onfido

The (r)evolution of generative models

Olivier Koch

olivier.koch@gmail.com

About me

onfido

VP of AI at Onfido

ex-Criteo (2015-2020)

MIT DARPA Urban Challenge Finalist

ENSTA 2002, MIT PhD 2010 (computer vision)

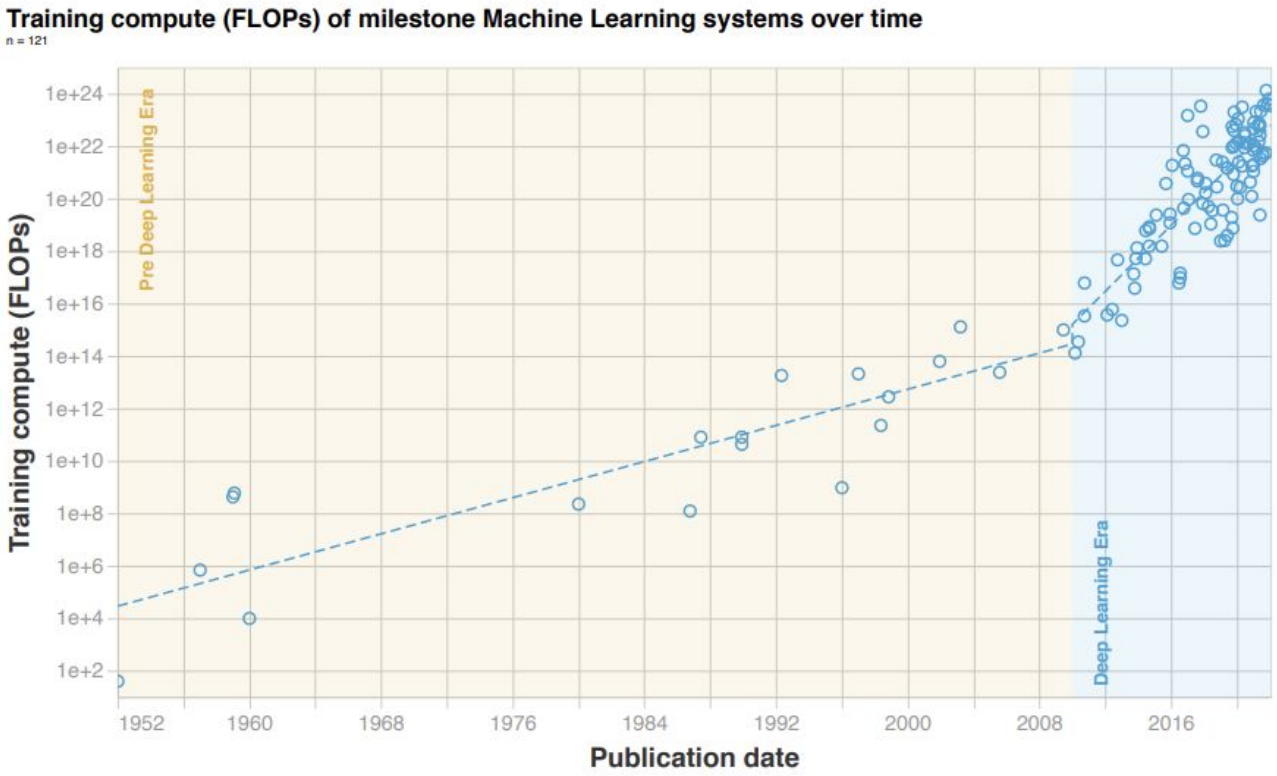
Published @ CVPR 2007, IJFR 2008, ICCV 2009, ICRA 2010

