



ETH Zurich + University of St.Gallen

Towards AI-based Products & Services

Alexander Ilic, Menna El-Assady



How 2025 started...

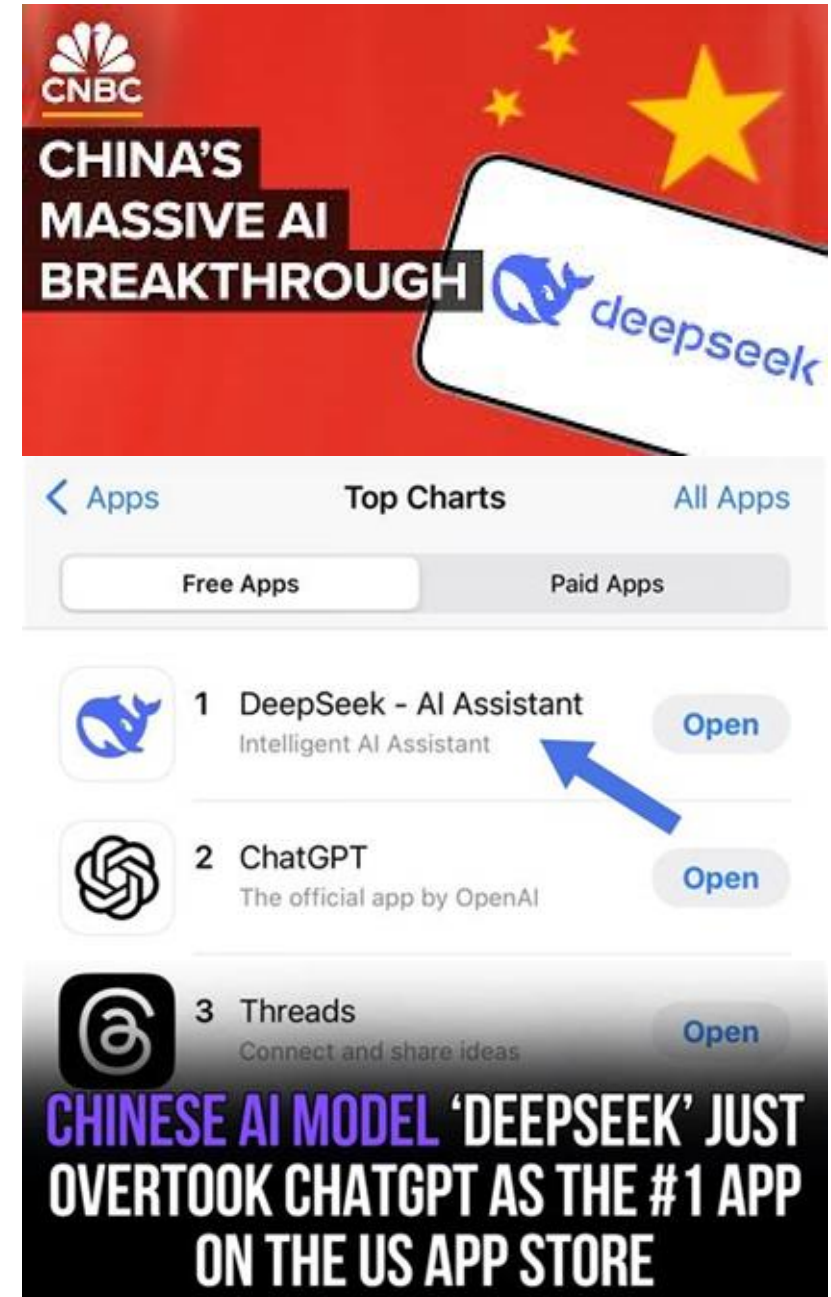
January 21, 2025



January 22, 2025



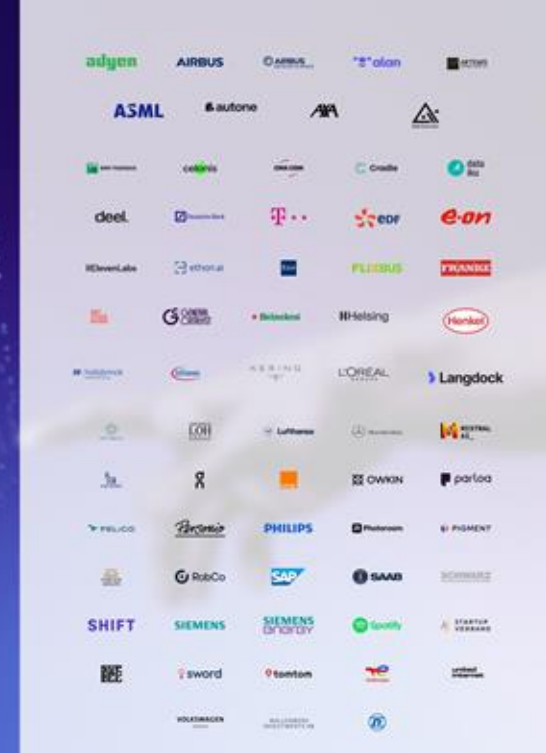
US announces \$500bn investment in AI



Europe's response



109bn investment by France



One year later – so what?

The AI sovereignty debate has become real



How tenaciously Palantir courted Switzerland

The controversial tech company Palantir provides surveillance technology to militaries and intelligence services around the world. In Switzerland, however, its pitches have been rejected by both government authorities and the army. Internal documents have now revealed why.

By [Adrienne Fichtes](#), Marguerite Meyer, Lorenz Naegeli, Balz Oertli, Jennifer Steiner (text) and Klawe Rzeczy (illustration), December 8, 2025; translation by [Priscilla Imboden](#), February 18, 2026

Teilen 2 Beiträge

15 Minuten

English | Deutsch

THE WALL STREET JOURNAL.

English Edition | Print Edition | Video | Audio | Latest Headlines | Puzzles | More

Business | U.S. | **Politics** | Economy | Tech | Markets & Finance | Opinion | Free Expression | Arts | Lifestyle | Real Estate | Personal Finance

TECHNOLOGY • ARTIFICIAL INTELLIGENCE

Anthropic Halts Access to Top AI Models After U.S. Ban on Foreign Use

All Fable 5 and Mythos 5 users lose access after Trump administration declares the models security risks

By [Amrith Ramkumar](#) and [Robert McMillan](#)

Updated June 13, 2026 12:34 am ET

195 Listen (2 min)



Anthropic CEO Dario Amodei DAVID PAUL MORRIS/BLOOMBERG NEWS

Who is Building AI?

Closed

OpenAI US
Anthropic US
Google US
xAI US
Mistral FR

Share outputs.

Open-Weight

Meta (Llama) US
DeepSeek CN
Alibaba (Qwen) CN
Moonshot (Kimi) CN
Google (Gemma) US
Z.AI (GLM) CN
Mistral FR

Share model weights.

Fully Open

SNAI (Apertus) CHEU 
AI2 (Olmo) US 
AI Singapore (SEA-LION) SG 

OpenEuroLLM EU 
EuroLLM EU 
(many smaller ones)

Share everything.

Large-scale AI is here to stay

(and what it means for your business)

What do these three fundamental papers have in common?

ImageNet Classification with Deep Convolutional Neural Networks

Alex Krizhevsky
University of Toronto
kriz@cs.utoronto.ca

Ilya Sutskever
University of Toronto
ilya@cs.utoronto.ca

Geoffrey E. Hinton
University of Toronto
hinton@cs.utoronto.ca

Abstract

We trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes. On the test data, we achieved top-1 and top-5 error rates of 37.5% and 17.0% which is considerably better than the previous state-of-the-art. The neural network, which has 60 million parameters and 650,000 neurons, consists of five convolutional layers, some of which are followed by max-over-time/pooling layers, and three fully-connected layers with a final 1000-way softmax. To reduce overfitting in the training set, we used data augmentation techniques, including random cropping of the image and color jittering. To speed up training, we used non-saturating neurons and a very efficient implementation of the convolution operation. To reduce overfitting in the test set, we employed a recently-developed regularization method called "dropout" that proved to be very effective. We also entered a variant of the network in the ILSVRC-2012 competition and achieved a winning top-5 test error rate compared to 26.2% achieved by the second-best entry.

