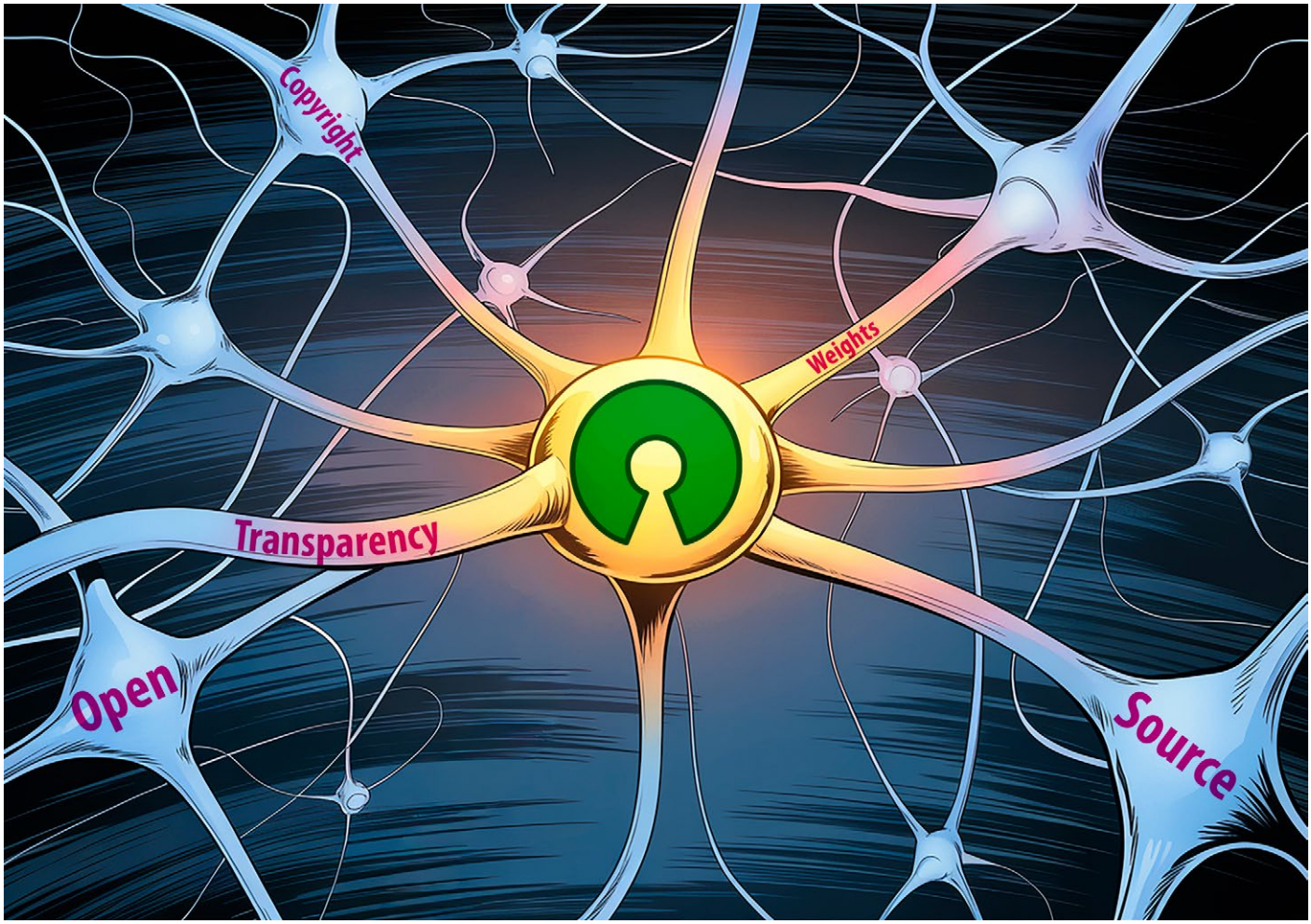

Bild: Martina Bruns / heise medien

Babylonische Verwirrung

Open-Source-KI: Welche Modelle es gibt und wie offen sie sind

Viele Sprachmodelle bezeichnen sich als Open Source, segeln aber unter falscher Flagge. Wer digitale Souveränität anstrebt oder strenge gesetzliche Anforderungen erfüllen muss, sollte genauer hinschauen. Wir geben einen Überblick über Lizenzen und Modelle.