oakfr

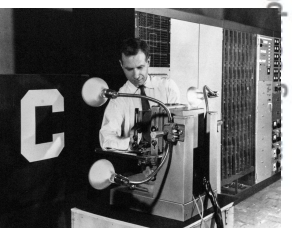
@olivkoch

olivierkoch.net

1. Historical perspective
2. The Transformer shake-up
3. The new landscape

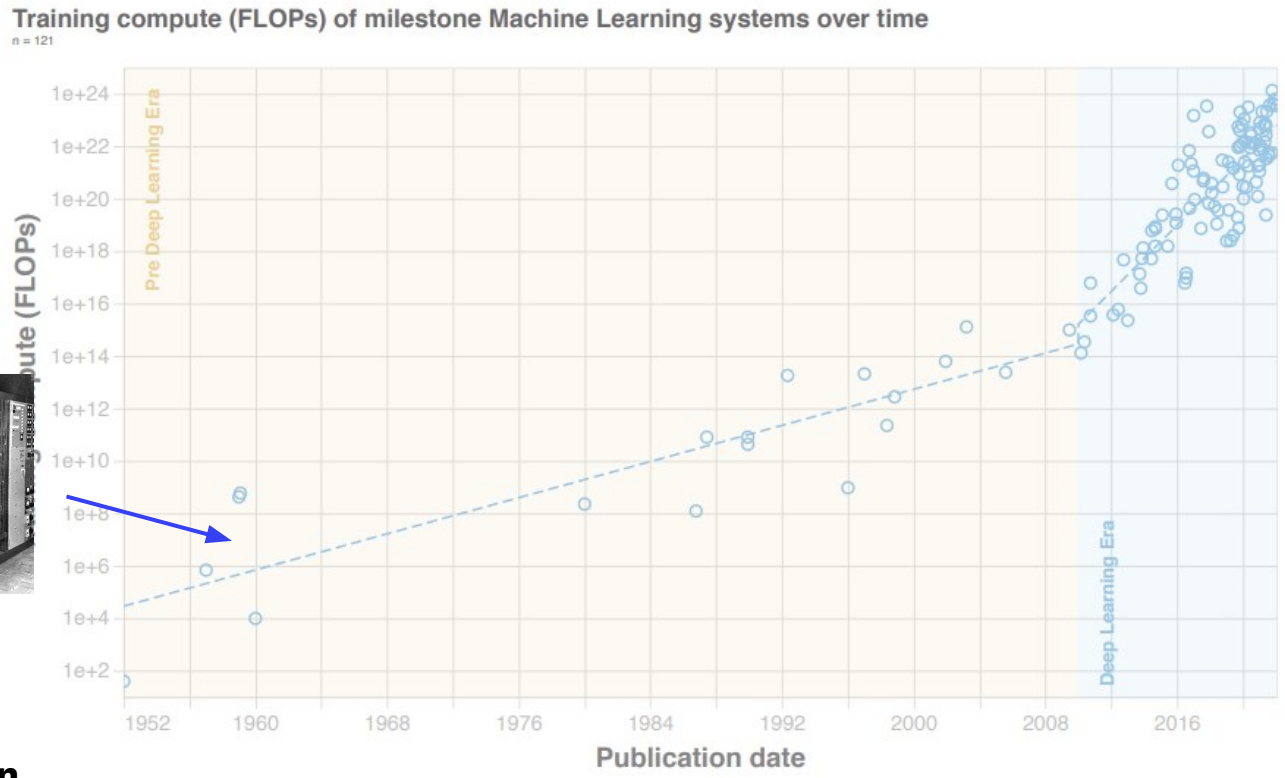

Training compute (FLOPs) of milestone Machine Learning systems over time
n = 121
Training compute (FLOPs)
1e+24
1e+22
1e+20
1e+18
1e+16
1e+14
1e+12
1e+10
1e+8
1e+6
1e+4
1e+2
Pre Deep Learning Era
Deep Learning Era
1952
1960
1968
1976
1984
1992
2000
2008
2016
Publication date

