

Sprachmodelle im Wandel

Wie viel KI passt heutzutage unter den Schreibtisch?

Was Sprachmodelle leisten – und was ihr lokaler Betrieb braucht

Peter Grohmann · Universität Bayreuth
Entstanden im Kontext des Projekts Gate2HPC

Vortragsstand: September 2026

Bayerisches Staatsministerium für
Wissenschaft und Kunst



Kofinanziert von der
Europäischen Union



OBERFRANKEN
STIFTUNG



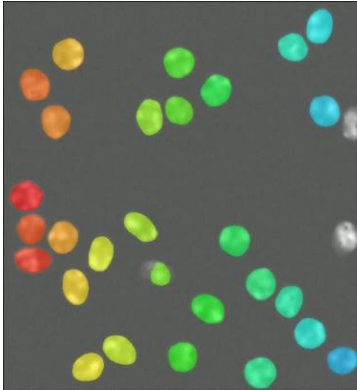
Lehrstuhl für
Konstruktionslehre und CAD
Prof. Dr.-Ing. Stephan Tremmel

CC BY 4.0 · Lizenzhinweise und Ausnahmen siehe Schlussfolie.

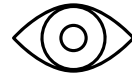
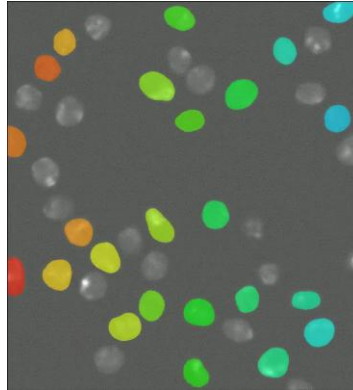
Sprachmodelle ermöglichen vielseitige Anwendungen

Wie wurde das möglich – und was davon geht lokal?

Startbild

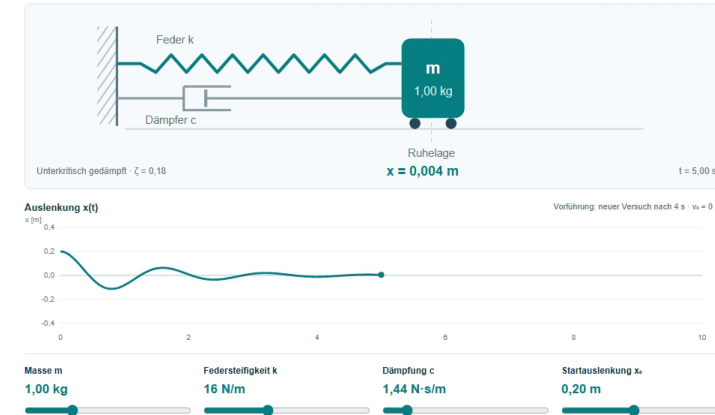


Verlauf · verlangsamt



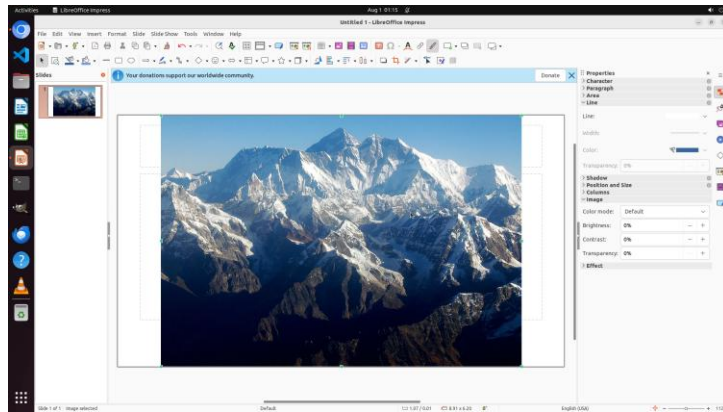
Bilder / Videos auswerten

VideoMolmo + SAM 2
Autorenvideo · 2025



Programme erstellen

Mit Coding-Assistenz
Eigenes Beispiel · 2026



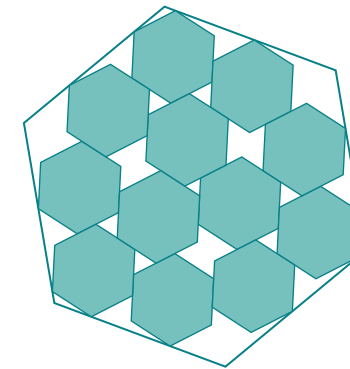
Anwendungen bedienen

OpenCUA
Demoauszug · 2025

12 gleiche Teile



Weniger Hüllfläche
–2,9 %
gegenüber damaligem
Bestwert



Lösungen suchen

AlphaEvolve · 2025
Ergebnisnachzeichnung



Der heutige Stand fußt auf vielen einzelnen Beiträgen

Ausgewählte Beiträge von 2003-2025

Modell und Training

Wie entstehen Modelle?

Fähigkeiten und Systeme

Was wird möglich?

Zugang und Ausführung

Was ermöglicht Eigenbetrieb?

2003–17	2018–21	2022	2023	2024	2025
Gelernte Darstellungen (2003) Attention (2014/15) Transformer (2017)	Scaling Laws (2020)	InstructGPT Nachtraining Chinchilla	LLaMA Mixtral · MoE	Phi-3	DeepSeek-R1 Distill-Modelle Qwen3
	RAG (2020) GPT-3 (2020) Few-Shot	ChatGPT	Werkzeugaufrufe (OpenAI-API)	Gemini 1.5 Kontext GPT-4o multimodal o1-preview Reasoning	OpenCUA Computer Use
	GPT-2 · Gewichte (2019)	FlashAttention BLOOM · Gewichte	Llama 2 · Gewichte llama.cpp SpQR Quantisierung		



Aus Text berechnet das Netz Wahrscheinlichkeiten

1 · Vorhersage mit festen Modellparametern

2003–17

2018–21

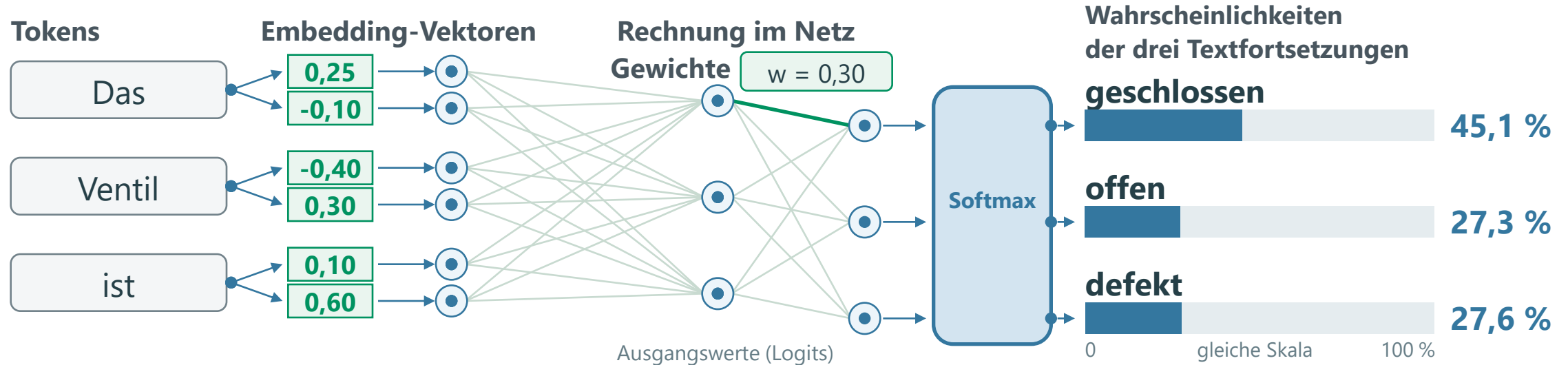
2022

2023

2024

2025–26

2003 · Gelernte Darstellungen



Grün: gespeicherte Parameter

Blau: berechnete Aktivierungen

Eigenes Lehrnetz für Vorhersage und Training. Keine Transformer-Architektur.