Von Andrea Trinkwalder

Digitale Souveränität, sensible Firmen- oder Gesundheitsdaten, strenge Transparenzvorgaben oder Rechtssicherheit: Es gibt viele Gründe, generative (Sprach-)Modelle lokal zu betreiben oder nach einer möglichst unabhängigen Lösung zu suchen. Maximalen Gestaltungsspielraum und Kontrolle verspricht der Einsatz von Open-Source-Software – insbesondere da diese den Anforderungen der europäischen KI-Verordnung eher entsprechen als die proprietären Systeme von OpenAI, Google & Co. Doch in der KI-Landschaft pflegen viele Akteure ihre eigene Lesart von Open Source – vermutlich, um den Begriff zu ihren Gunsten aufzuweichen. Das prominenteste Beispiel ist Meta, das sein Sprachmodell Llama unbeirrt als Open Source anpreist, obgleich es nur wenige Kriterien davon erfüllt. Wir skizzieren die wichtigsten Anforderungen und untersuchen, welche Modelle diesen überhaupt genügen.

Was klassische Open-Source-Software vom Office-Programm bis hin zum Betriebssystem auszeichnet, hat die Open Source Initiative klar definiert und listet auf ihrer Website genau auf, welche Lizenzen diesen Anforderungen entsprechen (alle im Artikel erwähnten Links siehe ct.de/yzne). Zu den Kriterien gehören unter anderem: freie Weiterverbreitung, offener Quellcode, Ableitungen sind erlaubt, keine Diskriminierung von Personen und Einsatzzwecken.

Bei Machine-Learning-Systemen und insbesondere bei generativen Modellen hingegen können Forscher und Entwickler mit dem Quellcode allein kaum etwas anfangen, denn der definiert nur die

Strukturen und Mechanismen, wie die Neuronen und Schichten miteinander interagieren. Damit die Funktionsweise moderner KIs so durchschaubar und beeinflussbar wird wie klassische Software, müssen auch alle wichtigen Parameter und Komponenten der Trainings-Pipeline veröffentlicht sein – von den Trainingsdaten und -gewichten über die Checkpoints bis hin zum Alignment und Finetuning.

Gekaperte Definition

Ein Vorreiter in puncto Offenheit war das von einem Forscherkollektiv unter der Ägide von Hugging Face entwickelte 176-Parameter-LLM Bloom, das allerdings aufgrund seiner schieren Größe nur schwer zu handhaben war. Den Durchbruch in der Community schaffte schließlich Meta mit Llama. Anders als von Meta behauptet, handelt es sich dabei allerdings nicht um ein Open-Source-, sondern um ein Open-Weights-Modell, bei dem lediglich die Trainingsparameter (Gewichte) dokumentiert sind.

Außerdem verstoßen die Lizenzbedingungen der dafür geltenden „Llama 3.1 Community License“ gegen die Anforderungen an Open Source, unter anderem, weil sie die kommerzielle Nutzung einschränkten, wie die Open Source Initiative (OSI) regelmäßig moniert. Ähnliches findet sich in vielen Lizenzbedingungen, etwa in der „Qwen License“, unter denen Alibaba die größeren Varianten seines Qwen – etwa Qwen2.5-72B – veröffentlicht. Lediglich die meisten kleineren Qwen-Varianten unterliegen der sehr freizügigen Apache-2.0-Lizenz. Damit vergleichbar ist die MIT-Lizenz, die ebenfalls uneingeschränkte Nutzung für private und kommerzielle Zwecke, Modifikation und Weitergabe des LLMs erlaubt, solange der Urheberrechtshinweis erhalten bleibt. Sie gilt beispielsweise für DeepSeek V3, allerdings mit einigen Einschränkungen etwa für die militärische Nutzung – was den Anforderungen der OSI widerspricht. Wer böse Überraschungen vermeiden möchte, muss also vorab prüfen, ob sich der geplante Einsatzzweck überhaupt mit dem Kleingedruckten vereinbaren lässt.

