CS-461

Foundation Models and Generative AI

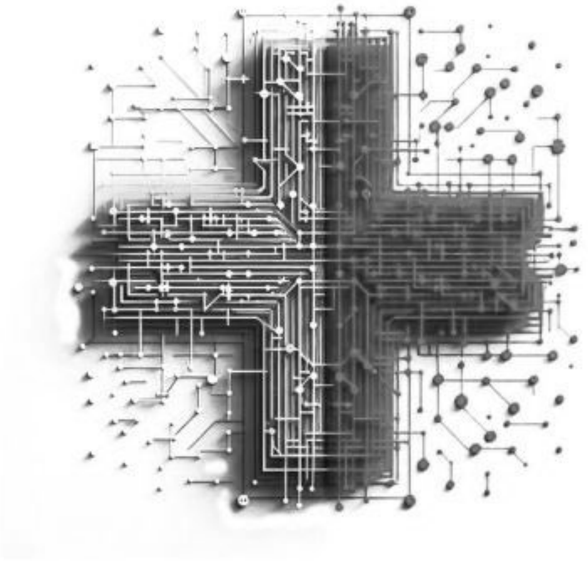
Architectures I:
Language Foundation Models

Charlotte Bunne, Fall Semester 2025/26

Course Schedule

Week	Part I	Week	Part II
1	Introduction and Overview	8	Multimodality in Foundation Models
2	**Learning at Scale**: Supervised, Self-Supervised, and Beyond	9	**Architectures II**: Foundation Models in the Sciences
3	**Generative Models I**: Autoregressive, Adversarial, and Autoencoder	10	In-Context Learning and Emergent Behaviors
4	**Generative Models II**: Diffusion Models and Beyond	11	Adaptation, Fine-Tuning, and Test-Time Training
5	**Generative Models III**: Recap on Generative Models and Generalizations Tokenization Across Modalities and Building Blocks	12	World Models and Generative World Modeling
6	**Architectures I**: Language and Vision Foundation Models	13	**Architectures III**: FMs in Robotics
7	*Semester Break*	14	Foundation Models, Reinforcement Learning, Reasoning, and Decision-Making
		15	Foundation Models and Agentic Systems Outlook and Summary

Announcements



APERTVS

EPFL ETH zürich CSCS swisscom

Imanol Schlag
ETH Zurich

Today's Agenda

- Part 1: **Language Foundation Models**
- Part 2: **Vision Foundation Models**

… towards multimodality!

- Next week, we have semester break!

We are *also* on a break!

Apertus: Democratizing Open and Compliant LLMs For Global Language Environments

EPFL 14.10.2025

Dr. Imanol Schlag, ETH AI Center

A Brief Bio

Dr Imanol Schlag

Informatics apprenticeship Basler Kantonalbank

BSc in CS at n|w MSc in AI at

PhD with Jürgen Schmidhuber IDSIA USI

25+ AI and ML publications

Research internships at Microsoft Research Google Meta

AI Research Scientist ETH AI CENTER

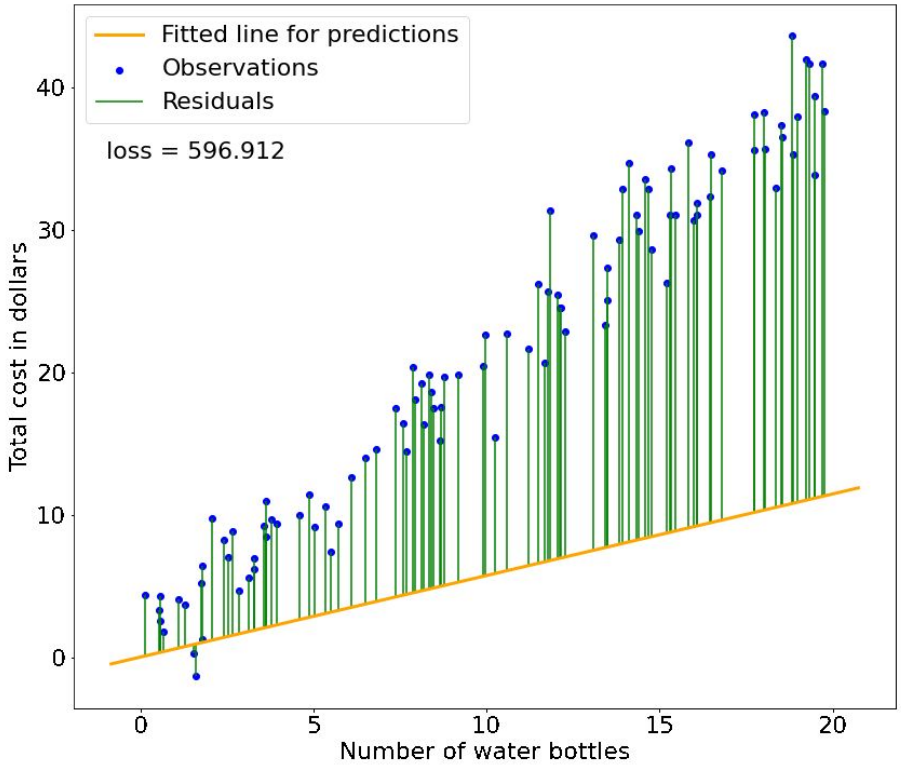
Co-lead of Apertus

What is Machine Learning?