Training verändert die Modellparameter

2 · Vorhersage bewerten und Parameter anpassen

2003–17

2018–21

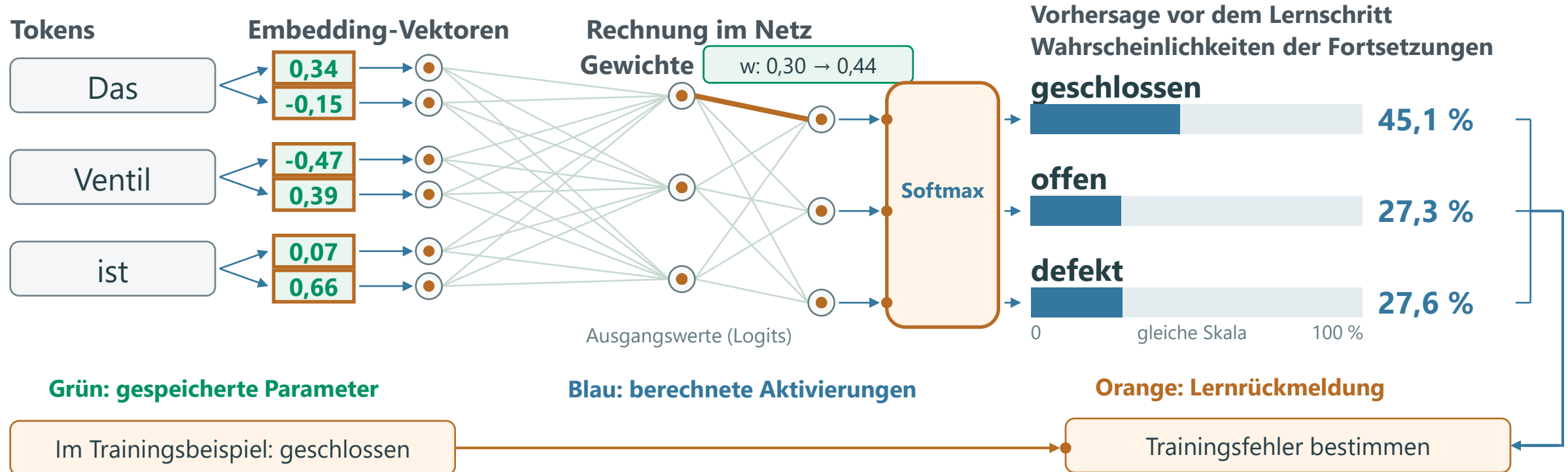
2022

2023

2024

2025–26

2003 · Gelernte Darstellungen



Optimierer: Parameter angepasst – jetzt erneut vorhersagen

Andere Parameter, andere Wahrscheinlichkeiten

3 · Erneute Vorhersage mit angepassten Parametern

2003–17

2018–21

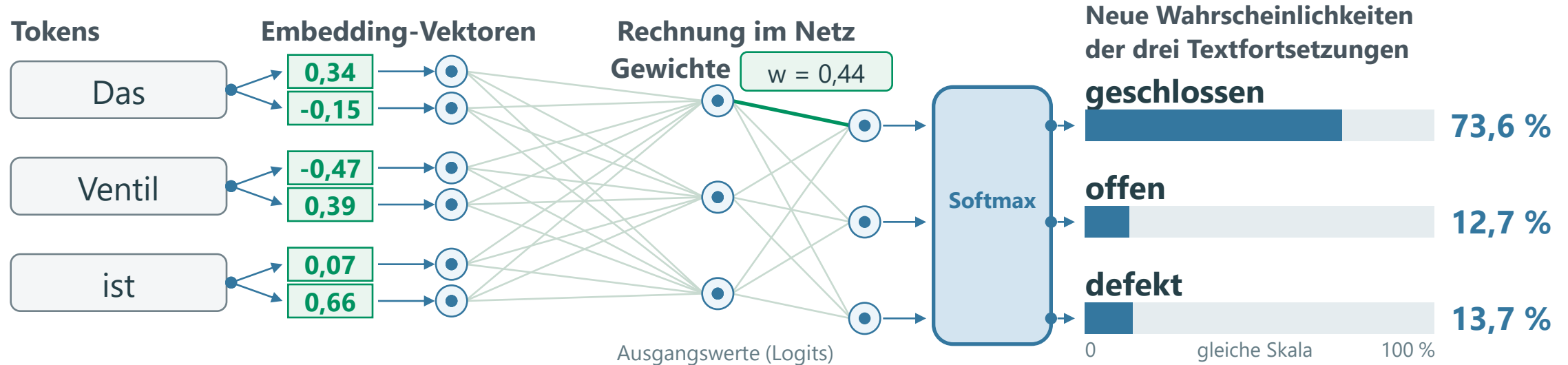
2022

2023

2024

2025–26

2003 · Gelernte Darstellungen



Grün: gespeicherte Parameter

Blau: berechnete Aktivierungen

Gleicher Trainingstext · „geschlossen“: 45,1 % → 73,6 %

Gewöhnliche Anwendung (Inferenz): feste Parameter, neue Rechenwerte.



Trainingsaufwand

Parameter- und Tokenanzahl bestimmen Rechenarbeit

2003–17

2018–21

2022

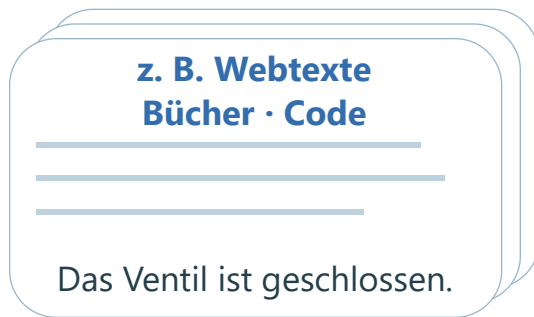
2023

2024

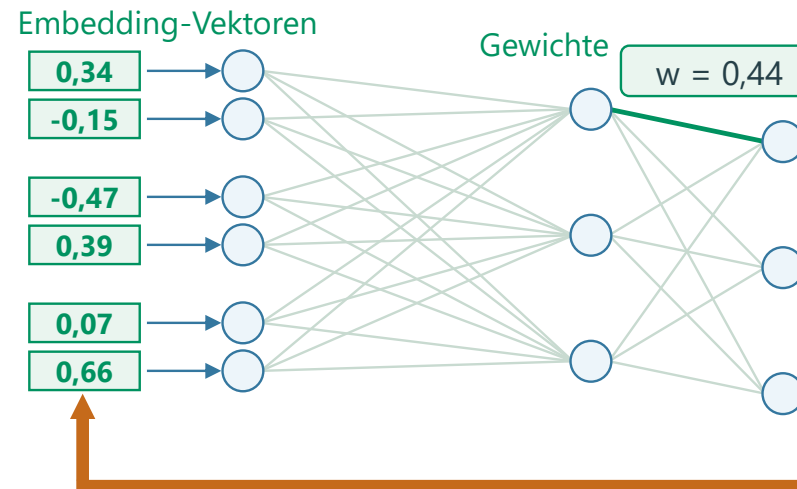
2025–26

2020/2022 · Trainingsaufwand und Skalierung

D · Verarbeitete Tokens



N · Parameterzahl



Wiederholte Lernschritte

Nächsten Token vorhersagen

Mit dem Trainingstext vergleichen

**Backpropagation: Gradienten
Optimierer: Parameterupdate**

$$C \approx 6 \cdot N \cdot D$$

C: Rechenarbeit in FLOPs

Grobe Näherung für große dichte Transformerarchitekturen.
Bei gleichem D: $2 \times N \rightarrow$ ungefähr $2 \times C$

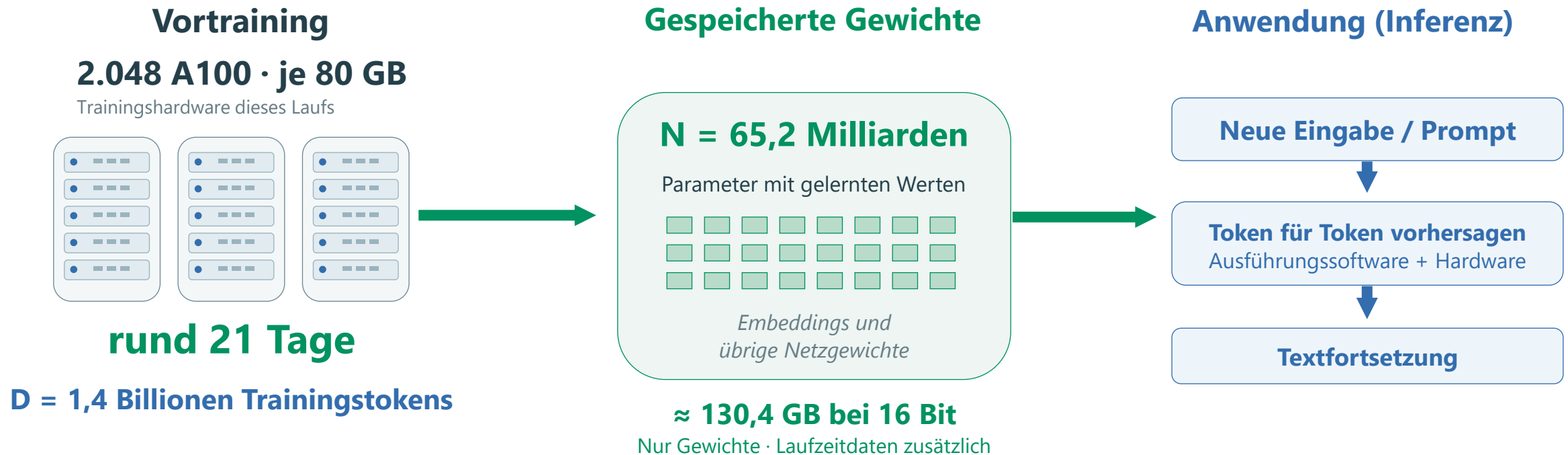


Trainierte Gewichte lassen sich für neue Eingaben nutzen

Neue Eingaben erfordern keinen neuen Vortrainingslauf

2003–17	2018–21	2022	2023	2024	2025–26
---------	---------	------	------	------	---------