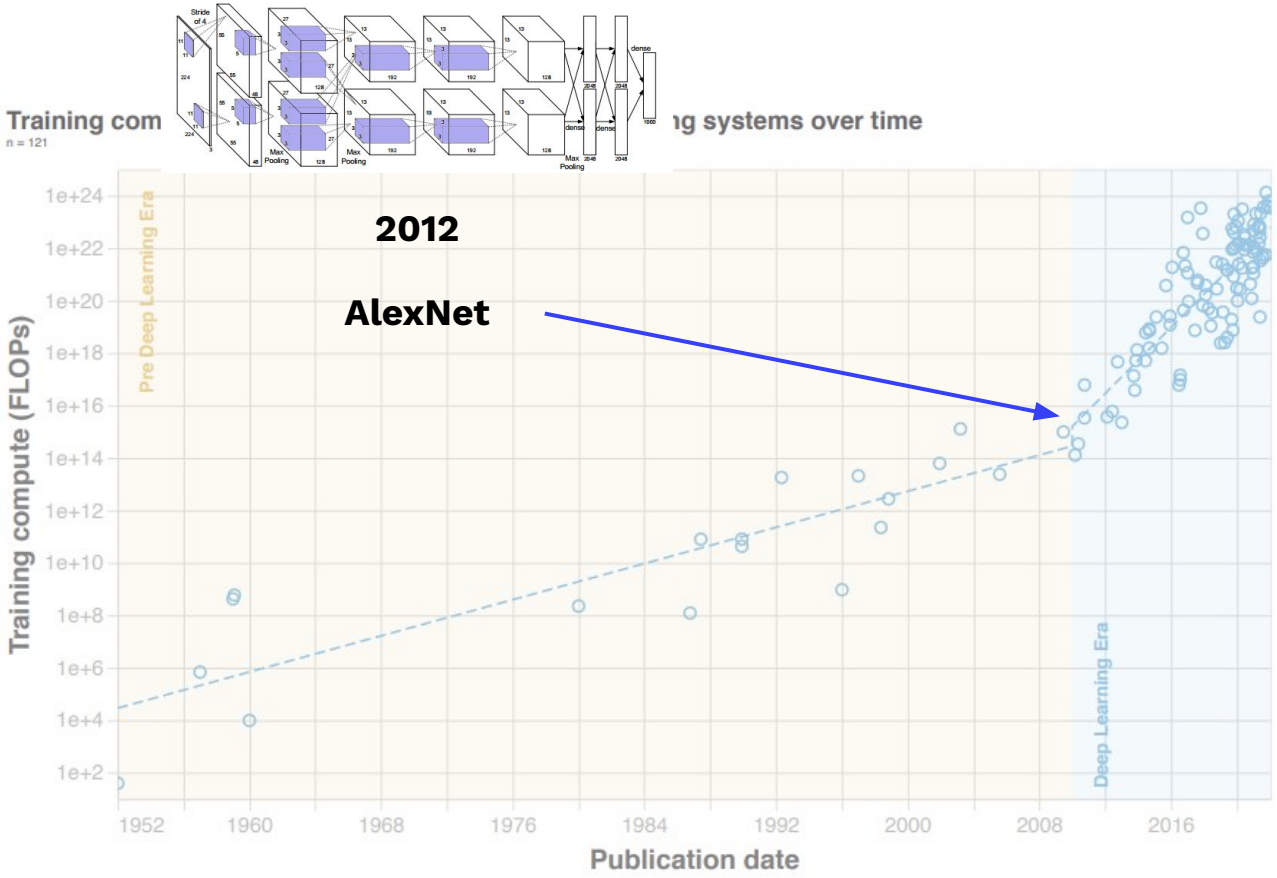
Source

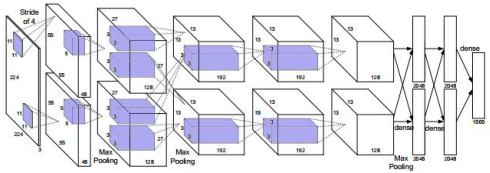


1958

The Perceptron

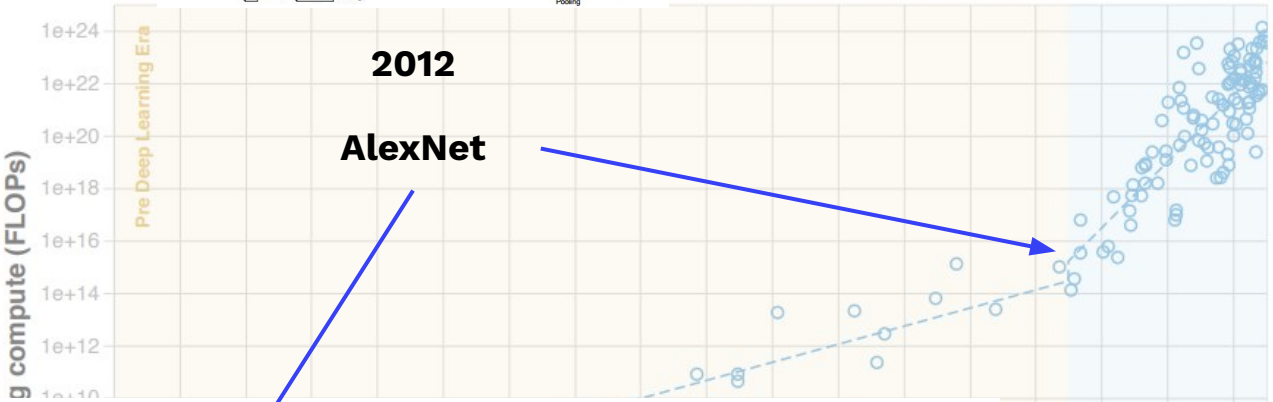



Training com
n = 121
ıg systems over time
2012
AlexNet
Pre Deep Learning Era
Deep Learning Era
Training compute (FLOPs)
1e+24
1e+22
1e+20
1e+18
1e+16
1e+14
1e+12
1e+10
1e+8
1e+6
1e+4
1e+2
1952
1960
1968
1976
1984
1992
2000
2008
2016
Publication date



