
# AI and Infrastructure

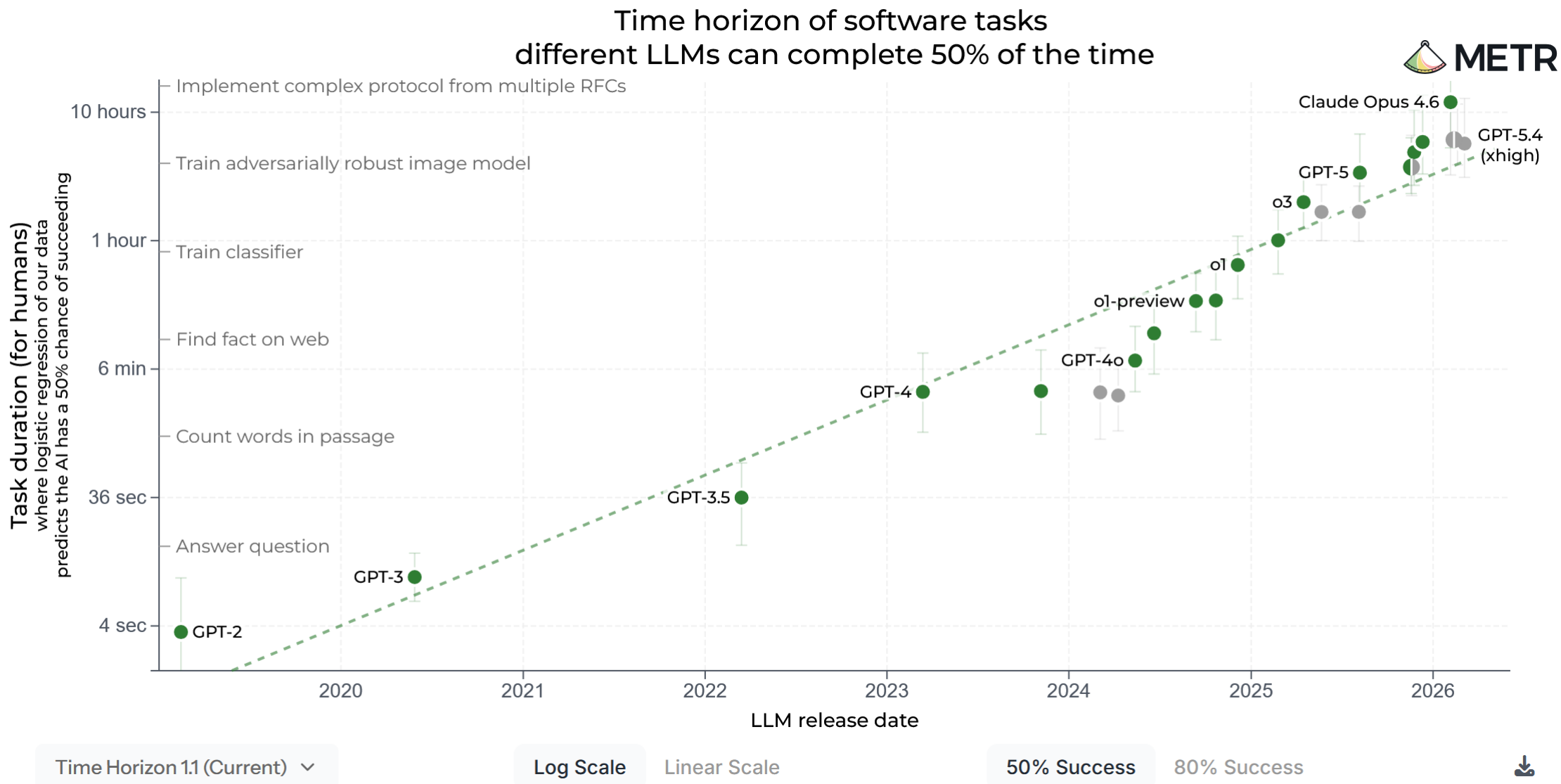
Prof. Fredrik Heintz
Institutionen för Datavetenskap, Linköpings universitet