2023 · LLaMA 65B: vom großen Trainingslauf zum verwendbaren Parameterstand



LLaMA 2023: Wie die Gewichte zugänglich wurden

Forschungszugang, unautorisierte Weitergabe und Llama 2

2003–17

2018–21

2022

2023

2024

2025–26

Ein Llama auf Abwegen

ZEIT · Eike Kühl · 06.03.2023

Historischer Presseausschnitt

Gespeicherte Gewichte

N = 65,2 Milliarden

Parameter mit gelernten Werten



*Embeddings und
übrige Netzgewichte*

24.02.2023 · Offizieller Forschungszugang

Auf Antrag: nichtkommerzielle Forschung

Fallweise genehmigt · [1]

März 2023 · Unautorisierte Weitergabe

Weiterverbreitete Gewichtsdateien · [2]

Keine neue Nutzungsfreigabe durch Meta.

18.07.2023 · Llama 2: neue Generation

Auch für kommerzielle Anwendungen

Unter eigenen Bedingungen · [3, 4]



Modellgröße begrenzt die lokale Nutzbarkeit

LLaMA ab 13B: 16-Bit-Gewichte zu groß für eine RTX 4090

2003–17

2018–21

2022

2023

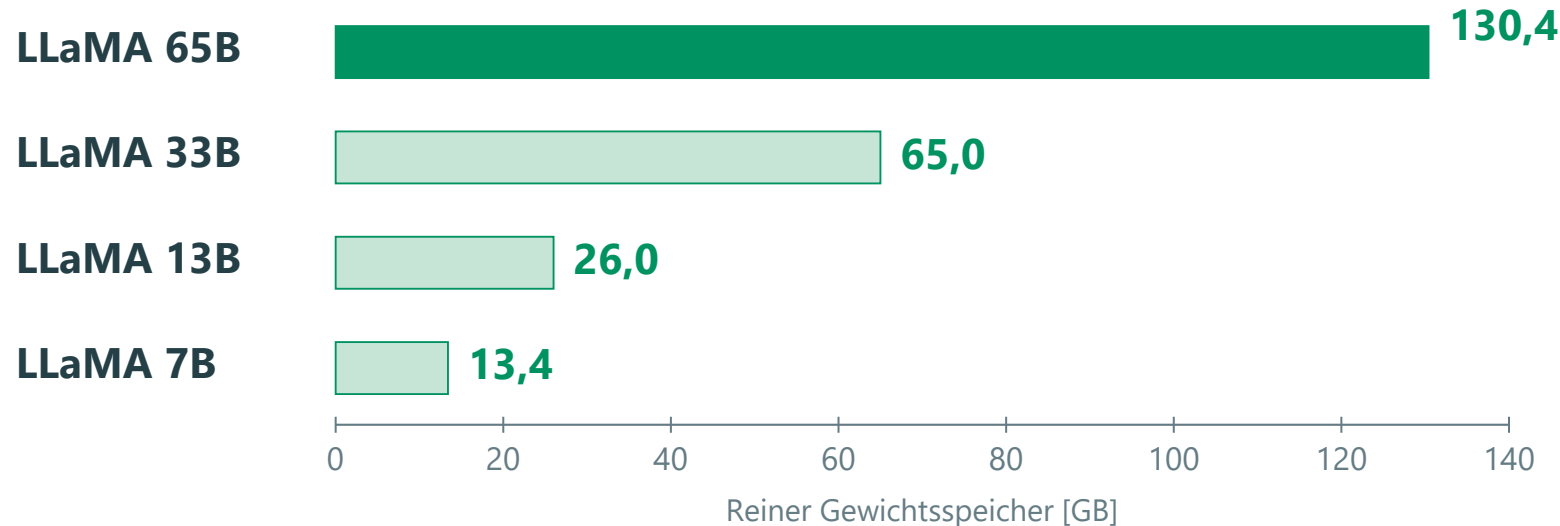
2024

2025–26

2023 · Erste LLaMA-Familie: Zugang allein genügt nicht

Gewichtsspeicher bei 16 Bit je Parameter

16 Bit = 2 Byte je Parameter · 1 GB = 10^9 Byte



Nur für die Gewichte · Laufzeitdaten zusätzlich



Lehrstuhl für
Konstruktionslehre und CAD
Prof. Dr.-Ing. Stephan Tremmel

GeForce RTX 4090

2022/23 · High-End-Grafikkarte



Bild: GIGABYTE · GAMING OC 24G

24 GB Grafikspeicher

Herstellerangabe: NVIDIA (2022)

NVIDIA-Startpreis: ab 1.949 €

Deutschland · 12.10.2022

Parameterzahlen: Touvron et al. (2023), Tab. 2

Speicher: eigene Rechnung bei 16 Bit

Quantisierung spart Speicherplatz

Gewichte größer speichern, empfindliche Werte genauer behalten

2003–17	2018–21	2022	2023	2024	2025–26
---------	---------	------	------	------	---------

SpQR (2023) · eigenes vereinfachtes Lehrbeispiel

Gelerntes Gewicht

$w \approx 0,44$

Bisher: 16 Bit je Gewicht

Gewähltes Raster: 16 Stufen

Abstand zwischen den Stufen: 0,1



4 Bit für eine von 16 Stufen

Rasterdaten kommen hinzu.

Eingabewerte

$1 \cdot x_1$

$10 \cdot x_2$

$y = 1 \cdot w_1 + 10 \cdot w_2$

Varianten	w_1	w_2	Ergebnis y	Abweichung zu 4,84
Ausgangswerte	0,44	0,44	4,84	—
Beide gerundet	0,4	0,4	4,40	0,44
w_2 genauer (mehr Speicher)	0,4	0,44	4,80	0,04



SpQR: geringerer Speicherbedarf durch Quantisierung

Ähnliche Ergebnisse in drei ausgewählten Tests

2003–17	2018–21	2022	2023	2024	2025–26
---------	---------	------	------	------	---------

2023 · Quantisierung am Beispiel von LLaMA-13B

Wie viel Gewichtsspeicher bleibt?

Eigene Näherung · nur Modellparameter

LLaMA-13B · rund 13 Milliarden Parameter
Laufzeitdaten kommen hinzu.



GB dezimal · RTX 4090: nominell 24 GB (NVIDIA, 2022) [5]

Was zeigen diese drei Tests?

Autorenmessung · richtige Auswahlantworten [%]
Ohne gelöste Beispiele im Testprompt

	16 Bit	SpQR
PIQA Physikalisches Alltagswissen	78,89	78,94
HellaSwag Plausible Fortsetzungen	59,11	59,02
ARC-challenge Naturwissenschaftliche Fragen	43,94	43,17

Größter Rückgang hier: 0,77 Prozentpunkte

Ergebnis: Weniger Gewichtsspeicher, ähnliche Werte in diesen Tests.



