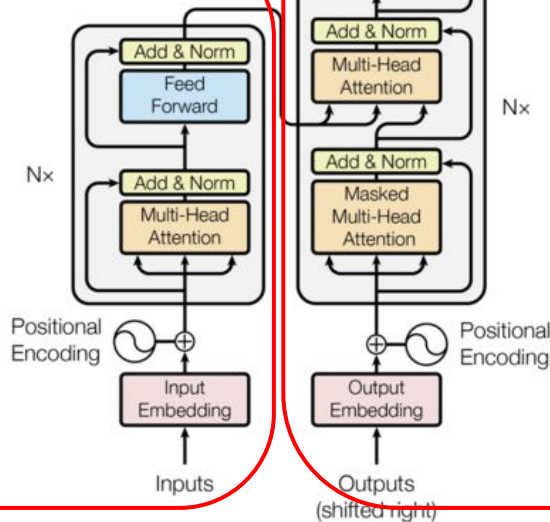


Wie offen
geht LLM?
Sprachmodelle
zwischen
Anspruch und
Realität

Worüber rede ich heute?

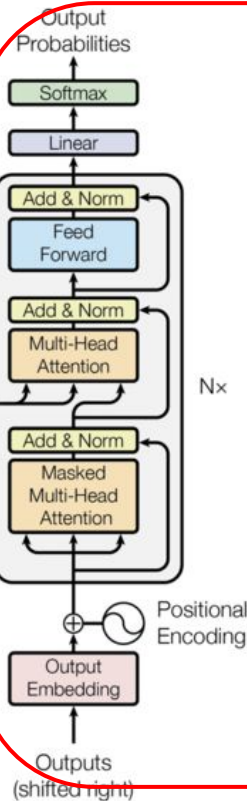
Encoder

- z.B. BERT
- relativ freie Lizenzen
- Verbreitung v.a. in Wissenschaft
- v.a. vor 11/2022



Decoder

- z.B. GPT
- hoch skaliert
- v.a. proprietäre Gouvernance
- Verbreitung auch über Wissenschaft hinaus
- v.a. nach 11/2022



Was ist das Problem bei Online-Schnittstellen zu LLMs im Kontext Forschung?

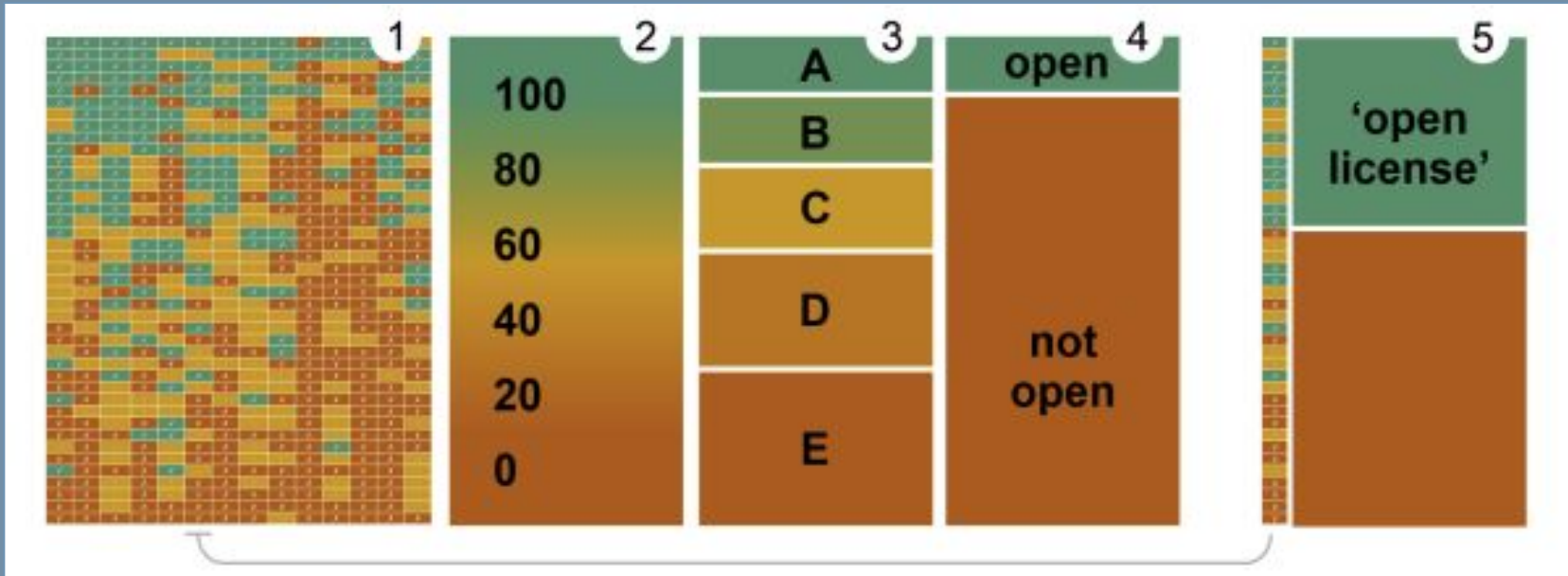
- Keine Kontrolle über den *Prozess* – Das Modell kann sich jederzeit ändern oder ich kann gar den Zugriff verlieren.
- Keine Kontrolle über den *Output* – Zwar bieten die meisten Frontier-LLMs Protokolle, aber diese sind evtl. nicht dauerhaft verfügbar.
- Keine Kontrolle über (*Test*)*Daten* – Alles, was ich als Input bereitstelle, wird ggfs. als Trainingsdatum genutzt und kontaminiert damit spätere Experimente (vgl. Balloccu et al. (2024): Leak, Cheat, Repeat <https://leak-llm.github.io/>)
- PLUS: Reihe von **Abhängigkeiten**

