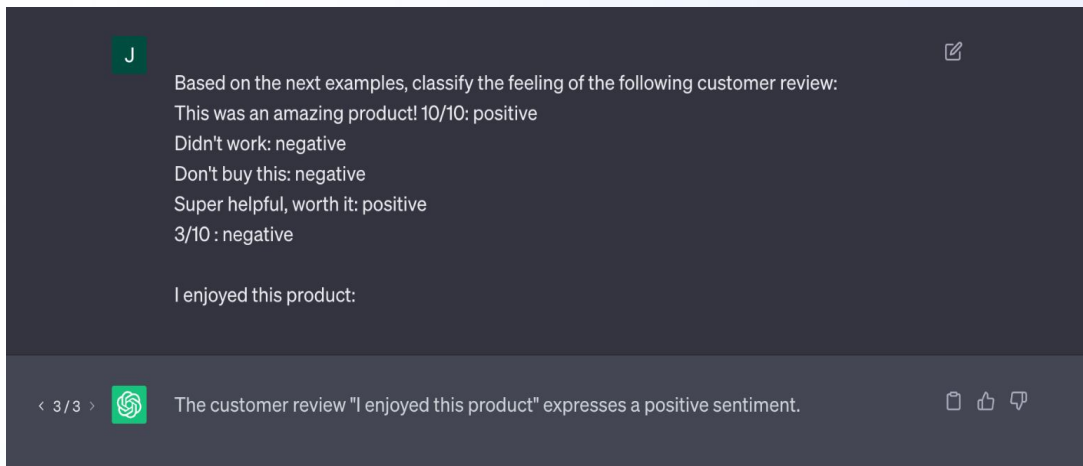
Few shot prompt tuning



Few shot prompt tuning

Naive prompt tuning or few shot tuning

Providing the models with
examples within the
prompt context



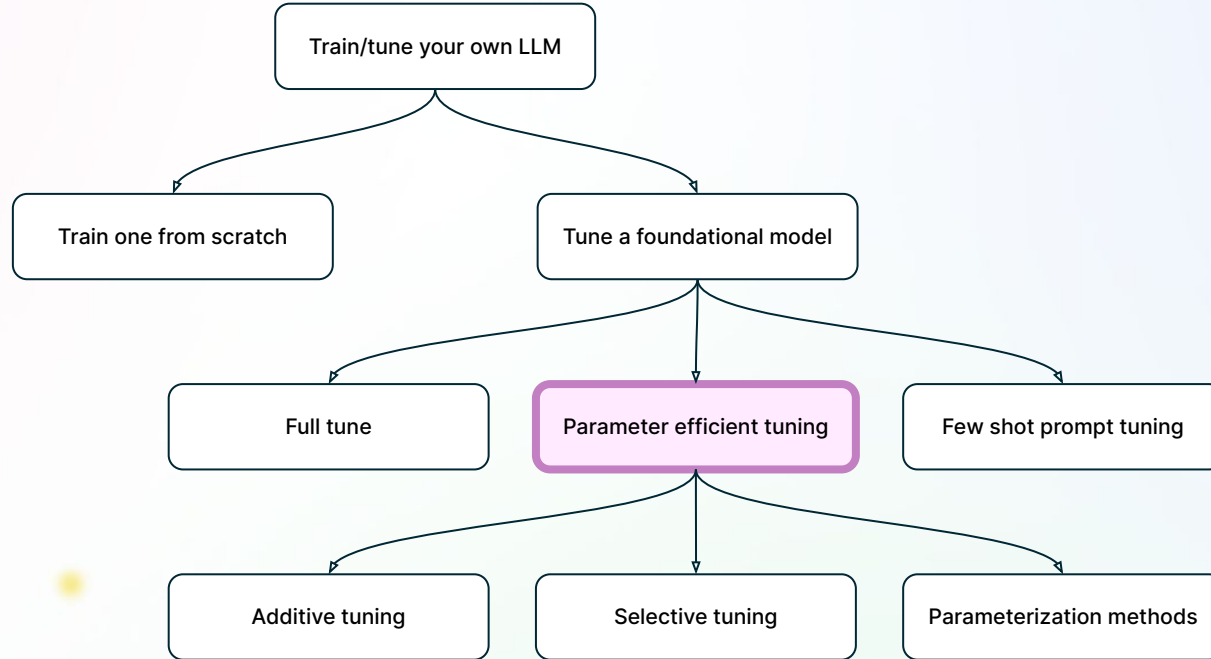
Pros:

- **No data needed**
- **No extra compute** (just inference)
- **Instant**

Cons:

