ETH zürich

ETH AI CENTER

Large Language Model Trends

Imanol Schlag

- Introduction
- Past Trends
- Future Trends
- Swiss AI Initiative
- Q&A

Introduction

- Imanol Schlag
- Neural network research since 2016
- PhD at the Swiss AI Lab (IDSIA) under Jürgen Schmidhuber
- 20+ publications on ML/AI
- Research Scientist at the ETH AI Center

Microsoft Research

Meta

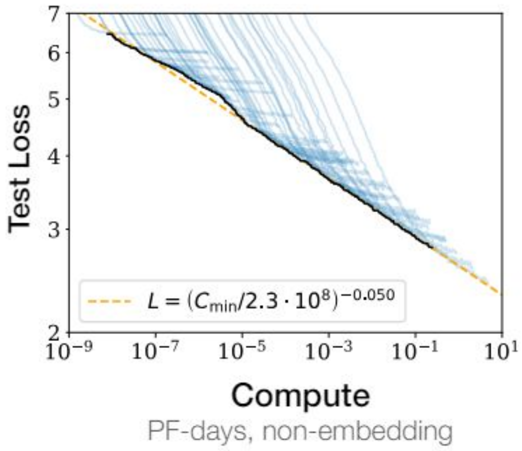
Google

ETH AI CENTER

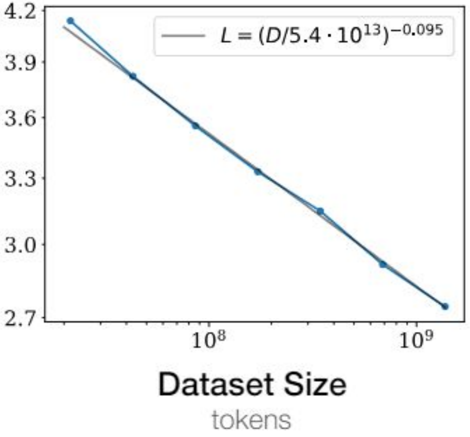
SOLA-Stafette 2024

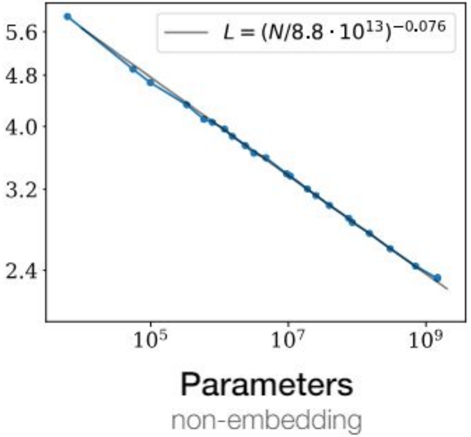
2020: GPT 3

- GPT 2, February 2019
- 175B Parameters, May 2020
- In-context learning -> excel at NLP tasks



Kaplan et al. (January 2020)



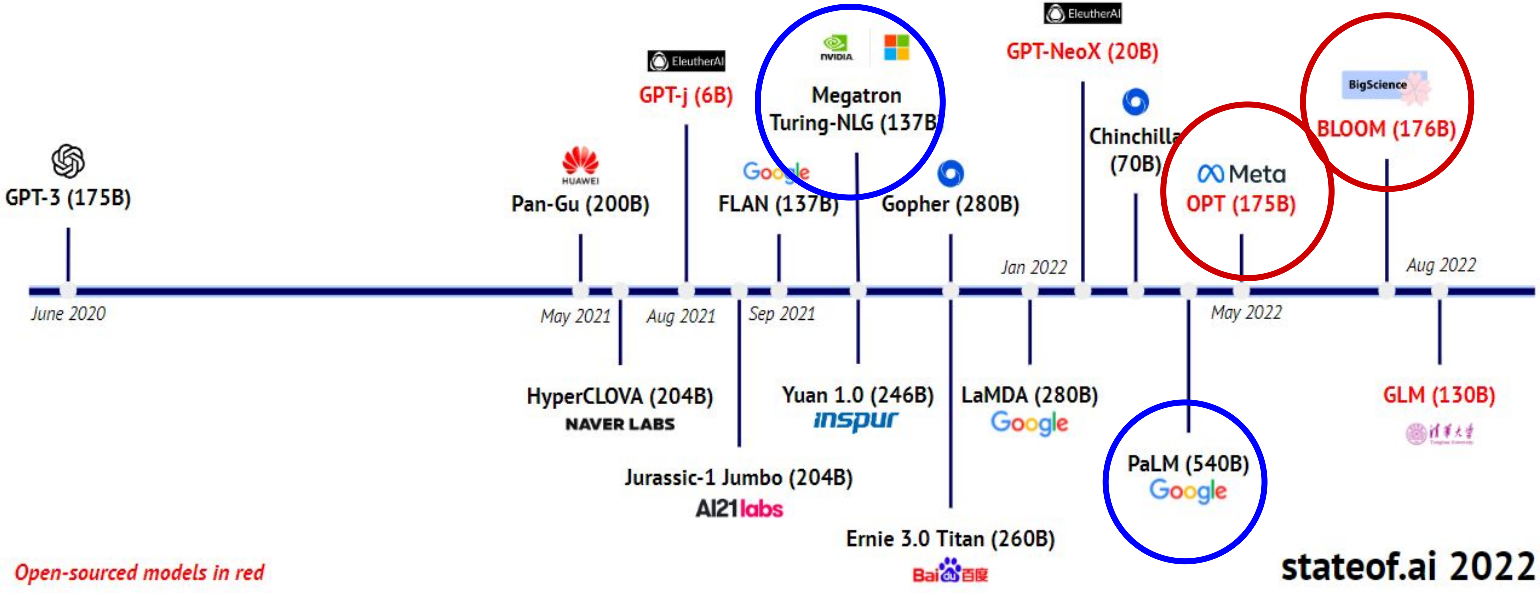


2021: First Copycats and API-based Research

- Jurassic-1, Yuan 1.0, Ernie, Gopher
- Sparse methods at a Trillion parameters: GShard, GLaM
- Research explosion based on OpenAI API
    - Many new benchmarks
    - GPT-3 is not truthful / hallucinations
    - Prompting is brittle and challenging
    - Increased discussion around ethics & safety

2022a: Bigger! Better?