Konzepte der Offenheit (Liesenfeld/Dingemanse)

- **Verfügbarkeit:** Code - Trainingsdaten - Gewichte - RL(HF)-Daten - RL(HF)-Gewichte - Lizenz
- **Dokumentation:** Code - Architektur - Preprint - Paper - DataSheet
- **Zugriffsmethoden:** (Webservice) - Package - API

[Link zum Paper](#) (2024) - [Link zur Open Source AI Initiative](#)

Konzepte der Offenheit (Liesenfeld/Dingemanse)



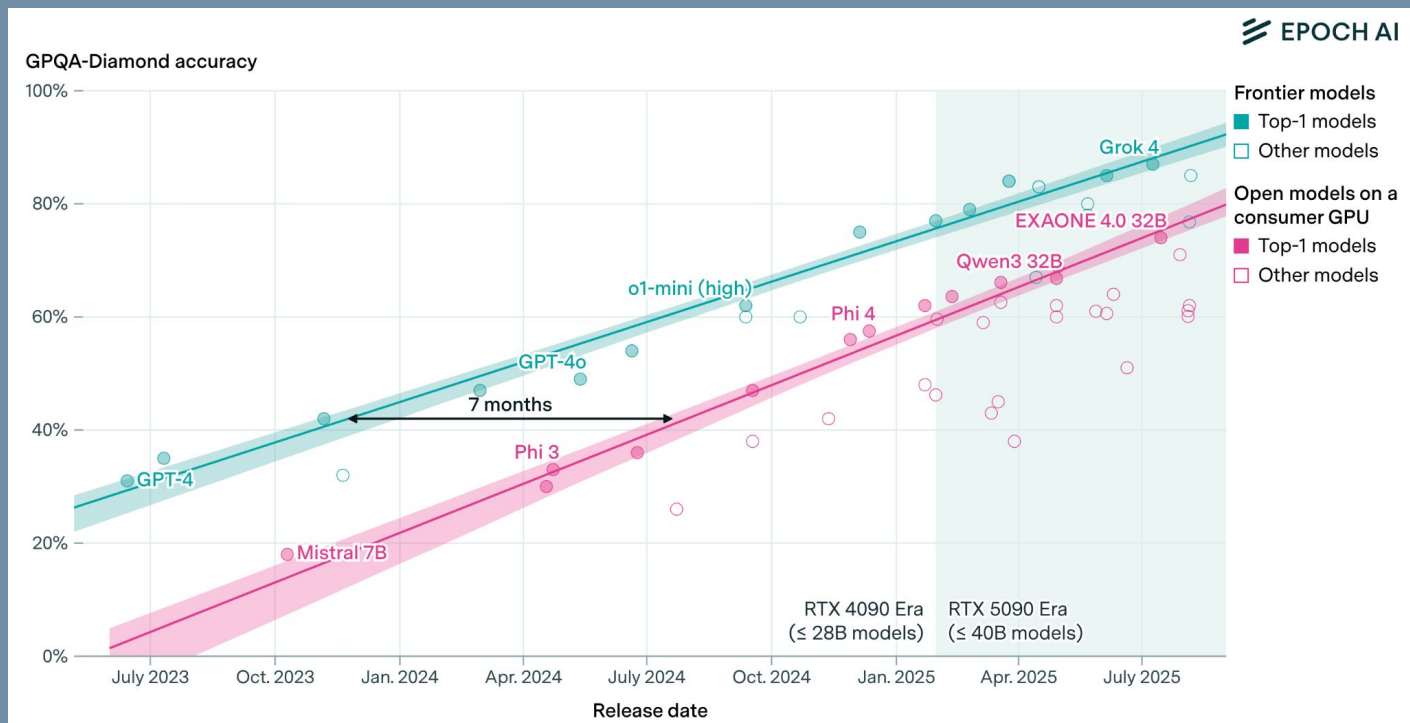
[Link zum Paper \(2024\)](#) – [Link zur Open Source AI Initiative](#)