Einen Überblick über konforme und nicht konforme Lizenzen hat die Free Software Foundation (FSF) auf ihrer Website zusammengestellt.

Daten sind der halbe Code

Doch die kommerziellen Entwickler schränken die Nutzung ihrer Modelle nicht nur mithilfe von Lizenzbedingungen ein, sondern auch, indem sie essenzielle Informationen zurückhalten, die eine eigenständige Weiterentwicklung oder unabhängige Forschung verhindern. Llama ist zwar in der Regel kostenlos und lässt sich lokal installieren, Meta veröffentlichte aber von Beginn an nur den Initialisierungscode und die Trainingsgewichte. In die Kategorie „Open-Weights-Modell“ fallen unter anderem auch DeepSeek R1, Qwen, Gemma, Kimi K2, GPT-OSS von OpenAI sowie die Phi-Serie von Microsoft.

Den Trainingsprozess und die Checkpoints dokumentieren Meta, Microsoft, Google, OpenAI, Anthropic & Co. ebenso wenig wie die Trainingsdaten – oder bestenfalls teilweise. Das hat zum einen urheberrechtliche Gründe. Die meisten generativen KIs wurden mit im Internet veröffentlichtem Material trainiert, darunter Zeitungsartikel, Songtexte et cetera; über die Rechtmäßigkeit dieses Vorgehens streiten und verhandeln Juristen weltweit noch immer.

In erster Linie geht es aber um Wettbewerbsvorteile. Denn die Trainingsdaten formen und verfeinern zusammen mit der Trainingsstrategie in einem mehrstufigen Prozess aus Pretraining, Posttraining und Alignment den Algorithmus, der am Ende den Text (oder Bilder, Videos et cetera) generiert. Darin liegt ein großer Teil des Betriebsgeheimnisses der proprietären Modelle und weniger im rohen Quellcode oder in den Gewichten.

Die OSI und die FSF werfen insbesondere Meta vor, den Open-Source-Begriff zu eigenen Gunsten umzudefinieren: „In einer Zeit, in der Meta versucht, Open Source zum eigenen Vorteil und auf Kosten unserer Freiheit neu zu definieren, rufen wir die gesamte Open-Source-Community dazu auf, sich zu vereinen und Metas Open Washing anzuprangern.“

Um dem entgegenzuwirken, hat die OSI im vergangenen Jahr eine eigene Open-Source-Definition für künstliche Intelligenz erarbeitet, die OSAID, um Unklarheiten auszuräumen und die Deutungshoheit zu behalten:

- Nutzung für jedwede Zwecke möglich, ohne eine Erlaubnis einholen zu müssen;
- Studieren des Systems, seiner Komponenten und deren Arbeitsweise möglich;
- Modifizieren für jedweden Zweck, einschließlich Änderungen des Outputs;
- Teilen des Systems mit anderen, verändert oder unverändert.

Eine lesenswerte Studie zum Thema haben die niederländischen Forscher Andreas Liesenfeld und Mark Dingemanse im Jahr 2024 verfasst – und eine differenzierte Metrik entwickelt, mit der sich die Offenheit der Systeme vergleichen lässt.

Transparenz mit Seltenheitswert

Unterm Strich bleibt eine Handvoll nennenswerter großer Sprachmodelle übrig, die sich so uneingeschränkt verwenden lassen, wie es die Open Source Initiative fordert – und wie es bei klassischer OSS durch das bloße Veröffentlichen des Quellcodes gegeben ist. Das eine heißt **OLMo und wird in den USA vom Allen Institute for AI (AI2)** entwickelt. Hinter dem gemeinnützigen Forschungsinstitut steckt

c't kompakt

- Viele kostenlos verfügbare Modelle wie Llama werben mit Open Source, schränken die vorgegebene Freiheit aber im Kleingedruckten ein.
- Die Open Source Initiative hat deshalb klare Richtlinien definiert.
- Derzeit genügen diesen nur zwei Modelle aus den USA und der Schweiz.

