Opening Keynote

Status Quo generativer KI vor dem Hintergrund realer Produktivitätsgewinne

KI-Ready 2025
Jonas Gottschalk
20.02.2025

DEYAN7

The GenAI industry is growing fast led by OpenAI

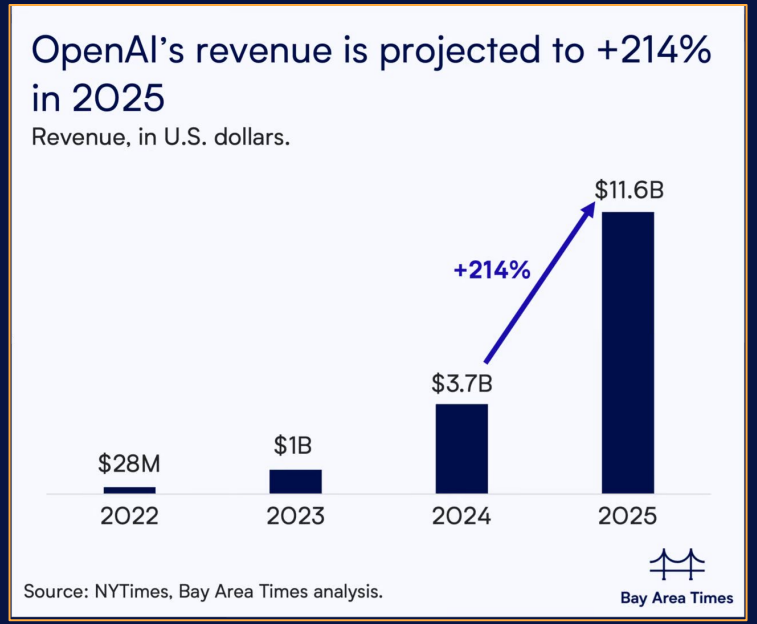
DEYAN7

OpenAI on track to hit $2bn revenue milestone as growth rockets

San Francisco-based start-up joins Google and Meta as one of the fastest-growing tech companies ever

Sam Altman says 92% of Fortune 500 companies use OpenAI products © FT montage/AP

Financial Times, February 9 2024



Bay Area Times, September 20 2024

High Demand for Gen AI consulting

DEYAN7

Accenture ushers in holiday season with $1.2 billion in Gen AI orders

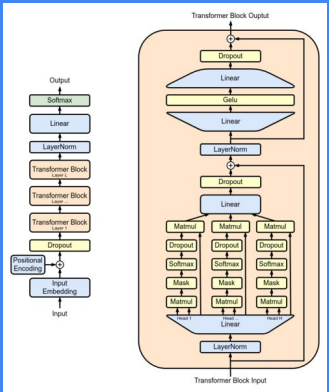
Jas Bardia | 4 min read | 19 Dec 2024, 10:58 PM IST

Accenture's Gen AI bookings is the highest the company has reported in a quarter. (REUTERS)

What is generative AI?