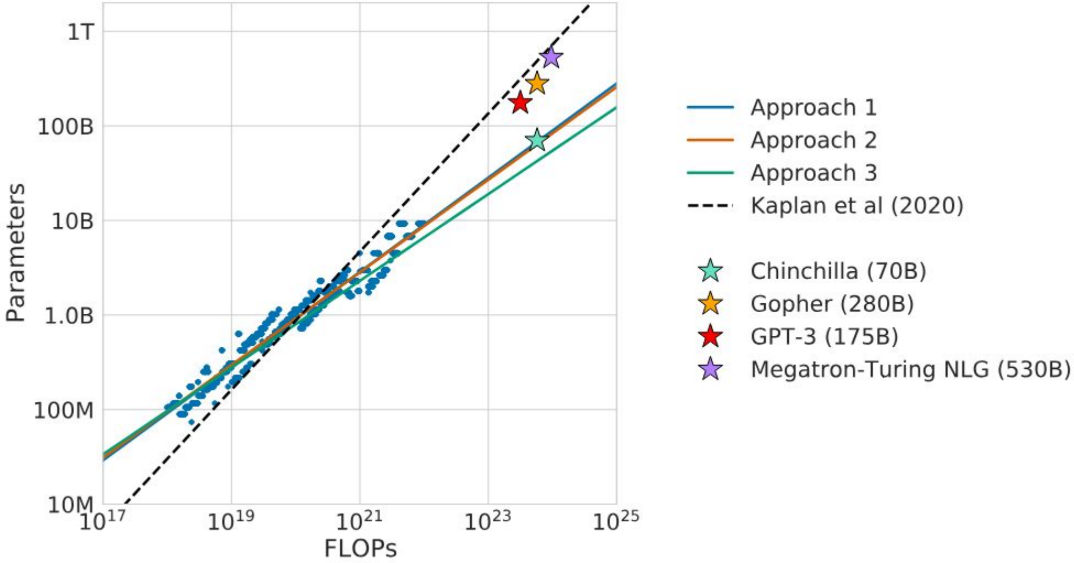
500B scale models / first GPT-3 level open source copycats



stateof.ai 2022

2022b: Largest Models are Undertrained

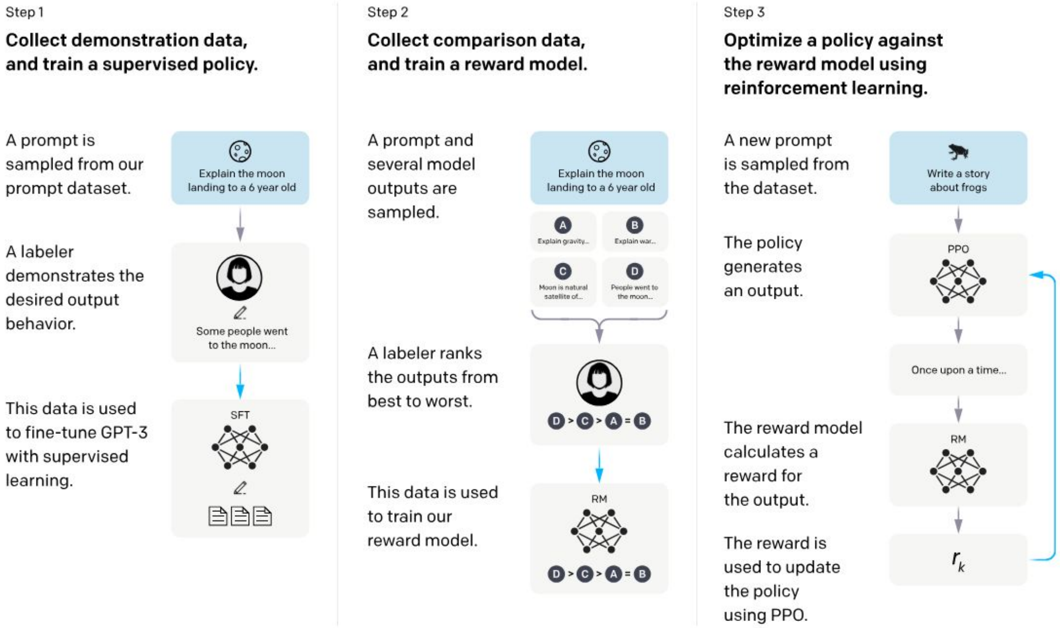
March: Chinchilla & Scaling Laws



Hofmann et al. (2022)

2022c: Instruction Tuning and Alignment

- InstructGPT, Jan 2022
- RLHF, March 2022



Step 1

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3 with supervised learning.

Step 2

Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.

Step 3

Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

The policy generates an output.

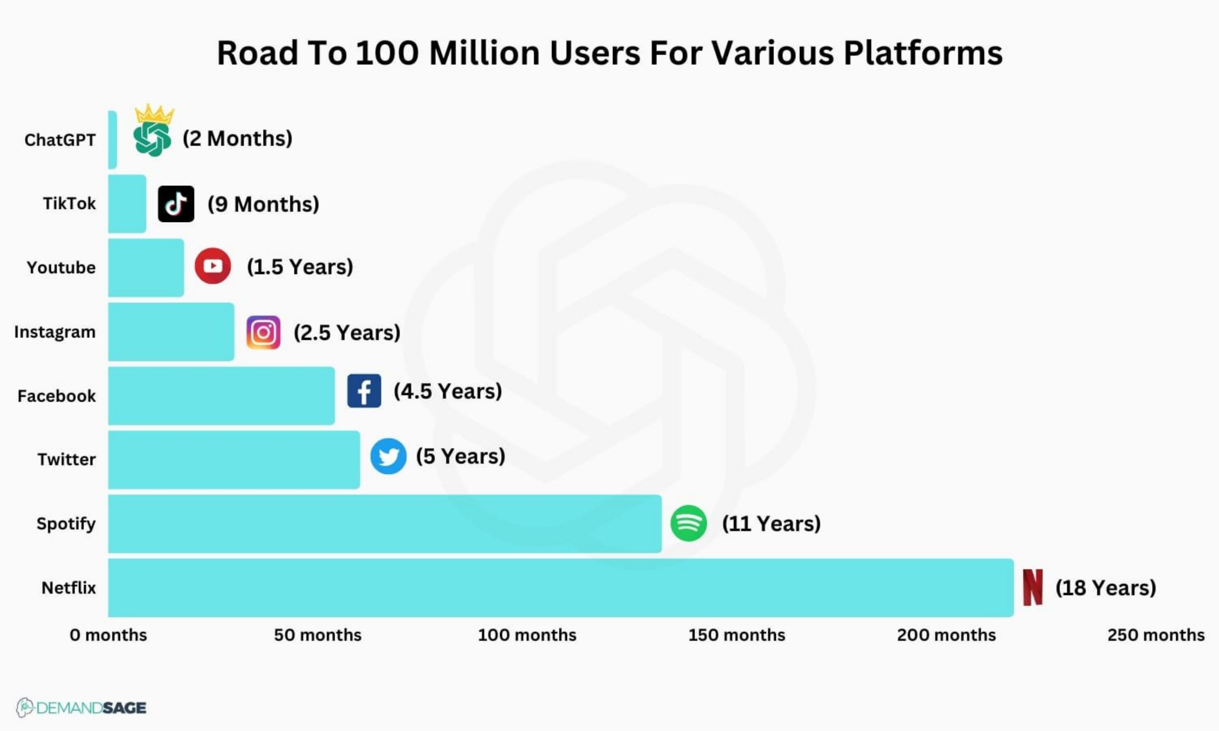
The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.