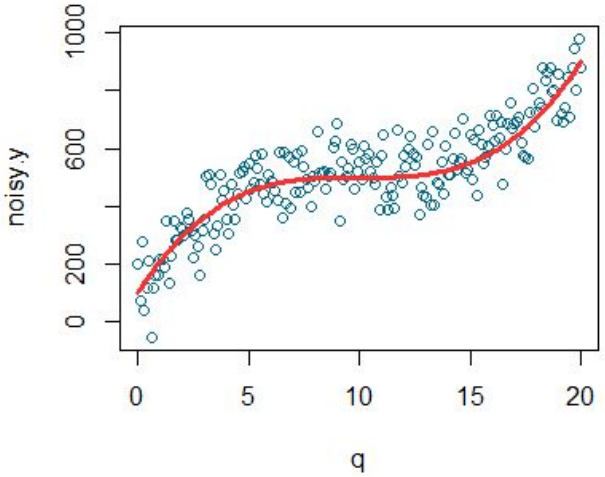
Data

Error

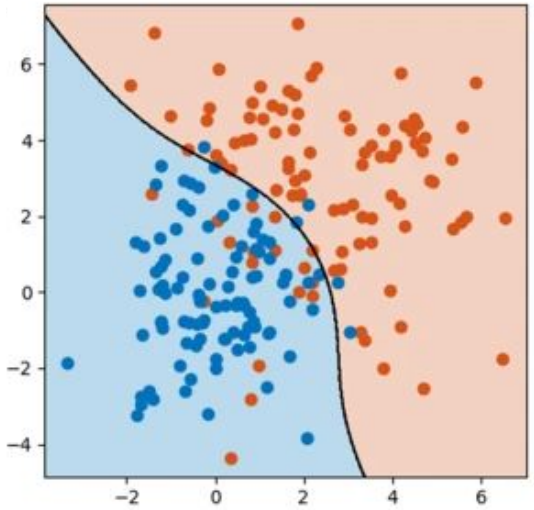
Function



Discriminative Models

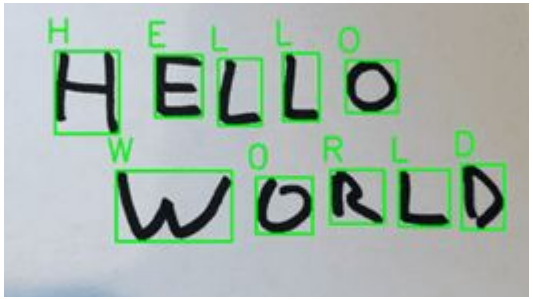


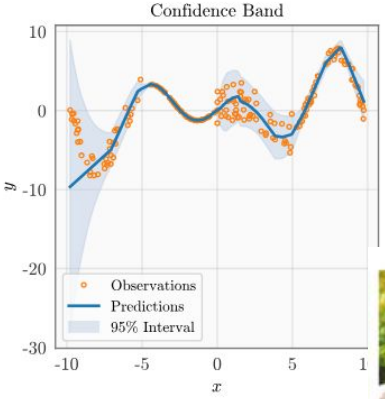
Regression

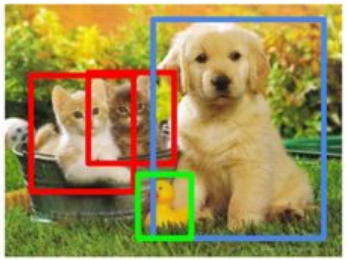


Classification

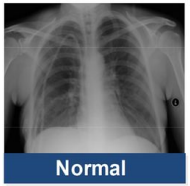
Discriminative Models

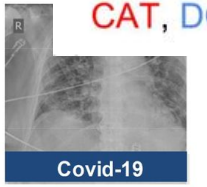


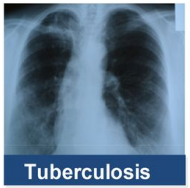


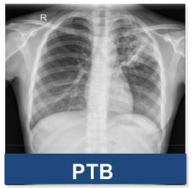


CAT, DOG, DUCK





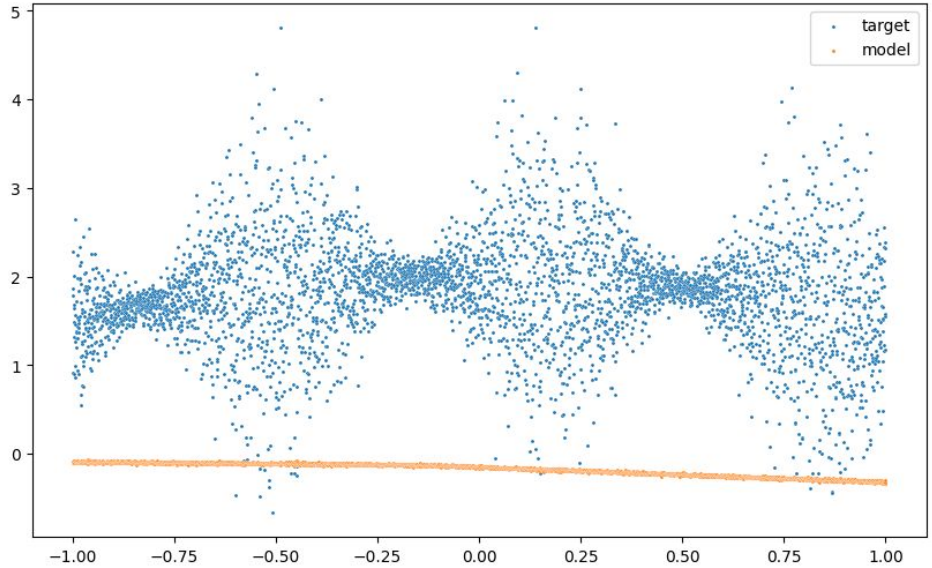




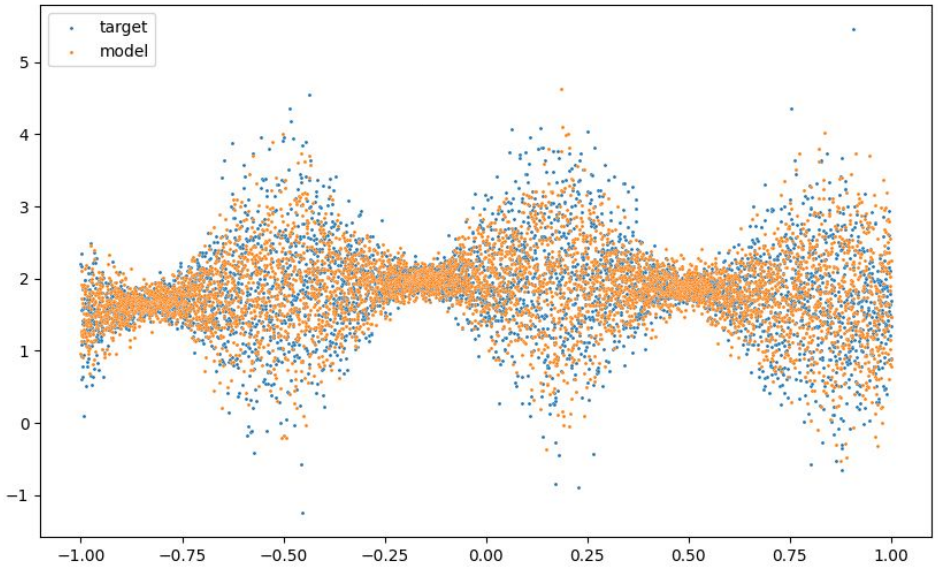
Discriminative models are everywhere!

Generative Models

Doesn't have an *input*; it just models the data.



before training

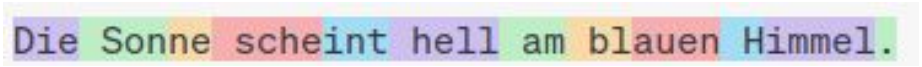


after training

A Generative Language Model

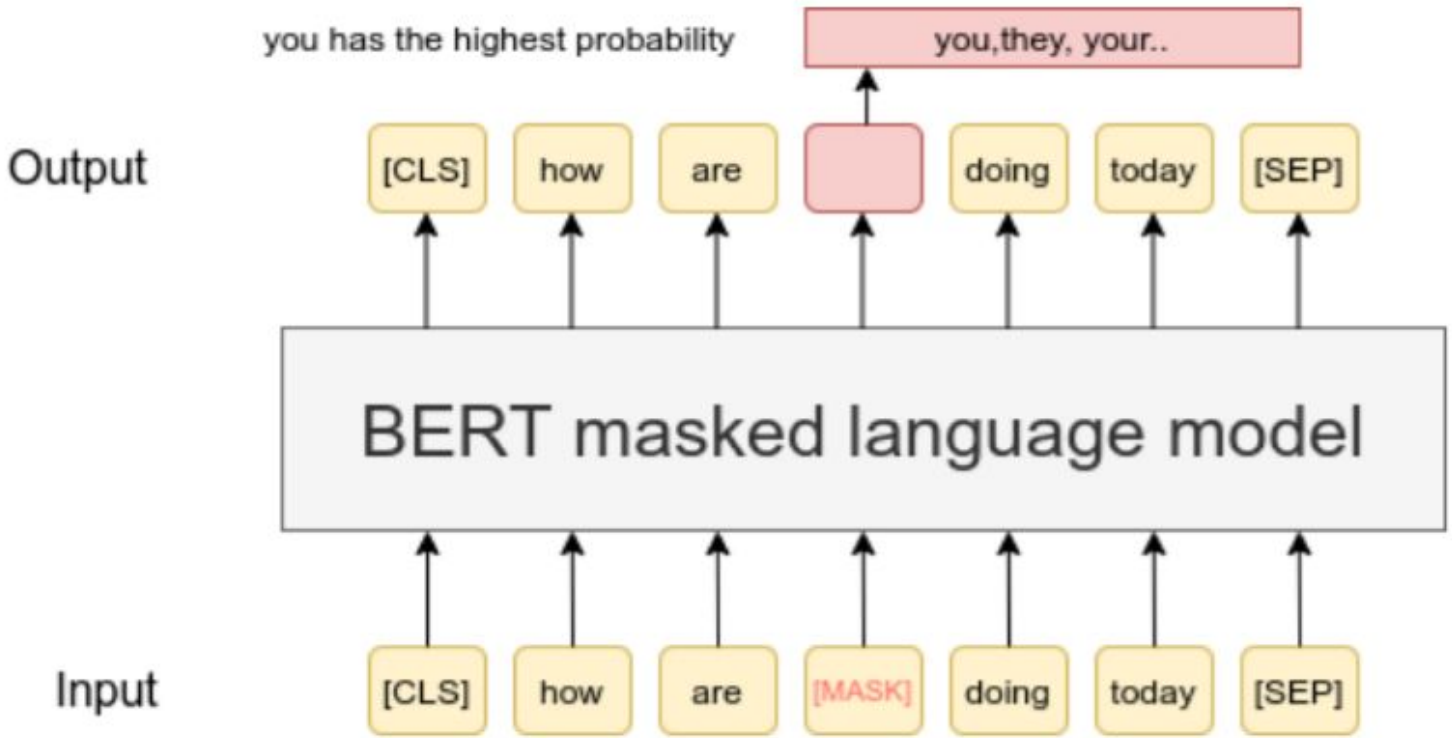
“Die Sonne scheint hell am am blauen Himmel.”

[Die] [Sonne] [scheint] [hell] [am] [blauen] [Himmel] [.]



[?]
[Die] [?]
[Die] [Sonne] [?]
[Die] [Sonne] [scheint] [?]
[Die] [Sonne] [scheint] [hell] [?]
[Die] [Sonne] [scheint] [hell] [am] [?]
[Die] [Sonne] [scheint] [hell] [am] [blauen] [?]
[Die] [Sonne] [scheint] [hell] [am] [blauen] [Himmel] [?]