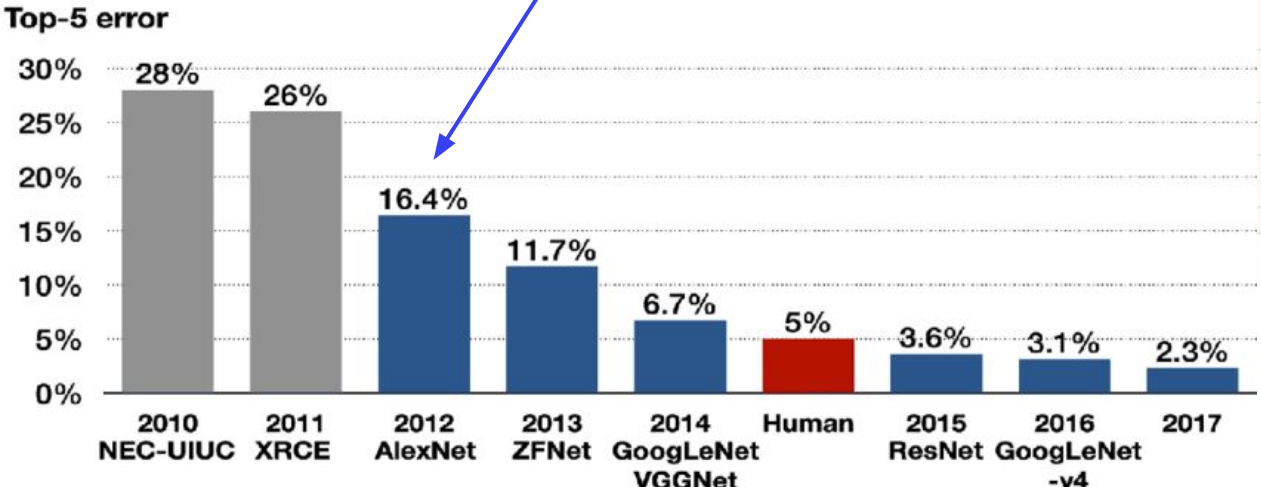
Training com ng systems over time

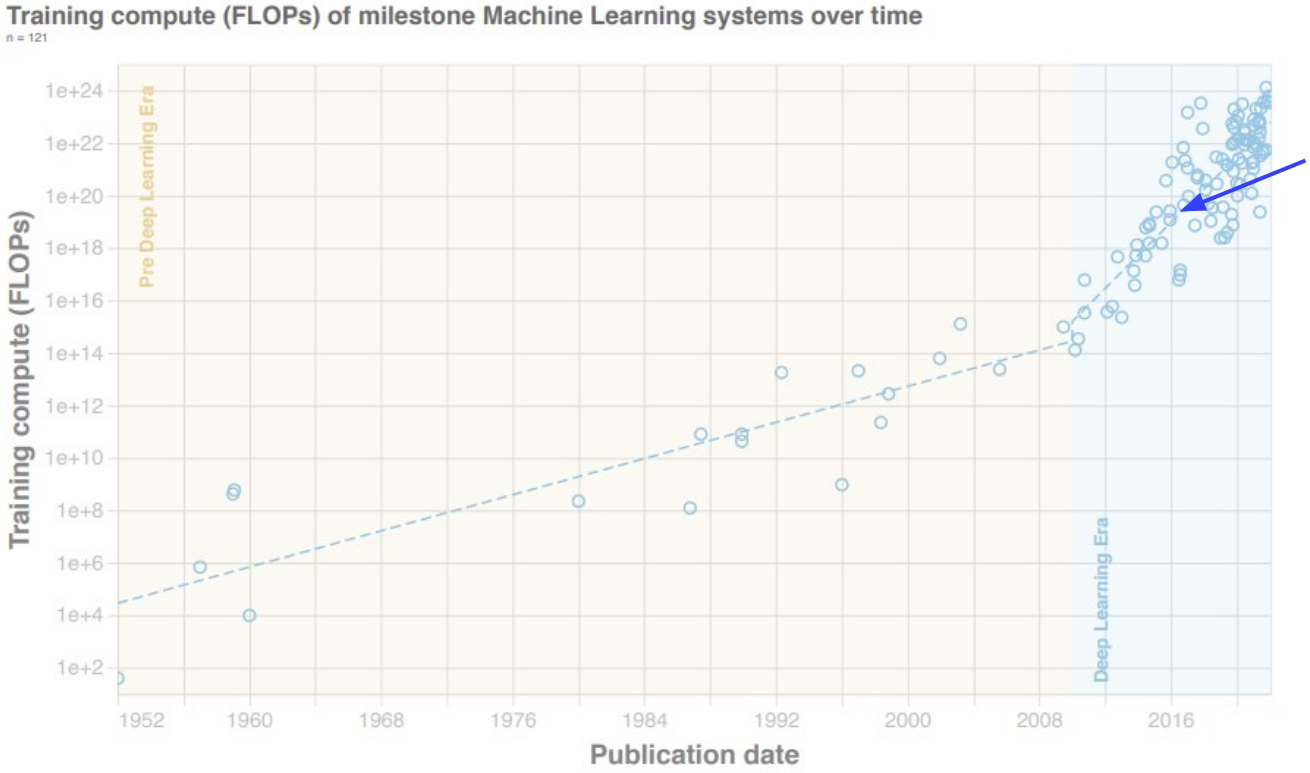
n = 121



2012

AlexNet

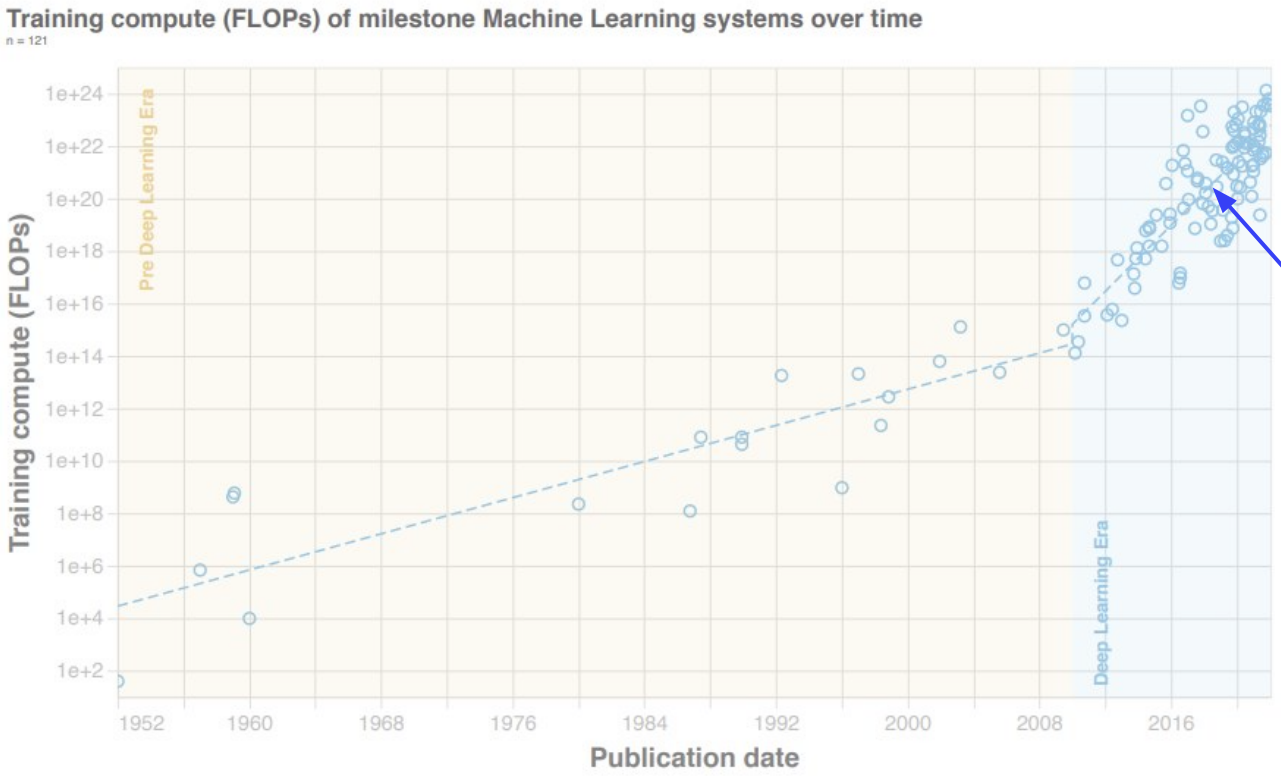


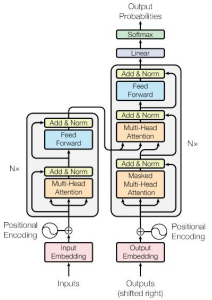


2016

AlphaGo

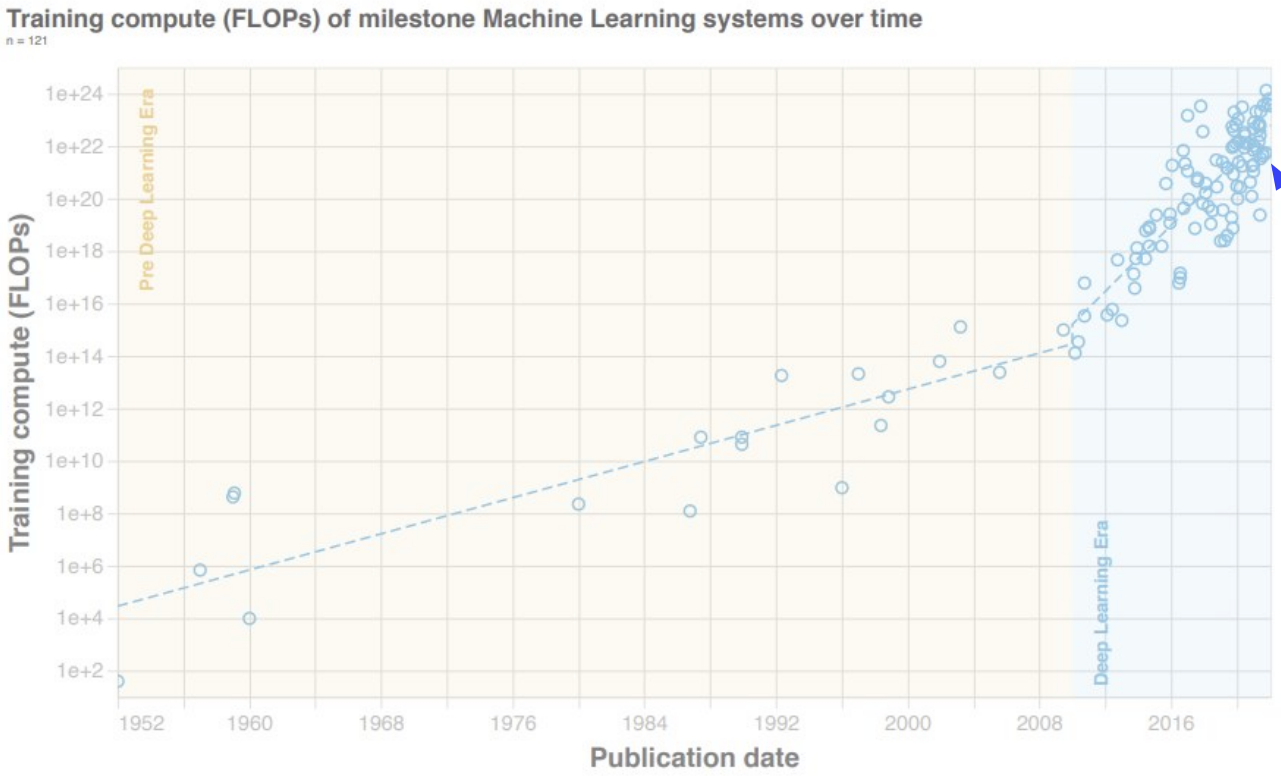
Reinforcement Learning (RL)