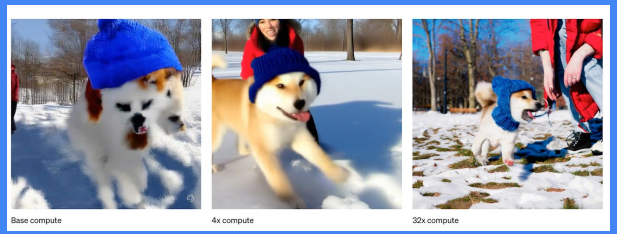
DEYAN7

New AI Architecture
Transformer

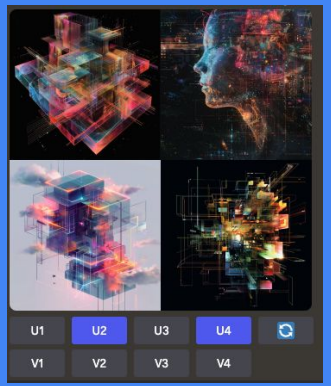


Scale
Massive Amounts of Data and Compute

All knowledge of the internet

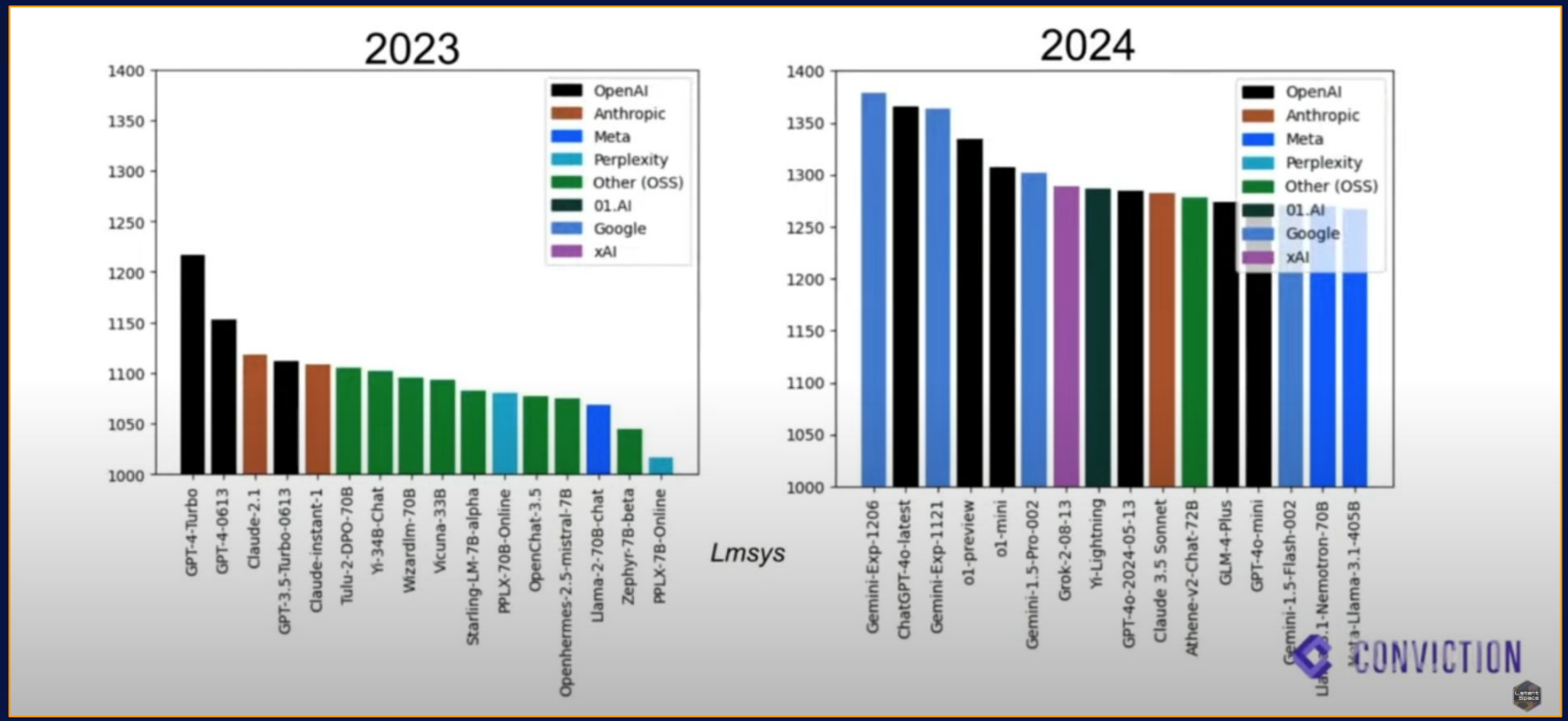


Algorithmic Breakthroughs
RLHF



The AI Race - 2024 was much closer than 2023

DEYAN7



The AI Race - 2025 is not only about the US labs…

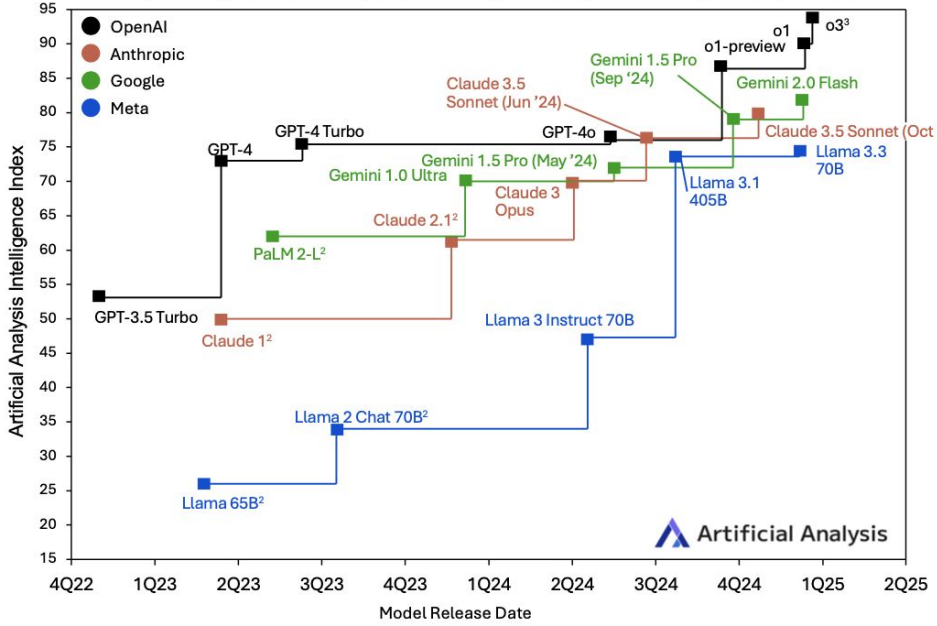
DEYAN7

LEADING US FRONTIER LANGUAGE MODELS

Since the launch of OpenAI's GPT-4 in early 2023, leading US AI labs have scrambled to catch up to OpenAI

Leading US AI Labs Frontier Language Model Intelligence, Over Time[1]

Artificial Analysis Intelligence Index: MMLU, GPQA Diamond, MATH-500, HumanEval. Top 5 Labs by Market Share.



Key Trends

Competing labs catch up to OpenAI's GPT-4: OpenAI started the language model race in November 2022 with the launch of GPT-3.5 in ChatGPT; leading US labs have largely caught up with frontier models from OpenAI.

Big Tech closes in on the frontier labs: Models from Google and Meta are rapidly closing in on frontier models, with Gemini 2.0 Flash exceeding Claude 3.5 Sonnet and GPT 4o capabilities.

Sparks of intelligence beyond GPT-4: The final months of 2024 have seen the emergence of the first major intelligence leaps beyond GPT-4, led by OpenAI's o3. Topics including reasoning models, data quality and new reinforcement learning techniques have joined pre-training compute scaling as dominant levers for improving models.

1. Artificial Analysis Intelligence Index: average across a range of language model intelligence and reasoning evaluation datasets. Currently includes MMLU, GPQA Diamond, MATH-500 & HumanEval. Release date is based on first public launch of the model. 2. Estimated based on company claims and comparable results where available, not yet independently benchmarked by Artificial Analysis. 3. o3 Intelligence Index estimated by scaling measured Intelligence Score of o1.

Artificial Analysis

The AI Race - … but also about China vs. US

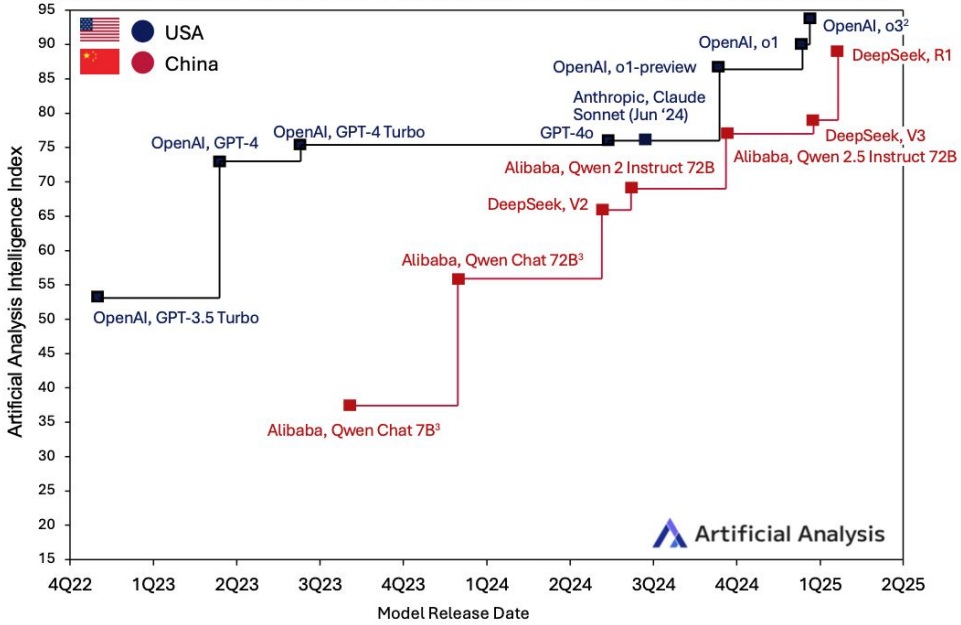
DEYAN7

FRONTIER LANGUAGE MODELS BY ORIGIN

Chinese AI labs have progressively caught up to US AI labs; models from Chinese labs are now approaching o1-level intelligence with the release of DeepSeek's R1 model

Frontier Language Model Intelligence, Over Time[1]

Artificial Analysis Intelligence Index: MMLU, GPQA Diamond, MATH-500, HumanEval. Top 5 Labs by Market Share.



Key Trends

Closing the gap: The final months of 2024 have seen the emergence of the numerous highly performant models from top Chinese AI labs. This has resulted in the delta between the level of intelligence offered by models from Chinese AI labs and US AI labs closing. Several Chinese models are now competitive with models from the top US labs.

Reasoning models quickly becoming commonplace: Reasoning models (that "think" before answering) were first introduced by OpenAI in 3Q24. Within months, Chinese competitors, led by DeepSeek, have largely replicated the intelligence of o1. Several AI labs in China now have a frontier-level reasoning model.

Open models close in on the frontier labs: Open weights models, led by those from DeepSeek and Alibaba, have approached o1level intelligence.

1. Artificial Analysis Intelligence Index: average across a range of language model intelligence and reasoning evaluation datasets. Currently includes MMLU, GPQA Diamond, MATH-500 & HumanEval. Release date is based on first public launch of the model. 2. o3 Intelligence Index estimated by scaling measured Intelligence Index of o1. 3. Estimated based on company claims and comparable results where available, not yet independently benchmarked by Artificial Analysis