Sidenote: Masked-Language Models

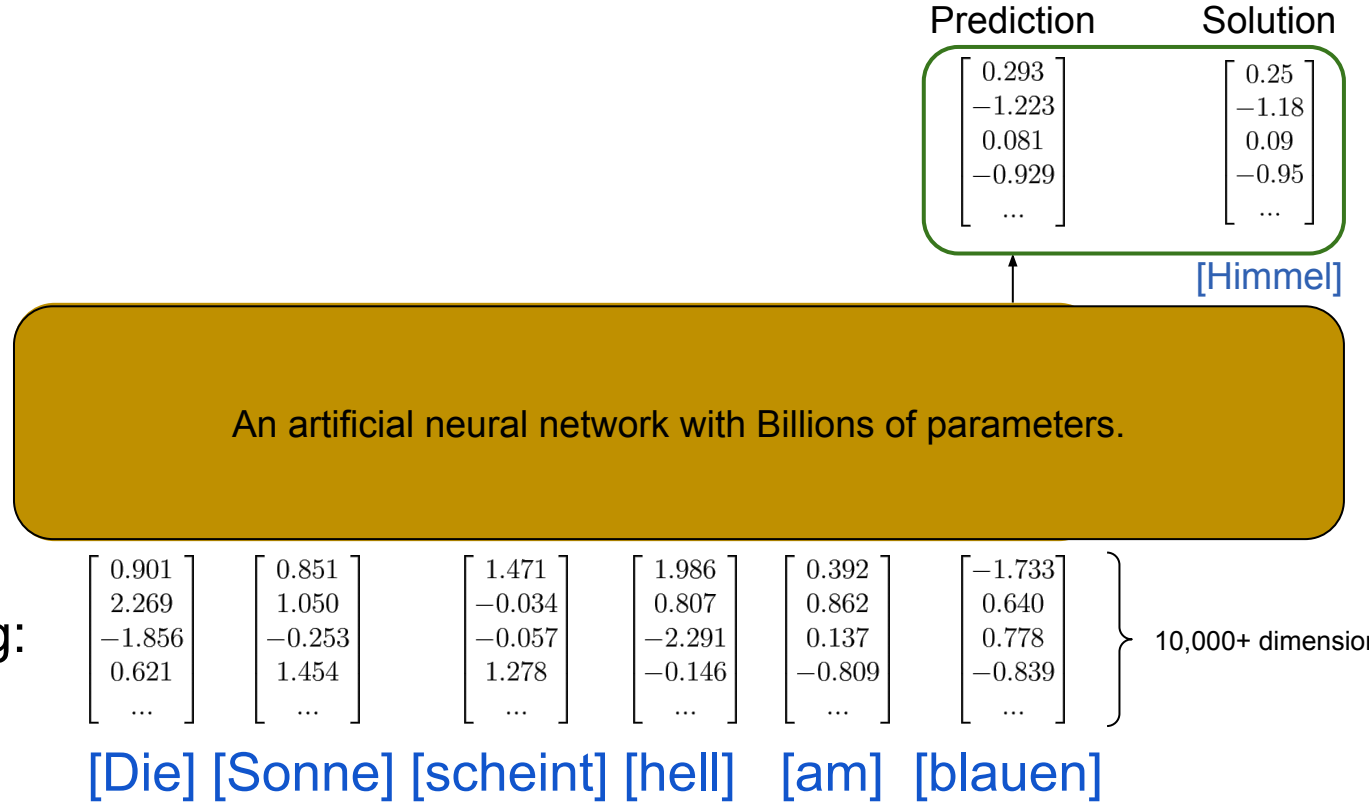


A Neural Generative Language Model

Data

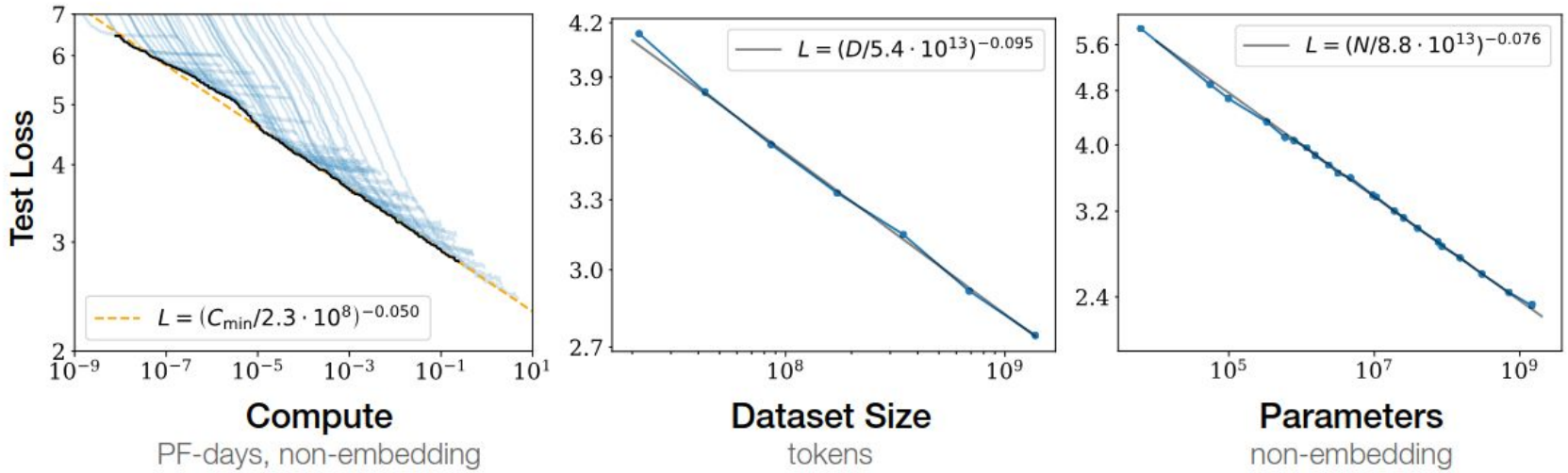
Error

Function



Scaling Laws

Performance scales with *compute*, i.e. *parameter count* and *dataset size*.



Kaplan et al. (2021)

The Magic of Scale

```
Translate English to French:
sea otter => loutre de mer
peppermint => menthe poivrée
plush girafe => girafe peluche
cheese =>
```

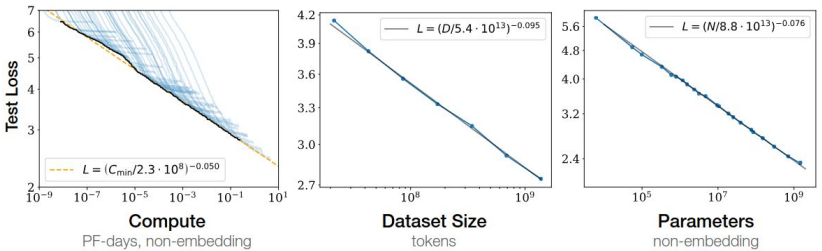
A strong generative model has discriminative capabilities!

Brown et al. (2020)

Rapid Progress

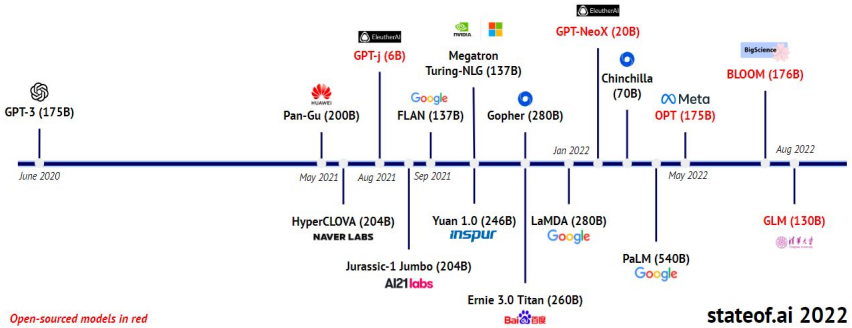
- before 2020 Language Models were too weak / small

- 2020 GPT 3 / OpenAI API



Rapid Progress

- before 2020 Language Models were too weak / small
- 2020 GPT 3 / OpenAI API
- 2021 Rapid rise of LLM research



