databricks

Large Language Models

Foundation Models from the Ground Up

Course Outline

Course Introduction

Module 1 – Transformers: Attention and the Transformer Architecture

Module 2 – Efficient Fine-Tuning: Doing more with less

Module 3 – Deployment Optimizations: Improving model size and speed

Module 4 – Multi-modal LLMs: Beyond text-based transformers

Course Outline

Course Introduction

Module 1 - Transformers: Attention and the Transformer Architecture

Module 2 - Efficient Fine-Tuning: Doing more with less

Module 3 - Deployment Optimizations: Improving model size and speed

Module 4 - Multi-modal LLMs: Beyond text-based transformers

Course Introduction

What are LLMs?

Matei Zaharia

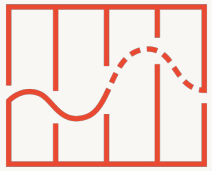
Co-founder & CTO of Databricks

Associate Professor of Computer Science at UC Berkeley

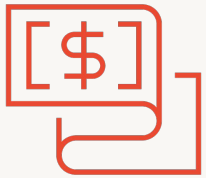
Generative AI state of the art is rapidly advancing

No single model to rule them all—trade-offs are required to find the best model for each use case

Privacy



Quality



Cost

Latency

Proprietary LLMs

ChatGPT

PaLM 2

ANTHROP\C

OpenAI

Open Source LLMs

databricks

Dolly

Hugging Face

mosaicML

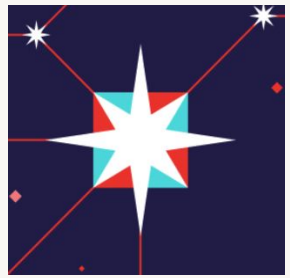
MPT

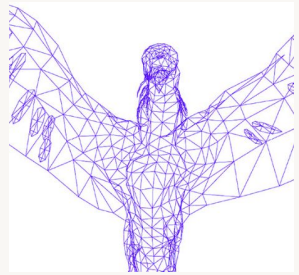
stability.ai

Stable Diffusion

Open Source quality is rapidly advancing – while fine tuning cost is rapidly decreasing

Dolly started the trend to open models with a commercially friendly license





Facebook Llama

"Smaller, more performant models such as Llama … democratizes access in this important, fast-changing field."

February 24, 2023

Stanford Alpaca

"Alpaca behaves qualitatively similarly to OpenAI … while being surprisingly small and easy /cheap to reproduce"

March 13, 2023

Databricks Dolly

"Dolly will help democratize LLMs, transforming them into a commodity every company can own and customize"

March 24, 2023

Mosaic MPT

"MPT-7B is trained from scratch on 1T tokens … is open source, available for commercial use, and matches the quality of LLaMA-7B"

May 5, 2023

TII Falcon

"Falcon significantly outperforms GPT-3 for … 75% of the training compute budget—and … a fifth of the compute at inference time."

May 24, 2023

Non Commercial Use Only | **Commercial Use Permitted**

OSS LLMs are getting better everyday

🤗 Open LLM Leaderboard

📐 The 🤗 Open LLM Leaderboard aims to track, rank and evaluate LLMs and chatbots as they are released.

🤗 Anyone from the community can submit a model for automated evaluation on the 🤗 GPU cluster, as long as it is a 🤗 Transformers model with weights on the Hub. We also support evaluation of models with delta-weights for non-commercial licensed models, such as the original LLaMa release.

Other cool benchmarks for LLMs are developed at HuggingFace, go check them out: 🤷🤖 human and GPT4 evals, 🖥️ performance benchmarks

- Constant development
- Supported by community and industry
- Rapid innovation cycle