1 Introduction

Current approaches to object recognition make essential use of machine learning methods. To improve their performance, we can collect larger datasets, learn more powerful models, and use better techniques for preventing overfitting. Until recently, datasets of labeled images were relatively small — on the order of tens of thousands of images (e.g., NORBB [16], Caltech-101/256 [8, 9], and CIFAR-10/100 [12]). Simple recognition tasks can be solved quite well with datasets of this size, especially if they are augmented with label-preserving transformations. For example, the current-best error rate on the MNIST digit-recognition task (<0.3%) approaches human performance [4]. But objects in realistic settings exhibit considerable variability, so to learn to recognize them it is necessary to use much larger training sets. And indeed, the shortcomings of small image datasets have been widely recognized (e.g., Pinto et al. [21]), but it has only recently become possible to collect labeled datasets with millions of images. The new larger datasets include LabelMe [23], which consists of hundreds of thousands of fully-segmented images, and ImageNet [6], which consists of over 15 million labeled high-resolution images in over 22,000 categories.

To learn about thousands of objects from millions of images, we need a model with a large learning capacity. However, the immense complexity of the object recognition task means that this problem cannot be specified even by a dataset as large as ImageNet, so our model should also have lots of prior knowledge to compensate for all the data we don't have. Convolutional neural networks (CNNs) constitute one such class of models [16, 11, 13, 18, 15, 22, 26]. Their capacity can be controlled by varying their depth and breadth, and they also make strong and mostly correct assumptions about the nature of images (namely, stationarity of statistics and locality of pixel dependencies). Thus, compared to standard feedforward neural networks with similarly-sized layers, CNNs have much fewer connections and parameters and so they are easier to train, while their theoretically-best performance is likely to be only slightly worse.

2012

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jacob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez*[†]
University of Toronto
aidan@cs.toronto.edu

Lukasz Kaiser*
Google Brain
lukasz.kaiser@google.com

Illia Polosukhin*[‡]
illia.polosukhin@gmail.com

Ideas amplified by massive scale at that time

ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.0 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature.

1 Introduction

Recurrent neural networks, long short-term memory [12] and gated recurrent [7] neural networks in particular, have been firmly established as state of the art approaches in sequence modeling and transduction problems such as language modeling and machine translation [29, 2, 5]. Numerous efforts have since continued to push the boundaries of recurrent language models and encoder-decoder architectures [31, 21, 13].

*Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention and started the effort to evaluate this idea. Ashish, with Illia, designed and implemented the first Transformer models and has been crucially involved in every aspect of this work. Noam proposed scaled dot-product attention, multi-head attention and the parameter-free position representation and became the other person involved in nearly every detail. Niki designed, implemented, tuned and evaluated countless model variants in our original codebase and tensor2tensor. Llion also experimented with novel model variants, was responsible for our initial codebase, and efficient inference and visualizations. Lukasz and Aidan spent countless long days designing various parts of and implementing tensor2tensor, replacing our earlier codebase, greatly improving results and massively accelerating our research.

[†]Work performed while at Google Brain.

[‡]Work performed while at Google Research.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

2017

Language Models are Few-Shot Learners

Tom B. Brown* Benjamin Mann* Nick Ryder* Melanie Sabblah*
Jared Kaplan[†] Prafulla Dhariwal Arvind Neelakantan Pranav Shyam Girish Sastry
Amanda Askell Sandhini Agarwal Ariel Herbert-Voss Gretchen Krueger Tom Henighan
Rewon Child Aditya Ramesh Daniel M. Ziegler Jeffrey Wu Clemens Winter
Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray
Benjamin Chess Jack Clark Christopher Berner

Alec Radford Ilya Sutskever Dario Amodei

OpenAI

Abstract

Large language models trained on substantial gains on many NLP tasks and benchmarks by pre-training followed by fine-tuning on a specific task. While typically task-agnostic, they still require task-specific fine-tuning datasets of thousands or tens of thousands of examples. By contrast, humans can generally perform a new language task from only a few examples or from simple instructions — something which current NLP systems still largely struggle to do. Here we show that scaling up language models greatly improves task-agnostic, few-shot performance, sometimes even reaching competitiveness with prior state-of-the-art fine-tuning approaches. Specifically, we train GPT-3, an autoregressive language model with 175 billion parameters, 10x more than any previous non-sparse language model, and test its performance in the few-shot setting. For all tasks, GPT-3 is applied without any gradient updates or fine-tuning, with tasks and few-shot demonstrations specified purely via text interaction with the model. GPT-3 achieves strong performance on many NLP datasets, including translation, question-answering, and cloze tasks, as well as several tasks that require on-the-fly reasoning or domain adaptation, such as unscrambling words, using a novel word in a sentence, or performing 3-digit arithmetic. At the same time, we also identify some datasets where GPT-3's few-shot learning still struggles, as well as some datasets where GPT-3 faces methodological issues related to training on large web corpora. Finally, we find that GPT-3 can generate samples of news articles which human evaluators have difficulty distinguishing from articles written by humans. We discuss broader societal impacts of this finding and of GPT-3 in general.

*Equal contribution
[†]Johns Hopkins University, OpenAI

Author contributions listed at end of paper.

2020

Scaling laws for LLMs: Size matters

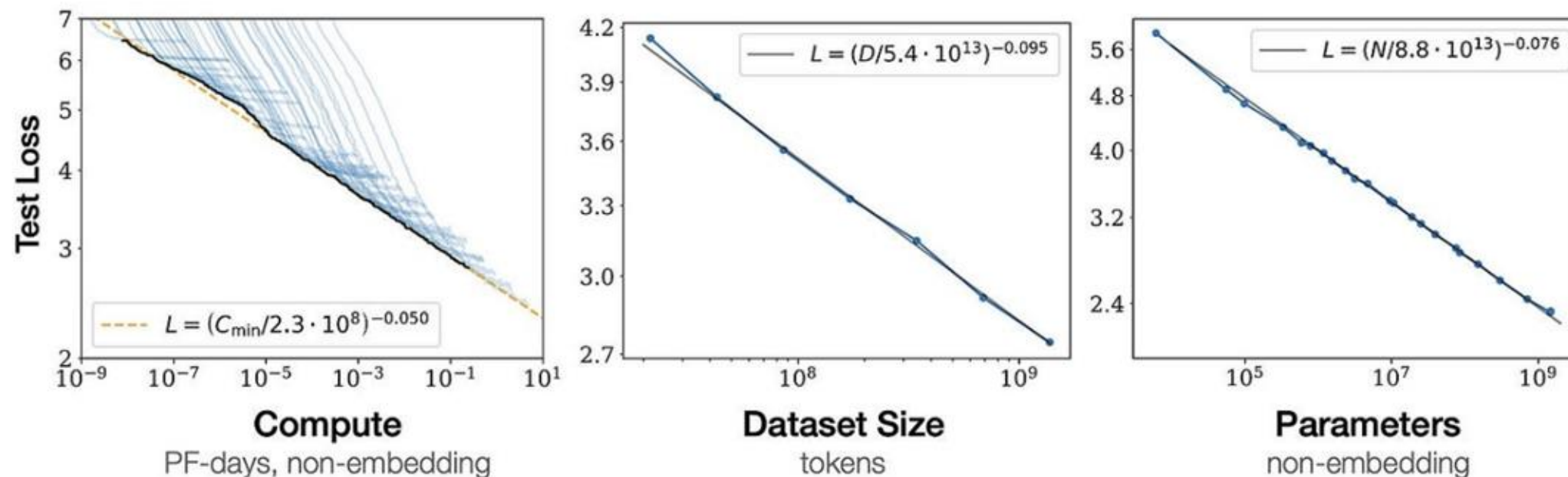
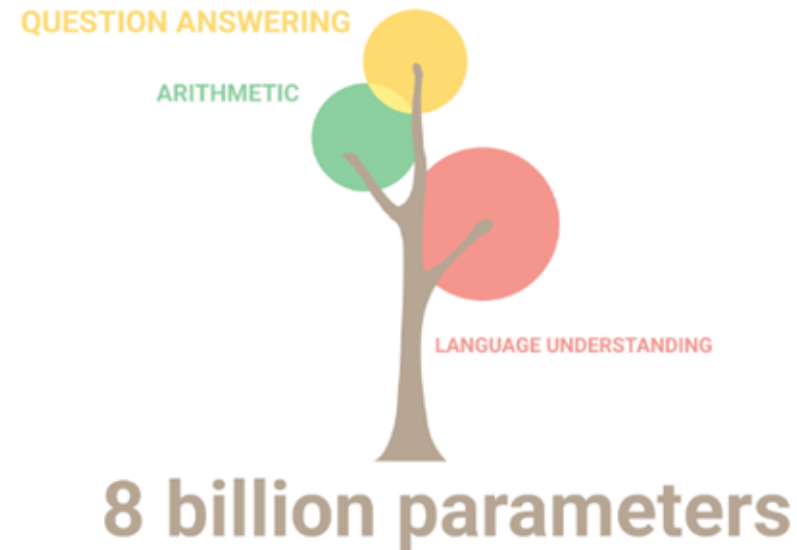