2017

Transformers

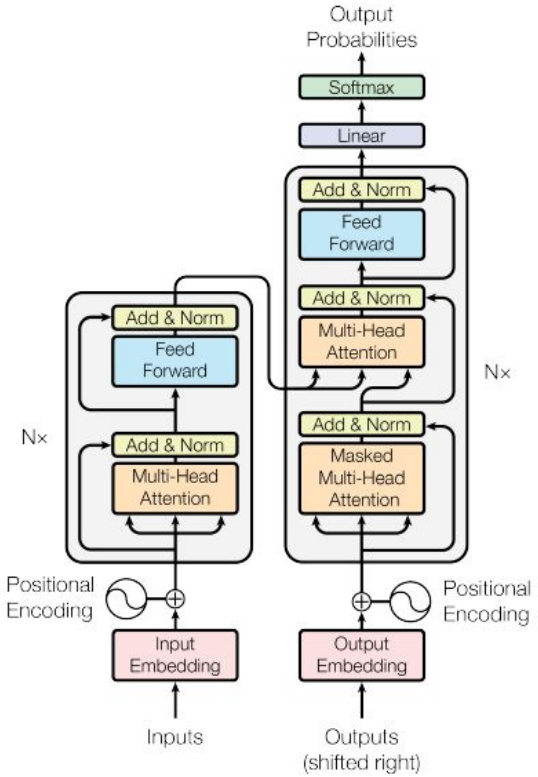

Training compute (FLOPs) of milestone Machine Learning systems over time
n = 121
Training compute (FLOPs)
1e+24
1e+22
1e+20
1e+18
1e+16
1e+14
1e+12
1e+10
1e+8
1e+6
1e+4
1e+2
Pre Deep Learning Era
Deep Learning Era
1952
1960
1968
1976
1984
1992
2000
2008
2016
Publication date

2023

GPT

1. Historical perspective
2. **The Transformer shake-up**
3. The new landscape

What is a Transformer?



No, seriously... what is a Transformer?

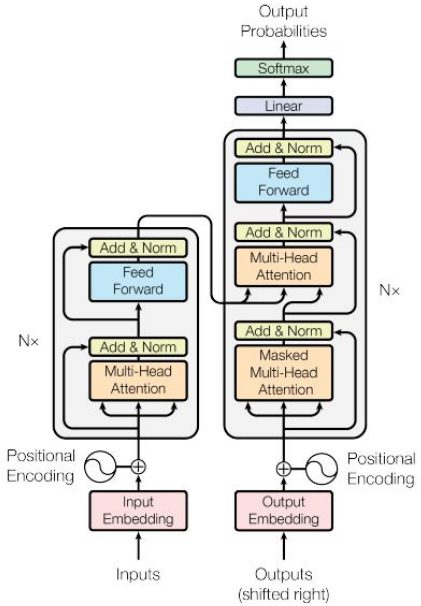
Take the (free) youtube course by Andrej Karpathy

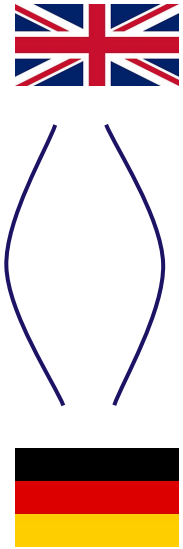
Read Lilian Weng's blog post

Attention is a communication mechanism between tokens in a document, with a positional embedding allowing to incorporate ordering into sets of tokens.

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}})\mathbf{V}$$

Transformers shake up translation

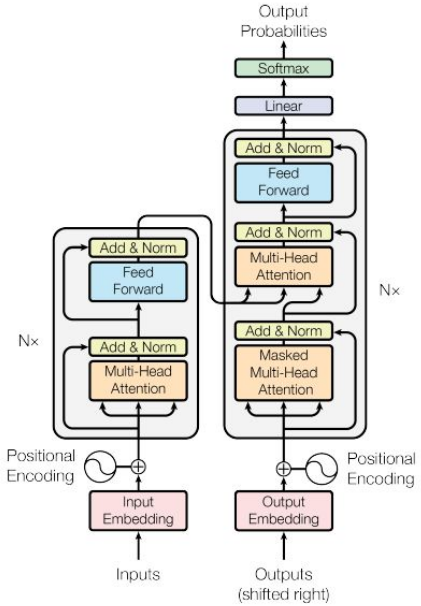




Abraham Lincoln (February 12, 1809 – April 15, 1865) was an American lawyer, politician and statesman who served as the 16th president of the United States from 1861 until his assassination in 1865. Lincoln led the Union through the American Civil War to defend the nation as a constitutional union and succeeded in abolishing slavery,...

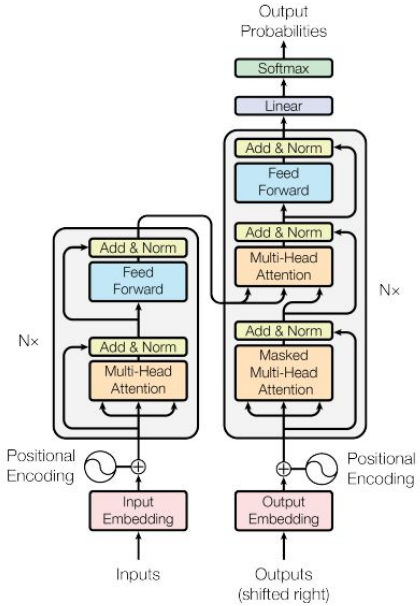
Abraham Lincoln (12. Februar 1809 – 15. April 1865) war ein amerikanischer Anwalt, Politiker und Staatsmann, der von 1861 bis zu seiner Ermordung im Jahr 1865 als 16. Präsident der Vereinigten Staaten diente. Lincoln führte die Union durch den Amerikanischen Bürgerkrieg, um die Nation als verfassungsmäßige Union zu verteidigen, und war erfolgreich bei der Abschaffung der Sklaverei.