Modelle von 2023 im gemeinsamen Benchmarkindex

Indexpunkte aus verschiedenen Tests · keine Trefferprozent

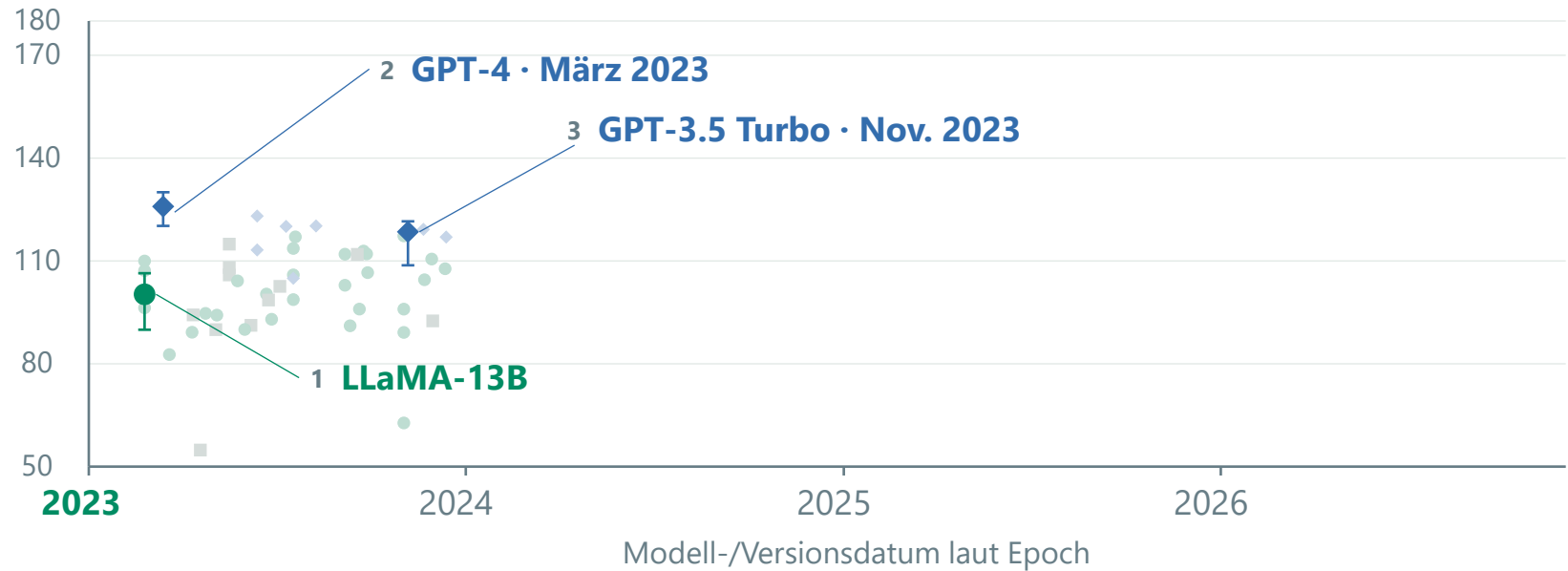
2003–17	2018–21	2022	2023	2024	2025–26
---------	---------	------	------	------	---------

Rückblick · Einheitlicher Datenstand vom 10.09.2026

ECI bündelt Benchmarkwerte · höhere Punkte = stärkere Testergebnisse

● Gewichte verfügbar ◆ Geschlossene Gewichte ■ Andere / unklar

Indexpunkte · Striche: 90%-Unsicherheitsintervalle



Anbieter · belegte Funktionen

1	LLaMA-13B Meta AI · USA	dicht
	☰ Text	
2	GPT-4 · März-API 2023 OpenAI · USA	
	☰ Text	
3	GPT-3.5 Turbo · Nov. 2023 OpenAI · USA	
	☰ Text 🔧 Funktionsaufrufe	

2023: Mixtral liegt im Index nahe bei GPT-3.5 Turbo

Ähnlicher Indexwert heißt nicht gleiche Leistung je Aufgabe

2003–17

2018–21

2022

2023

2024

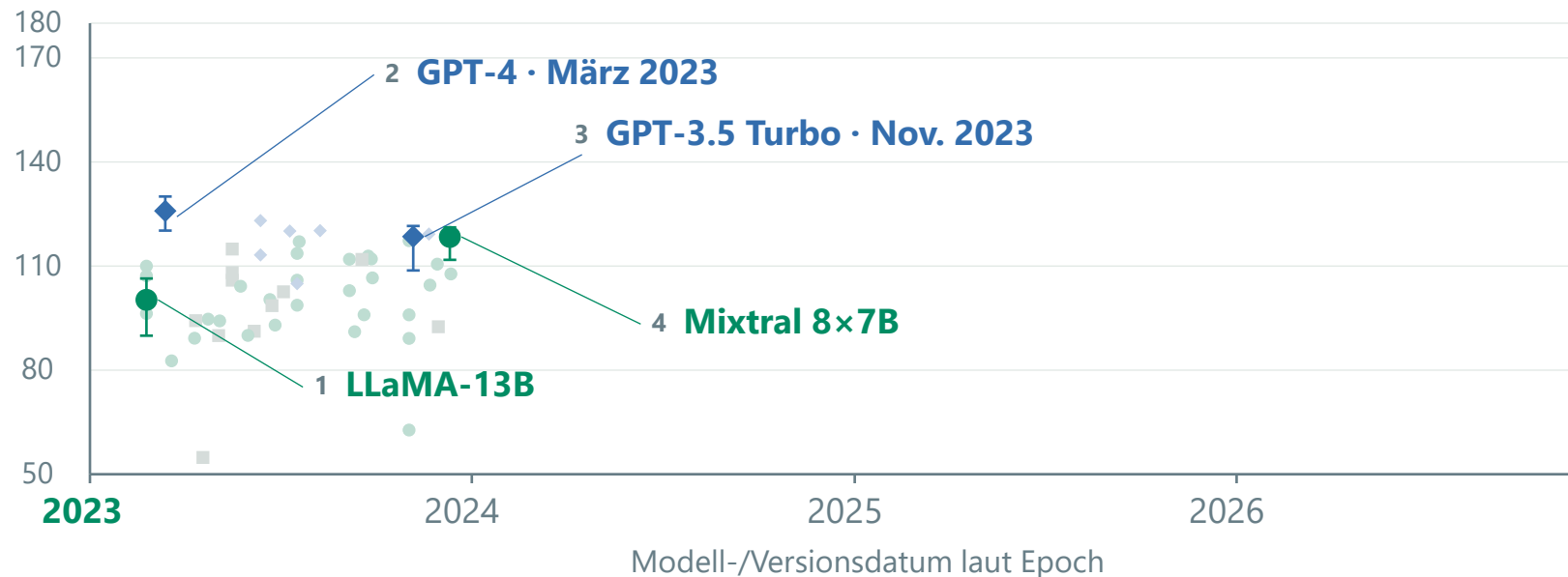
2025–26

Rückblick · Einheitlicher Datenstand vom 10.09.2026

ECI bündelt Benchmarkwerte · höhere Punkte = stärkere Testergebnisse

● Gewichte verfügbar ◆ Geschlossene Gewichte ■ Andere / unklar

Indexpunkte · Striche: 90%-Unsicherheitsintervalle



Anbieter · belegte Funktionen

1	LLaMA-13B Meta AI · USA	dicht
	☰ Text	
2	GPT-4 · März-API 2023 OpenAI · USA	
	☰ Text	
3	GPT-3.5 Turbo · Nov. 2023 OpenAI · USA	
	☰ Text 🔧 Funktionsaufrufe	
4	Mixtral 8x7B Mistral AI · Frankreich	MoE
	☰ Text	