Figure 1 Language modeling performance improves smoothly as we increase the model size, dataset size, and amount of compute² used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

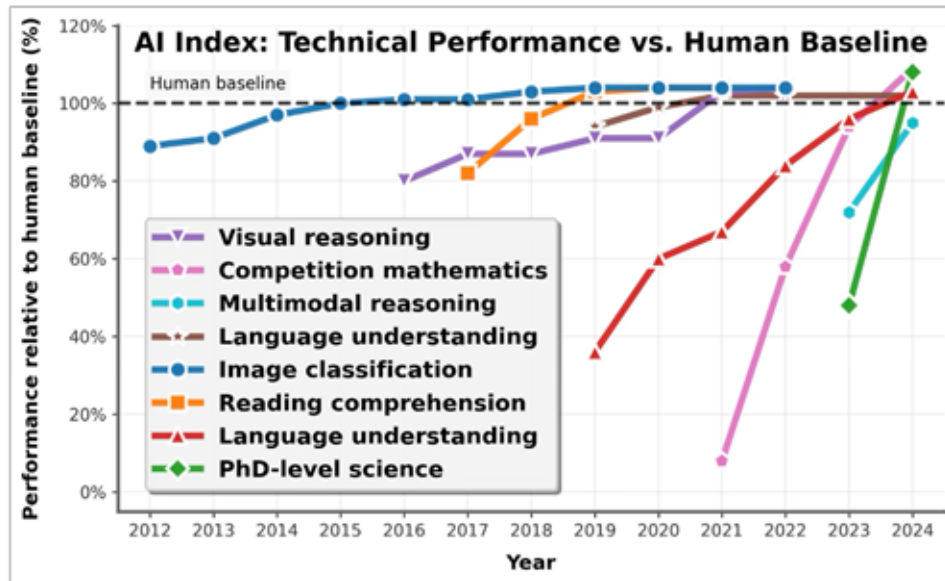
Example of capabilities with scale: Google PaLM



LLMs exceeding human capabilities

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

Now you understand why this is happening...

Zuckerberg's Meta Is Spending Billions to Buy 350,000 Nvidia H100 GPUs

In total, Meta will have the compute power equivalent to 600,000 Nvidia H100 GPUs to help it develop next-generation AI, says CEO Mark Zuckerberg.



By [Michael Kan](#) Jan 18, 2024

[f](#) [X](#) [in](#) [p](#)



Generative AI is here

Generieren

Was würden Sie

Präsentation

Import from URL

URL:

Cancel Import

Upload a file

Import from Drive

Import from URL

emba X: Your Alpine Journey to Leadership

Experience Europe's most innovative Executive MBA, jointly delivered by ETH Zurich and the University of St.Gallen. Designed to develop socially responsible leaders who integrate leadership and technology challenges to create positive impact.



Veo 3

embaX student walking on campus

Simple Custom v4

Describe your song

Write a rap song about emba X: A joint programme in general management focusing on technology and leadership by ETH Zürich and University of St Gallen in Switzerland

+ Audio + Lyrics Instrumental

SUNO



[Verse 3]
Blueprints drawn, but we color outside,
Strategy sharp, like a Swiss alpine slide.
Diving deep, tech trends we explore,
Disrupting the game, it's what we're here for.

[Bridge]
Global network, a squad so elite,
Crossing borders, where the worlds compete.
Ideas ignite, like a spark in the night,
EMBA X, we're the guiding light.

[Chorus]
EMBA X, where the leaders are born,

What You Will Study

Essential Courses

Our faculty has been tasked with rethinking these subject areas and 'teaching the future'. You'll not only relearn these topics but even unlearn some assumptions to make space for vital new knowledge.

1	2	3
Essentials 1 - Adaptive Cognitive Leadership - Organisational Leadership - Strategy and Innovation - Business Ethics	Essentials 2 - Geoeconomic Dynamics & Disruptions - Corporate Finance - Sustainability - Analytics (Business and Marketing)	Essentials 3 - Cognitive Technologies (AI & Machine Learning) - Energy Management, Technology & Economics - Product & Service Technologies - Technology Entrepreneurship

Programme Overview

At a Glance

- 18-month programme
- 55 days on campus plus online classes
- All teaching in English
- Cost: 110,000 CHF (with incentives available)

eXploration Weeks
Revisit essential concepts and prepare for the future

eXperiential Weeks
Five days of intense, faculty-driven immersion exposure

emba X Projects
Apply your learning in real-world impactful projects

Coaching & Mentoring
Innovative, customised Personal Development Plan

gamma

emba X: Europe's Innovative Executive MBA Journey

Sources

+ Add Discover

Select all sources

emba_X5_Brochure.pdf

Chat

emba X: Europe's Innovative Executive MBA Journey

1 source

The provided brochure outlines the emba X Executive MBA program, a collaborative effort between ETH Zurich and the University of St.Gallen, designed to cultivate socially responsible leaders who can integrate leadership and technology challenges. It details the program's innovative curriculum, which includes eXploration Weeks for essential concepts, eXperiential Weeks for practical application, and impactful projects like the Business Model Innovation Project and Company Impact Project. The program emphasizes holistic development, focusing on self-leadership, team leadership, and integrative leadership through workshops and personal

Start typing...

1 source

How does the emba X program uniquely integrate leadership and technology for

Studio

Share Settings

Audio Overview

Create an Audio Overview in more languages! [Learn more](#)

emba X: Europe's Innovati...