Explain the moon landing to a 6 year old
Some people went to the moon...
SFT
Explain the moon landing to a 6 year old
A
Explain gravity...
B
Explain war...
C
Moon is natural satellite of...
D
People went to the moon...
D > C > A = B
RM
D > C > A = B
Write a story about frogs
PPO
Once upon a time...
RM
r_k

2023a: Chatbots!

ChatGPT release: November 2022



2023b: GPT-4

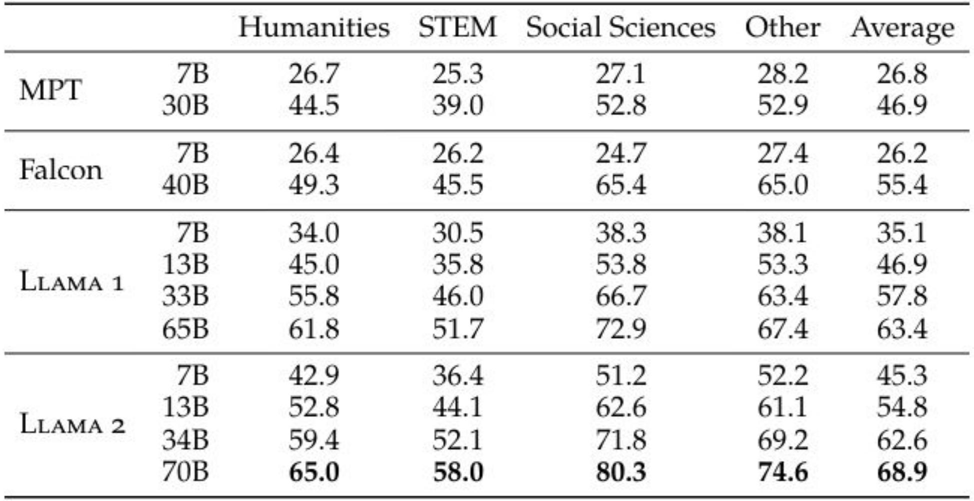
- Released in March
- Significant boost in performance veiled in secrecy
- Multimodal: image inputs + image generation through dall-e
- unofficial/leaked/rumoured details:
  - about 1.8T parameters, 120 layers, (x10 GPT-3)
  - 16 MLP-Experts each with ~111B parameters
  - 13T token training data

2023a: Chatbots!

- GPT-Turbo ChatGPT, March
- Claude 2, July
- Gemini, December

2023c: Competitive Open Weights LLMs

- Llama 1-65B, February
- MPT-30B, March
- Falcon-40B, June
- Llama 2-70B, July
- CodeLlama-70B, August



		Humanities	STEM	Social Sciences	Other	Average
MPT	7B	26.7	25.3	27.1	28.2	26.8
	30B	44.5	39.0	52.8	52.9	46.9
Falcon	7B	26.4	26.2	24.7	27.4	26.2
	40B	49.3	45.5	65.4	65.0	55.4
Llama 1	7B	34.0	30.5	38.3	38.1	35.1
	13B	45.0	35.8	53.8	53.3	46.9
	33B	55.8	46.0	66.7	63.4	57.8
	65B	61.8	51.7	72.9	67.4	63.4
Llama 2	7B	42.9	36.4	51.2	52.2	45.3
	13B	52.8	44.1	62.6	61.1	54.8
	34B	59.4	52.1	71.8	69.2	62.6
	70B	**65.0**	**58.0**	**80.3**	**74.6**	**68.9**

2023c: Competitive Open Weights LLMs

- Llama 1, February 2023

  ~5,500 citations

- Llama 2, July 2023

  ~4,300 citations

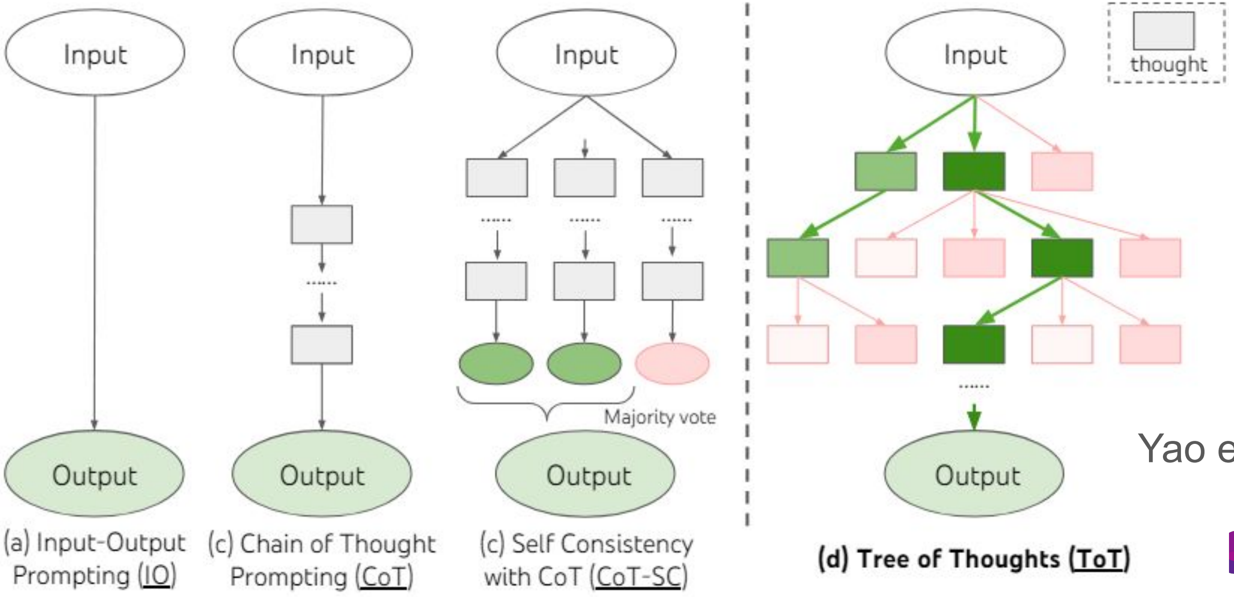
  7B: ~1.3M downloads

  70B: ~380k downloads

- 7,000 GitHub projects mentioning LLama

2023d: Explosion in LLM Projects/Research

- Agents using software tools
- Increasingly sophisticated prompting techniques, e.g. tree of thoughts


Input
Output
(a) Input-Output Prompting (IO)
Input
Output
(c) Chain of Thought Prompting (CoT)
Input
Majority vote
Output
(c) Self Consistency with CoT (CoT-SC)
Input
thought
Output
(d) Tree of Thoughts (ToT)

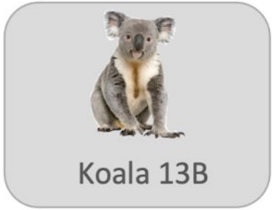
Yao et al. (2023)

2023d: Explosion in LLM Projects/Research

- Agents using software tools
- Increasingly sophisticated prompting techniques, e.g. tree of thoughts
- New alignment methods leads to many Llama derivatives:

Stanford Alpaca

Vicuna



Koala 13B

Orca

WizardLM

Llava

2024a: Next Generation General Purpose Assistants

- Multimodal (Voice, Video, Image)
- Long context (100k-1M tokens)
- Memory
- Strong coding and multilingual
- Tool use / execution environment
- Websearch / document upload

November 2023

February 2024

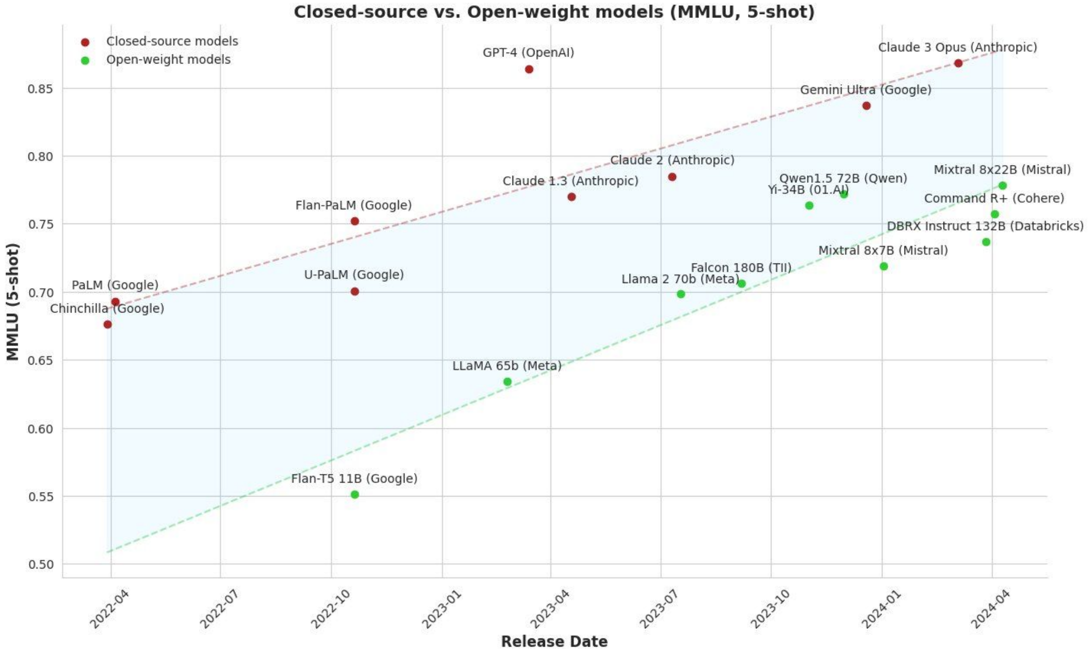
March 2024

yesterday

literally now?

2024b: Open Source/Weights Catching Up

- Grok-1, 314B, March
- DBRX, 132B, March
- Mixtral 8x22B, April
- Llama 3, 70B, April



Labonne (2024)

2024b: Open Source/Weights Catching Up

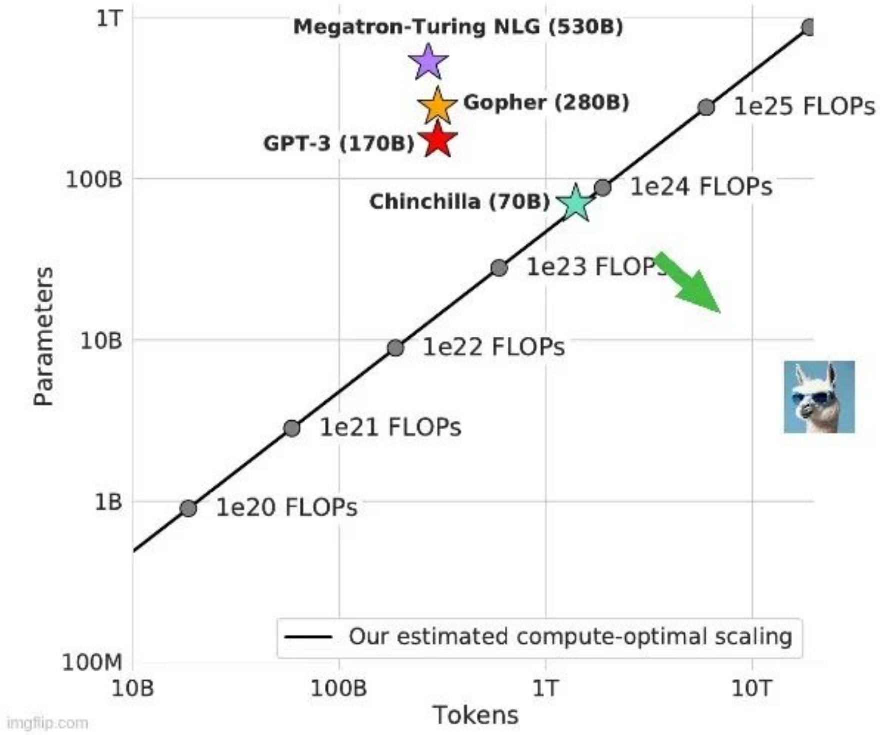
- Grok-1, 314B, March
- DBRX, 132B, March
- Mixtral 8x22B, April
- Llama 3, 70B, April

Inference cost matters!

1. 15T (!) tokens
2. Beats/competitive with Gemini Pro 1.5 and Claude 3 Sonnet

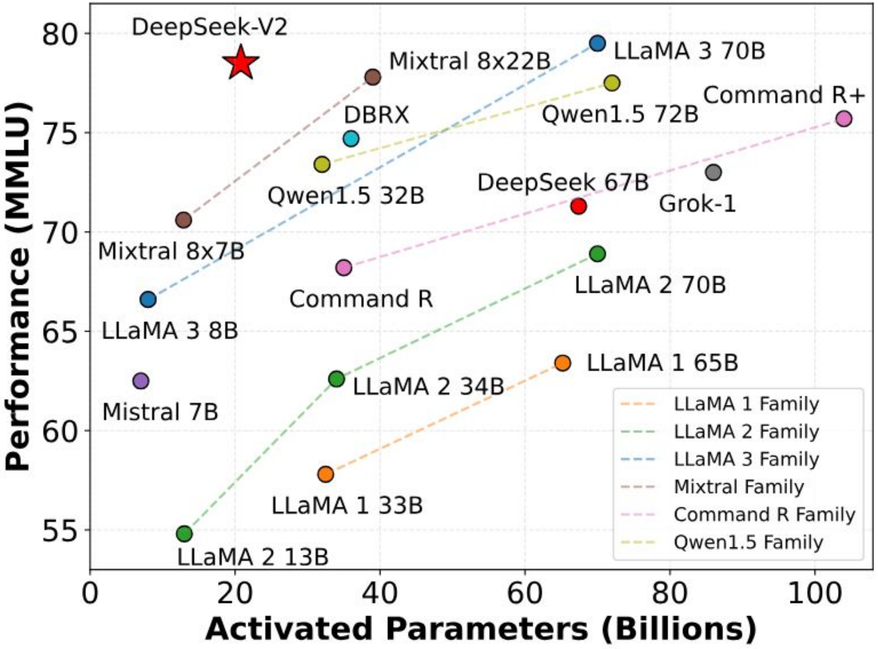
2024b: Open Source/Weights Catching Up

- Grok-1, 314B, March
- DBRX, 132B, March
- Mixtral 8x22B, April
- Llama 3, 70B, April