Source



Abraham Lincoln (February 12, 1809 – April 15, 1865) was an American lawyer, politician and statesman who served as the 16th president of the United States from 1861 until his assassination in 1865. Lincoln led the Union through the American Civil War to defend the nation as a constitutional union and succeeded in abolishing _____,...

Likelihood of “slavery”: 99.2%
Likelihood of “making”: 0.1%
Likelihood of “food”: 3%



Context length

Abraham Lincoln (February 12, 1809 – April 15, 1865) was an American lawyer, politician and statesman who served as the 16th president of the United States from 1861 until his assassination in 1865. Lincoln led the Union through the American Civil War to defend the nation as a constitutional union and succeeded in abolishing _____,...

Likelihood of “slavery”: 99.2%
Likelihood of “making”: 0.1%
Likelihood of “food”: 3%

Transformers for text generation

Language models are few-shot learners

Try it yourself!

Write With Transformer

Get a modern neural network to
auto-complete your thoughts.

This web app, built by the Hugging Face team, is the official demo of the
`/transformers` repository's text generation capabilities.

Star 93,671

https://transformer.huggingface.co/

Christopher Columbus discovered America in 1492. When he arrived, he saw a land full of unknown natives that would eventually become the United States . And while it's true that this discovery is one of the most famous events in our history, he didn't come across all the people he would meet , nor all the artifacts he would find .

OK, this is cool, but it hallucinates 100%.

The ChatGPT epiphany

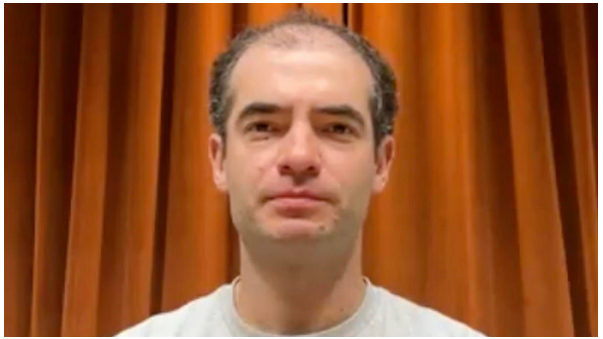
But we know how to train a machine to learn from feedback!

It is called Reinforcement Learning (RL).

Go World Champion Lee Sedol looking confused at move 37 by AlphaGo...

The ChatGPT epiphany

“This whole thing was just an experiment. We had no idea it would work so well”, Ilya Sutskever
(full interview)



The main breakthrough of LLM is the discovery that training a very large model on the simplest task (predict the next token) could yield remarkable intelligence.

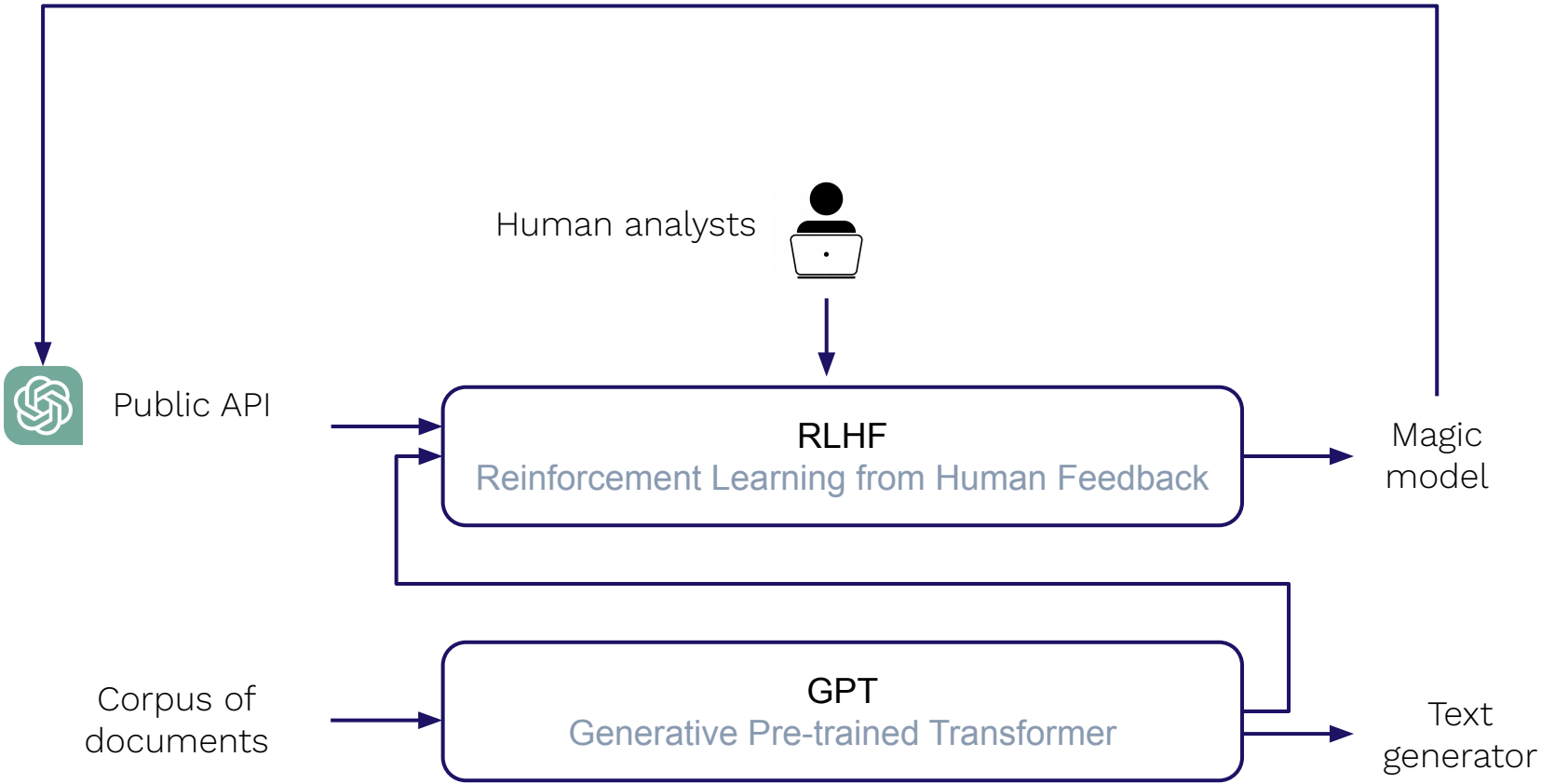
<philosophy>

Transformers are putting humankind through an existential crisis.