Director AI4x Excellens Center
Co-Director WASP

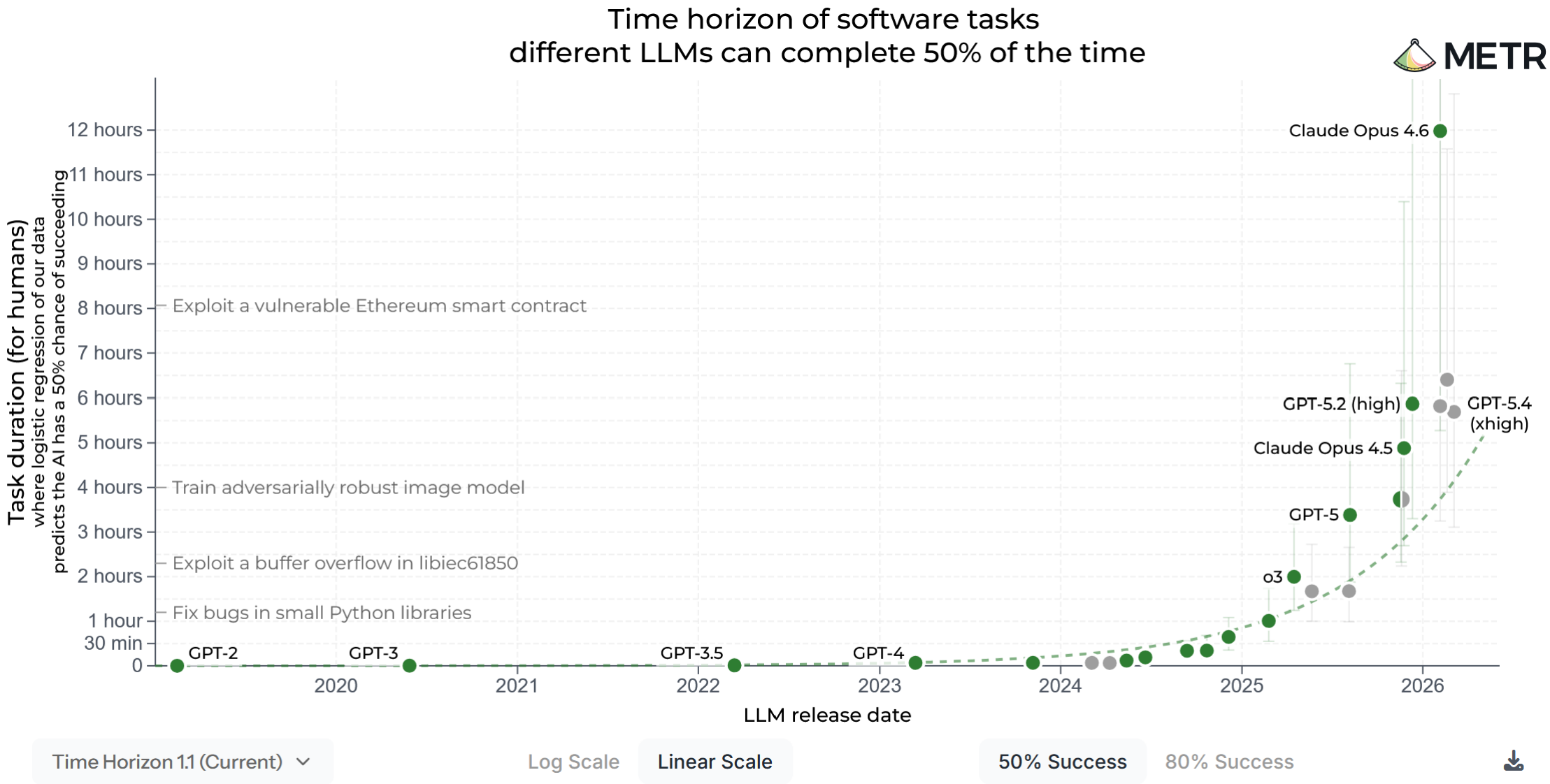
fredrik.heintz@liu.se @FredrikHeintz

LINKÖPING UNIVERSITY

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

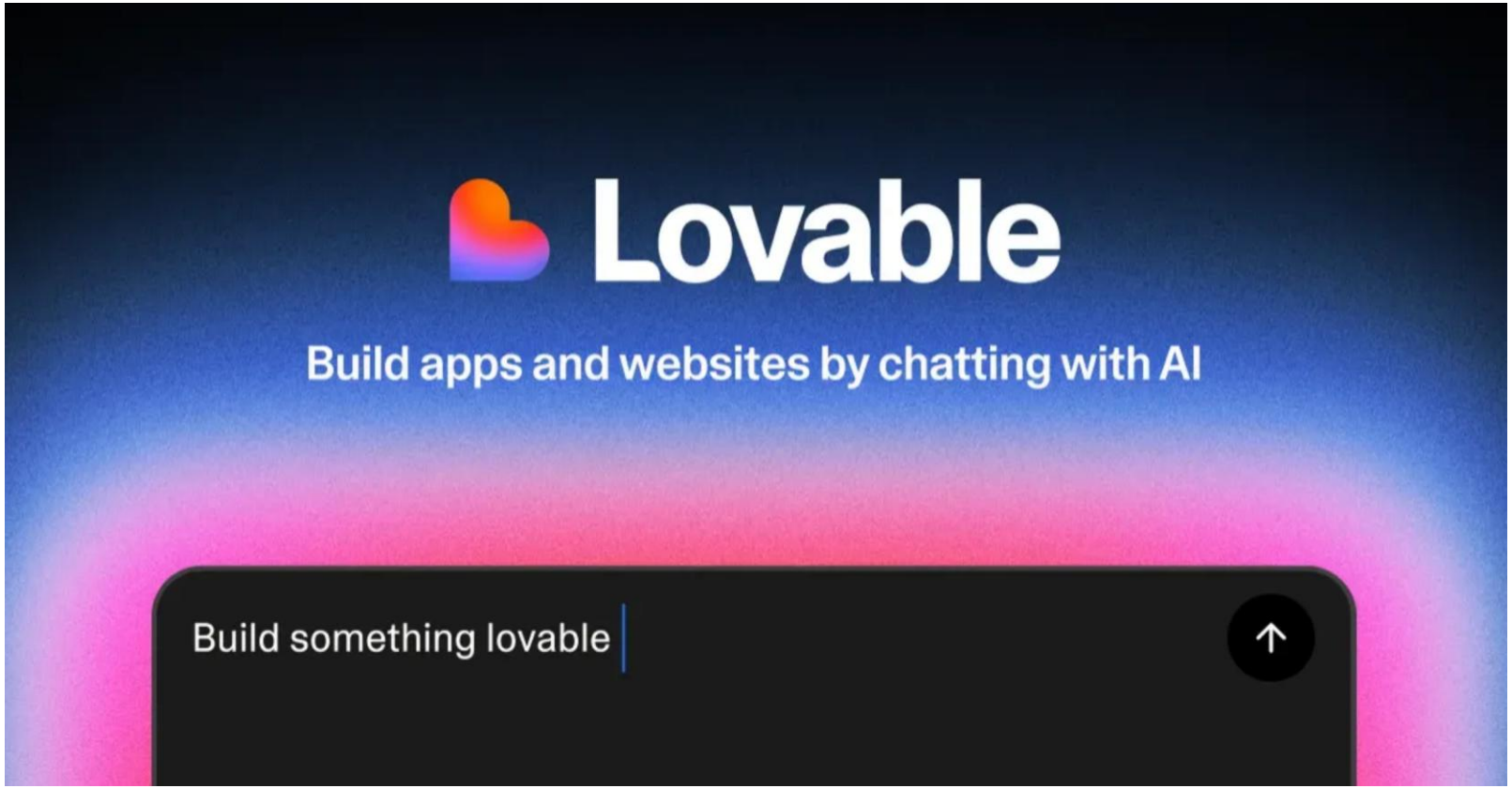



https://metr.org/time-horizons/

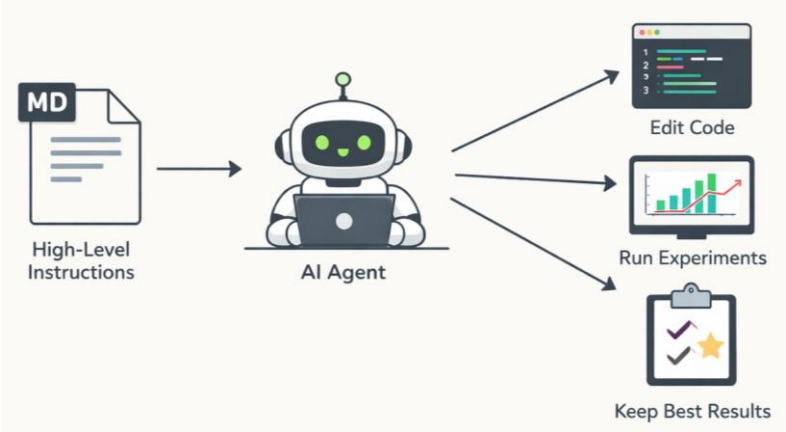



https://metr.org/time-horizons/

# Coding Assistants and Vibe Coding

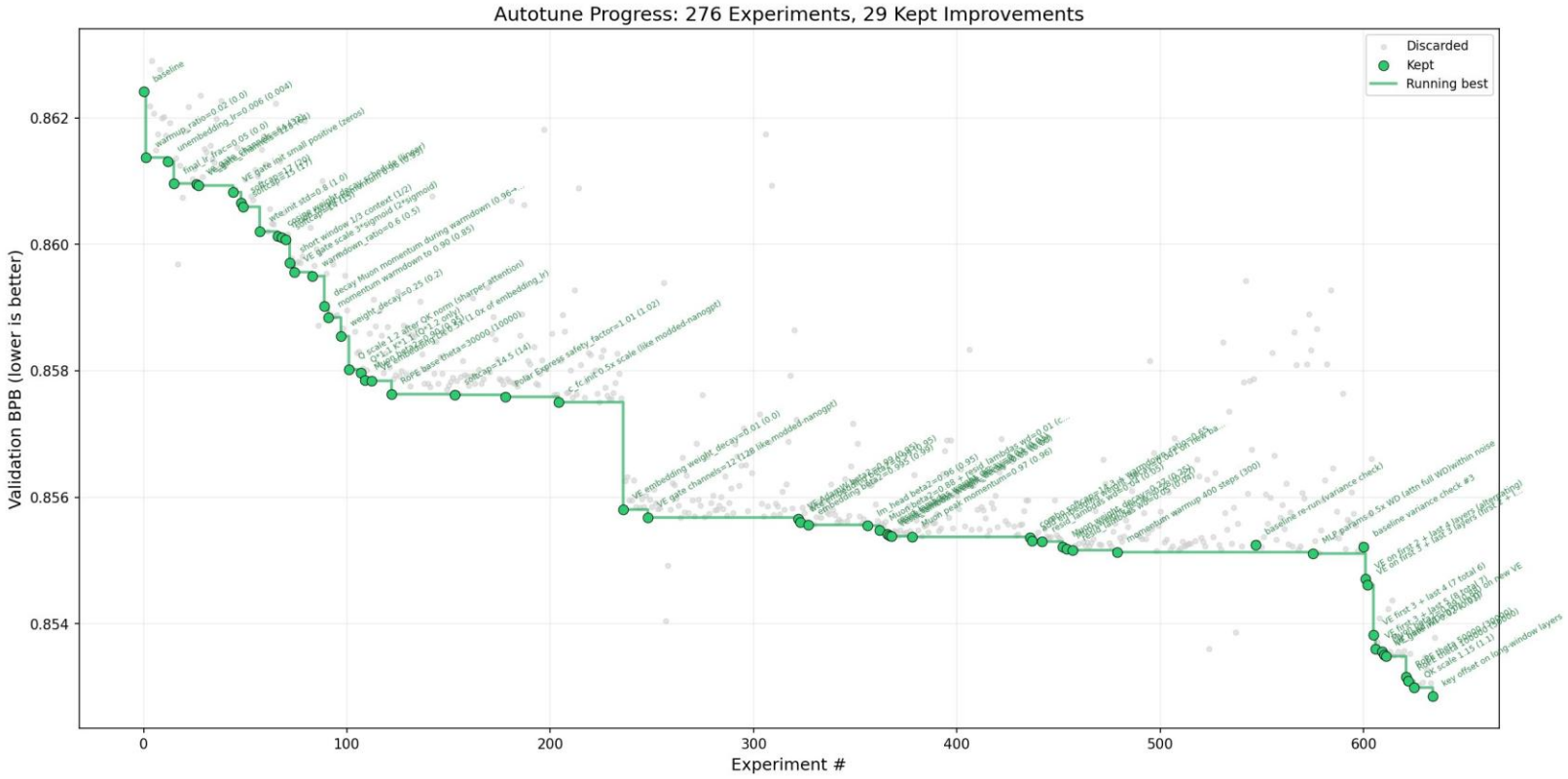