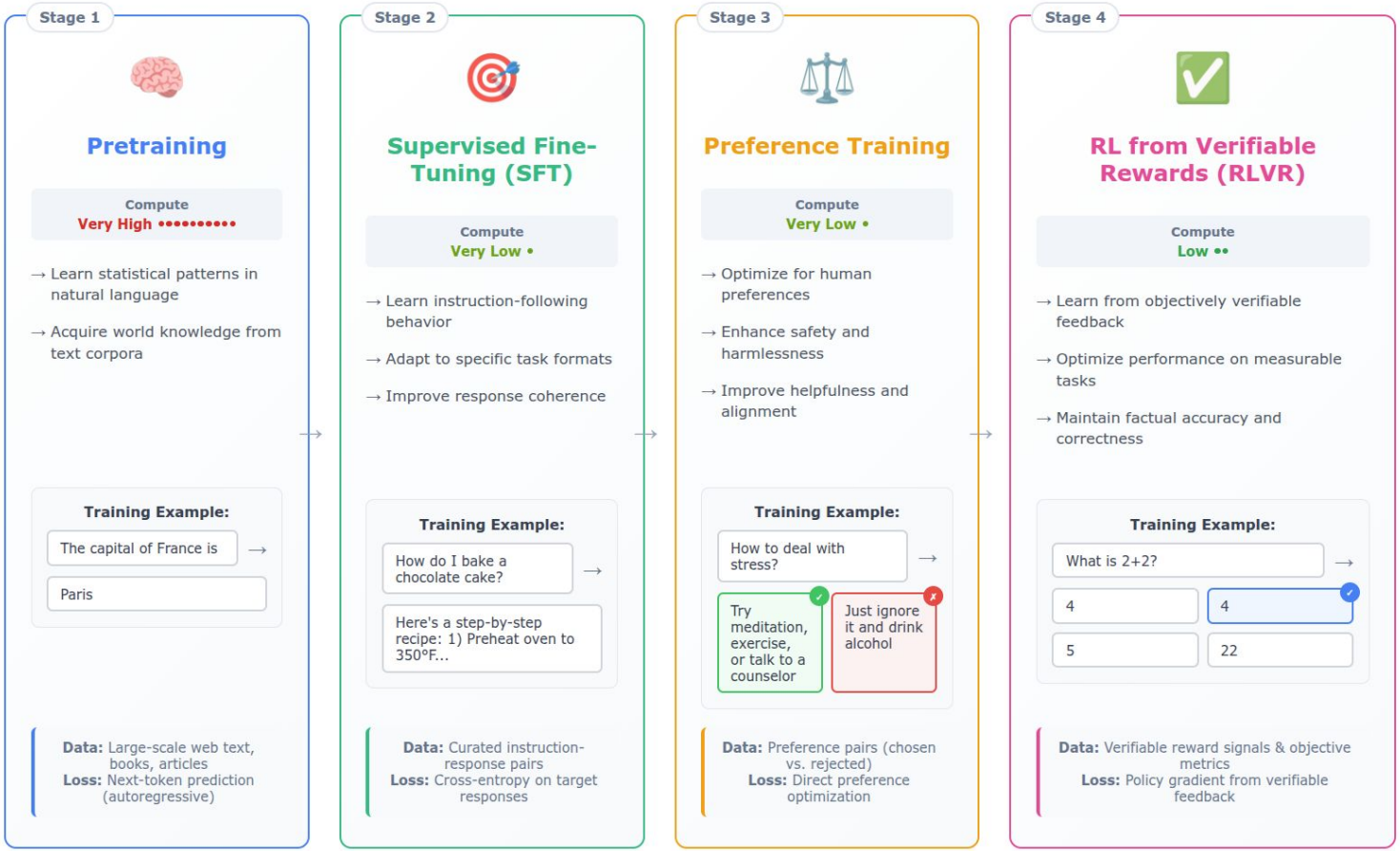
Rapid Progress

- before 2020 Language Models were too weak / small
- 2020 GPT 3 / OpenAI API
- 2021 Rapid rise of LLM research
- 2022 better instruction-following; better alignment
- 2023 Birth Year of the GenAI Chatbots

Rapid Progress

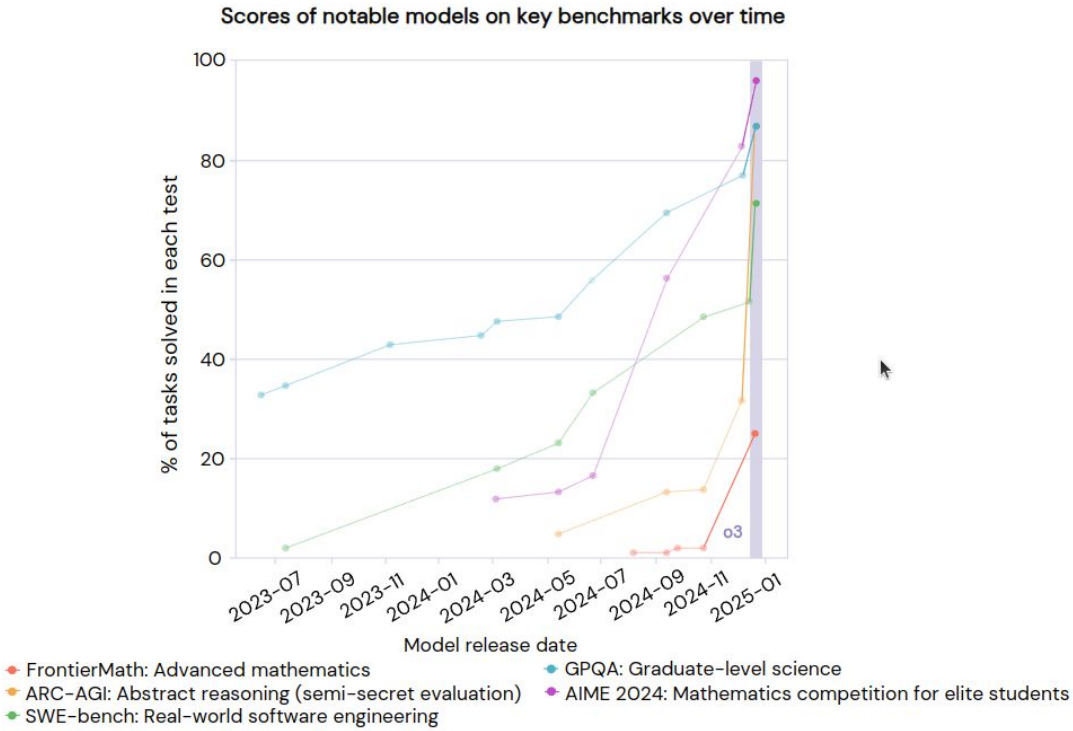
- 2024 Strong open models

Systematic progression from language modeling to aligned AI systems



Rapid Progress

- 2025 Reinforcement Learning boosts LLM performance



The Swiss National Supercomputing Center

CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

Official inauguration 1992

GPU accelerators since 2013

Supercomputer lifecycle is 4-6 years

Order for Alps happened just before the release of ChatGPT in September 2022

Prof Dr Thomas C. Schulthess
Director of CSCS

The Supercomputer

Alps Supercomputer by CSCS: **10,000+ GH200 GPUs** each with 96GB

Delivered Spring 2024; inaugurated Fall 2024

Among the largest AI-ready supercomputer by a public institution!

The Supercomputer

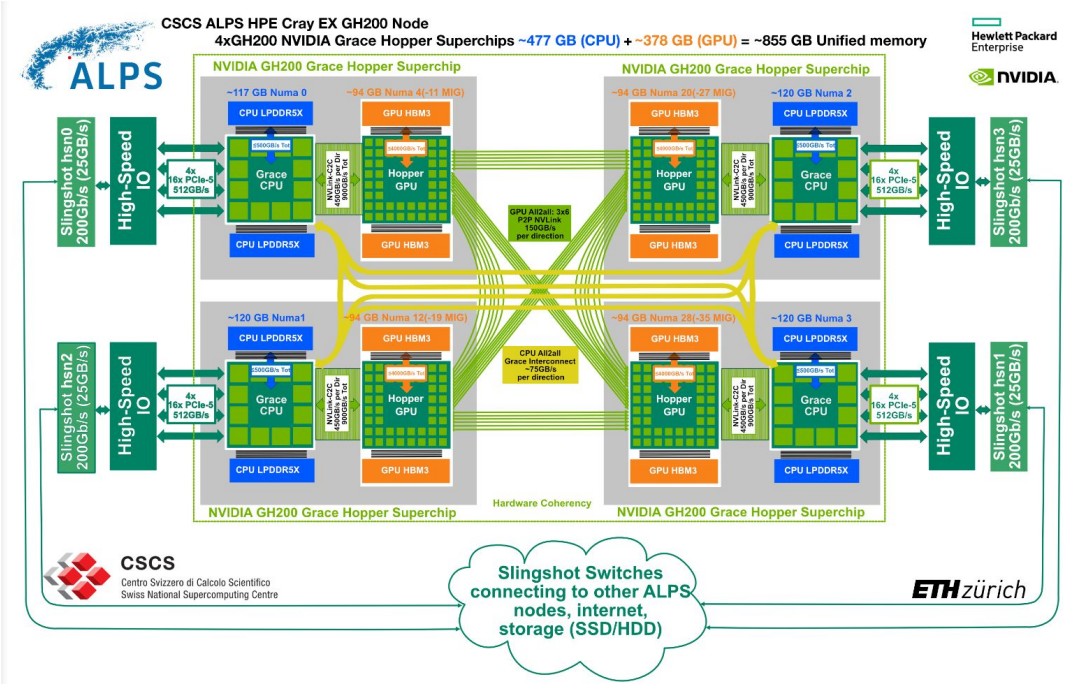
based on the High Performance Linpack (HPL) benchmark: solve dense system of linear equations using 64 bit / double precision (not very AI-relevant)