Artificial Analysis

The AI Race - Germany is not even on the map

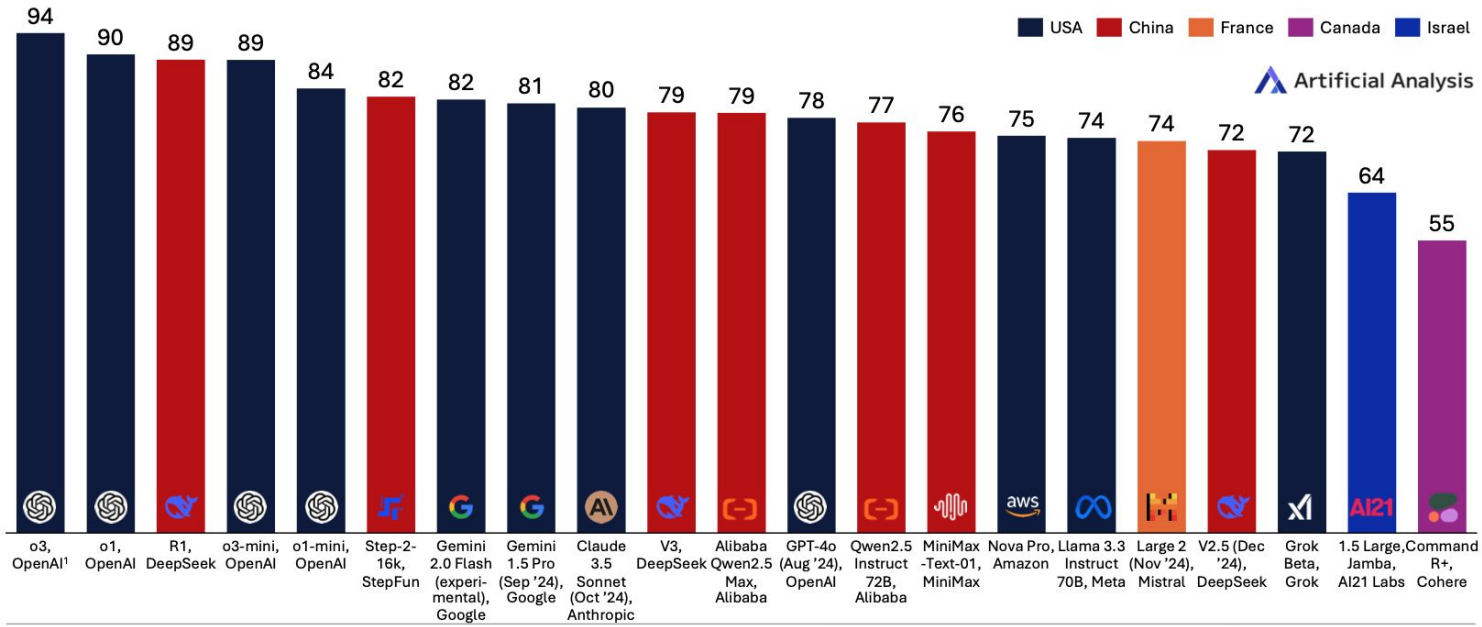
DEYAN7

LANGUAGE MODEL COUNTRY OF ORIGIN

While the US maintains an overall lead in the intelligence frontier, China is no longer far behind. Few other countries have demonstrated frontier-class training

The Language Model Frontier: Country of Origin

Artificial Analysis Intelligence Index, Selected Leading Models (Early 2025), Non-exhaustive



1. *Estimated based on company claims and comparable results where available, not yet independently benchmarked by Artificial Analysis*
2. *A number of leading models from Chinese AI labs are excluded due to limited access or evaluation data*

Artificial Analysis

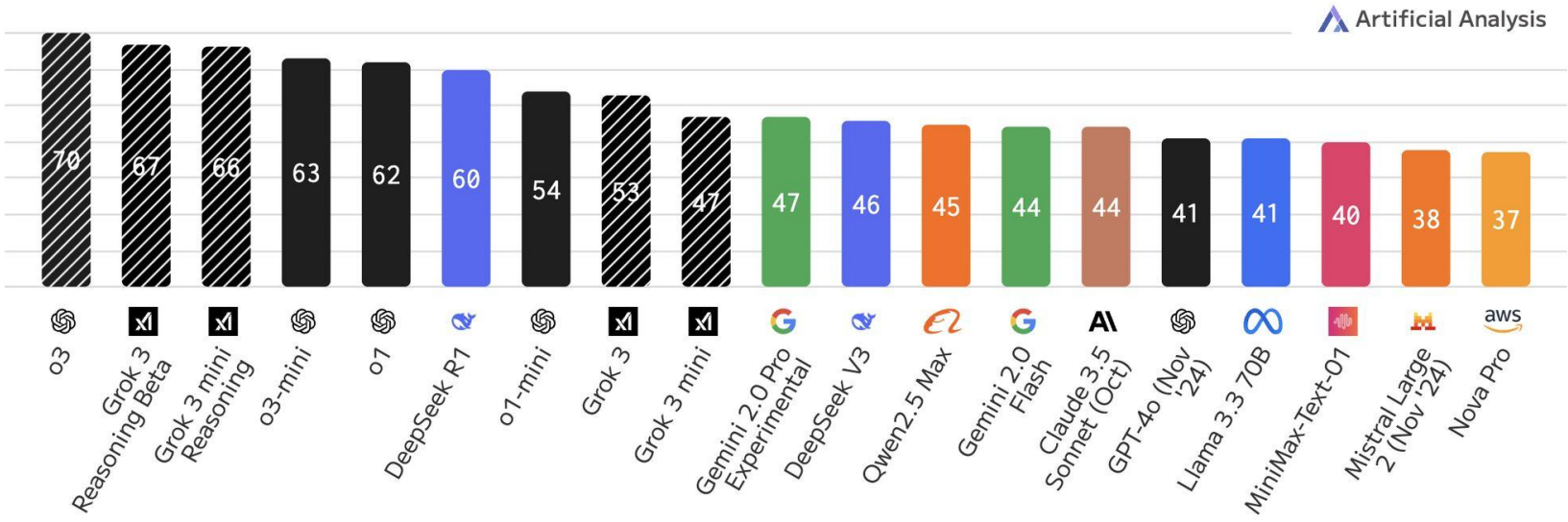
Hot update from Tuesday - xAI enters the chat

DEYAN7

Artificial Analysis Intelligence Index

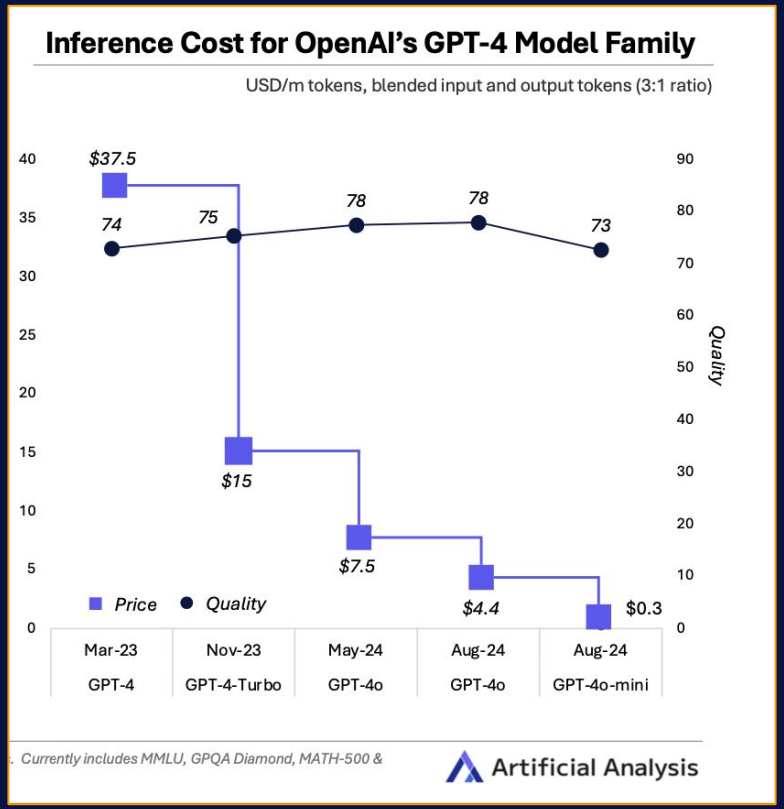
Artificial Analysis Intelligence Index (Version 2, released Feb '25); Higher is better

Estimate (independent evaluation forthcoming)