01:45 / 05:03 • English

Interactive mode (BETA)

Not

+ Add note

Study guide Briefing doc

FAQ Timeline

NotebookLM

Coding is disrupted (but not much else yet)

Home / Innovation / Artificial Intelligence

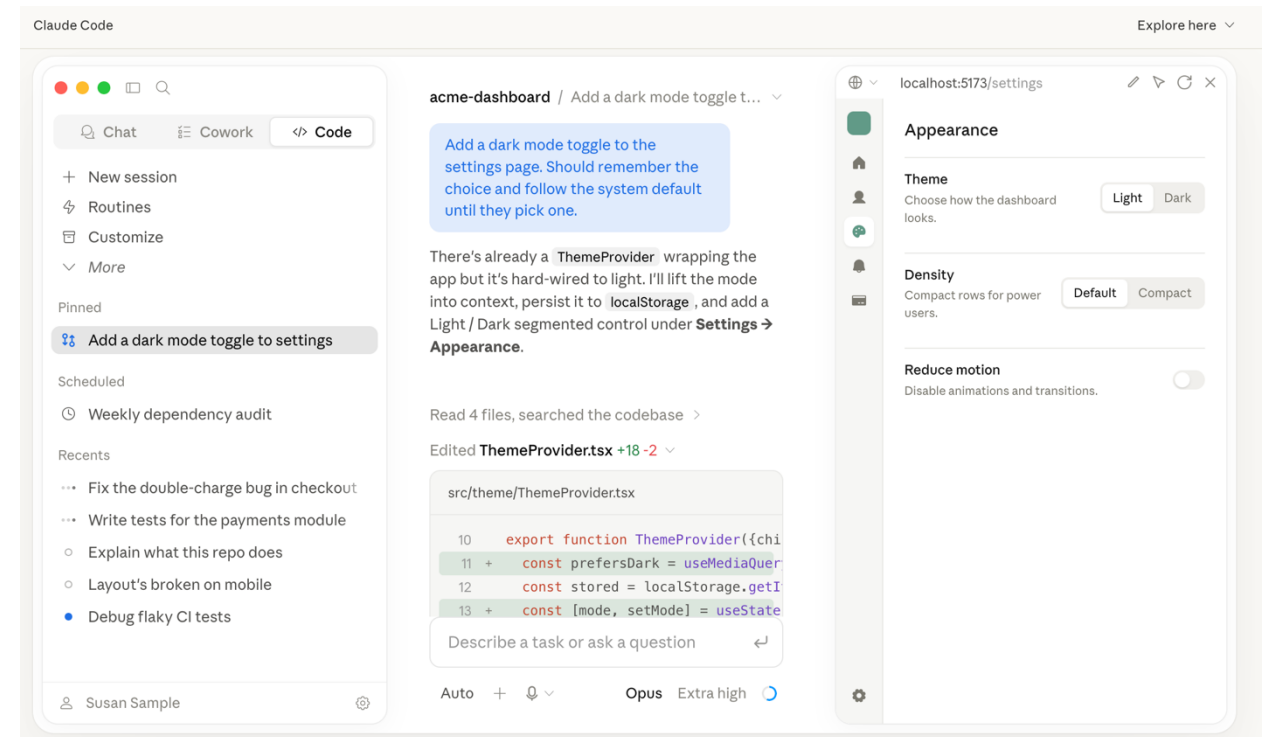
Linus Torvalds is 'a huge believer' in using AI to maintain code - just don't call it a revolution

Torvalds is sick of all the AI hype, but says AI is finally maturing to the point where it will be useful for Linux developers and maintainers.



Written by **Steven Vaughan-Nichols**, Senior Contributing Editor

Dec. 9, 2025 at 6:21 a.m. PT



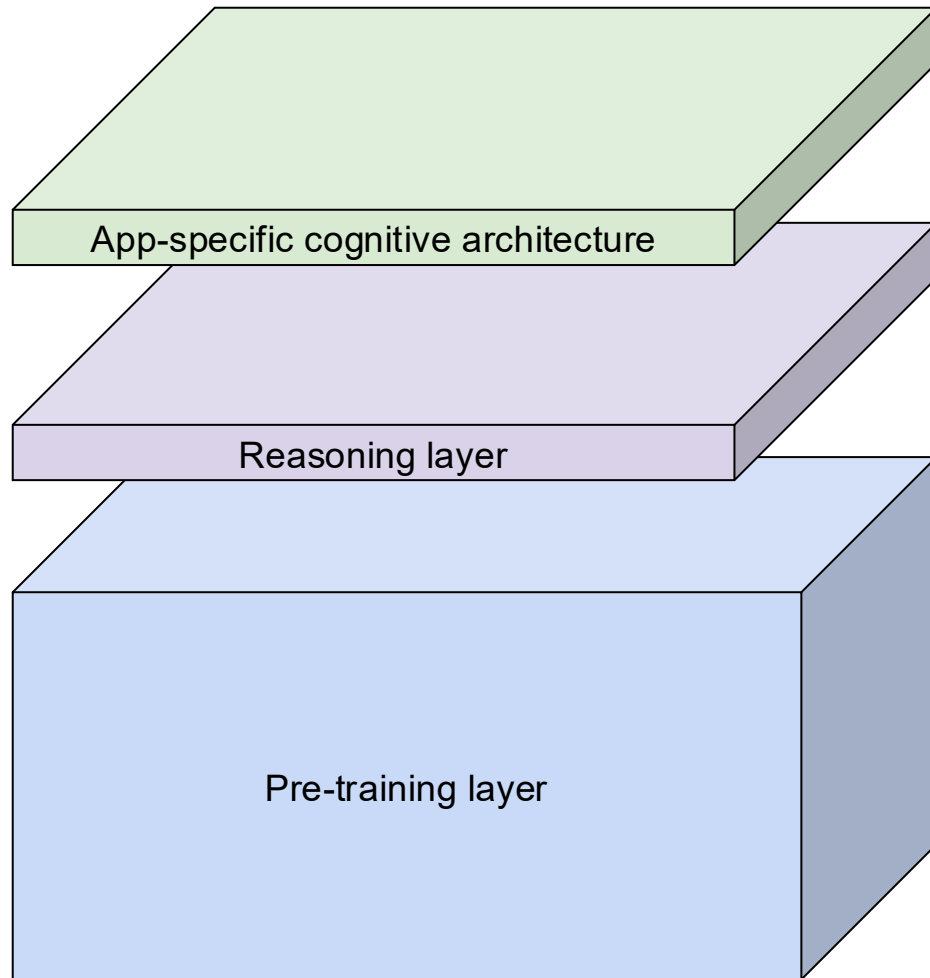
Two AI frontier labs will likely go public based on LLMs fuelled with public data

A pragmatic definition for AI Sovereignty

An organization must be able to embed **their knowledge** into a model **they control**

Why we built Apertus

What does it mean if the infrastructure layer becomes commoditized?



Application & fine-tuning layers

- Provides the performance
- High value capture
- Training costs <\$10-100k

Pre-training layer / large foundation models

- \$20-100m training costs
- Huge energy footprint
- Give basic capabilities

GPU access of leading universities



Source: <https://gpuserstudent.org/>

ETH & EPFL are currently the only universities in the world that can practically educate on foundation model development

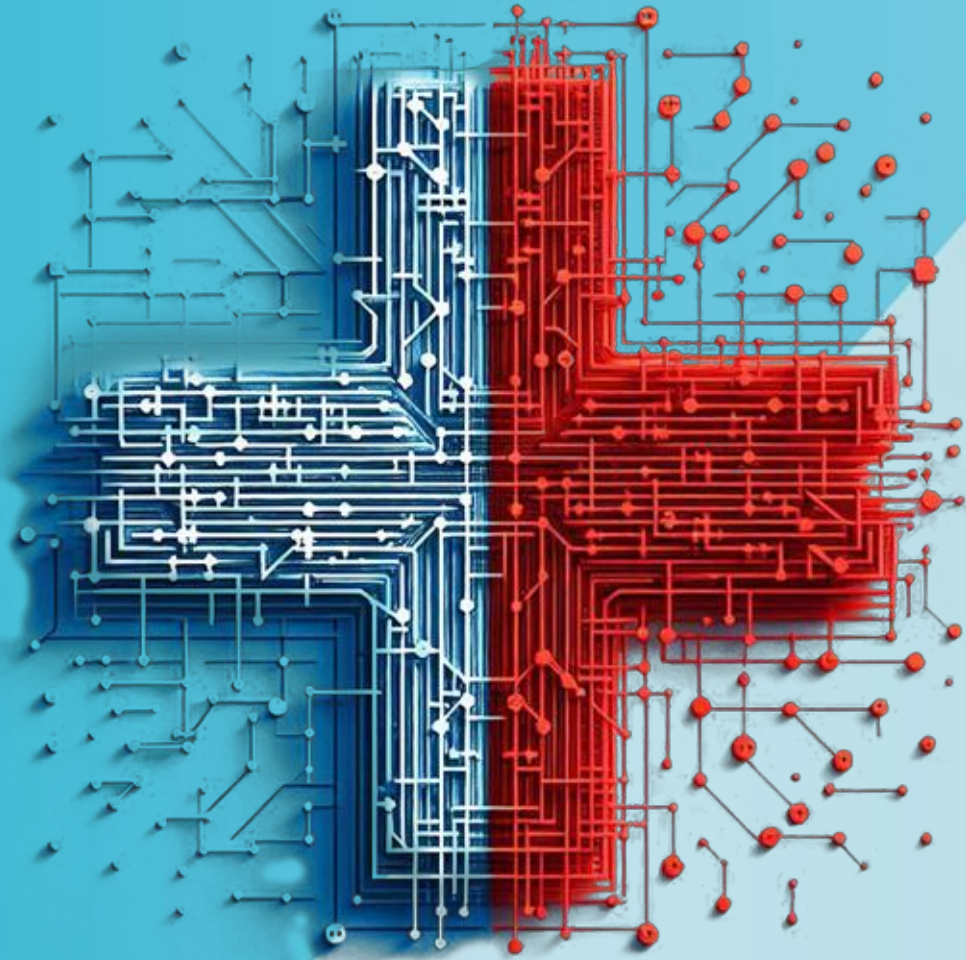
The Swiss National AI Institute (SNAI)