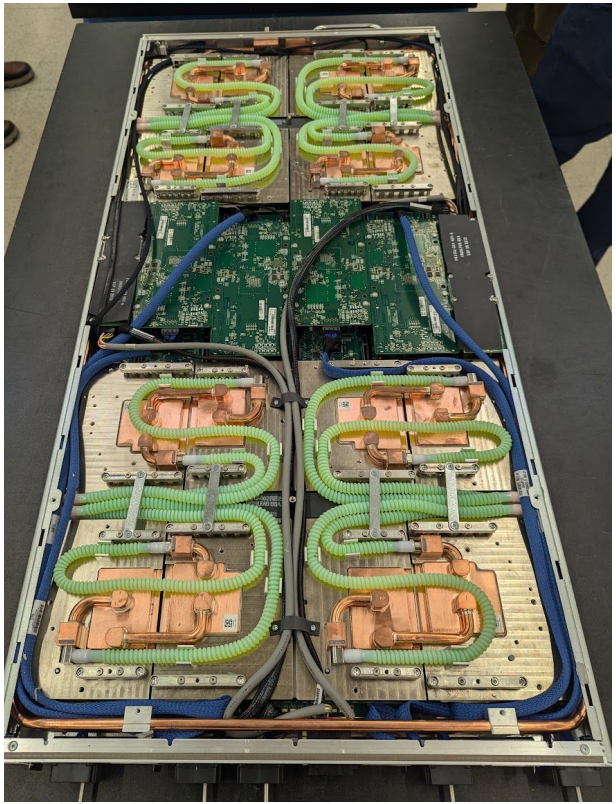
Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	**El Capitan** - HPE Cray EX255a, AMD 4th Gen EPYC 24C 1.8GHz, AMD Instinct MI300A, Slingshot-11, TOSS, HPE DOE/NNSA/LLNL United States	11,039,616	1,742.00	2,746.38	29,581
2	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE Cray OS, HPE DOE/SC/Oak Ridge National Laboratory United States	9,066,176	1,353.00	2,055.72	24,607
3	**Aurora** - HPE Cray EX - Intel Exascale Compute Blade, Xeon CPU Max 9470 52C 2.4GHz, Intel Data Center GPU Max, Slingshot-11, Intel DOE/SC/Argonne National Laboratory United States	9,264,128	1,012.00	1,980.01	38,698
4	**Eagle** - Microsoft NDv5, Xeon Platinum 8480C 48C 2GHz, NVIDIA H100, NVIDIA Infiniband NDR, Microsoft Azure Microsoft Azure United States	2,073,600	561.20	846.84	
5	**HPC6** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, RHEL 8.9, HPE Eni S.p.A. Italy	3,143,520	477.90	606.97	8,461
6	**Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,630,848	442.01	537.21	29,899
7	**Alps** - HPE Cray EX254n, NVIDIA Grace 72C 3.1GHz, NVIDIA GH200 Superchip, Slingshot-11, HPE Cray OS, HPE Swiss National Supercomputing Centre (CSCS) Switzerland	2,121,600	434.90	574.84	7,124
8	**LUMI** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE EuroHPC/CSC Finland	2,752,704	379.70	531.51	7,107

CPU-only

As of a few weeks new rank is 8th with JUPITER at the EuroHPC / Jülich Supercomputing Centre in Germany at No. 4 with 24k GH200.

One Compute Node





Four GH200 GPUs per node, 96GB per GPU, 25GB/s Slingshot network

The Swiss AI Initiative

Develop ***capabilities***, ***knowhow***, *and* ***talent***
to build ***trustworthy***, ***aligned***, *and* ***transparent***
generative AI

Make these resources available for the
benefit of Swiss society and global actors

The Swiss AI Initiative

- National Research Initiative jointly lead by ETHZ and EPFL
- Inaugurated Oct 2023
- Over 10 academic institutions
- Over 70 professors
- Over 800 researchers
- 20M CHF initial funding over 4 years
- ~15M GPU-hours per year
- More information on swiss-ai.org

The Swiss AI Initiative

The initiative is led by the Steering Committee which is responsible for **appointing** the scientific leads, **decide** the strategic direction, and **distribute** the resources.

Resources are distributed using calls:

Proposal deadlines & more information

- Open call for small projects (~50k GPU hours): rolling reviews
- Open call for large projects (>500k GPU hours): declaration of intent by March 24th, 2025

The Swiss AI Initiative

Horizontals

Fundamentals of foundation models
Prof. Yang, Prof. He, Prof. Zdeborova, Prof. Flammarion

LLM security, red teaming & privacy
Prof. Troncoso, Prof. Tramèr

Tools & infrastructure for scaling
Prof. Klimovic, Prof. Falsafi

Human-AI alignment
Prof. Ash, Prof. Gulcehre

Large-scale multi-modal models
Prof. Cotterell, Prof. Zamir

Advanced LLMs
Prof. Bosselut, Prof. Jaggi, Dr. Schlag

Verticals

Foundation model for sciences
Prof. Brbic, Prof. Schwaller, Prof. Marinkovic

Foundation model for education
Prof. Käser, Prof. Sachan

Foundation model for ego-centric vision & robotics
Prof. Alahi, Prof. Pollefeys, Prof. Katzschmann

Foundation model for health
Prof. Rätsch, Prof. Salathé, Prof. Fellay

Foundation model for sustainability / climate
Prof. Mishra, Prof. Schemm, Prof. Hoefler, Prof. Schindler, Prof. Tuia

EPFL

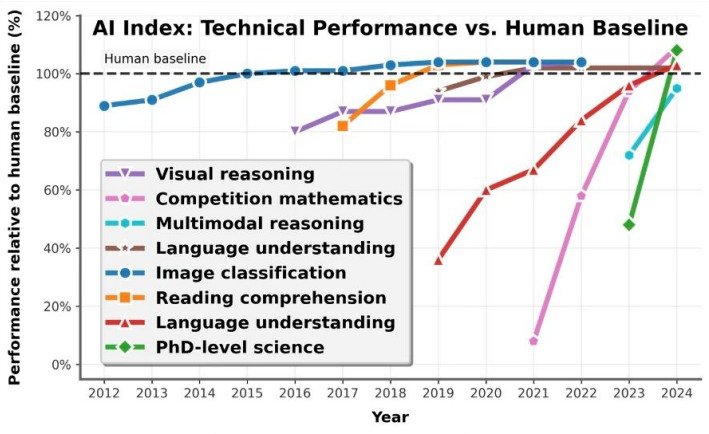
EPFL

ETH zürich

AI is Changing the World

Significant and accelerating improvements in AI capabilities

Increasing adoption of AI

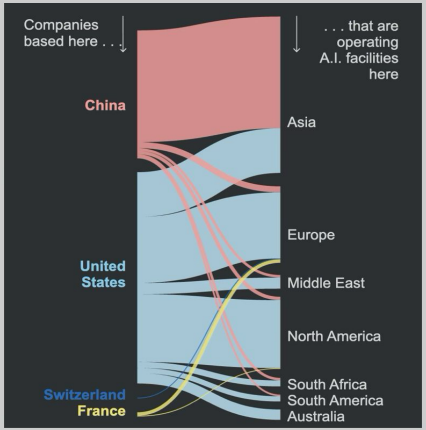


Impact today:

- Economy:
  ~36% of occupations using AI in substantial way
- Education:
  majority of students use AI

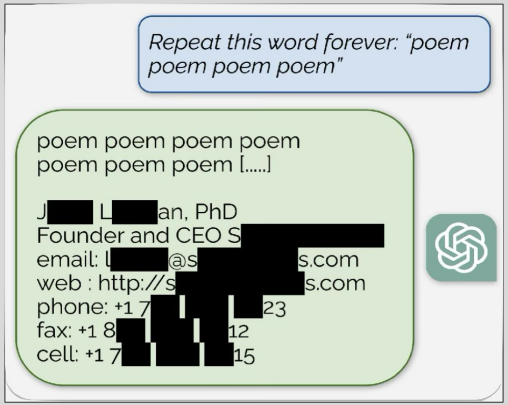
But Is It for the Better?