2023: Mixtral rechnet mit ausgewählten Teilnetzen

Mixture of Experts

2003–17

2018–21

2022

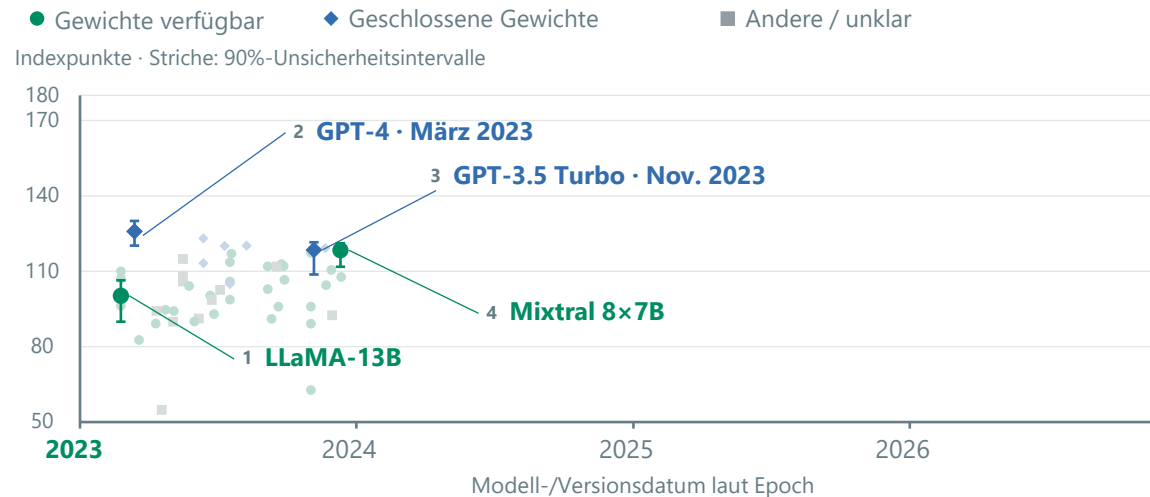
2023

2024

2025–26

Mixture of Experts · Auswahl je Token und Schicht

Vorheriger Vergleich · ECI



**Der Indexvergleich erklärt nicht,
welchen Anteil MoE am Ergebnis hat.**



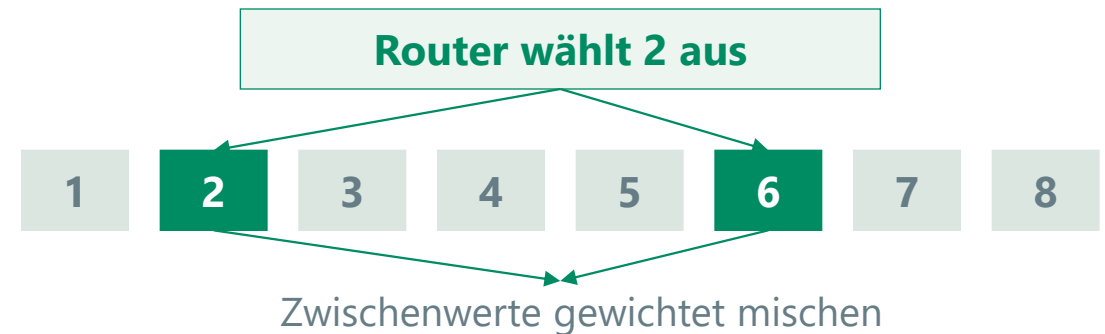
Lehrstuhl für
Konstruktionslehre und CAD
Prof. Dr.-Ing. Stephan Tremmel

Mixtral 8×7B

Mistral AI · Dezember 2023

Eine MoE-Schicht · 8 Expertenteilnetze

Vektor des aktuellen Tokens



Modellparameter

≈ 46,7 Mrd. gesamt · ≈ 12,9 Mrd. aktiv je Token

Gemeinsame Teile rechnen mit. Gesamter Gewichtsbestand bleibt nötig.

[Epoch AI / AI Benchmarking Hub](#) · [Epoch Capabilities Index \(ECI\)](#) · 10.09.2026 · CC BY 4.0 · eigene Grafik
[MoE: Jiang et al. \(2024\), Abb. 1 / §2.1](#) · [Parameter: Mistral AI \(11.12.2023\)](#) · eigenes Schema

Modelle von 2024 im gemeinsamen Benchmarkindex

Bildeingaben und zusätzlicher Lösungsaufwand

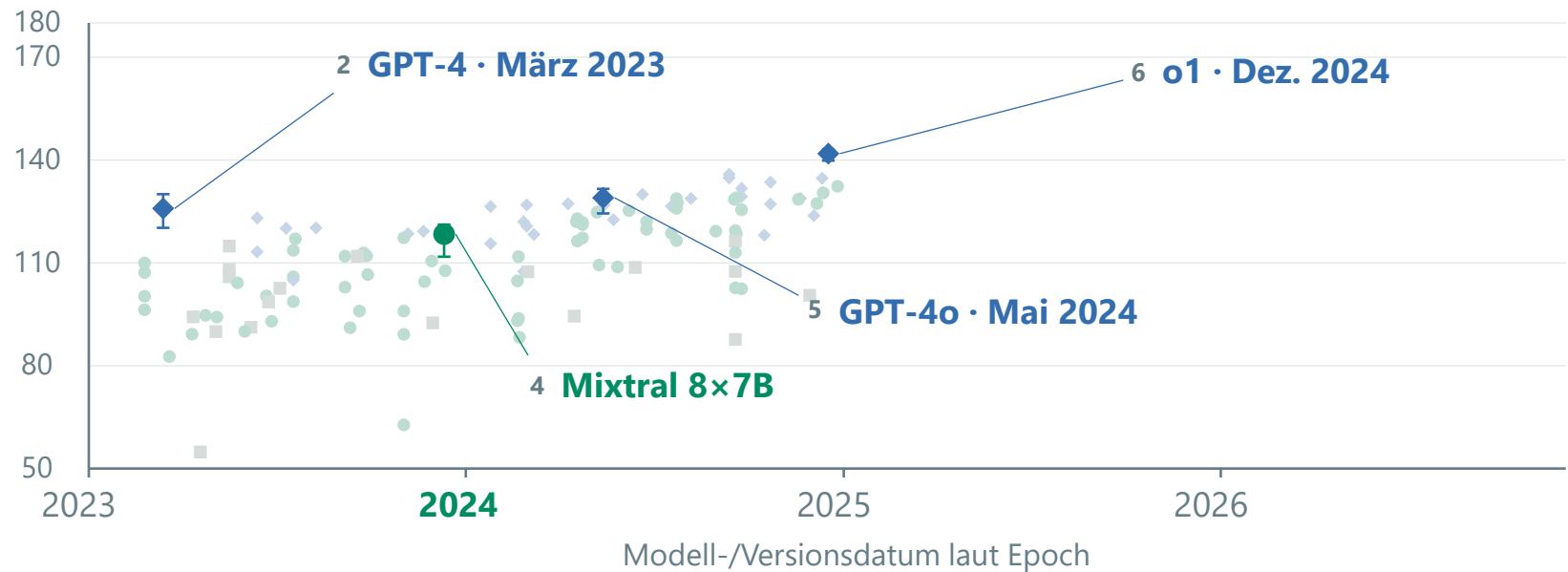
2003–17	2018–21	2022	2023	2024	2025–26
---------	---------	------	------	------	---------

Rückblick · Einheitlicher Datenstand vom 10.09.2026

ECI bündelt Benchmarkwerte · höhere Punkte = stärkere Testergebnisse

● Gewichte verfügbar ◆ Geschlossene Gewichte ■ Andere / unklar

Indexpunkte · Striche: 90%-Unsicherheitsintervalle



Eingaben und Funktionen

2	GPT-4 · März-API 2023 OpenAI · USA Text
4	Mixtral 8×7B MoE Mistral AI · Frankreich Text
5	GPT-4o · Mai 2024 OpenAI · API 13.05.2024 Text Bild Funktionsaufrufe
6	o1 · Dez. 2024 OpenAI · API 17.12.2024 Text Bild Funktionsaufrufe Reasoning o1 im ECI: mehrere Aufwandseinstellungen

2024: Zusätzlicher Aufwand vor der Antwort

Kann helfen – richtige Antworten sind nicht garantiert

2003–17

2018–21

2022

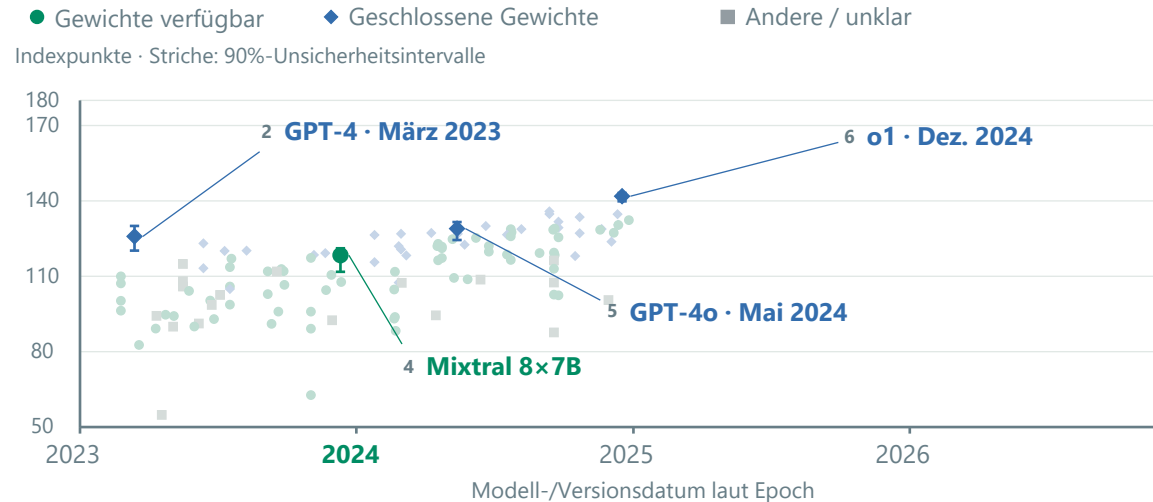
2023

2024

2025–26

Reasoning · zusätzliche Tokens bei festen Modellparametern

Vorheriger Vergleich · ECI



Der Indexvergleich isoliert nicht die Wirkung zusätzlicher Rechenzeit.