# Autoresearch



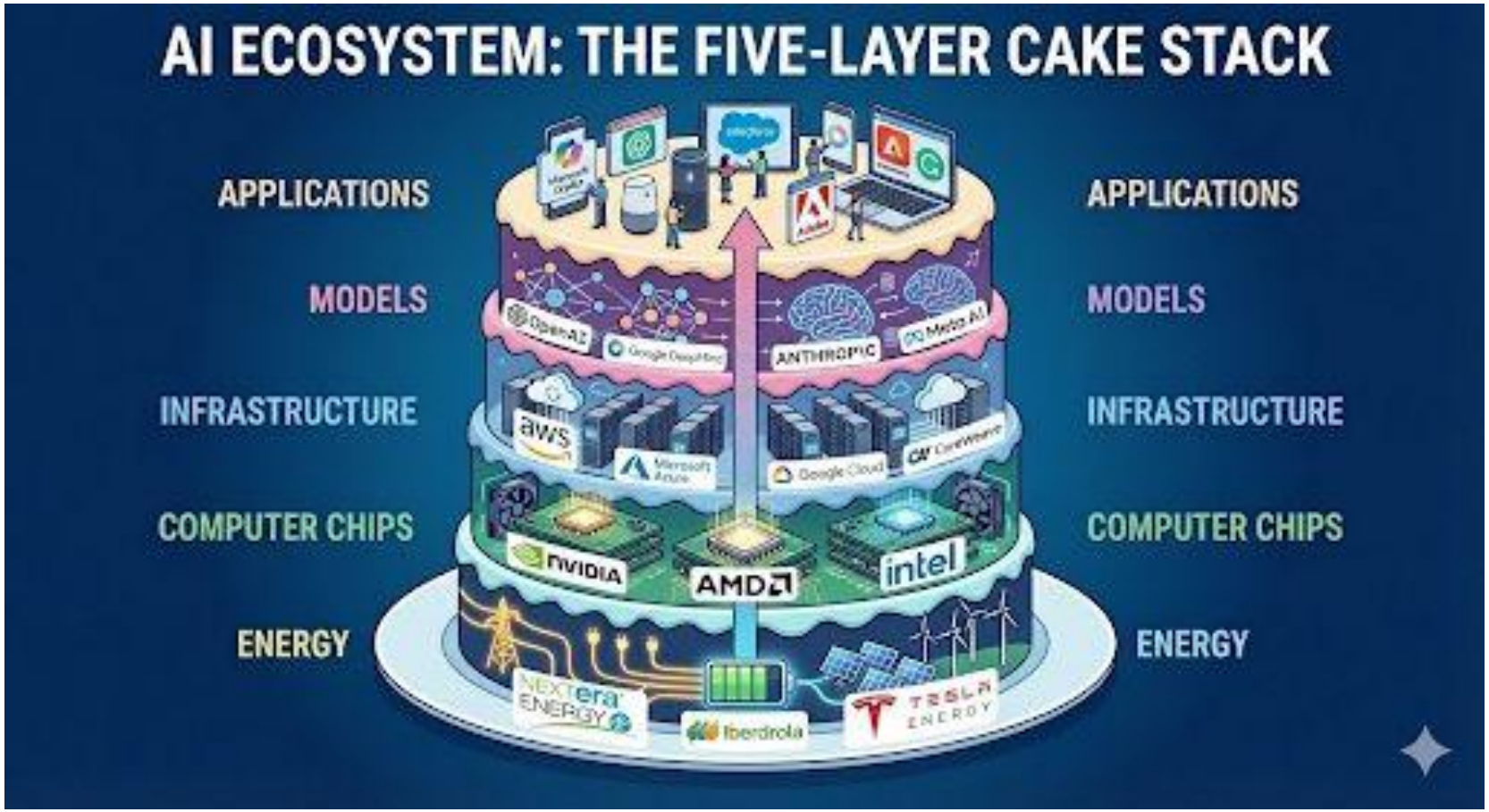

**Andrej Karpathy** @karpathy Mar 9

Three days ago I left autoresearch tuning nanochat for ~2 days on depth=12 model. It found ~20 changes that improved the validation loss. I tested these changes yesterday and all of them were additive and transferred to larger (depth=24) models. Stacking up all of these changes, today I measured that the leaderboard's "Time to GPT-2" drops from **2.02 hours to 1.80 hours** (~11% improvement), this will be the new leaderboard entry. So yes, these are real improvements and they make an actual difference. I am mildly surprised that my very first naive attempt already worked this well on top of what I thought was already a fairly manually well-tuned project.