A partnership among ETH Zurich and EPFL to concentrate our **national perspective** on AI-based research, education, and innovation.

SNAI drives forward high-impact research activities involving a national network of academic and industry partners. Its **flagship research activity is the Swiss AI Initiative** with a focus on advancing and democratizing large-scale AI for the benefit of society.

powered by





A P E R T V S

EPFL

ETH zürich



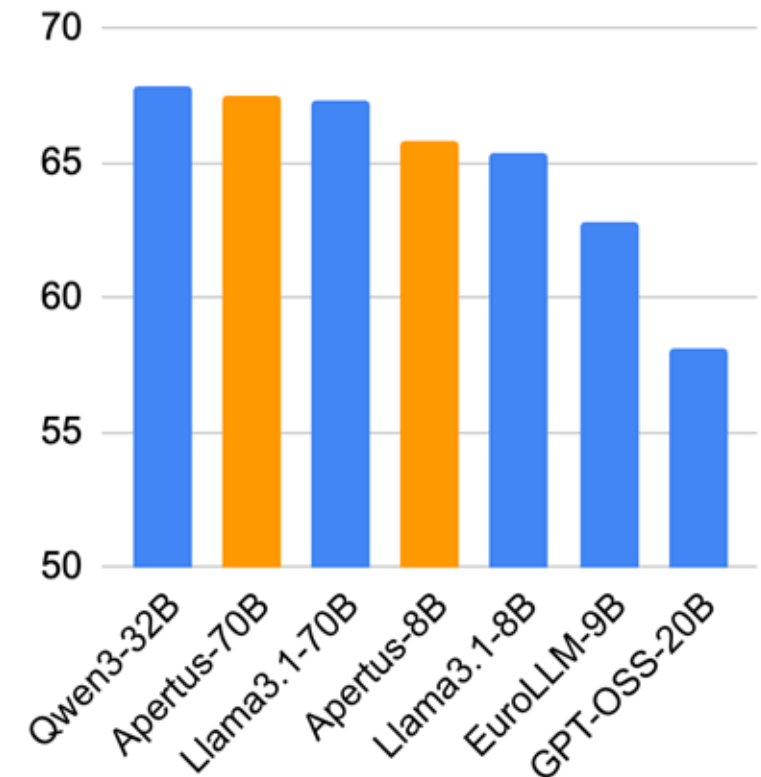
CSCS

**Democratizing the Foundation for
Research & Global Sovereign AI**

Q4/2025: World's first fully open frontier LLM

- **Two** models with **8B** & **70B** parameters trained on **15T tokens of text** using up to **4096 GPUs on Alps**.
- More than **2'500'000** downloads in first three months
- **Full transparency:** Open model to research and develop future AI systems
- **Responsible Data Practices:** Satisfies base model provisions of EU AI Act
- **Global representation:** Trained on data from over 1000 languages (40% non-English)

Strong benchmark performance within class of fully-open models



Apertus & EU AI Act



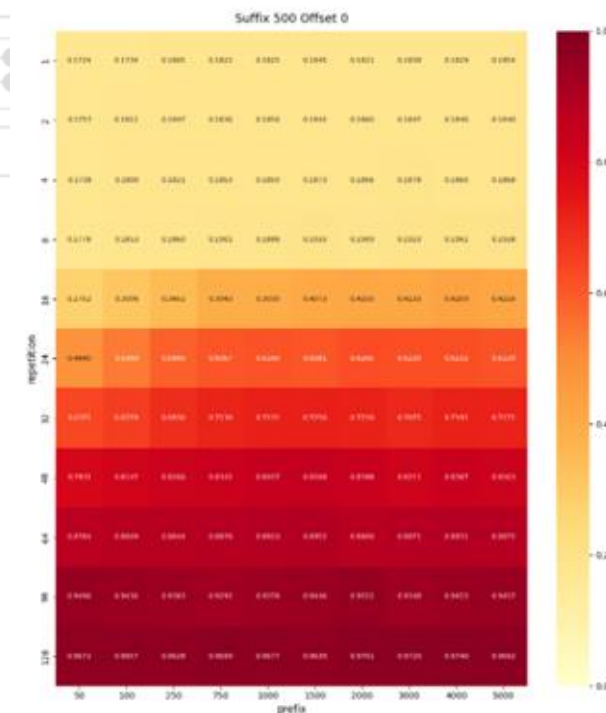
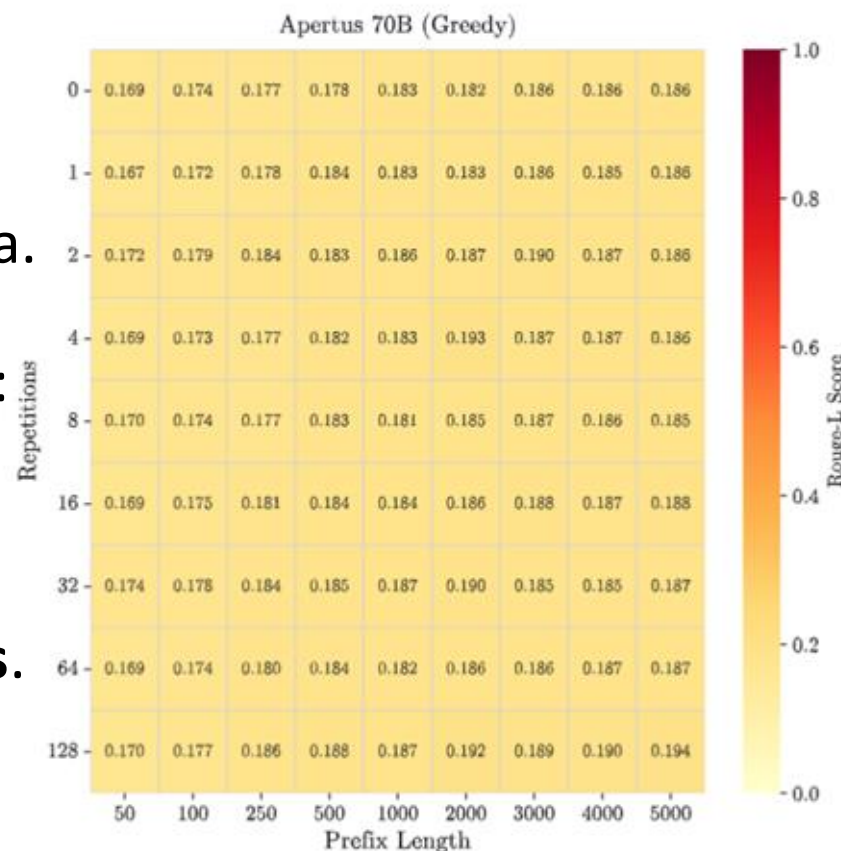
Responsible Data Practices

- **Public & versioned data only**, primarily sourced from CommonCrawl
- Respect **AI opt-out preferences** of authors (removes 8% English, 4% non-English data)
- **Prevent model memorisation** of data, specifically copyrighted content
- **Quality:** Remove PII, toxic & harmful content, and low quality data (e.g. spam, non-human text, etc.).