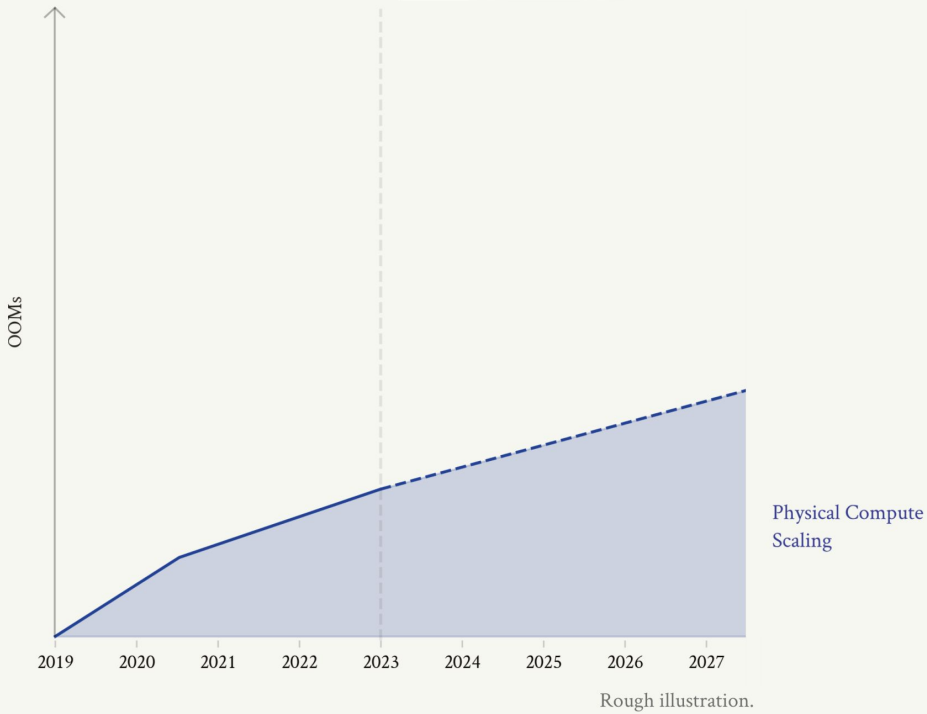
The AI Race - 100x price drop in 2024

DEYAN7



The decade ahead - Datacenter build-out

DEYAN7



The decade ahead - Datacenter build-out

DEYAN7

Year	OOMs	# of H100s-equivalent	Cost	Power	Power reference class
2022	~GPT-4 cluster	~10k	~$500M	~10 MW	~10,000 average homes
~2024	+1 OOM	~100k	$billions	~100MW	~100,000 homes
~2026	+2 OOMs	~1M	$10s of billions	~1 GW	The Hoover Dam, or a large nuclear reactor
~2028	+3 OOMs	~10M	$100s of billions	~10 GW	A small/medium US state
~2030	+4 OOMs	~100M	$1T+	~100GW	>20% of US electricity production

Scaling the largest training clusters, rough back-of-the-envelope calculations.

The decade ahead - US perspective

DEYAN7

The decade ahead - EU perspective

DEYAN7

AI | 12 FEB, 2025

‘Plug, baby, plug!’: Billions of AI investment heralded at European summit

At the AI Action Summit in Paris this week, French President Emmanuel Macron announced an investment of €109bn in the sector in Europe.

Ursula von der Leyen, president of the European Commission, also cited the figure of €200bn to be invested in the sector in Europe.

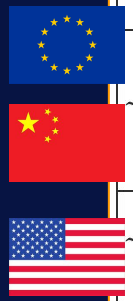
The decade ahead - Overview

DEYAN7

Projekt / Unternehmen	Geplante Investition	Kategorie	Region
Project Stargate (USA)	~500 Mrd. USD (über 4 Jahre; ca. 100 Mrd. sofort)	Staatl. Initiative (privat mitfinanziert)	USA
InvestAI (EU)	~200 Mrd. €	Überstaatl. Programm (öffentlich + privat)	Europa
Frankreich: Nationales KI-Programm	~110 Mrd. €	Nationale Initiative (staatl. + privat)	Frankreich (EU)
Microsoft	~80 Mrd. USD (CapEx 2025)	Unternehmen (Cloud / KI-Infrastruktur)	USA (global)
Alphabet (Google)	~75 Mrd. USD (CapEx 2025)	Unternehmen (Cloud / KI-Infrastruktur)	USA (global)
Meta	~60–65 Mrd. USD (CapEx 2025)	Unternehmen (KI-Infrastruktur)	USA (global)
OpenAI & Microsoft (Joint Venture)	~100 Mrd. USD (geplante Infra-Erweiterung)	Unternehmen (Rechenzentren, Modell-Training)	USA
Anthropic	~10 Mrd. USD (Finanzierungsrunde bis 2025)	Unternehmen (Startup-Finanzierung / F&E)	USA
USA – Gesamt (Prognose)	~800+ Mrd. USD (2025)	**Summe (Region)**	USA
Europa – Gesamt (Prognose)	~200 Mrd. € (InvestAI) + nationale Programme*	**Summe (Region)**	Europa
China – Geschätzt	~140+ Mrd. USD (geplante staatl. Programme u. Private)	**Summe (Region)**	China

The decade ahead - US leads the physical scale up

DEYAN7

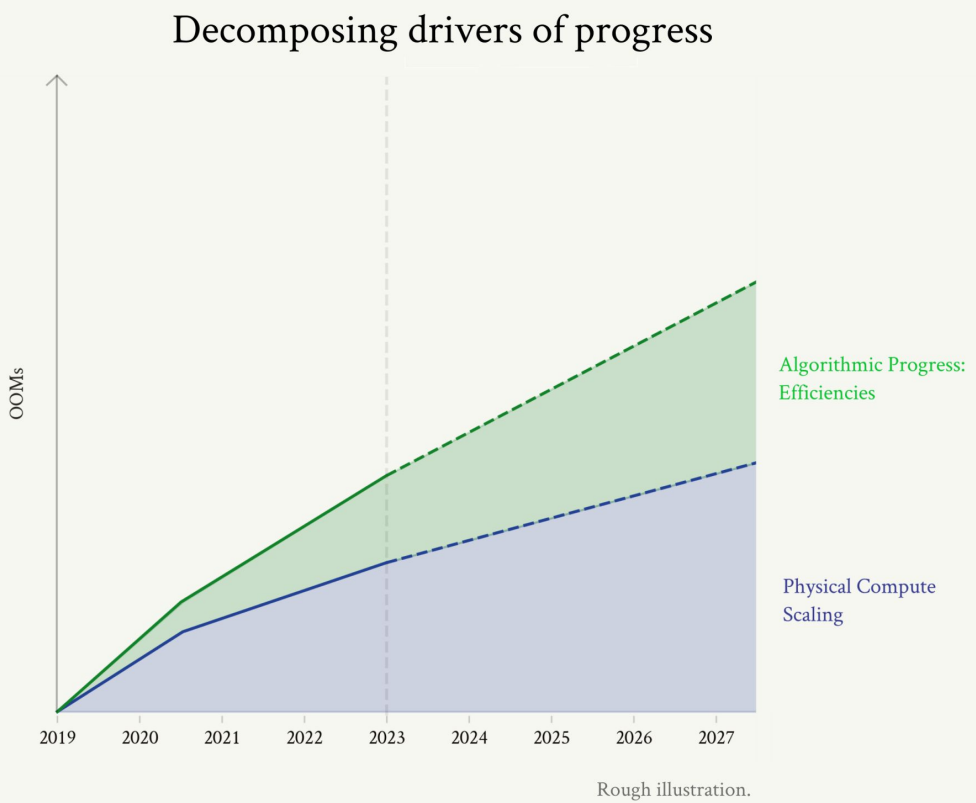


Year	OOMs	# of H100s-equivalent	Cost	Power	Power reference class
2022	~GPT-4 cluster	~10k	~$500M	~10 MW	~10,000 average homes
~2024	+1 OOM	~100k	$billions	~100MW	~100,000 homes
~2026	+2 OOMs	~1M	$10s of billions	~1 GW	The Hoover Dam, or a large nuclear reactor
~2028	+3 OOMs	~10M	$100s of billions	~10 GW	A small/medium US state
~2030	+4 OOMs	~100M	$1T+	~100GW	>20% of US electricity production

Scaling the largest training clusters, rough back-of-the-envelope calculations.