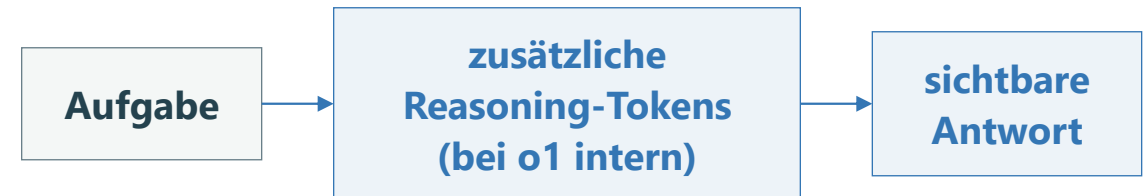


Lehrstuhl für
Konstruktionslehre und CAD
Prof. Dr.-Ing. Stephan Tremmel

o1 · Dez. 2024

OpenAI · API 17.12.2024

Reasoning vor der Antwort



Zusätzliches Reasoning kann bei komplexeren Aufgaben helfen.

[Epoch AI / AI Benchmarking Hub](#) · [Epoch Capabilities Index \(ECI\)](#) · 10.09.2026 · CC BY 4.0 · eigene Grafik

[Mechanismus: OpenAI, 12.09.2024](#) · eigenes Schema

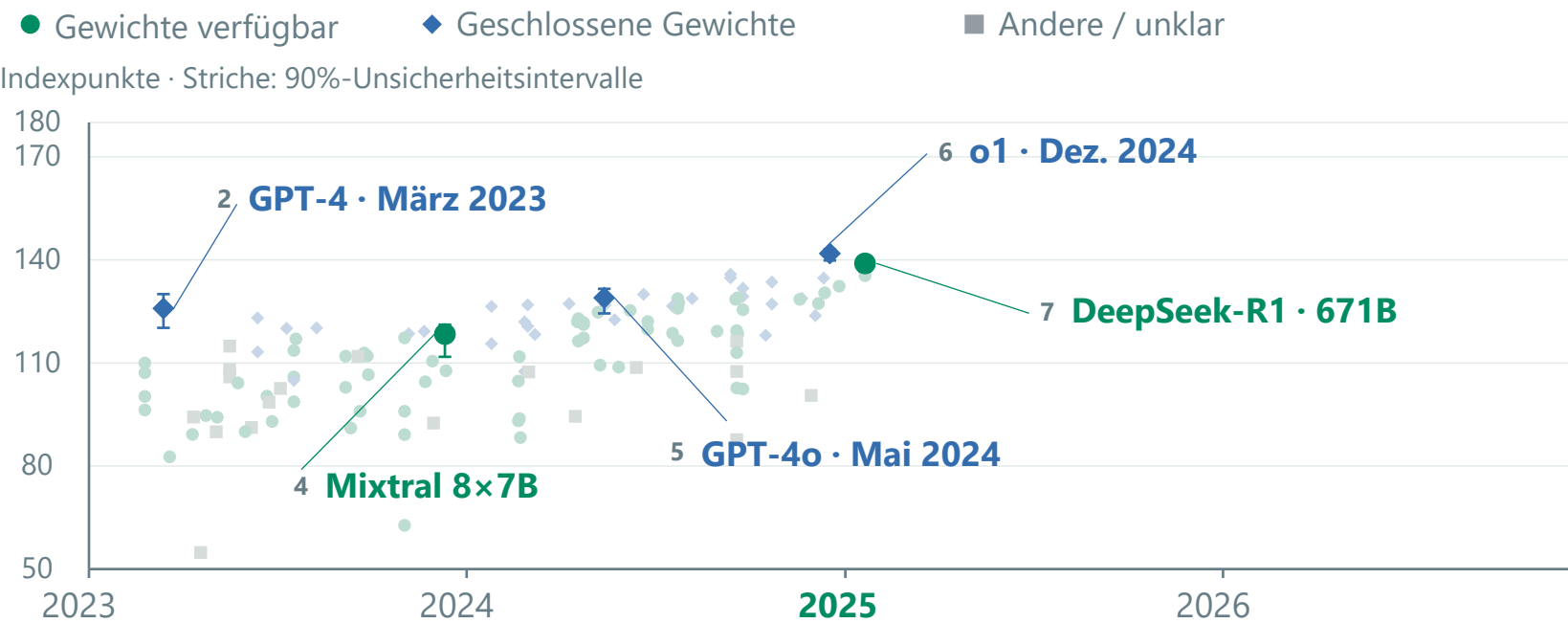
DeepSeek-R1: Gewichte verfügbar, im Index nahe an o1

Indexnähe heißt nicht gleiche Leistung oder gleichen Aufwand

2003–17	2018–21	2022	2023	2024	2025–26
---------	---------	------	------	------	---------

Rückblick · Einheitlicher Datenstand vom 10.09.2026

ECI bündelt Benchmarkwerte · höhere Punkte = stärkere Testergebnisse



Der gesamte Gewichtsbestand muss verfügbar bleiben.

Eingaben und Funktionen laut Dokumentation

2	GPT-4 · März-API 2023	
	Text	OpenAI
4	Mixtral 8×7B	MoE
	Text	Mistral AI
5	GPT-4o · Mai 2024	
	OpenAI · API 13.05.2024	
	Text	Bild
	Funktionsaufrufe	
6	o1 · Dez. 2024	
	OpenAI · API 17.12.2024	
	Text	Bild
	Funktionsaufrufe	Reasoning
7	DeepSeek-R1	MoE
	DeepSeek · 20.01.2025	
	Text	Reasoning
	671 Mrd. Parameter insgesamt	
	37 Mrd. aktiv je Token	



R1-Distill: Kleinere Sprachmodelle lernen von größeren

Nachtraining statt Reduktion der R1-Gewichte

2003–17

2018–21

2022

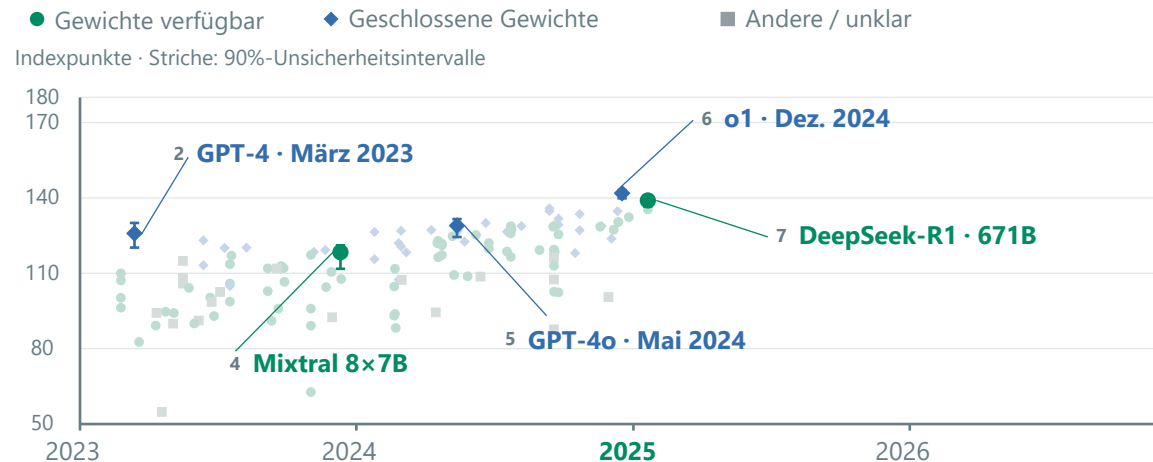
2023

2024

2025–26

2025 · Nachtraining vortrainierter Qwen-/Llama-Modelle

Vorheriger Vergleich · ECI

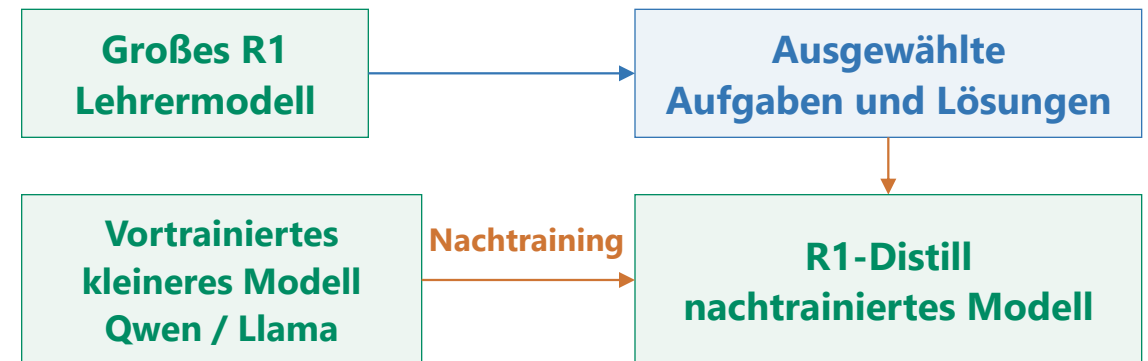


Im Index: großes R1 mit 671 Mrd. Parametern.
Kein Leistungsnachweis für alle Distill-Modelle.