	Apertus 8B / 70B	Qwen, DeepSeek	Mistral	Meta Llama	OpenAI GPT-4,5 series
Model Weights	✓	✓	✓	✓	✗
Inference code	✓	✓	✓	✓	✗
Training code	✓	●	●	●	✗
Training data transparency & availability	✓	✗	✗	✗	✗
Open License	✓	✓	✓	●	✗
EU AI Act Art 53 compliance	✓	✗	✗	✗	✗

Effective Memorisation Prevention

- Successfully **prevented the memorisation** of training data.
- Scientific experiment at scale: unable to extract **realistic control data** even when repeated **128 times** with a prefix/prompt of **5000 tokens**.



8B experiment **without** prevention shows significant memorisation

Figure 8: **Apertus Maintains Robustness Against Verbatim Memorization.** The heatmaps show average Rouge-L scores for suffixes of 500 tokens for Apertus-8B and Apertus-70B. The y-axis represents exposure frequencies during training (1–128), with unexposed Gutenberg bucket 0 serving as a baseline. The prefix length varied from 50 to 5,000 tokens at a fixed offset of 0. The results demonstrate successful mitigation of verbatim memorization in Apertus model, as the Rouge-L scores for both model scales remain at baseline levels regardless of exposure frequency or prefix length.

Watch out for Apertus 1.5

- **Continued Pretraining:** Additional high-quality text data. Adding 4T tokens to previous training
- **Multimodality:** Enable audio and image input
- **Focus Domains:** Data upsampling for strategic areas such as health, education, and justice
- **Agentic Skills:** Improved instruction-following, tool-use, and advanced reasoning capabilities with RL at scale

Beyond Apertus

Multimodal World Models for Physical AI



We develop novel multimodal World Foundation Models (WFM) that integrate video, text, 3D, proprioception, and environment signals to create robust, controllable world models. These models will power applications from autonomous mobility to robotics and assistive AI.

Beyond Apertus

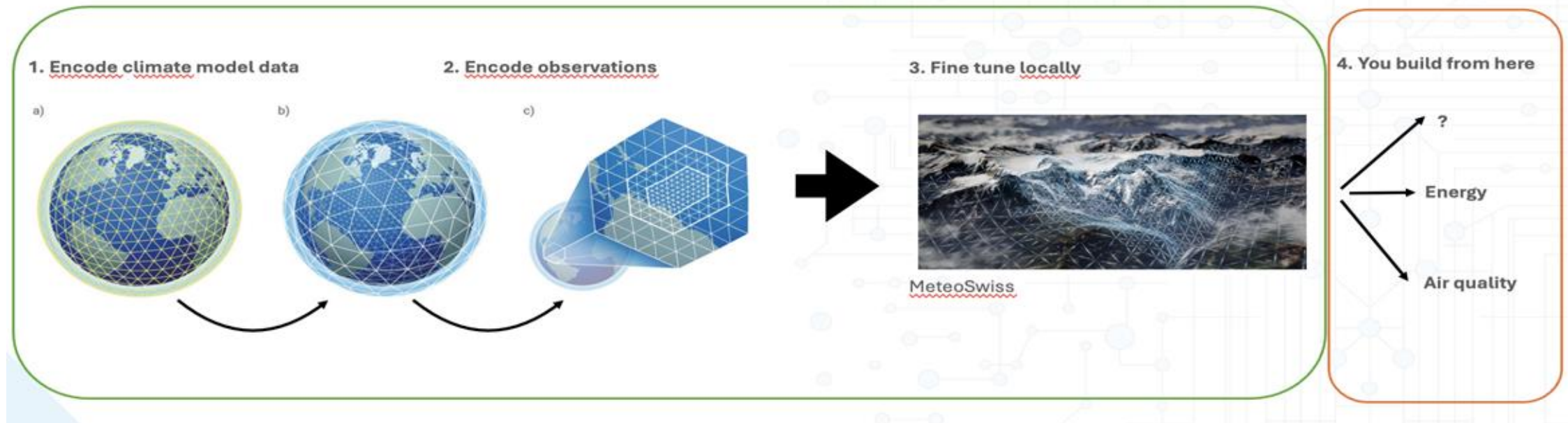
Foundation Models for Science



We develop a multimodal Catalysis foundation model that unifies data across chemistry domains to accelerate catalyst design and sustainable innovation, powered by large-scale compute and collaboration.

Beyond Apertus

Foundation Models for Weather and Climate



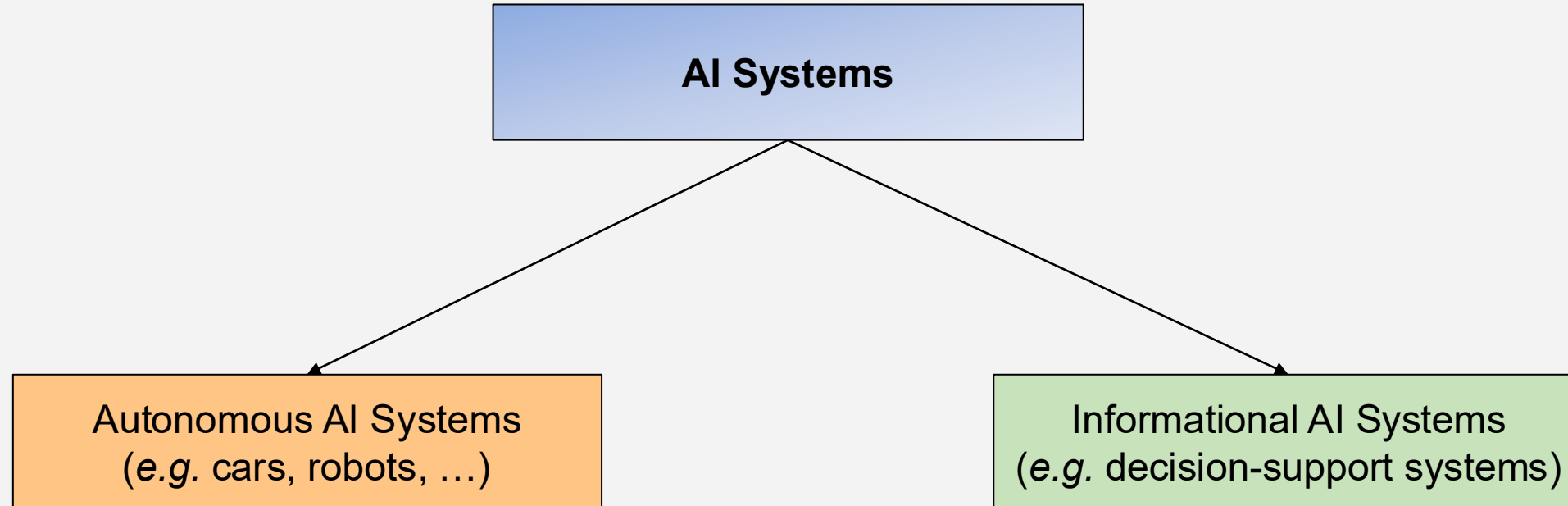
- ESFM (100M-1B) Released in April 26
- ClimLlama (3B-7B) Release in July 26
- Operationalization under Weather Generator

A model is like an engine block...