2024b: Open Source/Weights Catching Up

- Grok-1, 314B, March
- DBRX, 132B, March
- Mixtral 8x22B, April
- Llama 3, 70B, April
- DeepSeek-V2 67B, May

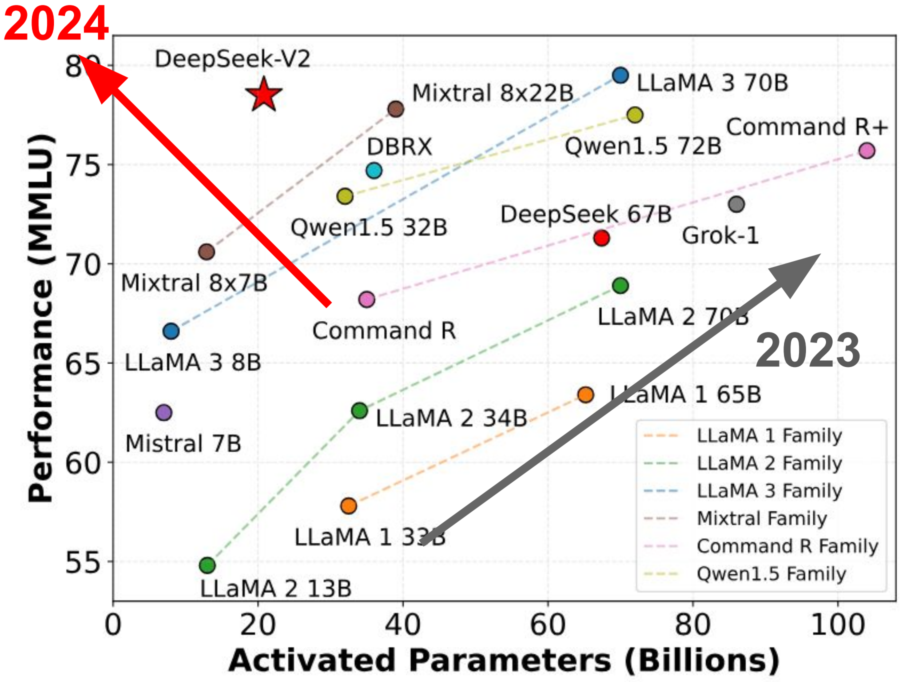


DeepSeek-AI (2024)

ETH AI CENTER

2024b: Open Source/Weights Catching Up!

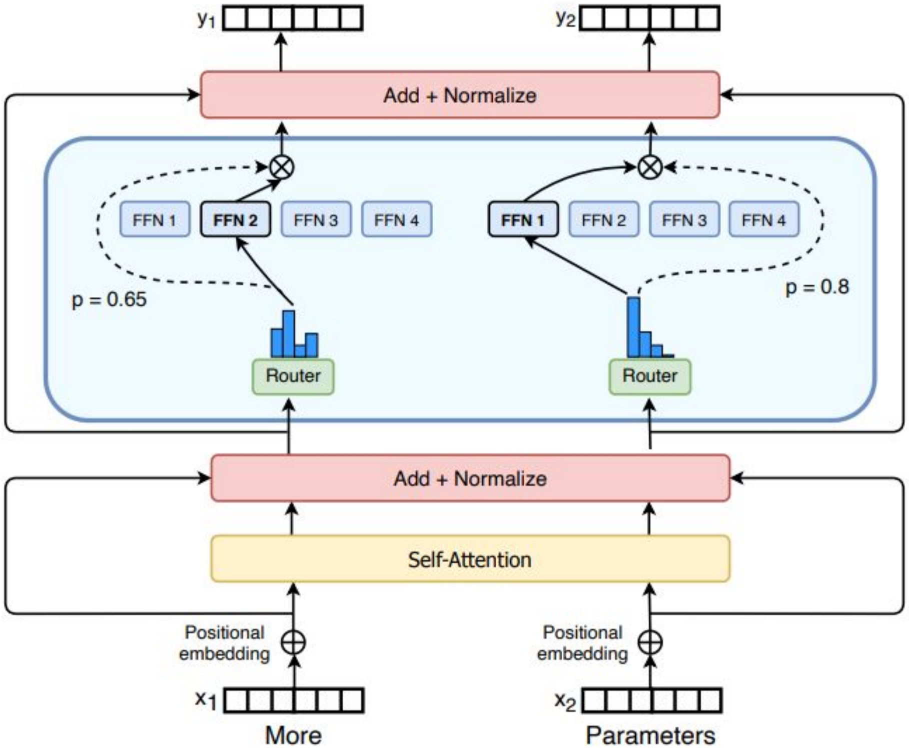
- Grok-1, 314B, March
- DBRX, 132B, March
- Mixtral 8x22B, April
- Llama 3, 70B, April
- DeepSeek-V2 67B, May
- Llama 3 405B, June?

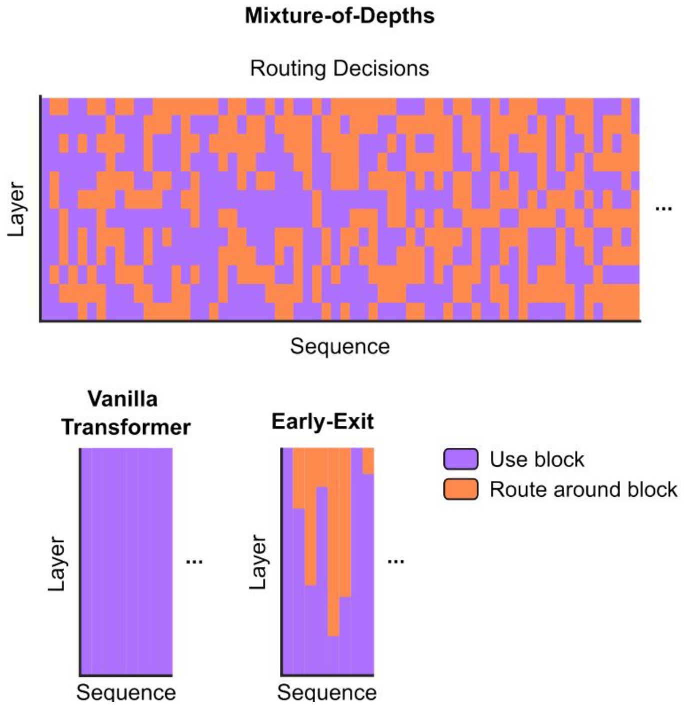


DeepSeek-AI (2024)