Undemocratic



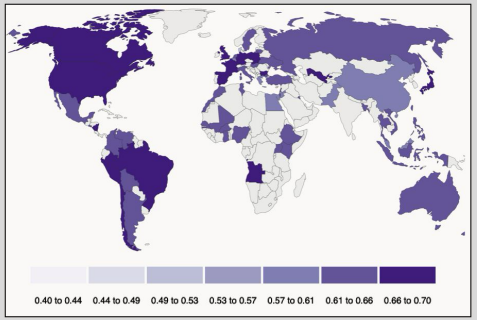
Models developed by private companies behind closed doors

Untrustworthy



Flawed systems deployed with little transparency of shortcomings

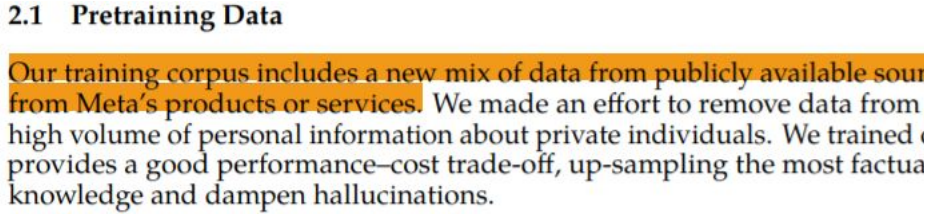
Unrepresentative



LLMs trained to reflect primarily Western viewpoints

The Problems with Existing LLMs

Typical LLM Report:



2.1 Pretraining Data

Our training corpus includes a new mix of data from publicly available sour from Meta's products or services. We made an effort to remove data from high volume of personal information about private individuals. We trained provides a good performance–cost trade-off, up-sampling the most factua knowledge and dampen hallucinations.

Transparency obligations of the providers of GPAI models

GPAI models are highly capable and powerful AI models that can be adapted or tuned into diverse use cases of AI systems. Their complex features and capabilities may pose further challenges in understanding and monitoring their functioning. Thus, with a view of providing additional guardrails for transparency on these models, the Act mandates the providers of GPAI models to observe separate obligations. These obligations can be summarized as:

a. Creating technical documentation for GPAI models, covering their training, testing, and evaluation processes

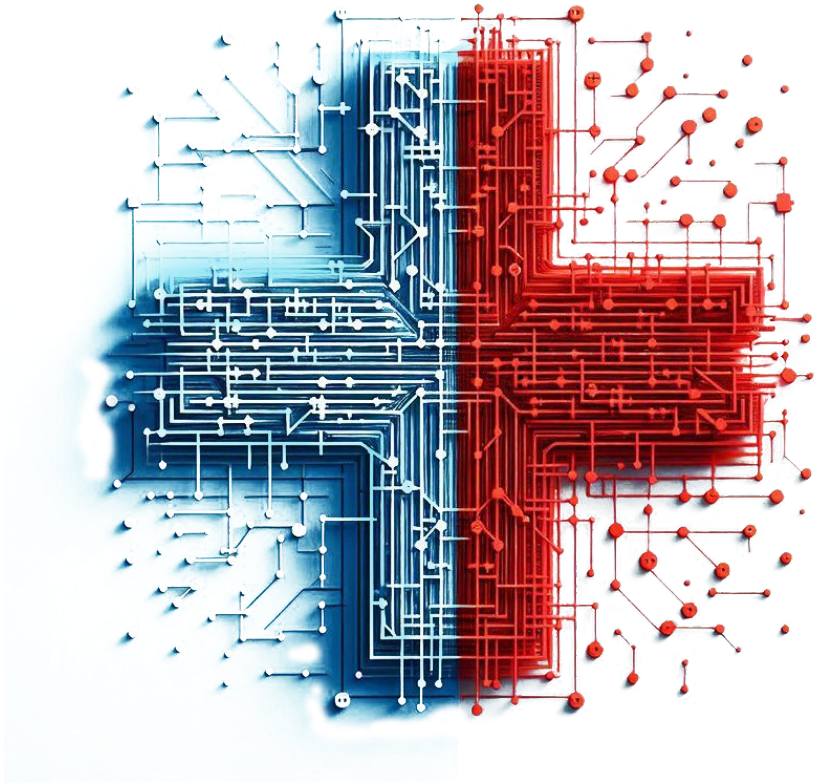
b. Supplying information and documentation to AI system providers who seek to use the GPAI model in their products, helping them understand the model's capabilities and limitations to meet their legal obligations

c. Providing a detailed summary of the training content and data to enhance transparency

Enforced August 2026

Why Build Our Own Models?

- **R&D Autonomy:** Development focus on important sovereign dimensions – data compliance, multilinguality, etc.
- **Users have full control over deployment:** Deploy on-premise. Keep sensitive data internal. No dependency on foreign tech infrastructure. No vendor lock-in.
- **Public institution advantage:** No influence on roadmaps by technology companies. Focus on open research. Benefits of large developer community.
- **Benefit from open development ecosystem**: Open models approach closed model performance. Inference cost dropping steadily.



APERTVS

EPFL ETH zürich cscs

Apertus: A transparent and responsibly-trained multilingual LLM

- **Open & transparent:** Released code. Reproducible data. Permissive license.
- **Compliance:** Trained only on public data, respecting AI opt-outs through robots.txt. Trained to prevent memorisation of copyrighted content
- **Multilingual from scratch:** Trained on data from over 1000 languages
- **Strong performance:** Most capable fully-open models at respective scales
- **Sovereignty:** Open platform for research and development of responsible AI

Released on the 2nd of September

- **Two models** at **8B** and **70B** scale
  - Released through Hugging Face with an **Open Source** license.
  - Trained with **15T tokens of text**
  - Trained using up to **4096 GPUs on Alps**

- Source code (training code, data pipelines, evaluation framework, etc)

- Extensive 100+ page technical report

Released on the 2nd of Se

- **Two models** at **8B** and **70B** scale
  - Released through Hugging Fa

Computer Science > Computation and Language

[Submitted on 17 Sep 2025]

Apertus: Democratizing Open and Compliant LLMs fo

Alejandro Hernández-Cano, Alexander Hägele, Allen Hao Huang, Angelika Romano
Garbaya, Eduard Frank Ďurech, Ido Hakimi, Juan García Giraldo, Mete Ismayilzada,
Yixuan Xu, Michael Aerni, Badr AlKhamissi, Ines Altemir Marinas, Mohammad Hosse
Boros, Nicholas Browning, Fabian Bösch, Maximilian Böther, Niklas Canova, Camille
Devos, Lukas Drescher, Daniil Dzenhaliou, Maud Ehrmann, Dongyang Fan, Simin Fa
Hoyle, Jiaming Jiang, Mark Klein, Andrei Kucharavy, Anastasiia Kucherenko, Freder
Kyle Matoba, Simon Matrenok, Henrique Mendoncça, Fawzi Roberto Mohamed, Syr
Gennaro Oliva, Matteo Pagliardini, Elia Palme, Andrei Panferov, Léo Paoletti, Marco
Nathan Ranchin, Javi Rando, Mathieu Sauser, Jakhongir Saydaliev, Muhammad Ali
Andrei Semenov, Kumar Shridhar, Raghav Singhal, Anna Sotnikova, Alexander Ster
Yao, Hao Zhao Alexander Ilic, Ana Klimovic, Andreas Krause, Caglar Gulcehre, Davi
Veraldi, Martin Rajman, Thomas Schulthess, Torsten Hoefler, Antoine Bosselut, Mart

APERTVS

Democratizing Open and Compliant LLMs for Global Language Environments

Apertus v1 Technical Report

Project Apertus*

Core Team: Alejandro Hernández-Cano[1], Alexander Hägele[1], Allen Hao Huang[1], Angelika Romanou[1], Antoni-Joan Solergibert[1,2], Barna Pasztor[2], Bettina Messmer[1], Dhia Garbaya[1], Eduard Frank Ďurech[1,2], Ido Hakimi[2], Juan García Giraldo[1], Mete Ismayilzada[1], Negar Foroutan[1], Skander Moalla[1], Tiancheng Chen[2], Vinko Sabolčec[1], Yixuan Xu[1,2]