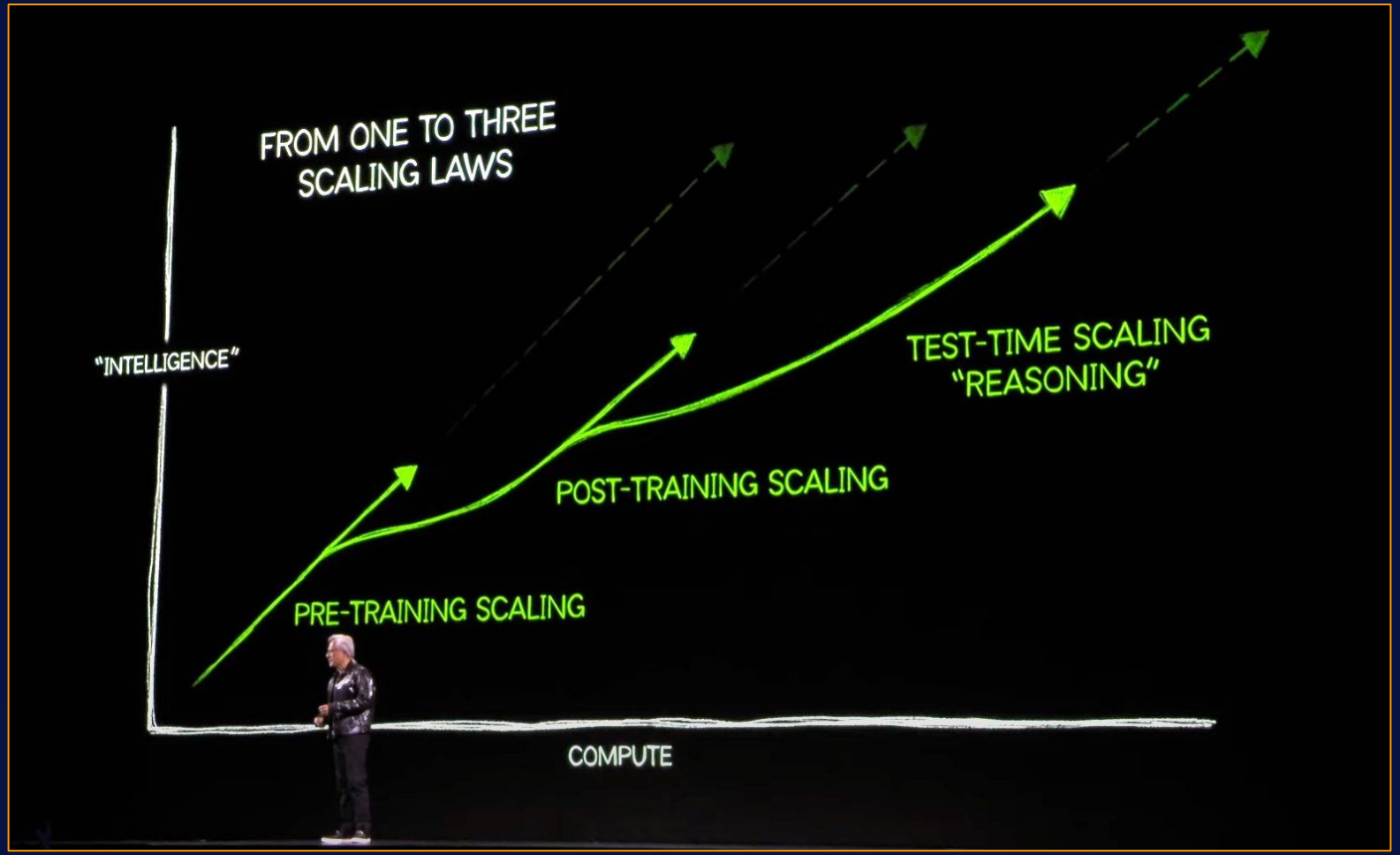
The decade ahead - AI research progress

DEYAN7



The decade ahead - From 1 to 3 scaling laws

DEYAN7

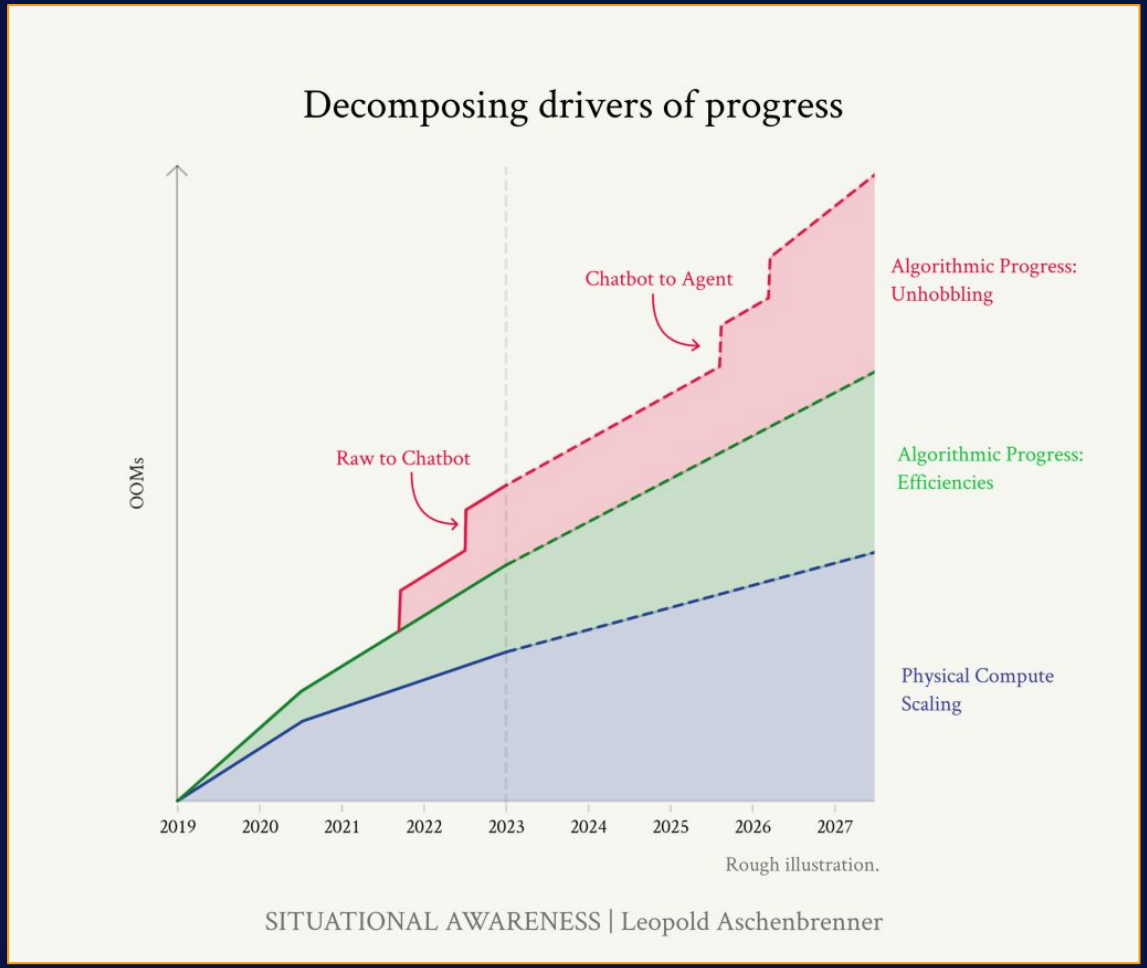


The decade ahead - Research still matters

DEYAN7

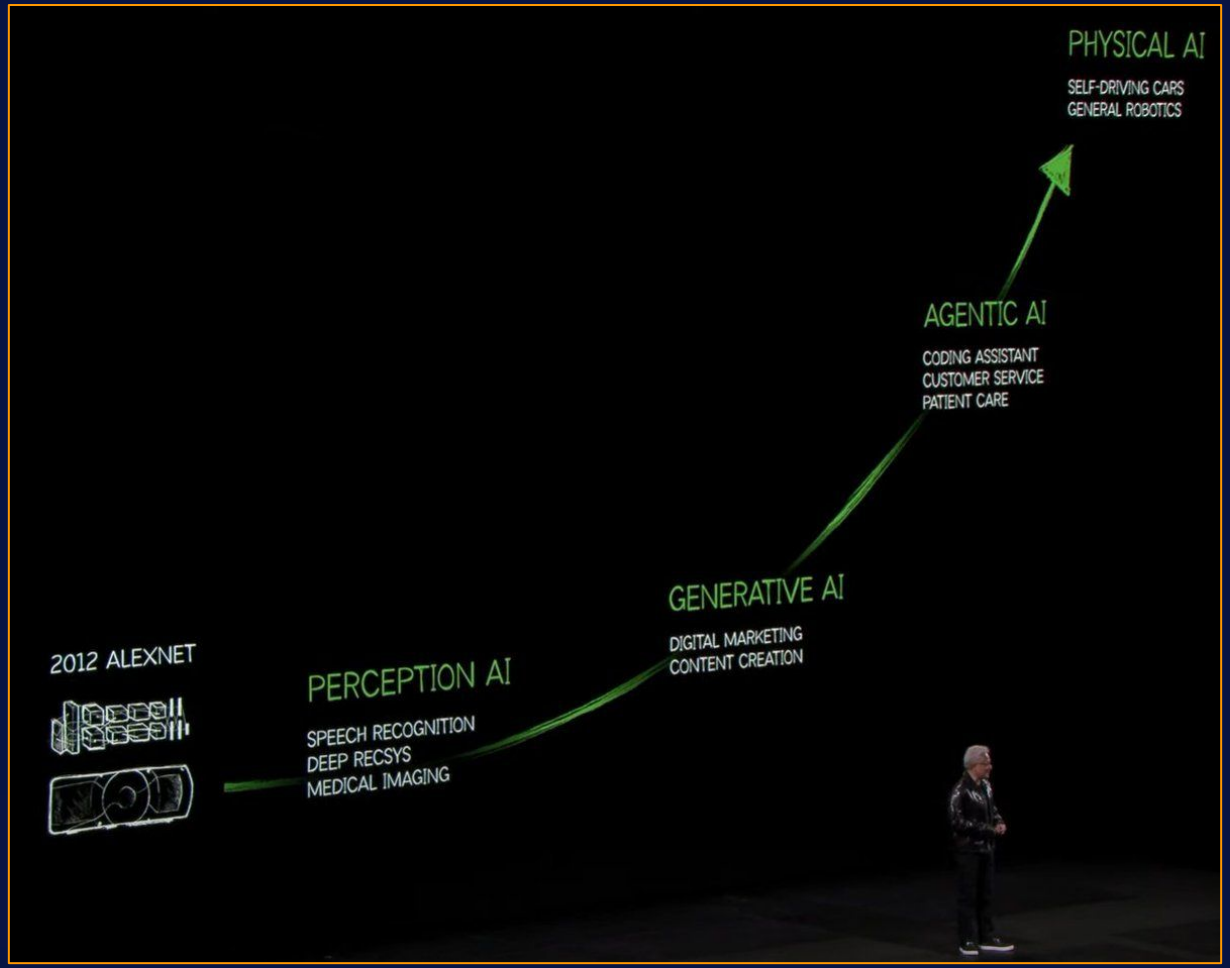
The decade ahead - Unhobble the LLMs

DEYAN7



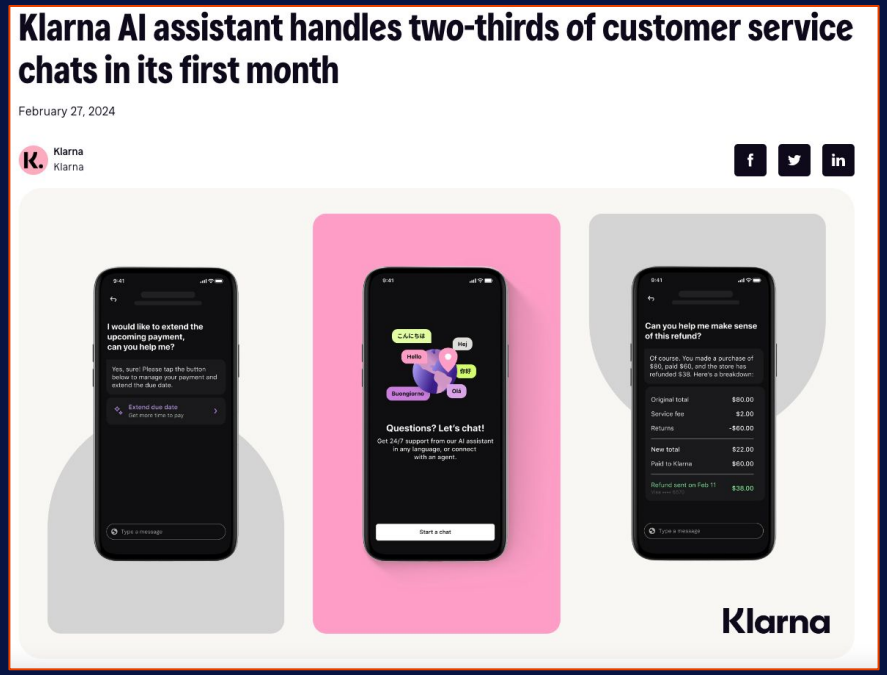
The decade ahead - Agentic AI on the horizon

DEYAN7



Klarna AI Assistant

- The AI assistant has had **2.3 million conversations, two-thirds of Klarna's customer service chats**
- It is doing the equivalent work of **700 full-time agents**
- It is on par with human agents in regard to customer satisfaction score
- It is more accurate in errand resolution, **leading to a 25% drop in repeat inquiries**