Open Source LLM Initiativen

- **BLOOMZ** ([GitHub](#)): Zusammenschluss von Wissenschaftler:innen, hervorgegangen aus BigScience Workshop, 2022 veröffentlicht, nicht mehr weiterentwickelt.
- **ATOM Project** ([Webseite](#)): “American Truly Open Models”, Initiative von amerikanischen Wissenschaftler:innen, in dem Kontext entwickelte *Allen Institute AI* OLMO-Versionen 2024 und 2025 (OLMO 2 32B).
- **EuroLLM** ([Webseite](#)): Konsortium europäischer Einrichtungen, Unterstützt von EU-Kommission, mit open-sci-ref 0.01 erstes Referenzmodell veröffentlicht. (utterproject/EuroLLM ist wieder was anderes)
- **Apertus** ([Webseite](#)): Schweizer Konsortium (u.a. EFPL, ETH), vor kurzem 8 und 70B-Modelle veröffentlicht.

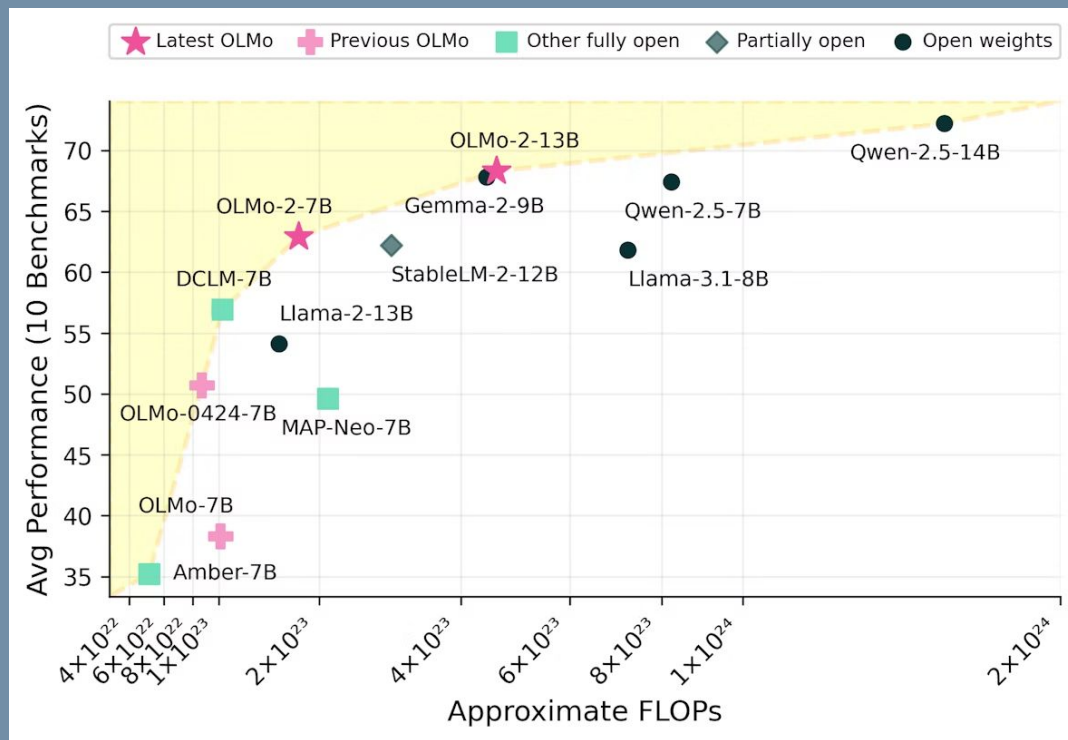
→ Wo sind diese Modelle zu finden im [Leaderboard](#)? ([OpenWeights LB](#))