...to actually drive, you need a whole **system**

Why it's not trivial to build AI-based products

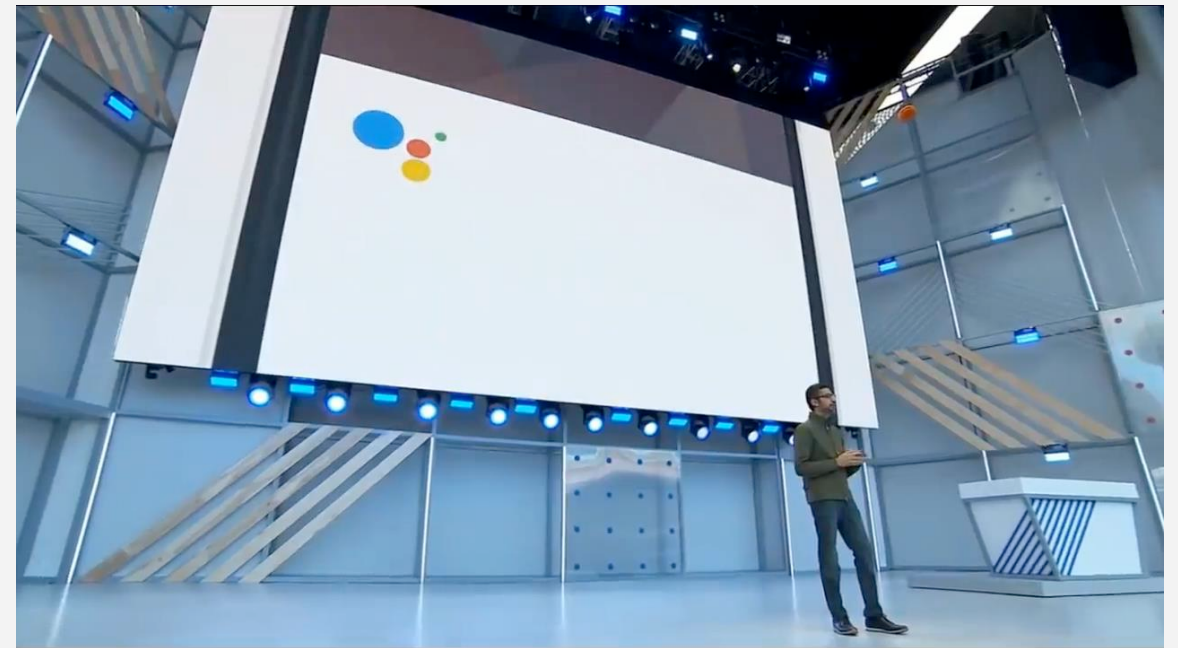
Two types of AI systems



Examples for autonomous AI systems



Source: youtube, id=V4PDTD2VHSU

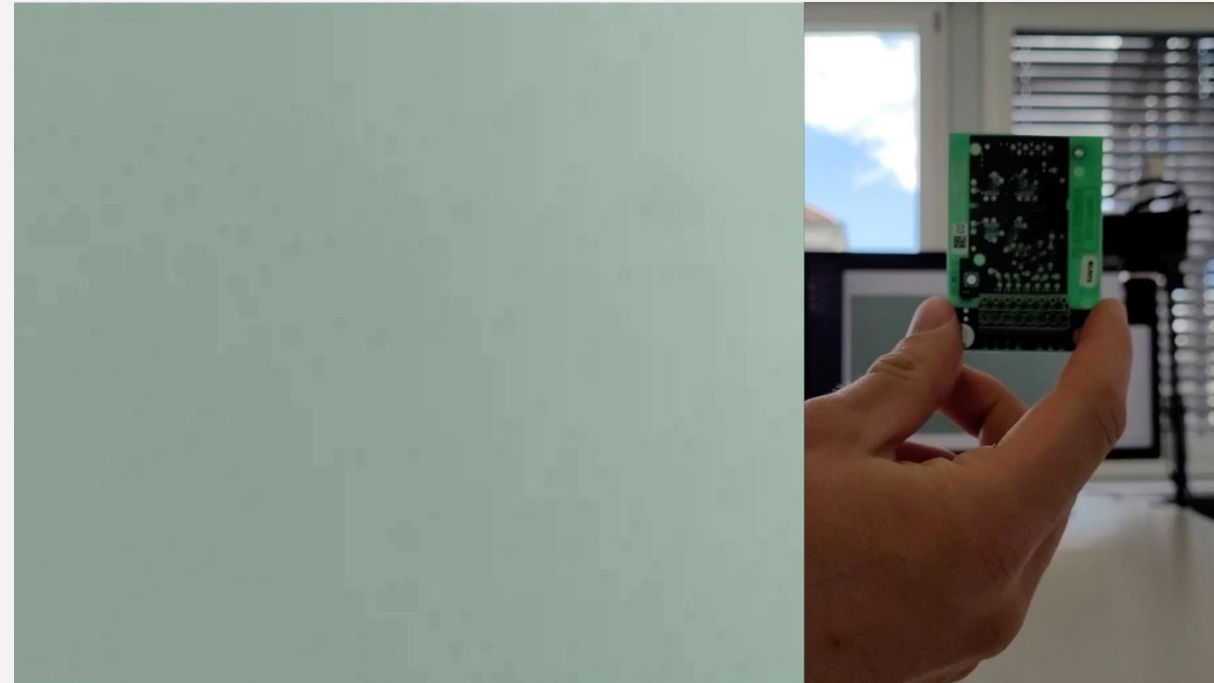


<https://youtu.be/D5VN56jQMWM?t=47>

Examples for informational AI systems



<https://www.brainlab.com/surgery-products/overview-neurosurgery-products/microscope-navigation/>



ethon.ai

Reflection

- ⌘ How are mistakes handled?
- ⌘ Who is responsible?
- ⌘ What is more difficult to roll-out?
- ⌘ In your industry, do you see predominately autonomous or informational AI systems?

What's different now
AI allows organizations to punch above their weight!

AI is a **force multiplier** where small teams can make a huge impact

Midjourney vs. Adobe

10 30'000 employees

OpenAI vs. Google

350 150'000 employees

What does it mean if you have employees that can make a >10x impact?

- High salary costs are not a problem anymore (bring back more parts of the value chain)
- Significant GDP uplift (and not just substitution)
- Trustworthiness in AI critical for adoption

How disruptive will AI be?

The unicorn challenge: AI enables a single person to create a billion dollar company?



<https://www.youtube.com/watch?v=Q5vsEUgxt3E>

A \$10tn+ service opportunity?