- Customers now **resolve their errands in less than 2 mins compared to 11 mins previously**
- It's available in **23 markets, 24/7** and communicates in more than **35 languages**
- It's estimated to drive a **$40 million USD in profit** improvement to Klarna in 2024



Zoom Virtual Agent handles 93% of inbound inquiries

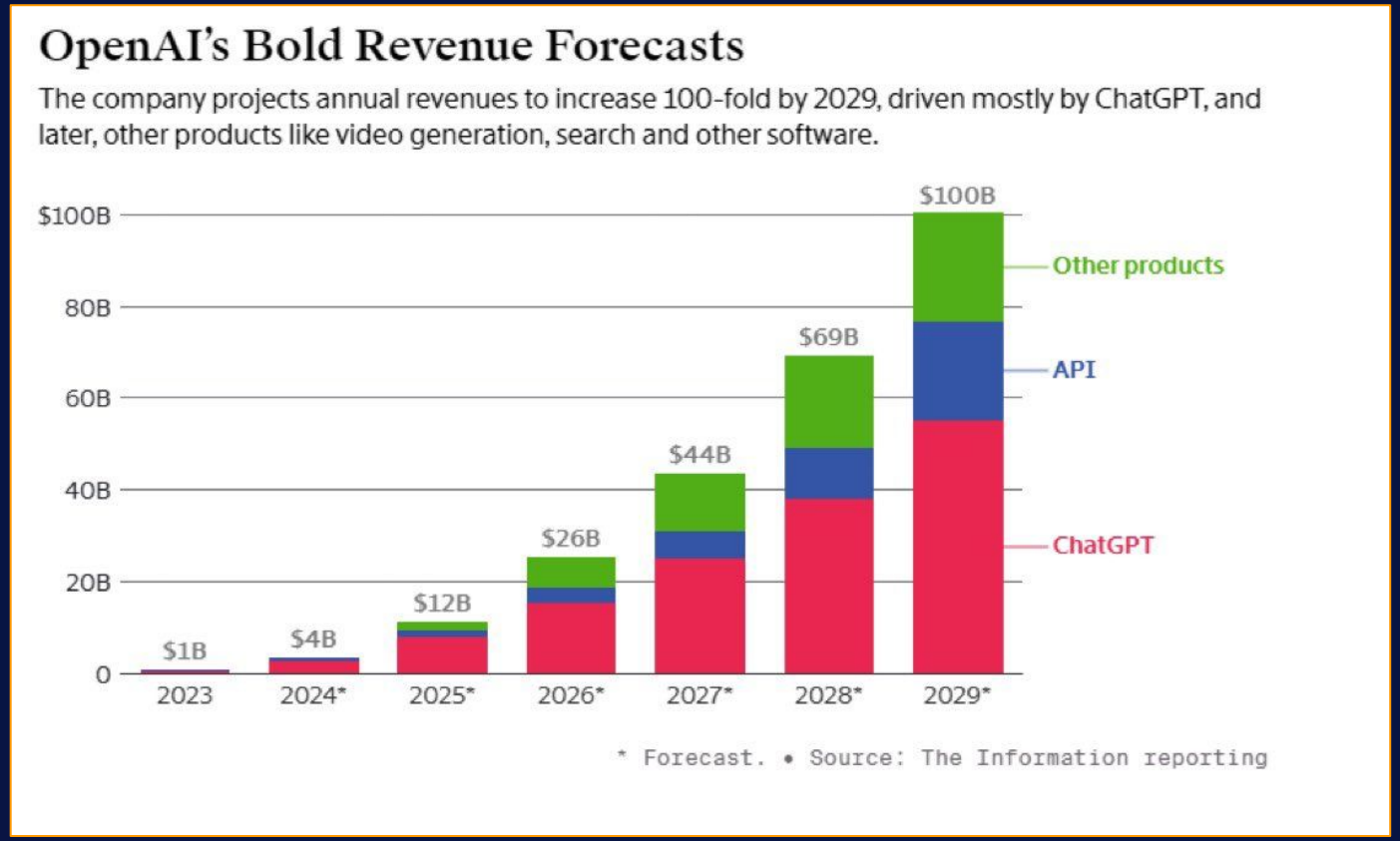
DEYAN7

Zoom Annual Report FY23

- **At the end of FY23,** we launched **Zoom Virtual Agent,** an intelligent conversational AI and chatbot solution for the Contact Center.
- After deploying ZVA on our own support site, **we found it could handle 93% of inbound inquiries and save us over 400,000 agent hours a month.**