Open Weights vs. Frontier



Quelle: Epoch AI ([Link](#))

Open Source vs. Open Weights

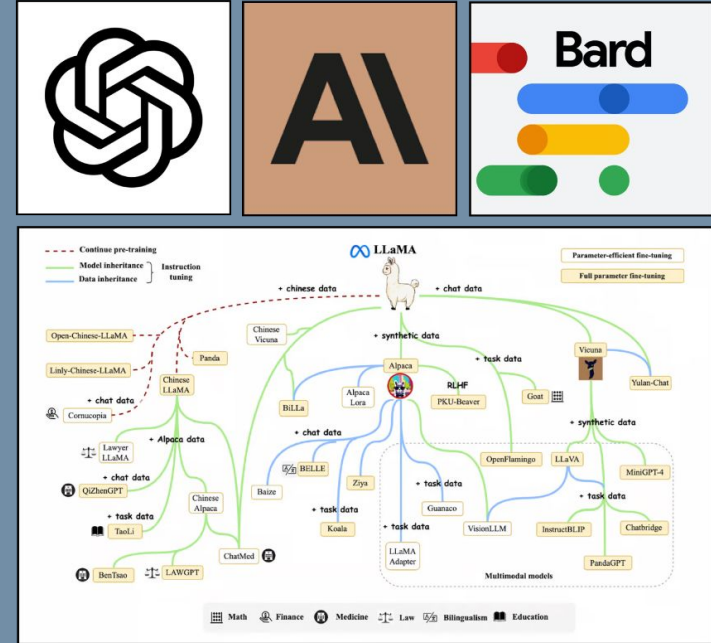


Quelle: AllenAI ([Link](#))

Closed vs. Open Weights vs. Truly Open

Kategorie	Online Frontier	Open Weights	Open Source
Wissenschaftliche Kontrolle / Reproduzierbarer Prozess	✗	✓	✓
Methodische Transparenz / Trainingsdaten, Modellarchitektur	✗	✗	✓
Datenschutz gewährleistet, keine externe Governance	✗	✓	✓
Interpretierbarkeit	✗	?	?

OS and LLM Ecosystems



Aus dem Vortrag von Andrej Karpathy:
Intro to Large Language Models
https://www.youtube.com/watch?v=zjkBMFhNj_g&list=WL&index=17

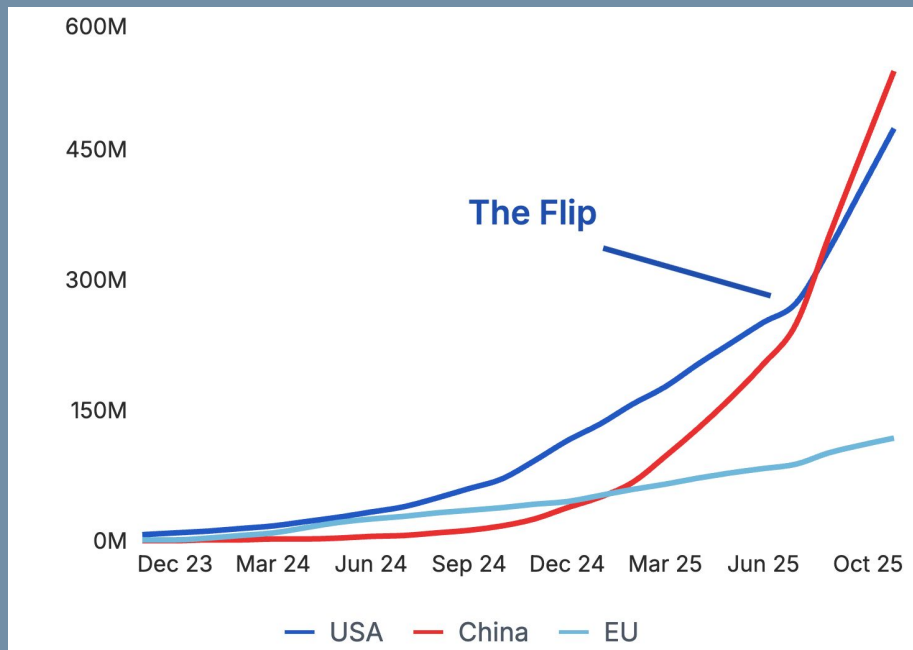
USA vs. China vs. EU

Kumulative Downloads von
open-weight-Modellen
führender LLM-Organisationen
in der Post-ChatGPT-Ära.