https://github.com/karpathy/autoresearch

# The Five Layer AI Stack (according to J. Huang)



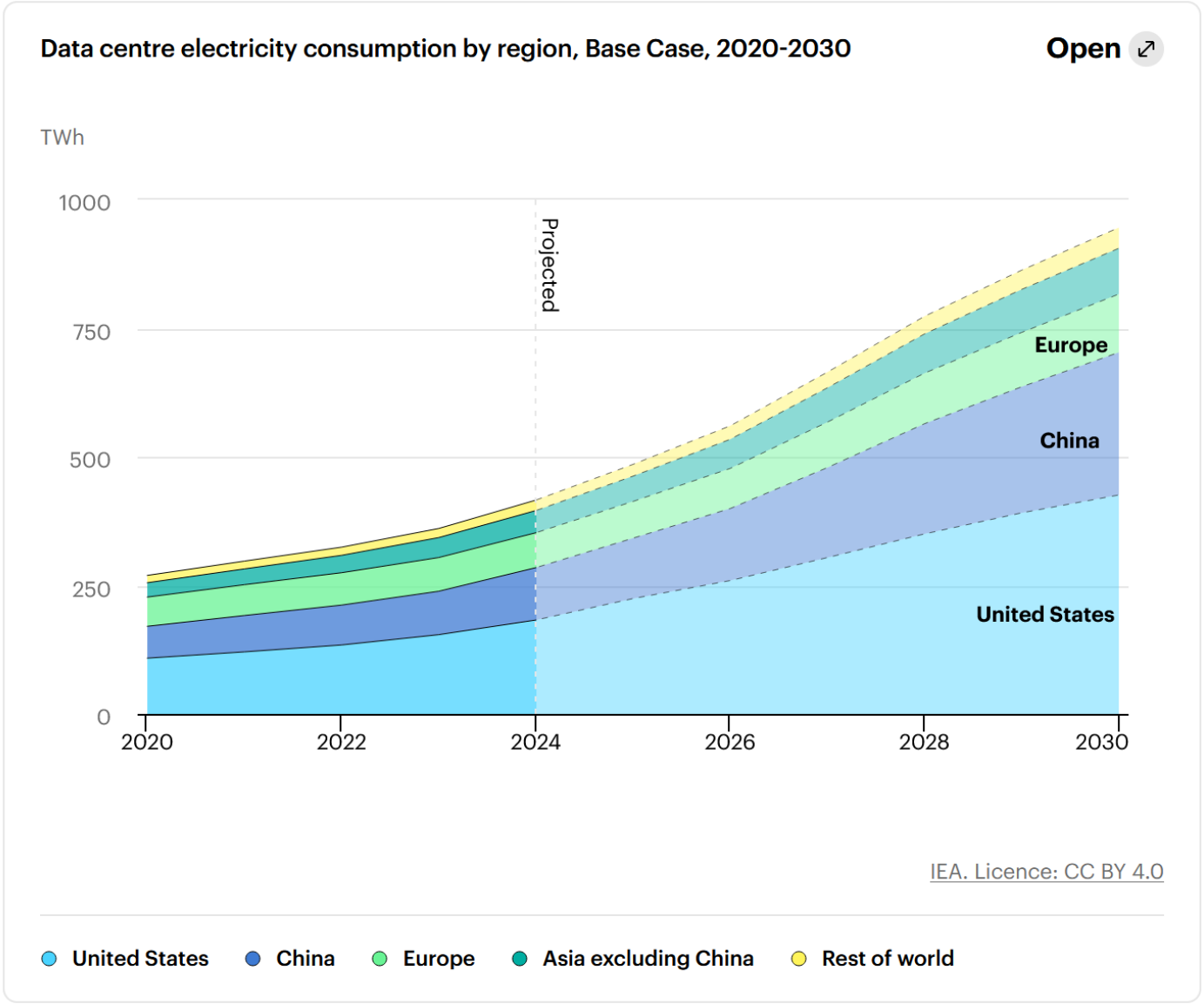

https://www.linkedin.com/posts/olawalerichardawe_ai-ecosystem-as-a-five-layer-cake-energy-activity-7402965341007667201-_B3R/

# Change in European power demand by sector 2030 vs 2024, TWh


Source: ICIS
Note: Hydrogen demand is based on EU RFNBO targets, which are likely overly optimistic