ChatGPT as the main revenue driver

DEYAN7



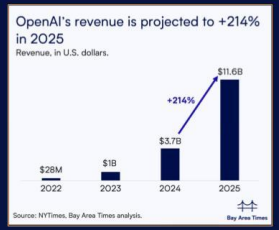
Recap

DEYAN7

The GenAI industry is growing fast led by OpenAI

DEYAN7

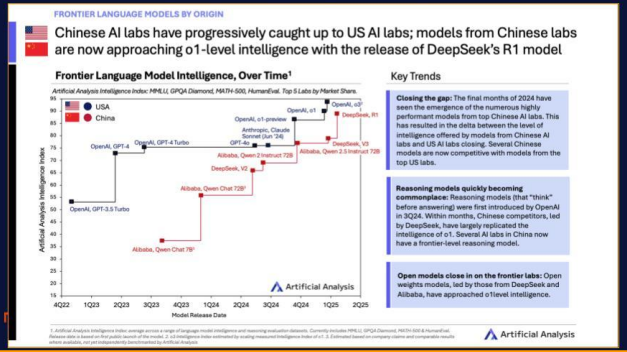
Financial Times, February 9 2024



Bay Area Times, September 20 2024

The AI Race - … but also about China vs. US

DEYAN7

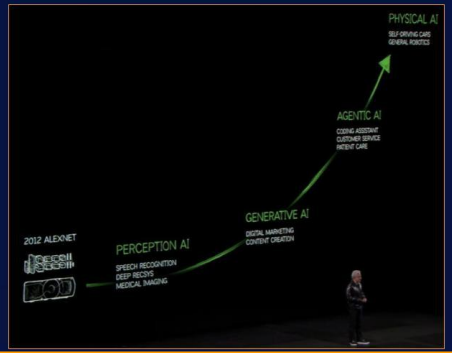


The decade ahead - US perspective

DEYAN7

The decade ahead - Agentic AI on the horizon

DEYAN7



DEYAN7

Welchen Mehrwert kann generative KI für mein Unternehmen bieten?

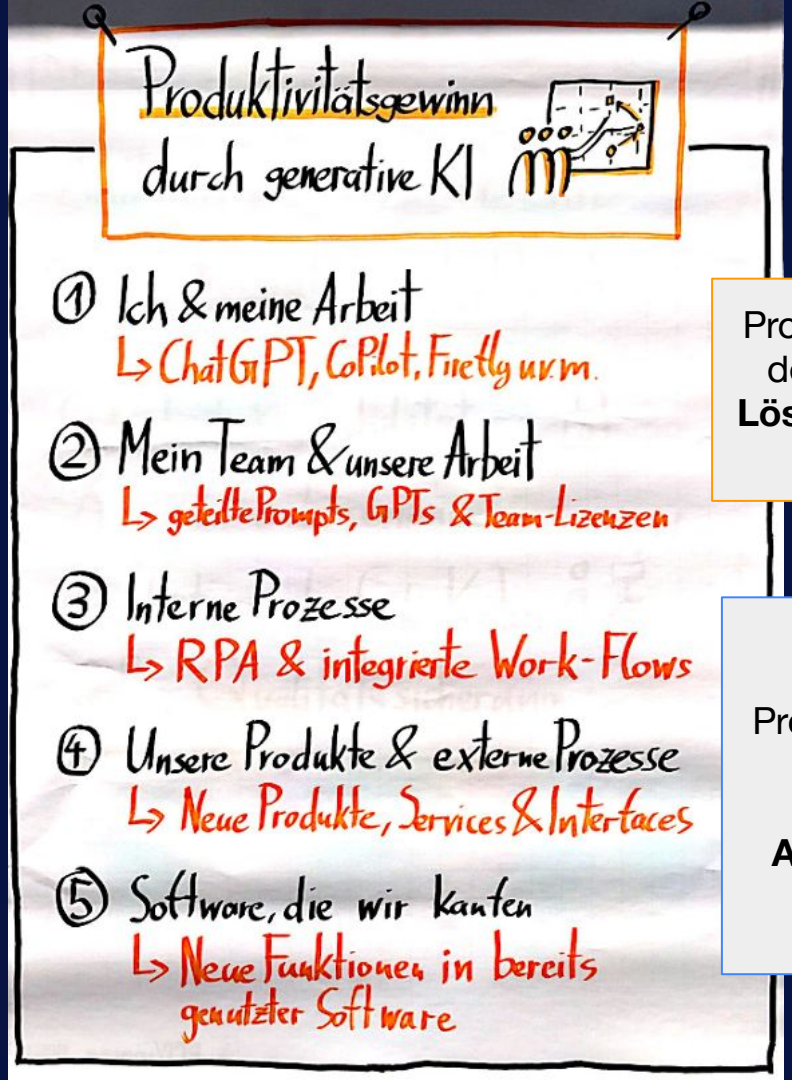
Von Stephan Barton

Produktivitätsgewinn durch generative KI

1. Ich & meine Arbeit
   - ↳ ChatGPT, CoPilot, Firefly u.v.m.
2. Mein Team & unsere Arbeit
   - ↳ geteilte Prompts, GPTs & Team-Lizenzen
3. Interne Prozesse
   - ↳ RPA & integrierte Work-Flows
4. Unsere Produkte & externe Prozesse
   - ↳ Neue Produkte, Services & Interfaces
5. Software, die wir kaufen
   - ↳ Neue Funktionen in bereits genutzter Software

DEYAN7

DEYAN7



Produktivitätsgewinn wird durch den Einkauf **bestehenden KI Lösungen** wie Chat GPT Teams, CoPilot, o.ä. gehoben

Produktivitätsgewinn wird durch die Entwicklung von spezialisierten **Agenten-Systemen** gehoben

Custom GPT Hackathon

In 10 min von der Idee zur KI Lösung

KI-Ready 2025
Jonas Gottschalk
20.02.2025

DEYAN7

ITO Modell - Write effective prompts

R	Role	Du bist Experte für Bier und Köln.
I	Information about the context of the task	Ich bin heute Abend in Köln und möchte in ein cooles Brauhaus gehen.
T	Task description	Mach mir vier Vorschläge
O	Output description	Liste die Vorschläge in einer Tabelle auf.

DEYAN7

KI-Agenten & agentische Workflows

Was brauche ich wirklich, um loszulegen?