Quellen:

Nathan Lambert: Open Models
in 2025 ([Slides](#))

ATOM: American Truly Open
Models atomproject.ai



Warum sollten wir versuchen, Kontrolle zurück zu bekommen?

- Offene Modelle sind auf einem guten Weg (ein Stück hinter den proprietären Modellen, aber noch in Bewegung, vielleicht sogar auf diese zu)
- Langfristiger Zugang für Forschung
- Reduzierung der Machtkonzentration und Abhängigkeiten im Sektor AI
- Projekt Bildung und offene Gesellschaft

Welche Ressourcen für welche LLMs?

- Nutzung <<< Finetuning
- Abhängig nicht nur von Größe, sondern auch von Kontextlänge, Quantisierung, Promptlänge...
- **Arbeitsplatz-PCs/MacBooks** mit aktueller Ausstattung: bis 7B-Modelle
- **High-End-Arbeitsplatz-PCs** (z.B. mit NVIDIA 5080, 64GB Speicher, 1+TB SSD): 13B, ggfs. 30B
- **Server mit professioneller Ausstattung/GPUs** (z.B. RTX 6000 ADA): 70B, ggfs. 120B
- Nicht-lokale Alternativen: [GWDG HPC/Cluster/Services](#); [HuggingFace Inference Endpoints](#)

LLMs auf dem eigenen Rechner

Ollama (Webseite)	HuggingFace / Transformers Library (GitHub)
kuratierte Modellliste (erweiterbar)	Zugriff auf 1000e Modelle
Größtenteils automatisiert (z.B. Quantisierung)	Durchgehend konfigurierbar
Einfache Installation, API und GUI-Chat	Python-Framework notwendig, API und Library
Für Einsteiger geeignet	Eher für Entwickler

Literatur

Balloccu, S., Schmidová, P., Lango, M., & Dusek, O. (2024). Leak, Cheat, Repeat: Data Contamination and Evaluation Malpractices in Closed-Source LLMs. In Y. Graham & M. Purver (Hrsg.), Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers) (S. 67-93). Association for Computational Linguistics. <https://aclanthology.org/2024.eacl-long.5>

Gibney, E. (2024). Not all 'open source' AI models are actually open: Here's a ranking. Nature. <https://doi.org/10.1038/d41586-024-02012-5>

Lambert, N. (2023a, November 24). Gemma 3, OLMo 2 32B, and the growing potential of open-source AI. <https://www.interconnects.ai/p/gemma-3-olmo-2-32b-and-the-growing>

Lambert, N. (2023b, November 24). On China's open source AI trajectory. <https://www.interconnects.ai/p/on-chinas-open-source-ai-trajectory>

Liesenfeld, A., & Dingemanse, M. (2024). Rethinking open source generative AI: Open-washing and the EU AI Act. Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, 1774-1787. <https://doi.org/10.1145/3630106.3659005>

Manchanda, J., Boettcher, L., Westphalen, M., & Jasser, J. (2025). The Open Source Advantage in Large Language Models (LLMs) (No. arXiv:2412.12004). arXiv. <https://doi.org/10.48550/arXiv.2412.12004>

Somala, V. (2025, August 15). Frontier AI performance becomes accessible on consumer hardware within a year. Epoch AI. <https://epoch.ai/data-insights/consumer-gpu-model-gap>

The Open Source AI Definition – 1.0-RC2. (o. J.). Open Source Initiative. Abgerufen 20. Oktober 2025, von <https://opensource.org/ai/drafts/the-open-source-ai-definition-1-0-rc2/>

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., ... Scialom, T. (2023). Llama 2: Open Foundation and Fine-Tuned Chat Models (No. arXiv:2307.09288). arXiv. <https://doi.org/10.48550/arXiv.2307.09288>