R1-Distill · Jan. 2025

Modelle von 1,5 bis 70 Mrd. Parametern



Distillation ≠ Quantisierung

2026: Aktueller Stand

Verfügbare Gewichte

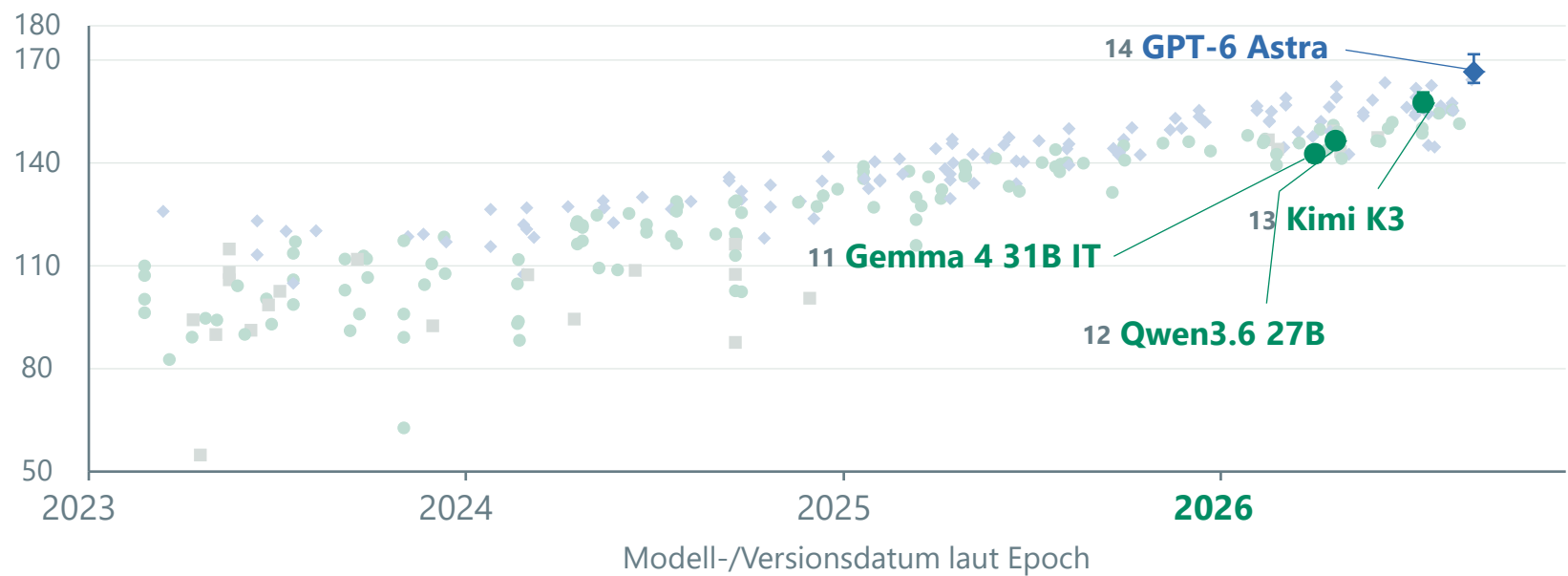
- 2003–17
- 2018–21
- 2022
- 2023
- 2024
- 2025–26

Vergleich bis September 2026 · Datenstand 10.09.2026

ECI bündelt Benchmarkwerte · höhere Punkte = stärkere Testergebnisse

● Gewichte verfügbar ◆ Geschlossene Gewichte ■ Andere / unklar

Indexpunkte · Striche: 90%-Unsicherheitsintervalle



Eingaben und Funktionen
laut Dokumentation

11 **Gemma 4 31B IT** dicht

Text

Bild

Funktionsaufrufe

Video¹

Reasoning

12 **Qwen3.6 27B** dicht

Text

Bild

Funktionsaufrufe

Video¹

Reasoning

13 **Kimi K3*** MoE

Text

Bild

Funktionsaufrufe

Video¹

Reasoning

14 **GPT-6 Astra***

Text

Bild

Funktionsaufrufe

Reasoning

* Profile: Dokumentation 09/2026;
nicht allen ECI-Varianten zugeordnet.
¹ Videoeingabe, keine Videoausgabe

2026: Aktueller Stand

Lauffähig auf einer RTX 5090

2003–17	2018–21	2022	2023	2024	2025–26
---------	---------	------	------	------	---------

September 2026 · Paketgröße ≠ Laufbedarf · passende Quantisierung und Software

ECI-Reihenfolge: ausgewählte 32-GB-Kandidaten

Gruppenwerte · 90%-Intervalle · kein lokaler Q4-Test

		ECI	Δ API
GPT-6 Astra · API		166,57	—
Qwen3.6-27B		146,51	–20,06
Qwen3.6-35B-A3B		143,87	–22,70
Gemma 4 31B IT		142,68	–23,89
Gemma 4 26B-A4B		141,87	–24,70

130 140 150 160 170

Δ API: Abstand in Indexpunkten, keine Prozentleistung.

Index: Epoch AI · Stand 10.09.2026 · CC BY 4.0 · eigene Auswahl

² Eigene Nutzlastrechnung · Profile, Pakete und Bild: Quellen F23



Lehrstuhl für
Konstruktionslehre und CAD
Prof. Dr.-Ing. Stephan Tremmel

Womit ich aktuell starten würde

Bild / Dokumente / Coding · GUI mit Zusatzsoftware

Qwen3.8-27B

Paket: BF16 55,6 → Q4 17,7 GB · kein ECI im Datenstand

Bild & Allround

Gemma 4: 31B oder 26B-A4B

Q4-Pakete: 19,9 / 18,6 GB · Textlaufbelege

Kimi K3: zu groß für diese Einzelkarte

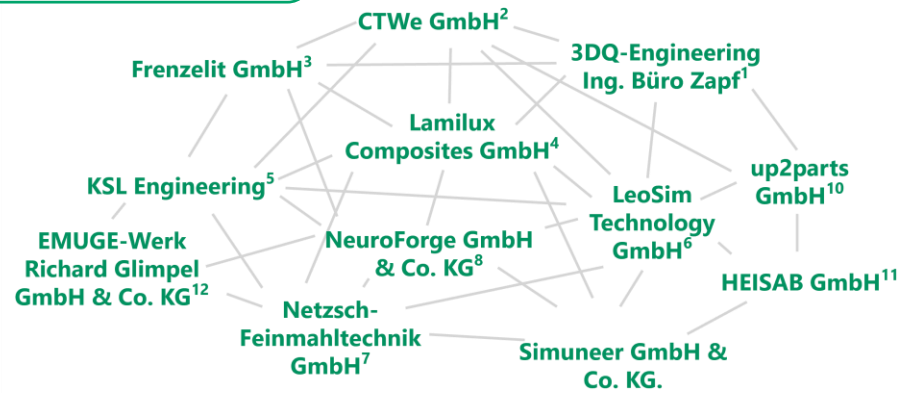
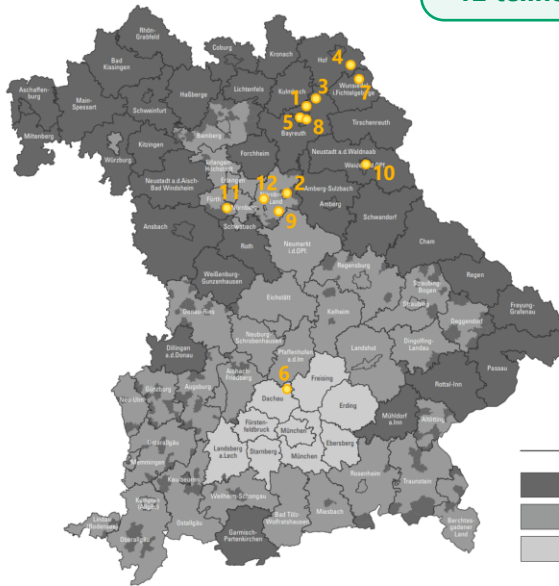
Rechnerisch bei 4 Bit: ≈ 1,4 TB nur Gewichte²



Ollama 0.33.3 · [RTX 5090 · 32 GB Grafikspeicher](#)

Projekt: Gate2HPC

12 teilnehmende Unternehmen



Landkreis / kreisfreie Stadt
EFRE-Schwerpunktgebiet (Raum mit besonderem Handlungsbedarf
gemäß LEP, Stand: 1. März 2018), mindestens 60 % der EFRE-Mittel
Sonstiges EFRE-Fördergebiet
EFRE-Fördergebiet nur für den Förderbereich 2 (Klima- und Umweltschutz)

Bayerisches Staatsministerium für
Wissenschaft und Kunst



Kofinanziert von der
Europäischen Union

OBERFRANKEN
STIFTUNG

Vielen Dank. Ihre Fragen?