https://www.icis.com/explore/resources/data-centres-hungry-for-power/




https://www.iea.org/reports/energy-and-ai/energy-demand-from-ai

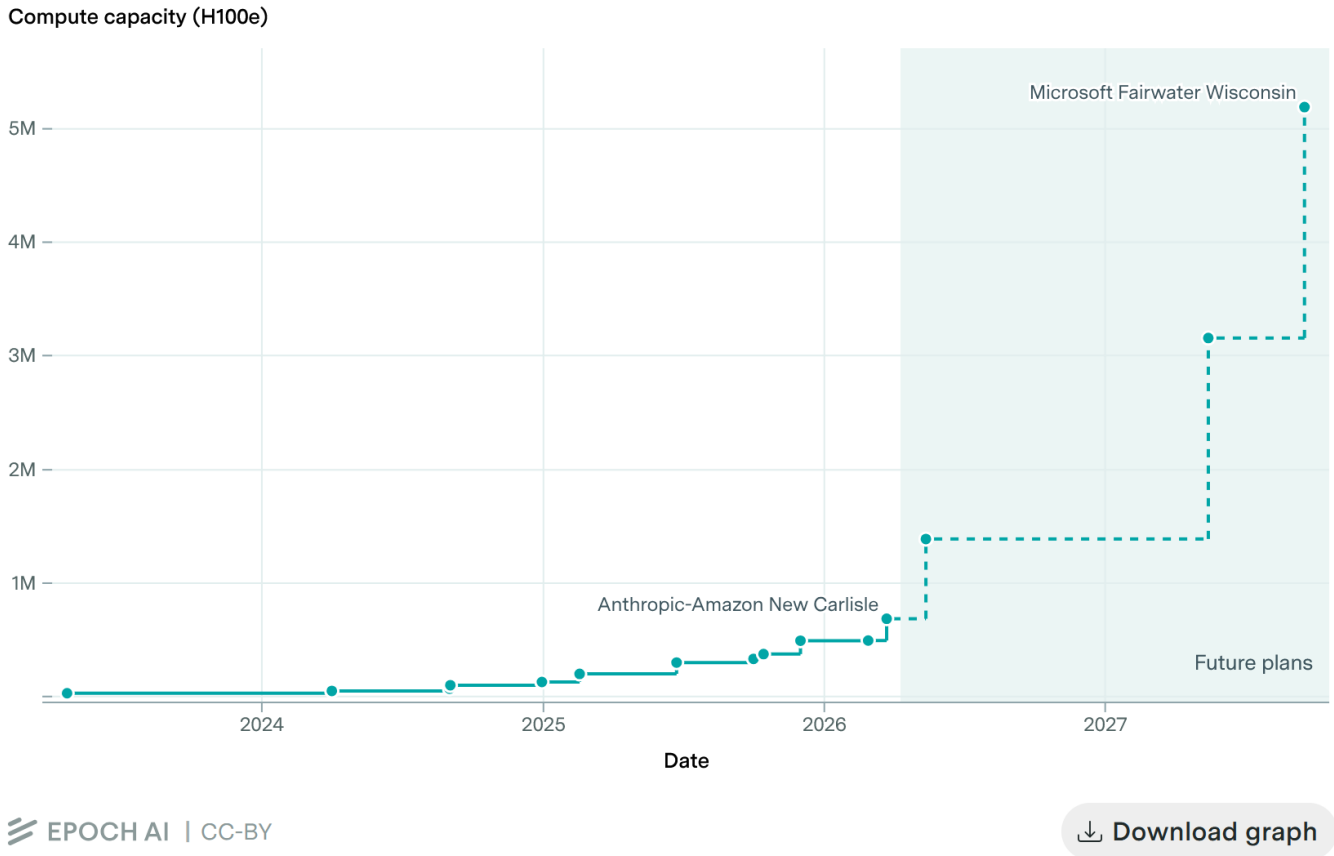
# Compute Build-Out

Data Centers

**The largest known AI data center has computing power equivalent to 700,000 NVIDIA H100 chips.**

Anthropic-Amazon New Carlisle has an estimated 1.1 GW of power capacity and $35 billion in capital costs, making it the largest known AI data center. Microsoft's Fairwater Wisconsin data center is expected to be nearly 8x more powerful, at 5.2M H100-equivalents by September 2027.

Learn more →






https://epoch.ai/trends#data-centers

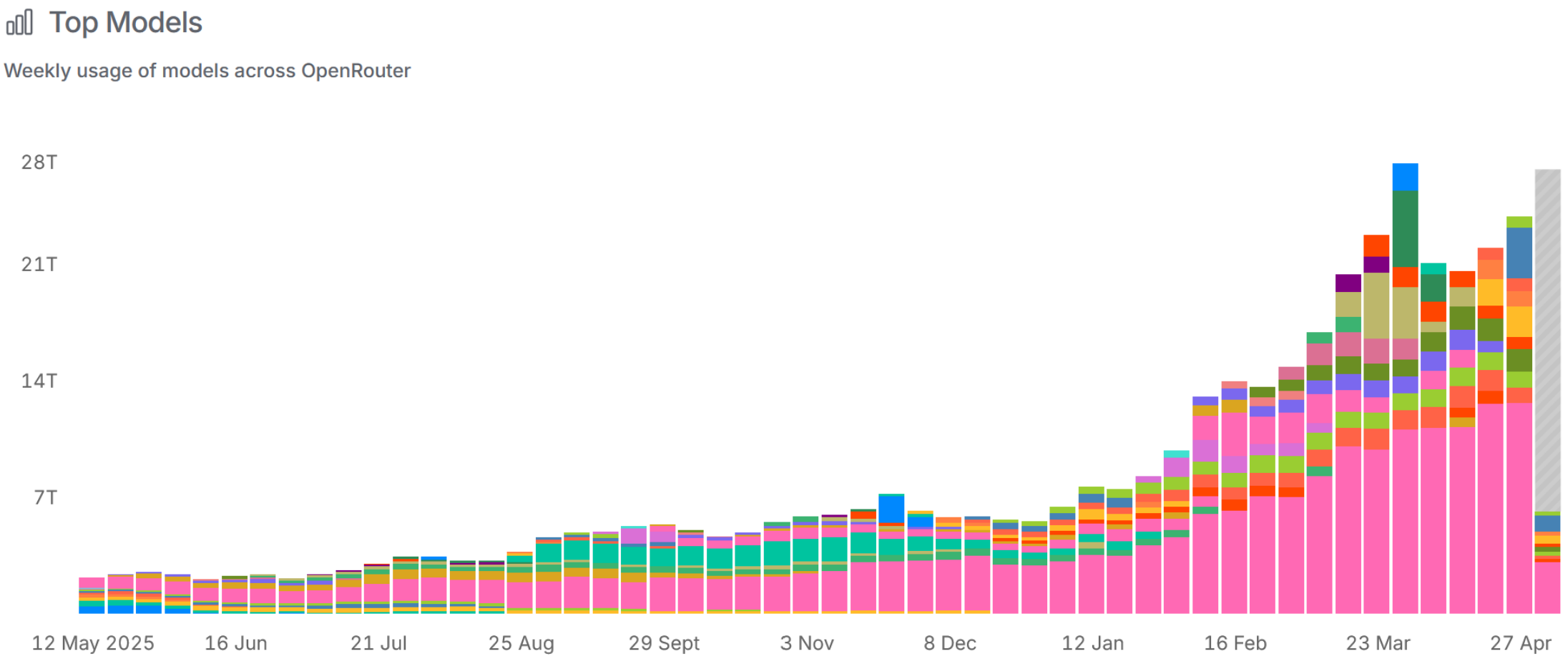
# Training vs Inference

- **2023:** Inference accounted for roughly **33%** of AI compute.
- **2025:** Inference reached **50%** of total AI compute.
- **2026 (Projected):** Inference will reach **65% to 70%** of all AI compute.
- **2029-2030 (Projected):** Inference is expected to account for **70-90%** of total compute and infrastructure spend

| IF YOUR WORKLOAD IS... | OPTIMIZE FOR | HARDWARE CHOICE | WHY |
|---|---|---|---|
| Training large models | Throughput | H100/H200, multi-node | Raw compute power matters |
| Production inference | Latency | B200/B300, specialized | User experience, cost per token |
| Variable inference load | Autoscaling | Cloud GPU instances | Match capacity to demand |
| Latency-critical inference | Edge deployment | Smaller GPUs distributed | Reduce network round-trip |
| Cost-sensitive inference | Efficiency | TPU, Trainium, AMD | 30-40% savings possible |

https://introl.com/blog/ai-inference-vs-training-infrastructure-economics-diverging