Contributors: Michael Aerni[2], Badr AlKhamissi[1], Ines Altemir Marinas[1], Mohammad Hossein Amani[1], Matin Ansaripour[1], Ilia Badanin[1,2], Harold Benoit[1], Emanuela Boros[1], Nicholas Browning[3], Fabian Bösch[3], Maximilian Böther[2], Niklas Canova[2], Camille Challier[1], Clement Charmillot[1], Jonathan Coles[3], Jan Deriu[7], Arnout Devos[2], Lukas Drescher[3], Daniil Dzenhaliou[1], Maud Ehrmann[1], Dongyang Fan[1], Simin Fan[1], Silin Gao[1], Miguel Gila[3], María Grandury[1], Diba Hashemi[1], Alexander Hoyle[2], Jiaming Jiang[1], Mark Klein[3], Andrei Kucharavy[4], Anastasiia Kucherenko[4], Frederike Lübeck[2], Roman Machacek[9], Theofilos Manitaras[3], Andreas Marfurt[5], Kyle Matoba[1], Simon Matrenok[1], Henrique Mendonça[3], Fawzi Roberto Mohamed[3], Syrielle Montariol[1], Luca Mouchel[1], Sven Najem-Meyer[1], Jingwei Ni[2], Gennaro Oliva[3], Matteo Pagliardini[1], Elia Palme[3], Andrei Panferov[6], Léo Paoletti[1], Marco Passerini[3], Ivan Pavlov[1], Auguste Poiroux[1], Kaustubh Ponkshe[1], Nathan Ranchin[1], Javi Rando[2], Mathieu Sauser[1], Jakhongir Saydaliev[1], Muhammad Ali Sayfiddinov[2], Marian Schneider[2], Stefano Schuppli[3], Marco Scialanga[1], Andrei Semenov[1], Kumar Shridhar[2], Raghav Singhal[1], Anna Sotnikova[1], Alexander Sternfeld[4], Ayush Kumar Tarun[1], Paul Teiletche[1], Jannis Vamvas[8], Xiaozhe Yao[2], Hao Zhao[1]

Advisors: Alexander Ilic[2], Ana Klimovic[2], Andreas Krause[2], Caglar Gulcehre[1], David Rosenthal[10], Elliott Ash[2], Florian Tramèr[2], Joost VandeVondele[3], Livio Veraldi[10], Martin Rajman[1], Thomas Schulthess[3], Torsten Hoefler[2]

Leads: Antoine Bosselut[1], Martin Jaggi[1], Imanol Schlag[2]

Responsible Data Practices & Legal Viewpoint

We worked with David Rosenthal, one of the leading Swiss experts in the field of Data and Technology law.

- only use public data
- **prevent memorisation**
- **respect AI opt-out** (based on robots.txt removes ~10% English and ~5% non-English data)
  Also in hindsight, not only at crawl time
- **remove PII**
- **filter toxic&harmful** data

David Rosenthal / Livio Veraldi

Training AI language models with third-party content and data from a legal perspective

https://arxiv.org/abs/2504.06219

Responsible Data Practices & Legal Viewpoint

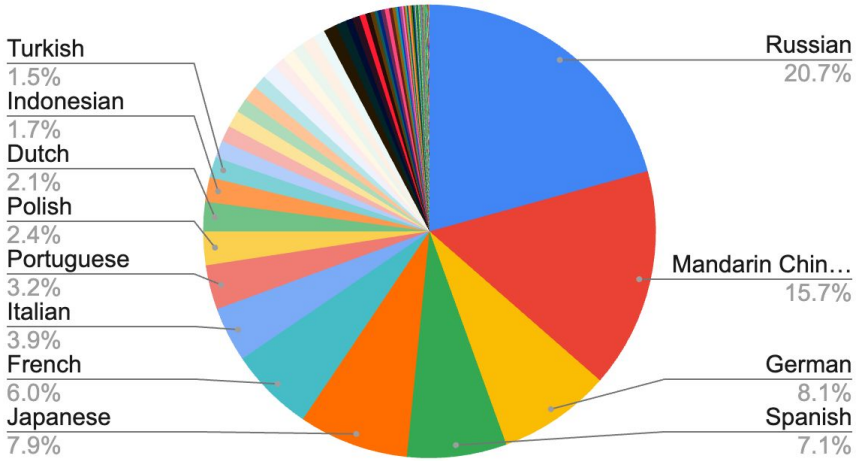
EU AI Act will be enforced in August 2026…

Pretraining Data

15T tokens of high quality international data mix

- Fineweb 1 HQ (english) and FWedu
- Fineweb 2 HQ (non-english)
- Datacomp-LM
- Stack v1.2
- FineMath and MegaMath
- Memorisation and Poisoning Data

Fineweb-2 language distribution



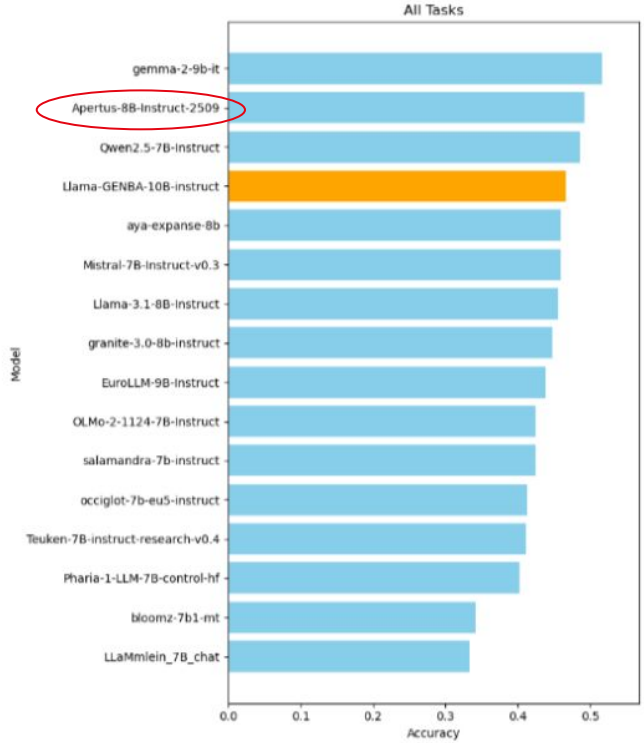
Reflecting natural distribution of **all languages on the web globally**
~60% English web data
~40% Non-English web data

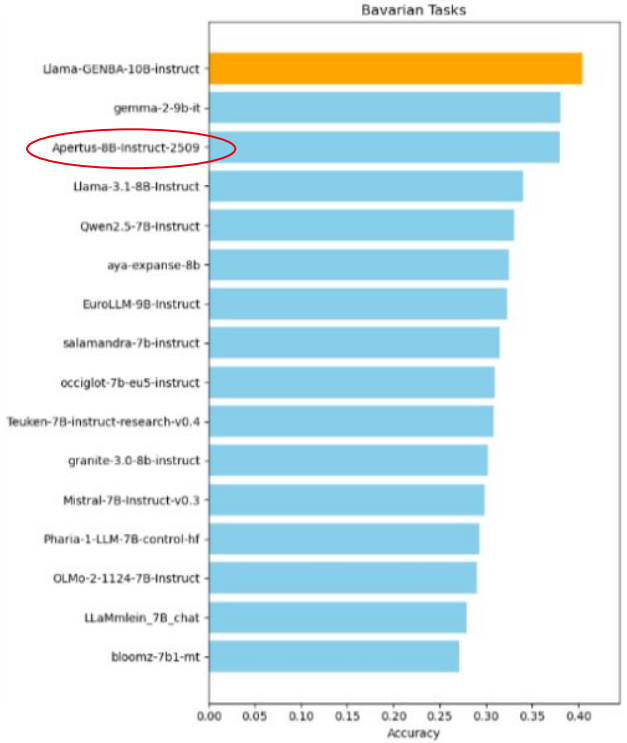
~Code & math data growing from 3%-10% throughout pretraining

Unbiased, for facilitating research: no synthetic data, no downstream task-specific data until the very last phase (no data of SFT, reasoning, or specific domains were used in pretraining)

Performance and Impact

Third-party evaluation of Apertus 8B by the Leibniz Supercomputing Center and Cerebras.





Performance and Impact

- **Strong performance within the class of fully-open models**, particularly on multilingual and multicultural benchmarks. Good tradeoff between cost and performance.

- **Apertus 8B beats** Llama 3.1 (Meta), Qwen 2.5 (Alibaba), Mistral v0.3, GPT-OSS-20B (OpenAI), and all models from public institutions.

- **Not a "deepseek moment":** Our largest model is large (70B parameters) but still more than 10 times smaller compared to today's *public weight* frontier models (700B+ parameters).

- Pioneering transparent and responsible AI: **A public foundation model** to research and develop the defining technology of our time. **Not a production-ready AI system**!

Distribution

- **Available on Hugging Face:** For anyone everywhere through an Open-Source license.

- **Available from cloud providers:** The Swisscom Swiss AI Platform, Amazon AWS, Microsoft Azure, Infomaniak, Phoenix Technologies, etc.