T	Model	Average ⬆️	HellaSwag	MMLU	TruthfulQA	#Params (B)
🟢	meta-llama/Llama-2-70b-hf	67.3	87.3	69.8	44.9	70
🟢	huggyllama/llama-65b	64.2	86.1	63.9	43.4	65.286
🟢	llama-65b	64.2	86.1	63.9	43.4	65
🟢	tiiuae/falcon-40b	61.5	85.3	57	41.7	40
🟢	llama-30b	61.7	84.7	58.5	42.3	30
🟢	mosaicml/mpt-30b	56.2	82.4	47.9	38.4	30
🟢	TheBloke/Llama-2-13B-fp16	58.6	82.2	55.7	37.4	13
🟢	meta-llama/Llama-2-13b-hf	58.7	82.1	55.8	37.4	13
🟢	llama-13b	56.1	80.9	47.7	39.5	13
🟢	huggyllama/llama-13b	56	80.9	47.6	39.5	13.016
🟢	dvruette/llama-13b-pretrained-sft-do2	58.5	80.3	47.2	47.4	13
🟢	dvruette/llama-13b-pretrained-sft-epoch-1	56.8	80	45.5	44.5	13
🟢	dvruette/llama-13b-pretrained	57.8	79.3	47	48.4	13
🟢	dvruette/llama-13b-pretrained-dropout	57.7	79.3	46.6	48.6	13
🟢	meta-llama/Llama-2-7b-hf	54.4	78.6	46.9	38.8	7
🟢	tiiuae/falcon-7b	47	78.1	27.8	34.3	7
🟢	llama-7b	49.7	77.8	35.7	34.3	7

Source: Hugginface.co

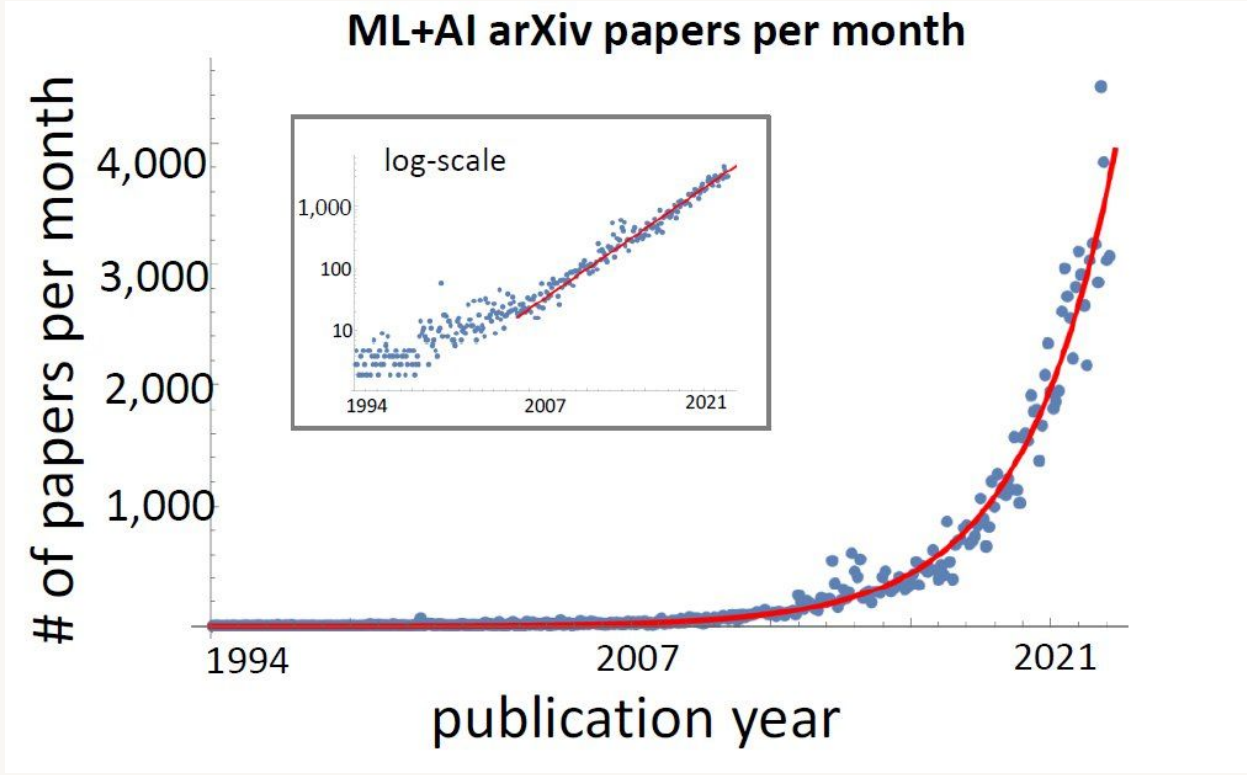
“A strong foundation… is all you need”

What and who this course is for

If you want to keep up:

- The **fundamentals** of LLMs have not changed since 2018
- Most of the innovation are variations of the original.

Research and innovation has exploded around LLMs.



Source: Reddit

Enjoy the course!