KI-Ready 2025
Jonas Gottschalk & Marius Jeskulke
21.02.2025

DEYAN7

Auswahl des richtigen Use Case

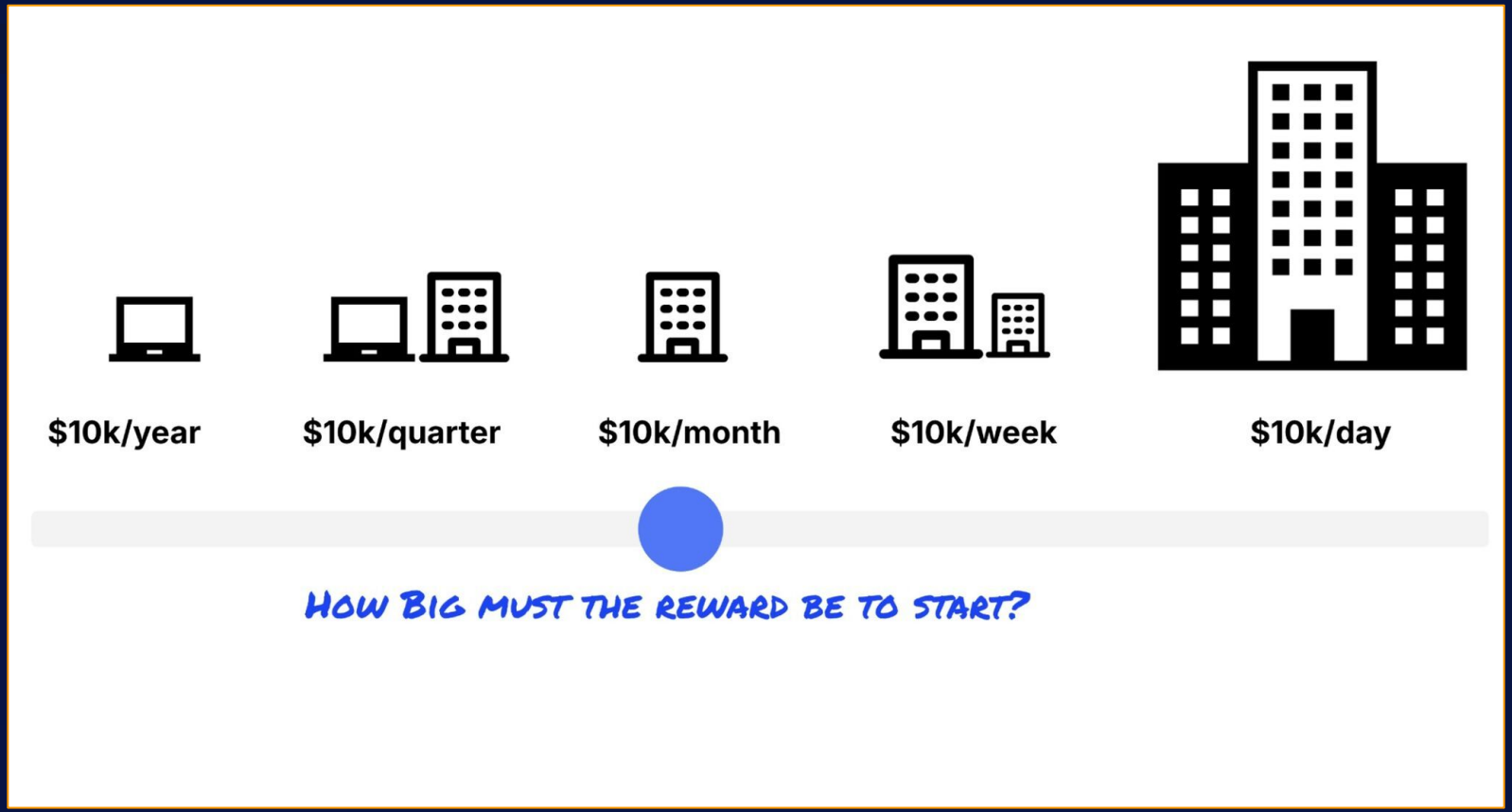


Bild von Tobias Zwingmann, Linkedin

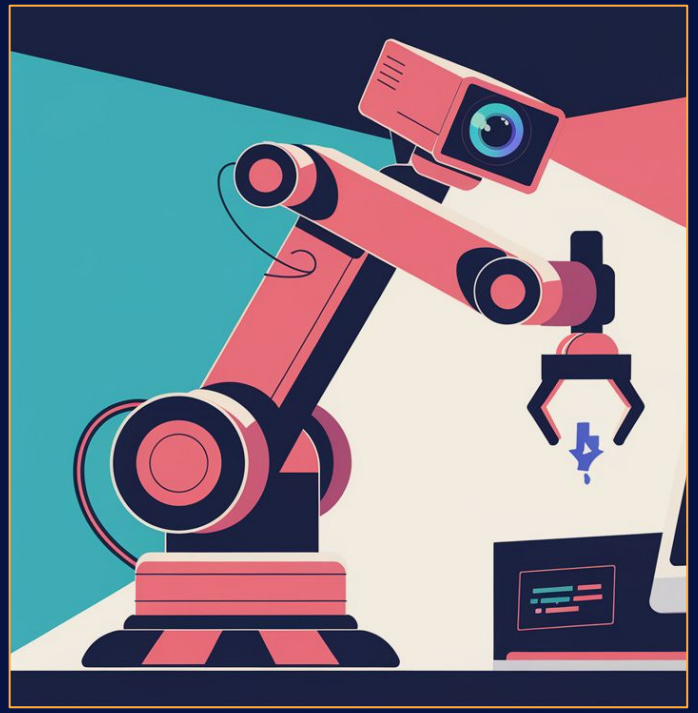
Auswahl des richtigen Use Case

DEYAN7

Häufig

Kompliziert

Mühselig



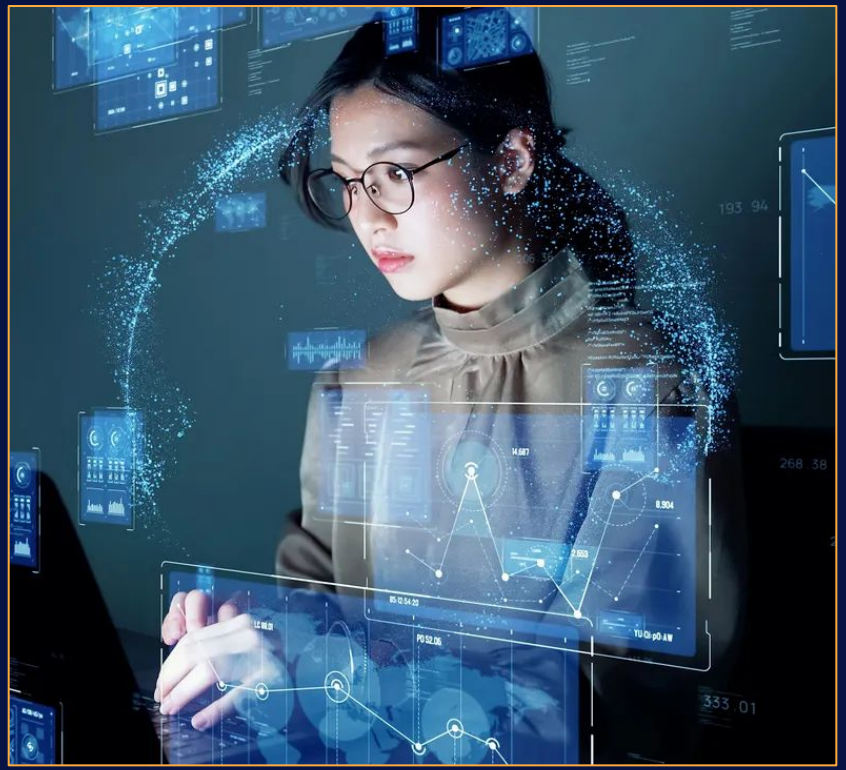
Die neue Rolle des Agent Operators

Verantwortlichkeiten:

- Fachliche Leitung des agentischen Workflows
- Anpassung und Optimierung von Prompts
- Auswahl und Bewertung sinnvoller Inputs (z. B. Dokumente, Fachwissen)
- Definition und Qualitätsbewertung des Outputs

Rolle im Prozess:

- Kontinuierliches Monitoring und „Grading“ der Ergebnisse
- Mitwirkung an der Weiterentwicklung der Features



Anatomie eines KI Agenten

DEYAN7

Integrationen

Wissensdatenbanken
Mediendatenbanken
Systems of Record (ERP, CRM)
APIs (schreiben/lesend)

Kernfunktionalität

Versionierter Prompt
RAG-System
Automatische Quellupates
Prozessablauf
Datenmodelle

Interface

Chat
Email/Ticketsystem
Artifact Mode
Realtime Audio
Siri/Gemini
Word-Plugin
NotebookLM

Anatomie eines KI Agenten

DEYAN7

Integrationen

Wissensdatenbanken
Mediendatenbanken
Systems of Record (ERP, CRM)
APIs (schreiben/lesend)

Je nach Anwendungsfall und IT Landschaft mehr Aufwand als der Agent selbst

Kernfunktionalität

Versionierter Prompt
RAG-System
Automatische Quellupates
Prozessablauf
Datenmodelle

Seit Sommer 2024 günstig und stabil zu entwickeln

Interface

Chat
Email/Ticketsystem
Artifact Mode
Realtime Audio
Siri/Gemini
Word-Plugin
NotebookLM

Extrem dynamisch aufgrund des hohen Kapitaleinsatzes

Was brauche ich um loszulegen?

DEYAN7

Use Case

Agent Operator

Integration