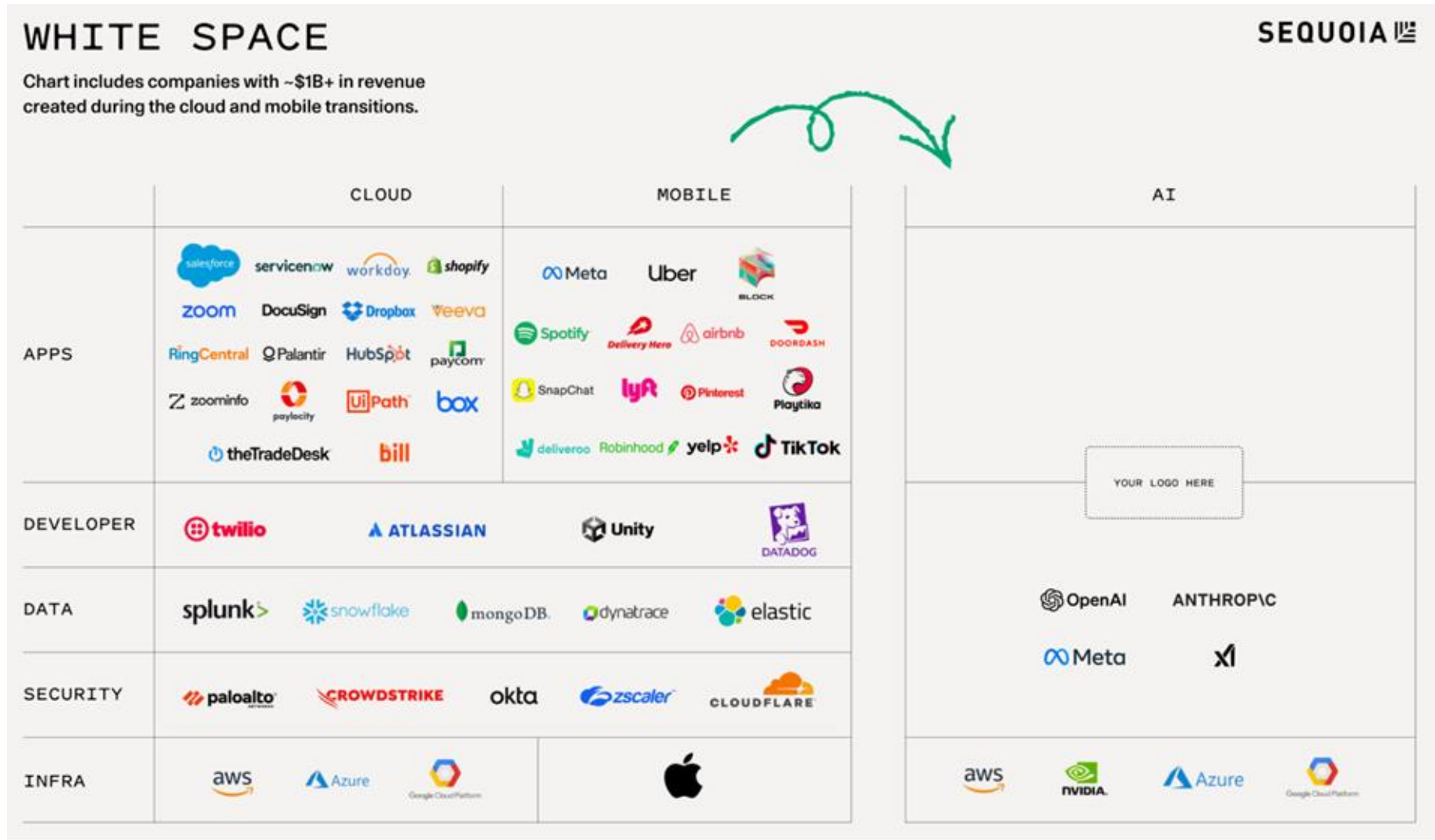
Application layer AI companies are not just UIs on top of a foundation model. Far from it. They have sophisticated cognitive architectures that typically include multiple foundation models with some sort of routing mechanism on top, vector and/or graph databases for RAG, guardrails to ensure compliance, and application logic that mimics the way a human might think about reasoning through a workflow.

Service-as-a-Software

The cloud transition was software-as-a-service. Software companies became cloud service providers. This was a \$350B opportunity. Thanks to agentic reasoning, the AI transition is service-as-a-software. Software companies turn labor into software. That means the addressable market is not the software market, but the services market measured in the *trillions* of dollars.

<https://www.sequoiacap.com/article/generative-ais-act-o1/>

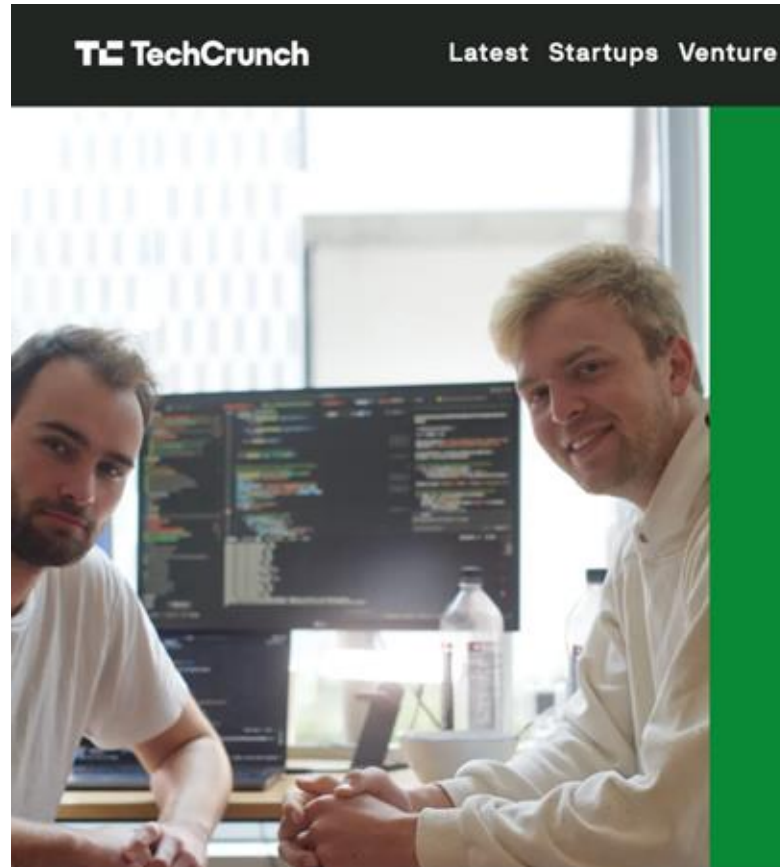
Service-as-a-Software: A \$10T+ industry in the making?



browser-use.com: From ETH AI Student Project to round at 100m

Series of events in 2024-2025:

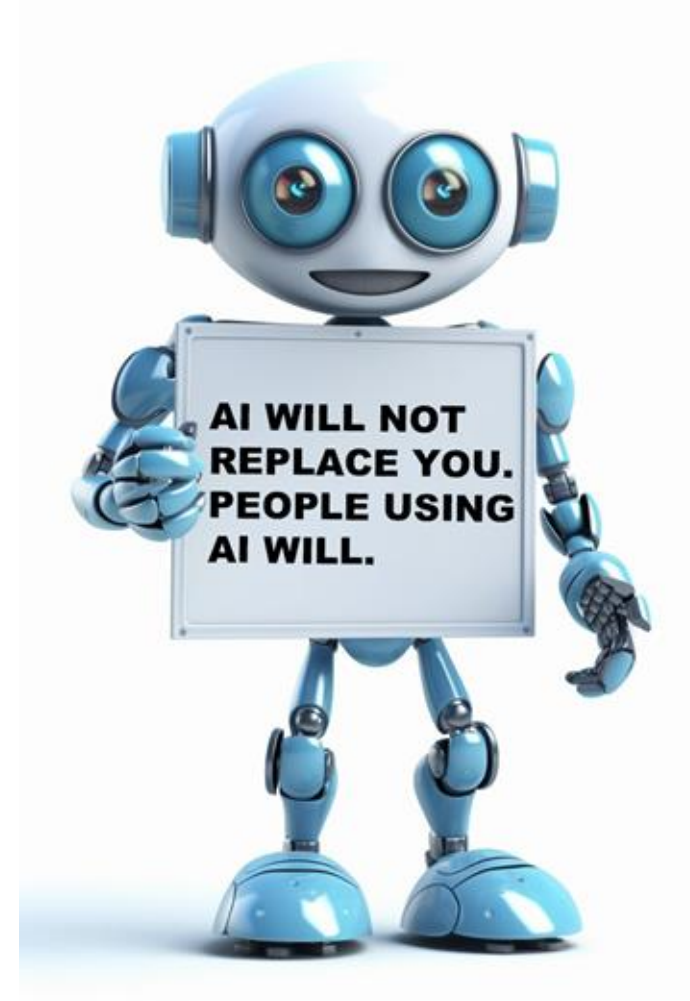
- 11-26 Oct: Co-Founder takes *Technology Investing* VC course
- Oct 31: Browser Use goes open-source
- Nov 12: Y Combinator (YC) deadline
- Jan → Mar: YC acceleration W25
- Mar: \$20M SAFE @ \$100M



Browser Use, the tool making it easier for AI 'agents' to navigate websites

Conclusion

- Think of AI as a having access to 1000 interns
- You are still responsible for the decisions
- Focus on core process innovation
- Need to rethink processes & business models
- Use AI as a force multiplier
- Leverage the Swiss ecosystem





ETH Zurich + University of St.Gallen

Open discussion