- Bad performance (non robust + non scalable)
- Extensive prompt engineering

Parameter efficient tuning

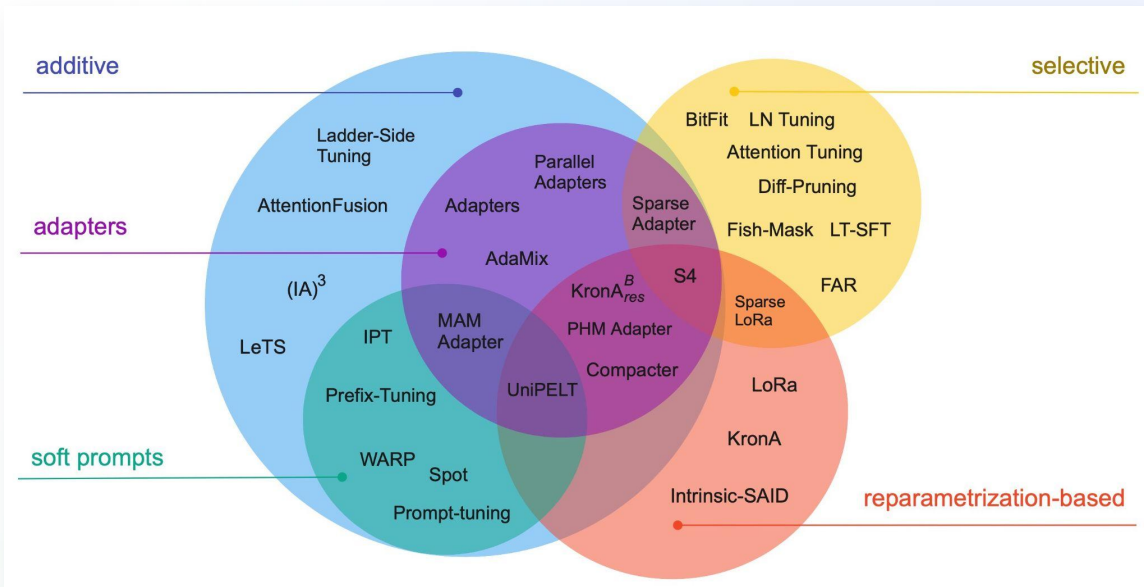


Parameter Efficient Fine Tuning (PEFT)

Make more with less!
Use less data/compute power achieve comparable performance



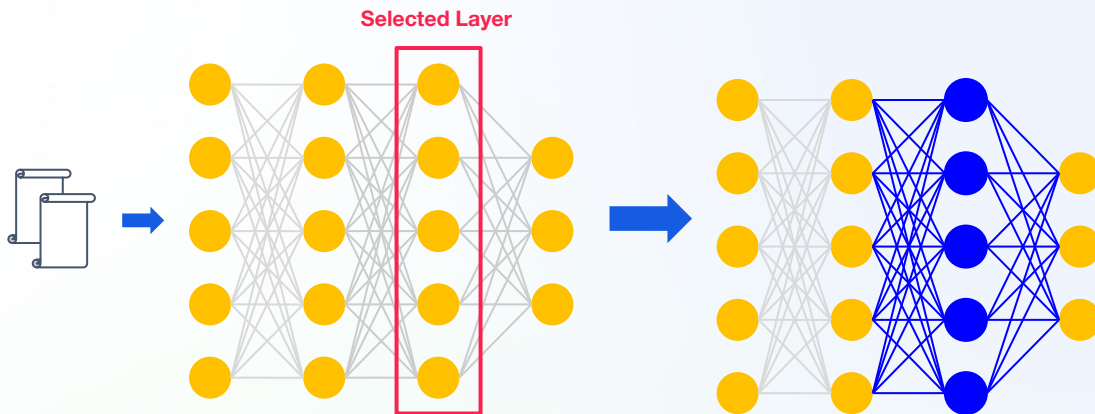
Reduce costs!