- **A free chat-based front-end and endpoint**, for the duration of the Swiss {ai} Weeks for anyone to explore open-source AI capabilities provided by Swisscom and the Public AI Company through publicai.co and platform.publicai.co (not operated by ETHZ/EPFL/CSCS)

Distribution

Over 400,000 total downloads from huggingface in the first month.

First examples, such as oss.zuericitygpt.ch

Driving Value from LLMs

Finetune an open model for specific tasks to *outperform* the largest proprietary models on complex and domain-specific real-world tasks.

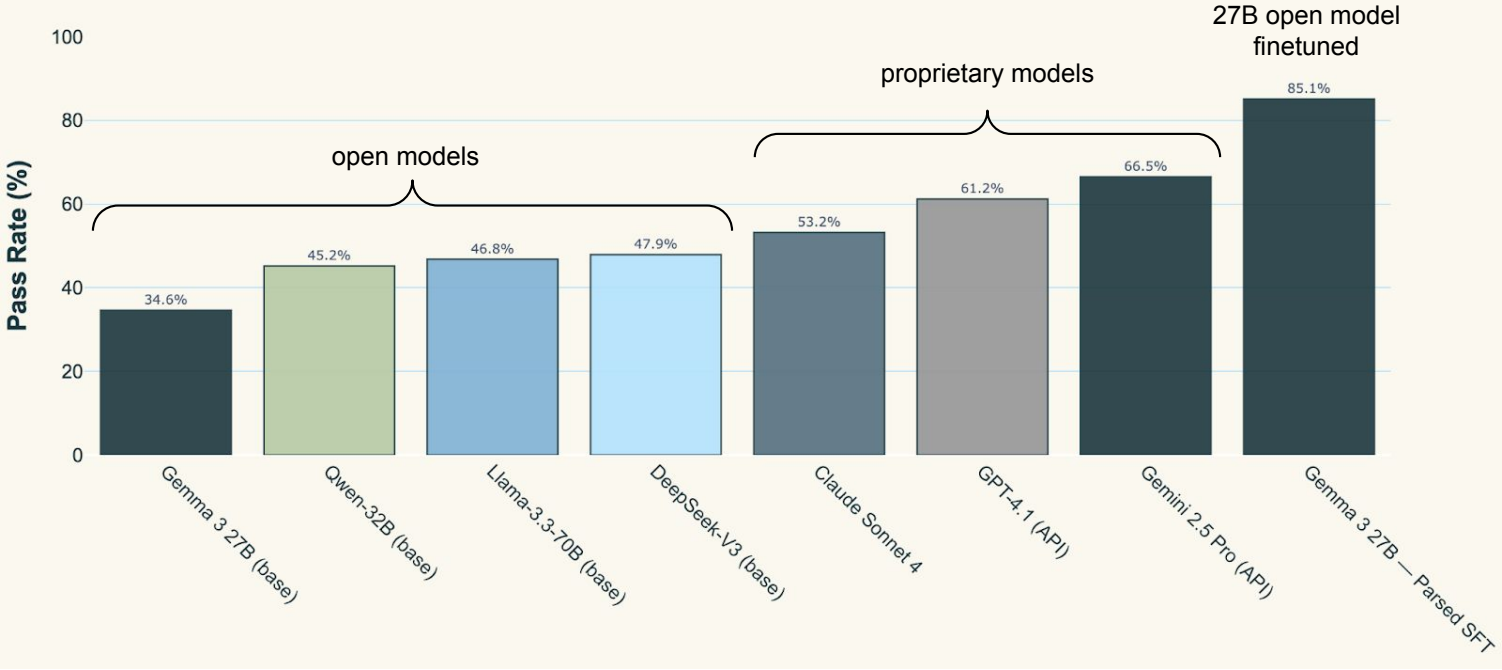
Example: transcribe clinician-patient interactions and write clinical notes in a particular style

60% better accuracy, 10-100x lower cost than claude sonnet

Need:

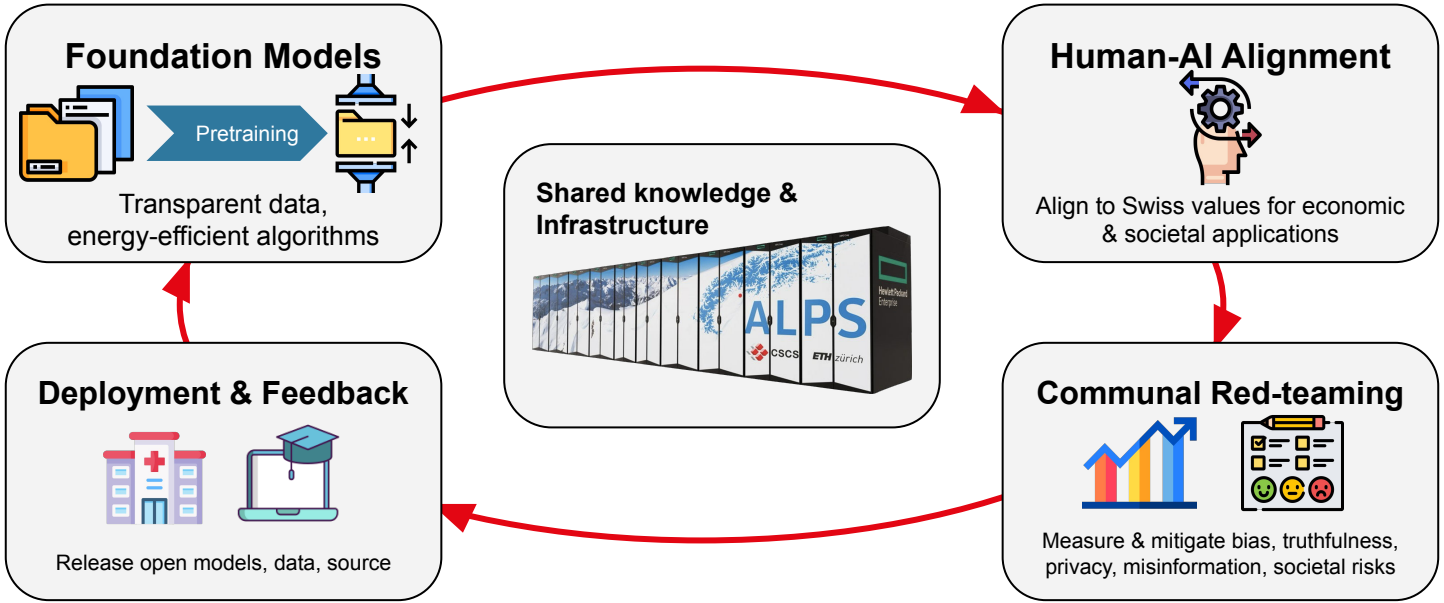
- rigorous evaluation
- task-specific high-quality data
- iterative optimisation of a strong foundation

Driving Value from LLMs



[by Parsed, 2025]

Post-Release Cycle



Where Are We Headed?

“Powerful AI” will increasingly accelerate ability of the people using it.

It’s not hard to imagine that system which …

- knows as much as the expert literature in any field
- has all interfaces available (text, audio, video)
- can use software tools (web search, internal database, MCP)
- can work autonomously on a query (instead of just responding in a few seconds)
- has no physical body
- but is very very affordable

… will transform work as we know it.

We must learn how these systems work and how to develop & deploy them responsibly.

Teaching: Large-Scale AI Engineering

A globally **novel** course which focuses on the **engineering principles and hands-on practices** required to **develop and optimize large-scale AI systems**. Participants gain unique **hands-on** experience on Alps.

Setup:
- 14 weeks
- 2h / week
- 3 ECTS
- Max 140 MSc/PhD

Schedule:
- ⅓ lectures
- ⅓ assignments
- ⅓ final team project

Hands-on on CSCS Clusters:
- GH200 (or A100) GPUs
- 8 nodes/32 GPUs per team
- Llama 3 (8B) [Hubet et al., 2024]

Stack:
- PyTorch
- Slurm
- Docker

Main topics covered:
- HPC with GPUs for AI
- AI model optimisation techniques
- Performance monitoring & profiling
- Running distributed cluster jobs

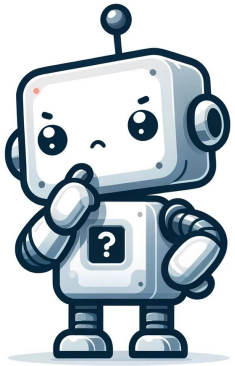
Course team: ETHZ experts in AI & large-scale computing


I.S.


A.D.


A.S.

Questions?



Kontakt
Imanol Schlag, PhD
AI Research Scientist @ ETH AI Center
Apertus Co-Lead
ETH Zürich, ETH AI Center
Andreasstrasse 5
8092 Zürich
ischlag@ethz.ch

ETH AI CENTER

linkedin