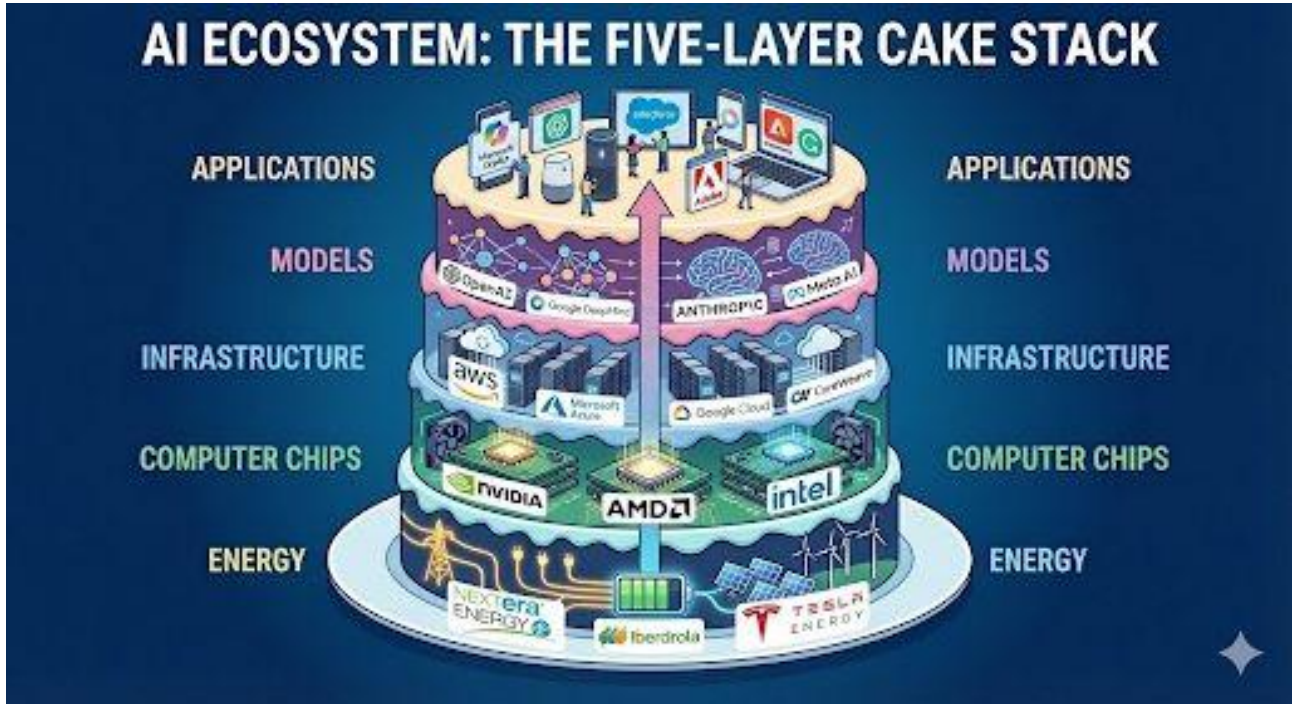
# Token Usage




https://openrouter.ai/rankings

# Alternatives to the Current Paradigm

- Small models
- Distributed / federated models
- Neurosymbolic AI




penligent

**Claude Mythos and Cyber Security, What the Leak Actually Tells Defenders**

# RESIST CRAI Center

**Center Management** (Director/PI, Manager, co-PIs) | **Board** | **SAC**

## Research Program
Trustworthy and Verifiable AI
Runtime Security Assurance
Robust and Secure AI-Supported Development
Resilient Distributed and Agentic AI

## Validation Program
AI-driven critical Infrastructure
AI-empowered software services
AI-automated robots and vehicles

## Center Activities
Strategic research agenda
Education and training
Policy and strategy advise
Internationalization
Dissemination

**Academia**

LINKÖPINGS UNIVERSITET
RISE
ÖREBRO UNIVERSITET
LUNDS UNIVERSITET

**Industry**

COMBITECH
COMBIENT
SECTRA
Epiroc
SIEMENS energy
ERICSSON
SAAB

**Society**

FOI
MSB Swedish Civil Contingencies Agency
CYBERCAMPUS SVERIGE

**Internat.**

IMPERIAL
Inria
The Alan Turing Institute
UMEÅ UNIVERSITET
UNIVERSITY OF WASHINGTON
UNIVERSITY OF CALIFORNIA SANTA BARBARA
ACTION AI INSTITUTE

# Where are the Bottlenecks?

- Energy – Compute – Tokens
- Domain/Application Optimised Models
- Regulation, legal certainty
- Competence