With Copernicus, we discovered that we were not at the center of the universe.

With GPT, we discover that (maybe) we are not that intelligent after all.

</philosophy>

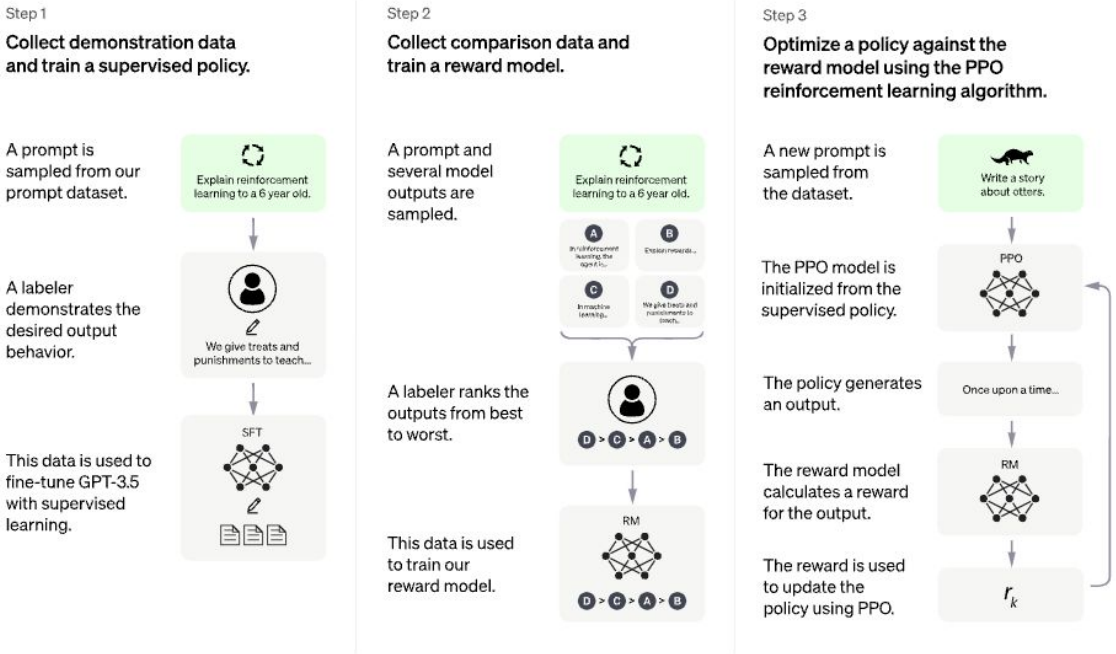

Human analysts
Public API
RLHF
Reinforcement Learning from Human Feedback
Magic model
Corpus of documents
GPT
Generative Pre-trained Transformer
Text generator

Source

Midjourney

“Man hoarding tigers with a whip, black and white drawing”

The science of human feedback (alignment)



[Source]

1. Historical perspective
2. The Transformer shake-up
3. **The new landscape**

The Cambrian explosion of foundational models

Mistral AI

Experimentation is getting much easier

2010

Write your own backprop manually

2018

`loss.backward()`

2023

```
from transformers
import SwinModel
```

Data is (relatively) small

Crowd thinking: "GPT-3 was trained on a huge chunk of the internet"

The reality:

GPT-2 was trained on 40GB of text
GPT-3 was trained on 570GB of text (filtered from 50TB of data)
Web2Text weighs 65GB

Stable Diffusion was trained on LAION-400M (240TB). Relatively small.

Raw data is cheap, clean data is expensive

Lots of data cleaning tricks[1]

Clean data beats more data

Data requires significant ML effort

(1) DALL-E pretraining mitigations, OpenAI

Complex models are getting cheaper

New GPUs provide 5-10x speed-up every few years (H100 >> A100)

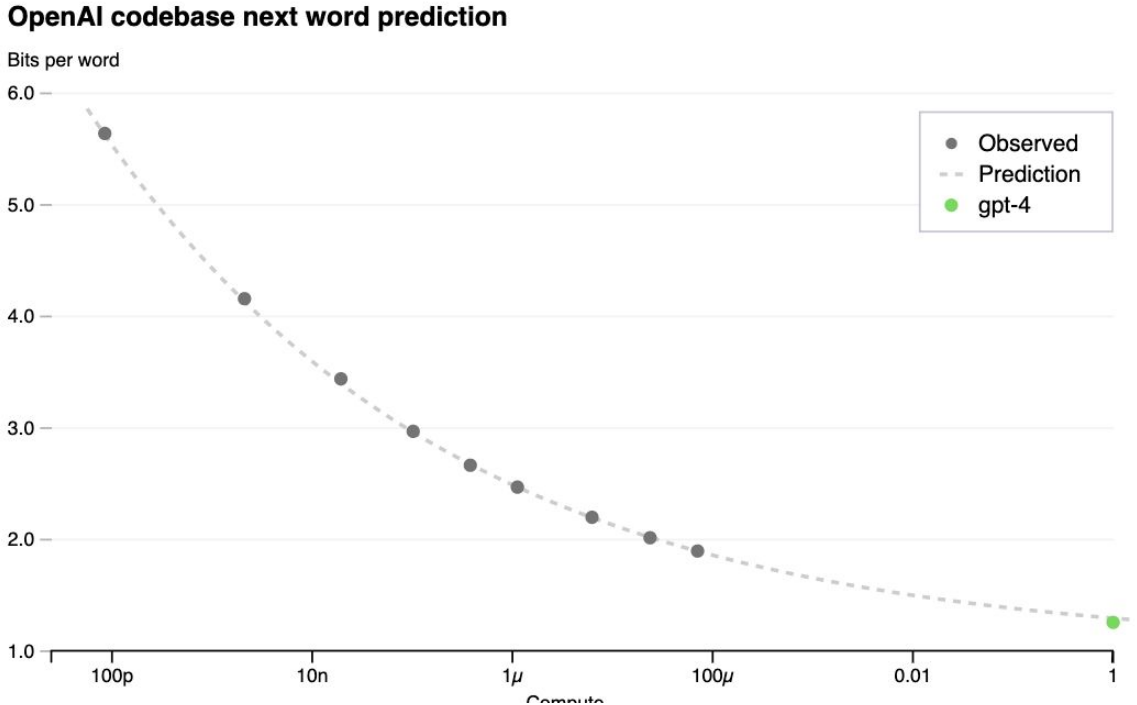
- Stable Diffusion cost $600K to train in 2021.
- Today, the same model would cost $60K.