ETH AI CENTER

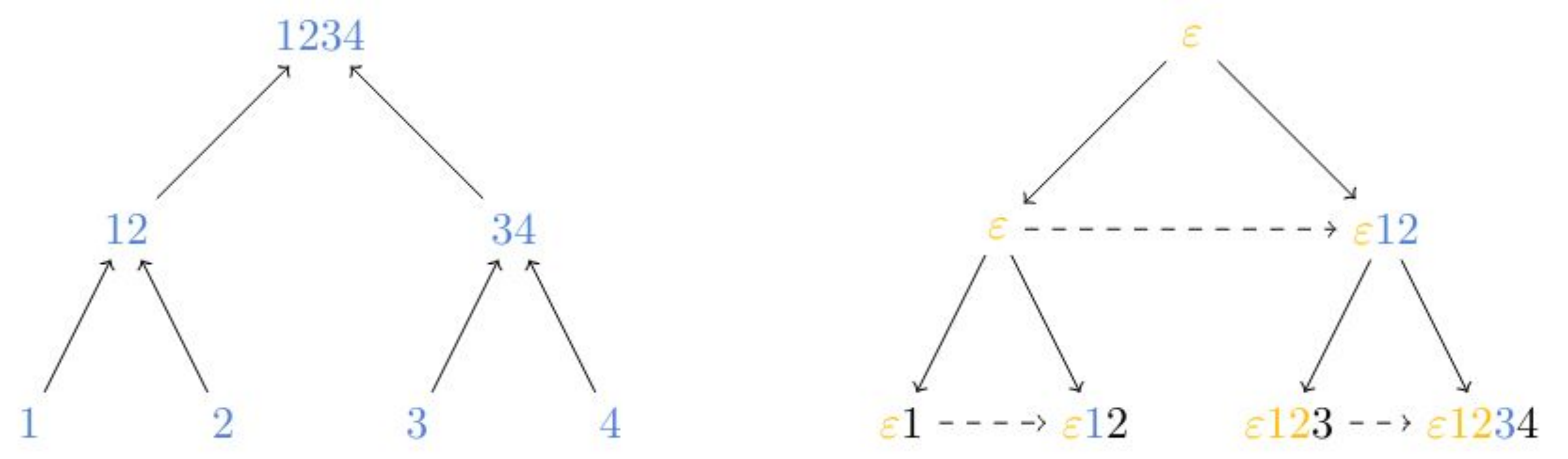
Exciting Trend 1/6: MoE & Adaptive Computation





Exciting Trend 2/6: The Return of Recurrent Nets

Associative RNN cell allows us to “scan” the sequence



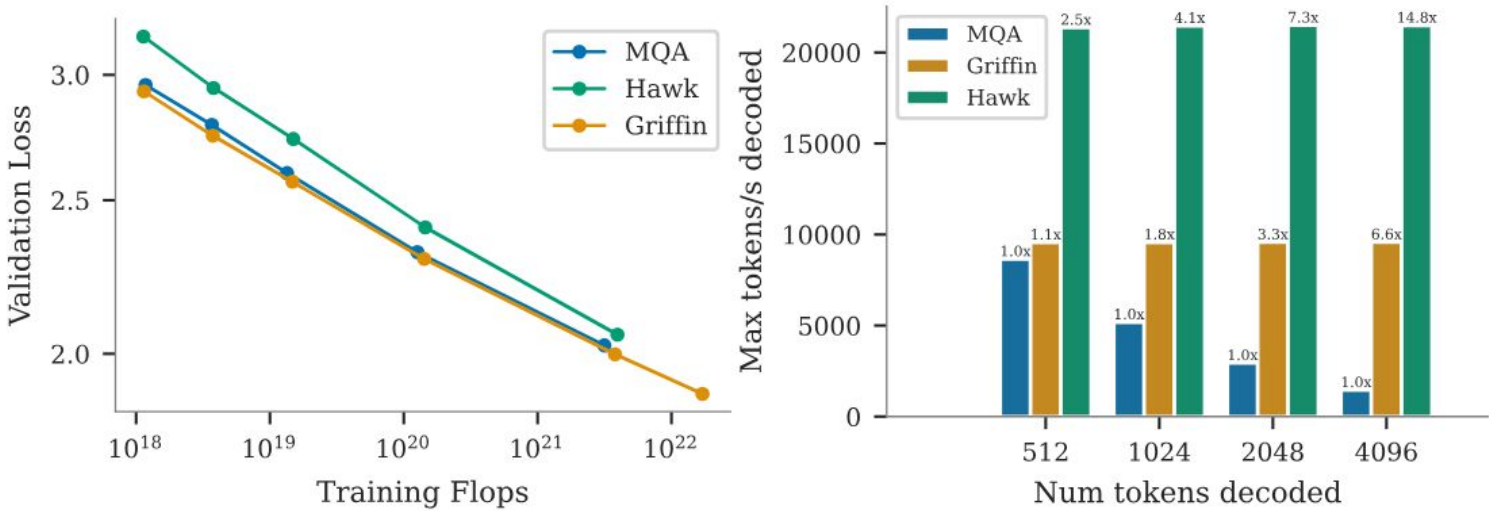
(a) Up: parent combines values of its children

(b) Down: right child combines statistics of its parent with the left sibling.

Blelloch (1990)

Exciting Trend 2/6: The Return of RNNs

Associative RNNs increasingly competitive with Transformers



(a) Scaling curve during training

(b) Maximum throughput at 1B parameter scale.

De et al. (2024)

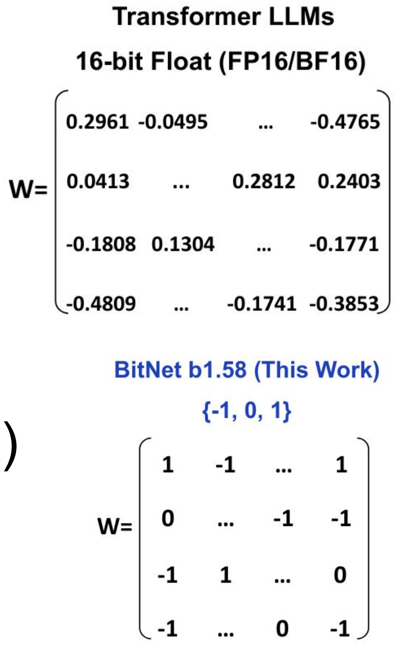
Exciting Trend 3/6: Low Precision & New Hardware

Precision

- fp16 (standard) -> fp8 training (ongoing) -> fp4 (soon)
- Inference with 4 bits, 2 bits, teneray (-1, 0, +1)