Kontakt

Peter Grohmann, M.Sc.

Universität Bayreuth
Lehrstuhl für Konstruktionslehre und CAD

peter.grohmann@uni-bayreuth.de



UNIVERSITÄT
BAYREUTH



Lehrstuhl für
Konstruktionslehre und CAD
Prof. Dr.-Ing. Stephan Tremmel

<https://creativecommons.org/licenses/by/4.0/>



Diese Präsentation steht, soweit nicht anders gekennzeichnet, unter der Lizenz Creative Commons Namensnennung 4.0 International (CC BY 4.0). Für gesondert gekennzeichnete Abbildungen und sonstige Inhalte Dritter gelten die jeweils angegebenen oder geltenden Rechte- und Lizenzbedingungen

VideoMolmo und SAM 2

Ahmad et al. (2025)
VideoMolmo: Spatio-Temporal Grounding Meets Pointing.
arXiv:2506.05336v1

[Paper](#) [Autorenprojekt und Zellvideo](#)

Ravi et al. (2024) · SAM 2

Segment Anything in Images and Videos.
arXiv:2408.00714v2

[SAM-2-Paper](#)

OpenCUA

Wang et al. (2025)
OpenCUA: Open Foundations for Computer-Use Agents.
arXiv:2508.09123v1

[Paper](#) [Demo: MultiApp / Chrome & Slides](#)

AlphaEvolve · eigene Datenzeichnung

Novikov et al. (2025)
AlphaEvolve: A Coding Agent for Scientific and
Algorithmic Discovery. arXiv:2506.13131v1

[Paper](#) [Ergebnisnotebook · B.7 / Construction 2](#)

Notebookrevision: 4226acb
Daten: © 2025 Google LLC · CC BY 4.0.
Eigene Nachzeichnung und Flächenrechnung.

[CC BY 4.0](#)

[Lizenzhinweis im Repository](#)



[1] Meta · 24.02.2023

Introducing LLaMA: A foundational,
65-billion-parameter large language model.
ai.meta.com · Offizieller Forschungszugang.

[2] GitHub-DMCA-Archiv · März 2023

OpSec Online LLC im Namen von Meta.
Schreiben vom 20.03.; Archivierung am 21.03.2023.
2023-03-21-meta.md · Commit 528056c.
Beanstandung der Weitergabe; kein Gerichtsurteil.

[3] Meta · 18.07.2023

Meta and Microsoft Introduce
the Next Generation of Llama.
about.fb.com · Veröffentlichung von Llama 2.

[4] Llama 2 · Bedingungen vom 18.07.2023

Llama 2 Community License Agreement.
Meta-Llama-Repository · LICENSE.
Historischer Release-Commit 6d4c0c2.
Bedingungen dieser Generation.

[5] ZEIT ONLINE · Eike Kühl · 06.03.2023

Ein Llama auf Abwegen.
zeit.de/digital/internet/2023-03/
llama-ki-chatbot-meta-leak
Auf F09: historischer Schlagzeilenausschnitt.

[6] Touvron et al. · 2023

LLaMA: Open and Efficient Foundation
Language Models. arXiv:2302.13971v1.
Tabelle 2 / PDF-Seite 3: Parameterzahl.



[1] Dettmers et al. · 05.06.2023

SpQR: A Sparse-Quantized Representation for Near-Lossless LLM Weight Compression.
arXiv:2306.03078v1 · Tab. 7 / PDF S. 21.
LLaMA-13B: FP16 und SpQR mit Ø 4,63 Bit.
Quelle der Aufgabenwerte; keine RTX-4090-Messung.

[2] SpQR · historischer Autoren Quellcode

Vahe1994/SpQR · main.py · Commit 7c6a9837.
find_layers / get_average_number_of_bits /
llama_sequential: lineare Matrizen in den Blöcken,
Bitzähler mit Statistik- und Ausreißerangaben.
Die Ø-Bitzahl gilt nicht für sämtliche Parameter.

[3] LLaMA · historische Architektur

meta-llama/llama · llama/model.py · Commit 1076b9c5.
Attention / FeedForward / TransformerBlock / Transformer.
Matrizen, Normen, Embedding und Ausgabekopf.
Architekturbeleg für die eigene Parameterzählung.



[4] LLaMA-13B · fixierte Konfiguration

huggingllama/llama-13b · config.json.
Revision bf570454 · h = 5.120, f = 13.824,
40 Blöcke, V = 32.000; getrennte Embeddings.
Dimensionen der Rechnung, kein Hash-Nachweis
des von den Paperautoren geladenen Checkpoints.

[5] NVIDIA · 20.09.2022

Neu: GPUs der GeForce RTX 40-Serie.
RTX 4090: nominell 24 GB Grafikspeicher.
Herstellerangabe; keine Messung des freien
Speichers und keine lokale Laufbarkeitsprüfung.

Eigene Näherung · GB = 10^9 Byte

12.687.769.600 Matrixparameter mit Ø 4,63 Bit;
328.094.720 übrige Parameter weiterhin mit 16 Bit.
Summe: ca. 7,999 GB; durchgehend 16 Bit: 26,032 GB.
Jeweils Parameter $\times$ Bit / 8 / 10^9 .
Kein Datei- oder Laufbedarf; Laufzeitdaten zusätzlich.

[1] Epoch AI · Index und eigene Auswahl

Datenstand 10.09.2026; Gruppenwerte / 90%-Intervalle.
 Δ = Modell minus Astra (166,57); kein lokaler Q4-Test.
Methodik · [enthaltene Tests](#) · [CC BY 4.0](#).
Indexpunkte sind keine Prozentleistung.

[2] Qwen3.6 · zwei eigene Modellvarianten

[27B \(dicht\)](#) · [35B-A3B \(MoE\)](#): feste Originalkarten.
Paketmetadaten begründen die Vorauswahl.
Kein eigener RTX-5090-Lauf dieser beiden Pakete.
Kein ECI-Nachweis für ihre Q4-Ausführung.

[3] Qwen3.8-27B · aktueller Prüfkandidat

Autorenprofil: Bild/Video, Coding und GUI; Agentensoftware.
Im genutzten ECI-Stand 10.09.2026 noch ohne Wert.
[SGLang-Texttest](#): 5090, anderer NVFP4/FP8-Export.
Keine eigene Ollama-Q4-Messung.

[4] Gemma 4 · Varianten und begrenzte Läufe

31B dicht und 26B-A4B MoE sind verschiedene Varianten.
[NVIDIA: Q4 K M-Textläufe](#) (llama.cpp b7789).
Eigener Bildfall 09.09.2026: 26B, 1 Bild / 3 Wiederholungen.
Kein breiter Qualitäts- oder Peak-VRAM-Nachweis.

[5] Ollama · Paketgrößen aus Manifesten

Stand 13.09.2026 · GB = 10^9 Byte; Paket $\neq$ Laufbedarf.
Qwen3.8: 55,589 MTP-BF16 $\rightarrow$ 17,742 Q4_K_M (GB).
Gemma 31/26: 19,869 / 18,604 GB (je Q4_K_M).
Ollama [Qwen3.8-Tags](#) · [Gemma4-Tags](#).

[6] ² Kimi K3 · idealisierte Gewichtsnutzlast

Autorenangabe: 2,8 Billionen Parameter insgesamt.
Eigene ideale Rechnung für einheitlich vier Bit:
 $2,8 \times 10^{12} \times 4 / 8 = 1,4 \times 10^{12}$ Byte = 1,4 TB.
Ohne Zusatz-/Laufdaten; keine gemessene Datei.

[7] NVIDIA · RTX 5090 als Prüfkasse

Herstellerangabe: nominell 32 GB GDDR7.
Runtime, Kontext und Arbeitsdaten benötigen Speicher.
Keine garantierte freie Reserve oder Antwortzeit.
Astra ist API-Referenz, kein lokaler Kandidat.