Selective tuning

Selective methods

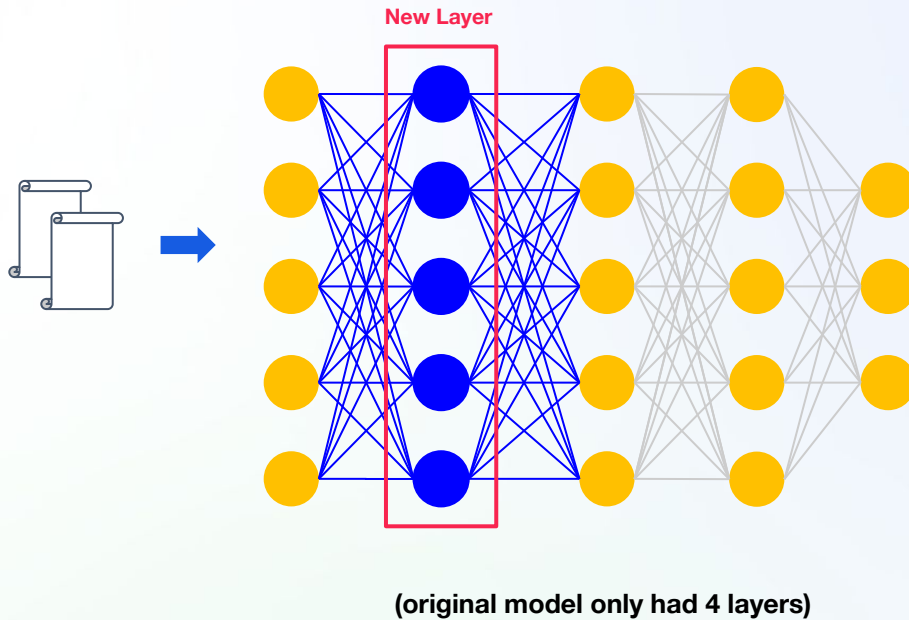
Re-train only a few layers/parameters, keeping all other parameters fixed



Additive tuning

Additive methods

Add a few new parameters and train them from scratch, keeping all other parameters fixed



Parameter- ization tuning

Parameterization methods

Decrease foundational model size before tuning

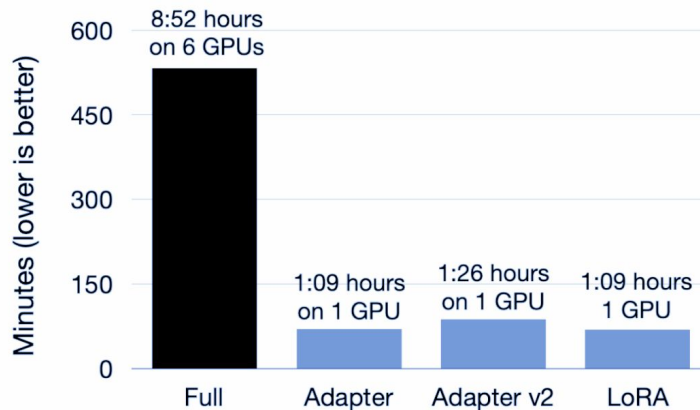
- Low rank approximation/adaptation
- Lowering precision



Parameter efficient tuning

Parameter efficient fine tuning (PEFT)

Time to complete 52k training iterations for Falcon 7B



Lightning.ai - Fine tuning Falcon LLMs more Efficiently with LoRA and adapters

Parameter efficient tuning

PEFT Pros:

- Significantly reduces compute power
- Significantly reduces training time (**hours**)
- **Reduce costs!**
- **Comparable accuracy**

PEFT Cons:

- Lacks extensive accuracy evaluation for every method
- Small downgrade in accuracy

Monitoring LLMs

Integrity

- Prompts not understood
- Readability match
- Sentiment match
- Language match
- User feedback
- Cutoff responses

Drift

- Changes in language
- Changes in vocabulary
- Topic drift
- Prompts out of training data

Governance

- Bias and profanity indicators
- Governance & compliance rules
- Personal information
- Privacy preservation

Hallucinations

- Response outliers
- Similarity scoring
- Reference (URL) validation

Attacks

- Adversarial attempts to extract data
- Bypass safety controls
- Prompt injection
- Prompt leakage

Monitoring LLMs with Elemeta

Image



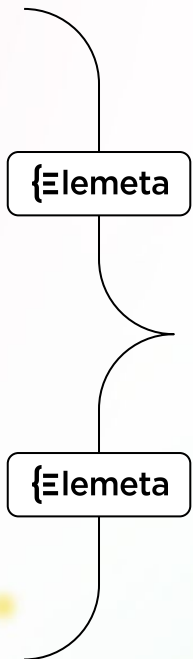
Audio



Text



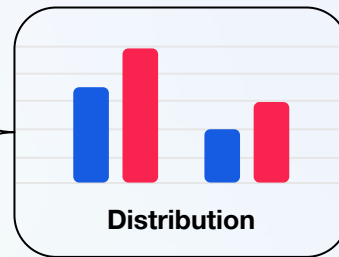
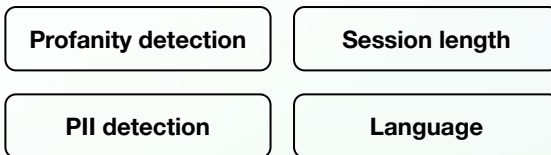
Prompt