Hardware

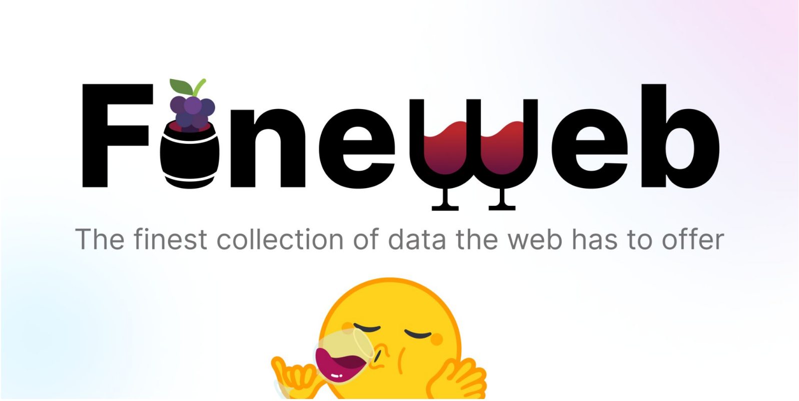
- Specialised hardware for cheaper inference (groq.com)
- custom silicon (Apple, Meta, Google, Amazon, …)
- Nvidia Blackwell (FP4)
- Local LLMs -> embedded devices



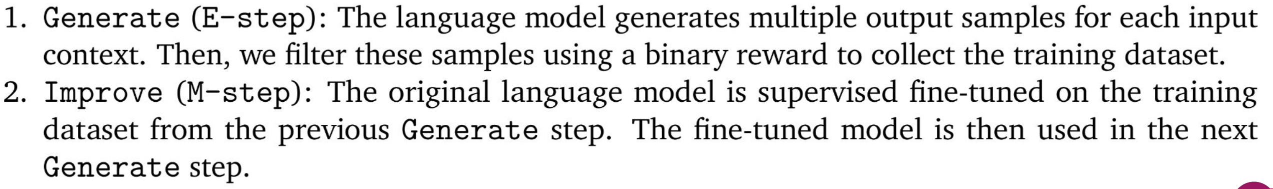
Ma et al. (2024)

Exciting Trend 4/6: Quality Data & Synthetic Data

- High-quality public data for pretrain: FineWeb (15T tokens; webcrawl)



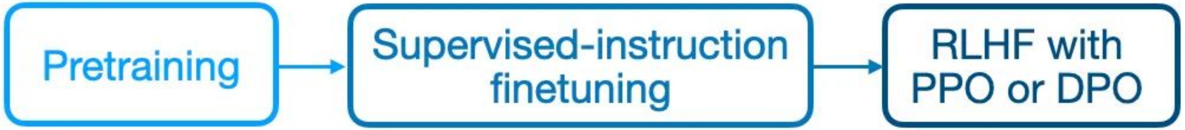
- Synthetic data for reasoning and alignment


1. Generate (E-step): The language model generates multiple output samples for each input context. Then, we filter these samples using a binary reward to collect the training dataset.
2. Improve (M-step): The original language model is supervised fine-tuned on the training dataset from the previous Generate step. The fine-tuned model is then used in the next Generate step.

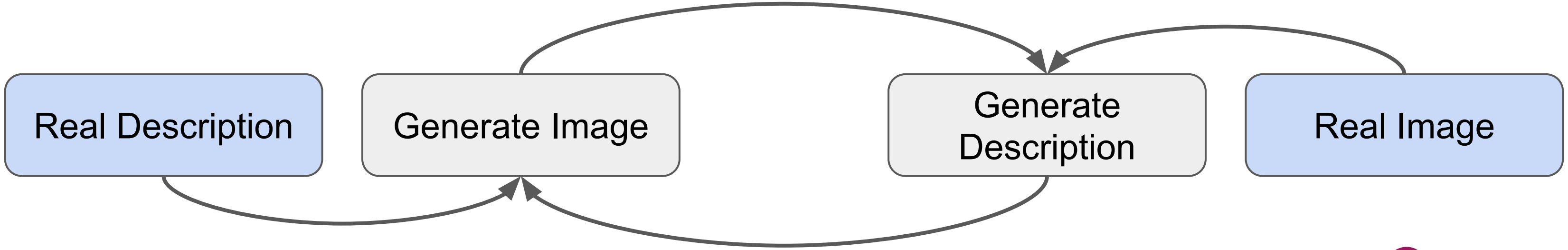
Singh et al. (2024)

Exciting Trend 5/6: Better Alignment

- RLHF: aligning LLMs via reinforcement learning with human feedback

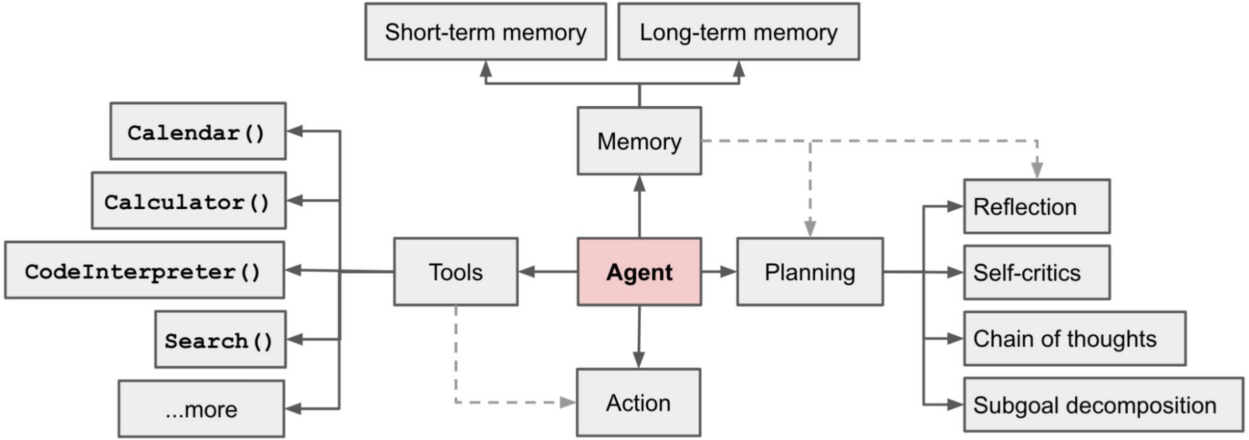


- Alignment across modalities: images, videos, audio



Exciting Trend 6/6: System Interactions & Agents

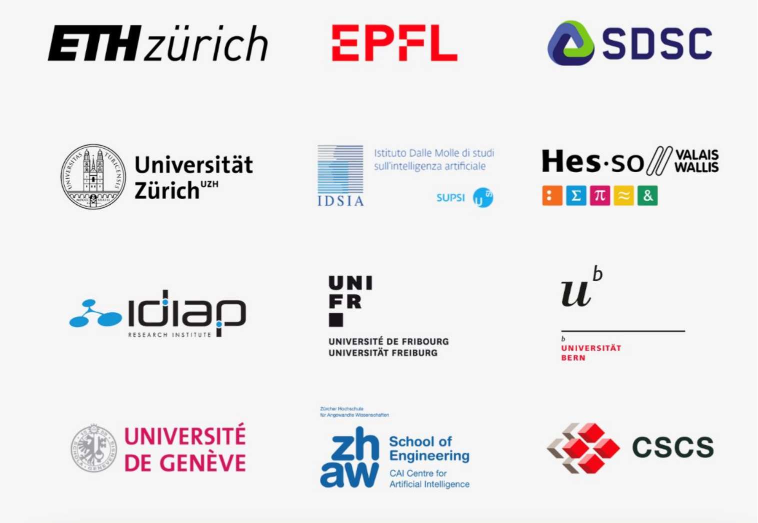
- Agents interacting / Tool use
- Web/Database Search & RAG: Retrieval Augmented Generation
- Coding & execution environment