More efficient Transformers are coming.

- Transformers are quadratic in complexity.
- n.log n Transformers are on their way.
- A $50M job becomes a $100K job.

You can finetune a 7B LLM on a single GPU in 1-2 h using techniques like low-rank adaptation (LoRA).

Model performance can be predicted with 1/1,1000th of the compute.



Pre-training is still very hard

Goal: get a 175B dense model up and running by any means necessary.

Solution:
114 pages of OPT175B logbook

Research is far from over

Durk Kingma @dpkingma · Jun 9

Why do LLMs hallucinate? TL;DR:
- LLM pretraining generally results in a well-calibrated distribution p(completion|prompt)
- Obviously, a single sample from this distribution doesn't express the uncertainty contained in it
- This is fixable, e.g. through sufficient RLHF/RLAIF

23 36 260 82.4K

Yann LeCun @ylecun · Jun 10

Replying to @dpkingma

I disagree. I don't think it's fixable within the auto-regressive prediction paradigm.

15 6 143 21.2K

Summary

Data is small. Clean data is expensive.

Experimentation is (much) easier

Models are getting cheaper

Pre-training is (still) hard and expensive

Research is far from over

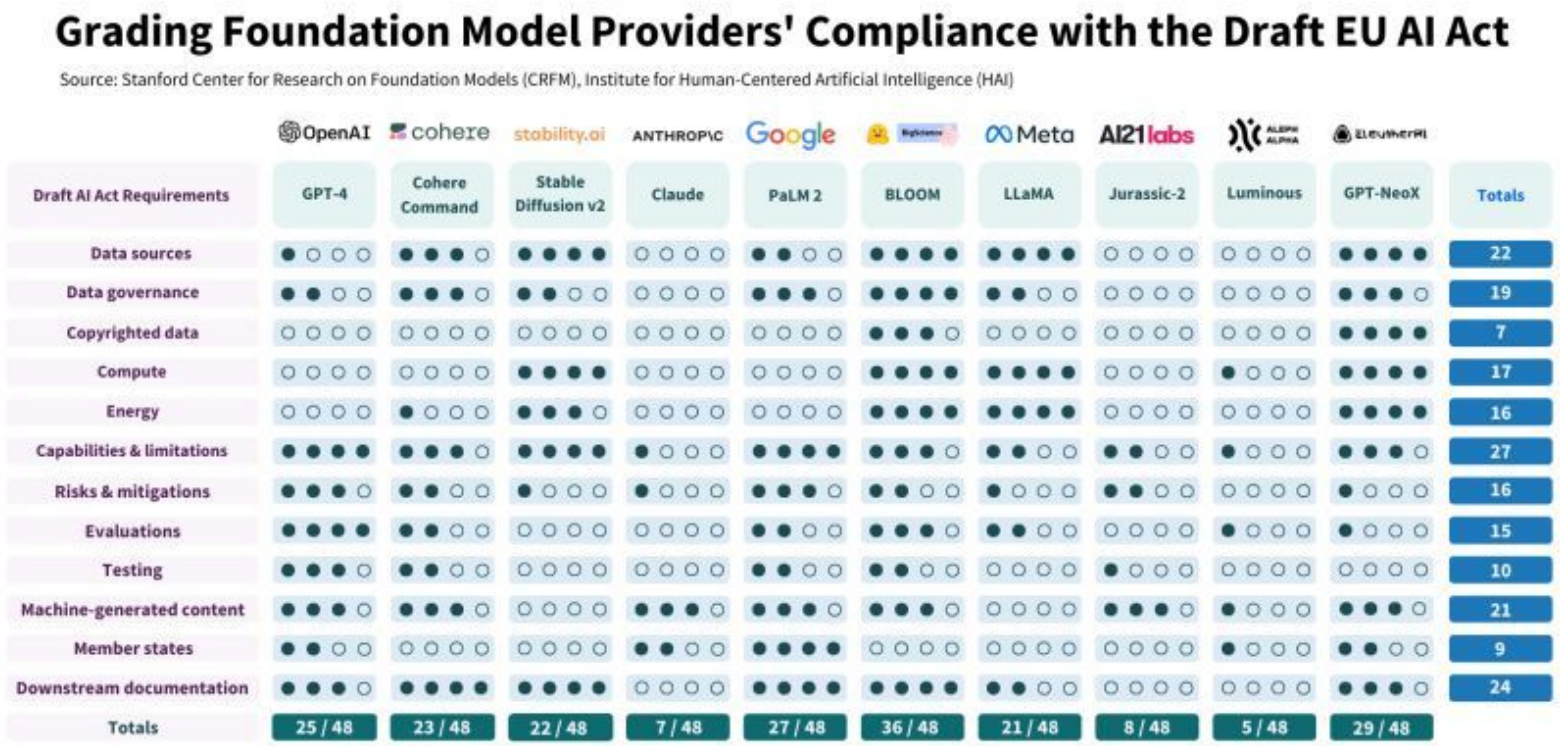
Take-aways for CTOs

1. System complexity
2. Hallucinations
3. Pre-training is hard
4. What is the ROI?


Thank you!
olivier.koch@gmail.com

Foundational models vs EU AI Act

Stanford Institute for Human-Centered Artificial Intelligence (HAI)



Source

Close vs open source

Two drivers for close source:

1. competitive pressure
2. safety

Open source is usually safer

Open source is going strong

The code is not enough (data matters)

The right question: is it safe?

Carbon footprint of generative AI?

Gen. AI has a significant CO2 footprint

Can the footprint be compensated with the benefits of generative AI for the planet?