Embedding

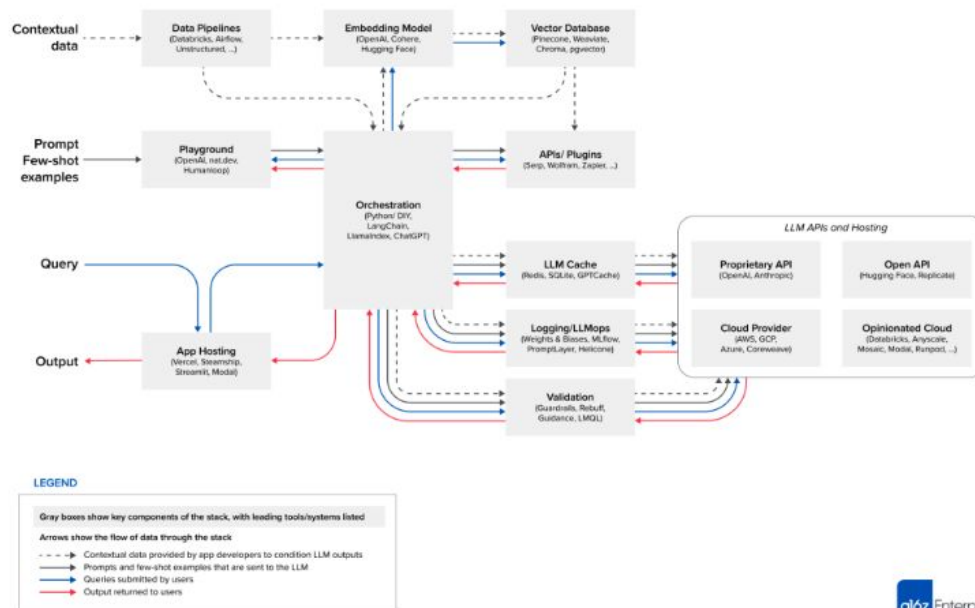
0.45	0.11	0.11	0.98	0.4	0.5
------	------	------	------	-----	-----

Meta-features



Beyond tuning

Emerging LLM App Stack



LLM Garden (<https://llm.garden/>)

LLM Garden Add an LLM

Welcome to the LLM Garden

With so many Large Language Models (LLMs) released daily, we put together a list of everything available so we can easily search and compare our options.
Have an LLM to add or update? [Let us know!](#)

LLM	Released	Maintainer	License	Commercial	Accessible via	Description
1 AutoGPT	2023-03-01	OpenAI	MIT	Yes	GitHub	AI tech
2 BERT	2018-10-01	Google	Apache 2.0	Yes	Google Cloud	Bidirec
3 BLOOM	2022-11-01	BigScience	RAIL v1.0	No	Hugging Face	An aut
4 BLOOMChat	2023-05-01	SambaNova & T...	BLOOMChat-1...	Yes	Hugging Face	BLOO
5 Cerebras-GPT	2023-03-01	Cerebras	Apache 2.0	Yes	Hugging Face	Includ
6 Claude	2023-03-01	Anthropic	N/A	Yes	Anthropic	Claude
7 DLite (v2)	2023-05-01	AI Squared	Apache 2.0	Yes	GitHub , Hugging Face	Family
8 Dolly 2.0	2023-04-01	Databricks	Apache 2.0	Yes	Hugging Face	An ope
9 Falcon-40B	2023-05-01	Technology Inn...	TII Falcon LLM...	Yes	Hugging Face	Open-
10 FastChat-T5	2023-04-01	LMSYS	Apache 2.0	Yes	GitHub , Hugging Face	Open :
11 FinLLM	2023-06-01	AI4Finance Fou...	MIT	Yes	GitHub (FinGPT) & GitHub (FinNLP)	Open-
12 GPT2	2019-02-01	OpenAI	MIT	Yes	GitHub , Hugging Face	Gener



Q&A

Oren Razon, CO-Founder & CEO @ Superwise | oren.razon@superwise.ai | [linkedin/oren-razon](https://www.linkedin.com/in/oren-razon)

Gad Benram, Founder & CTO @ TensorOps | gad@tensorops.ai | [linkedin/gad-benram](https://www.linkedin.com/in/gad-benram)