Lilian Weng (2023)

Swiss AI Initiative: swiss-ai.org

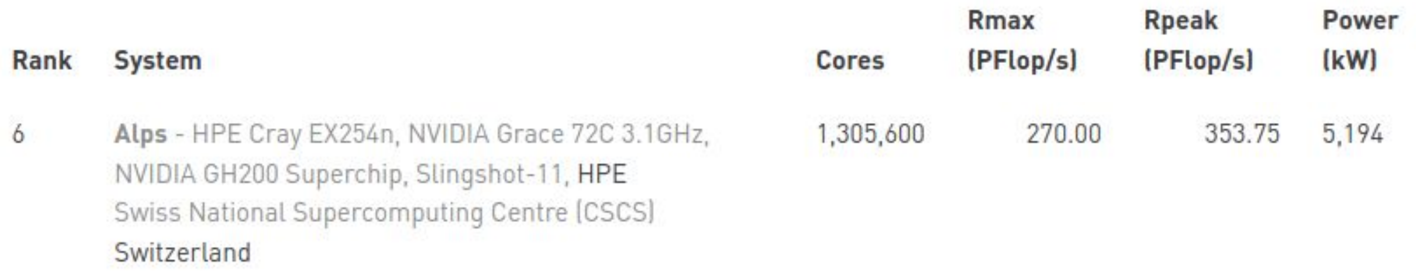
- National Research Initiative jointly lead by ETHZ and EPFL
- Scientific Council: 26 professors / researchers
- Assembly: >100 researchers
- 10M GPU hour commitment on Alps



Swiss AI Initiative: swiss-ai.org

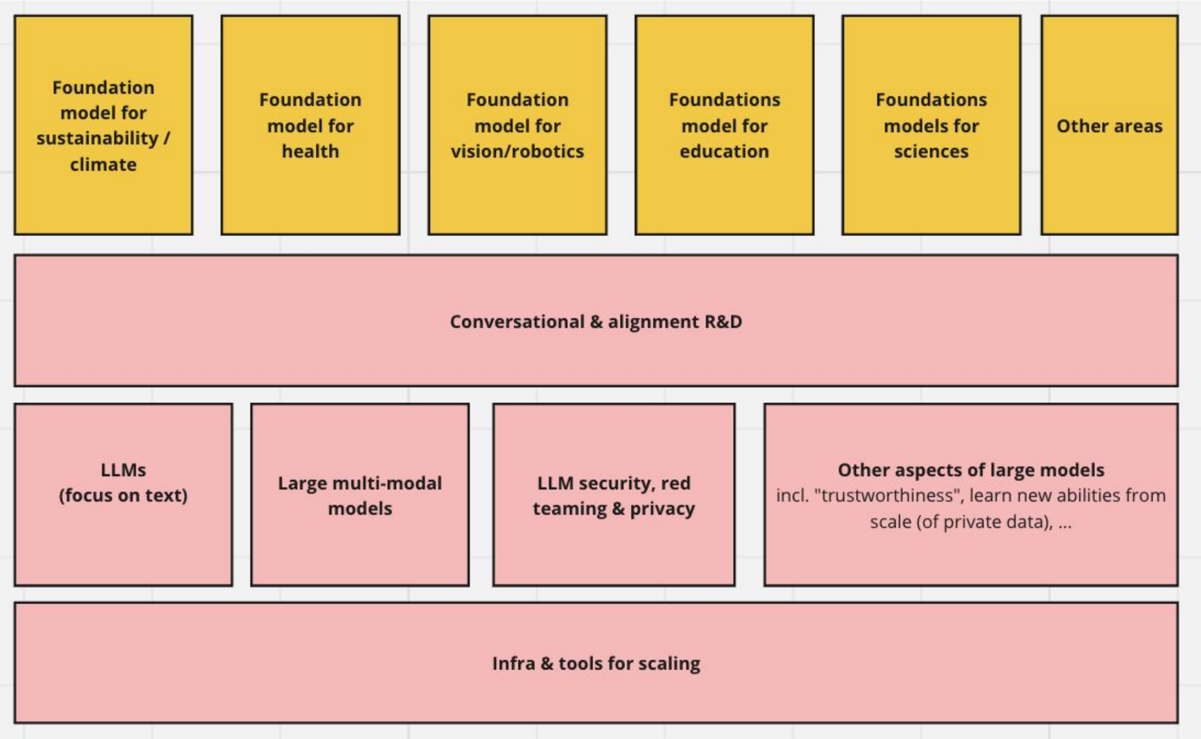
Alps Supercomputer: 10’000 GH200 GPUs

TOP 500 The List.



Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
6	**Alps** - HPE Cray EX254n, NVIDIA Grace 72C 3.1GHz, NVIDIA GH200 Superchip, Slingshot-11, HPE Swiss National Supercomputing Centre (CSCS) Switzerland	1,305,600	270.00	353.75	5,194

Swiss AI Initiative: swiss-ai.org



Foundation model for sustainability / climate

Foundation model for health

Foundation model for vision/robotics

Foundations model for education

Foundations models for sciences

Other areas

Conversational & alignment R&D

LLMs (focus on text)

Large multi-modal models

LLM security, red teaming & privacy

Other aspects of large models
incl. "trustworthiness", learn new abilities from scale (of private data), ...

Infra & tools for scaling

Swiss AI Initiative: swiss-ai.org

LLM Area:

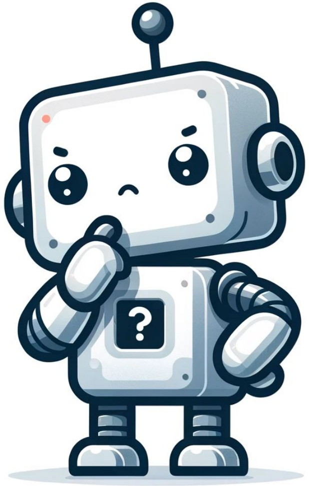
- An LLM for Switzerland
- Trustworthy and Responsible
- Transparent and compliant (open source / open weights)
- Multilingual with Swiss societal values

- Attract and develop talent
- Startup fuel
- Teaching and sharing lessons, code, models, …

- Collaborations: users, developers, legal, ….

Questions?

Thank you for your attention.



Feel free to get in touch:

Imanol Schlag
ischlag@ethz.ch