Project	Availability						Documentation						Access	
	Open code	LLM data	LLM weights	RL data	RL weights	License	Code	Architecture	Preprint	Paper	Modelcard	Datasheet	Package	API
OLMo 7B Instruct	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	~
BLOOMZ	✓	✓	✓	✓	~	~	✓	✓	✓	✓	✓	✓	✗	✓
AmberChat	✓	✓	✓	✓	✓	✓	~	~	✓	✗	~	~	✗	✓
Open Assistant	✓	✓	✓	✓	✗	✓	✓	✓	~	✗	✗	✗	✓	✓
OpenChat 3.5 7B	✓	✗	✓	✗	✓	✓	~	✓	✓	✓	~	✗	✓	~
Pythia-Chat-Base-7…	✓	✓	✓	✓	✗	✓	✓	✓	~	✗	~	~	✓	✗
Cerebras GPT 111…	~	✓	✓	✓	✓	~	✗	✓	~	✗	✗	✓	✗	✓
RedPajama-INCITE…	~	✓	✓	✓	✓	~	~	~	✗	✗	✓	✓	✗	~

Bild: Liesenfeld, Dingemanse

Niederländische Forscher haben 2024 ein Schema entwickelt, um die Offenheit von generativen Sprachmodellen zu messen. OLMo belegte Rang 1.

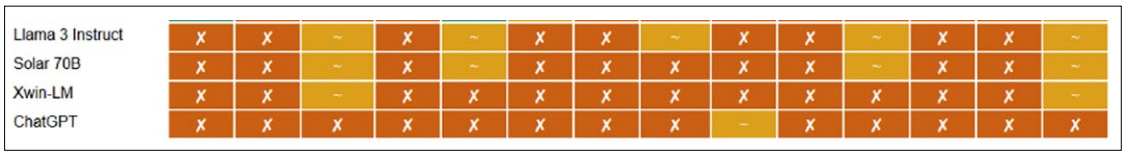

Bild: Liesenfeld, Dingenmanse

Metas Llama hingegen rangiert sehr weit unten, wird aber offensiv als Open Source vermarktet.

eine gut finanzierte Stiftung des Microsoft-Gründers Paul Allen.

Das zweite heißt **Apertus, existiert in einer 8- und einer 70-Milliarden-Parameter-Version** und stammt von einem Schweizer Forscherkollektiv der ETH Zürich sowie dem EPFL Lausanne. Es wurde multilingual trainiert und finanziert sich aus Steuergeldern und Investitionen; die immensen Rechenressourcen liefert der landeseigene Alps-Supercomputer in Lugano. Eine Besonderheit ist, dass die schweizerischen Forscher eine Filtertechnik entwickelt haben, um urheberrechtlich kritisches Material auch nachträglich aus dem Trainingskorpus zu entfernen.

Dass dies sinnvoll ist, zeigt der Fall des **OpenELM von Apple**. Der Konzern sieht sich derzeit mit einer Klage konfrontiert, weil er sein für den Betrieb auf kleinen Geräten optimiertes Open-Source-Modell mit der RedPajama-Datenbank trainiert hatte, die wiederum E-Books aus illegalen Quellen enthielt.

Schon etwas älter sind das 2023 veröffentlichte **12-Milliarden-Parameter-Modell Pythia von EleutherAI** sowie das darauf basierende **Dolly 2.0 von Databricks**. An dem arbeitsintensiven Aufbau eines eigenen Instruction-Tuning-Datensatzes beteiligten sich über 5000 Mitarbeiter von Databricks. Warum dies notwendig war, beschreibt das Team in seinem Blog: Die für das Training der Version 1.0 verwendeten Daten, die das Alpaca-Entwicklerteam der Uni Stanford automatisiert mithilfe des OpenAI-API erstellt hatten, hätten lizenzrechtliche Probleme aufgeworfen: Denn deren Nutzungsbedingungen verbieten, damit Modelle zu trainieren, die mit OpenAI konkurrieren. Demzufolge dürfen potenzielle Anwender die erste Dolly-Inkarnation höchstwahrscheinlich nicht kommerziell einsetzen – eine gravierende Einschränkung, der auch Alpaca, Koala, GPT4All und Vicuna unterliegen.

Auch das Together.ai-Team hat für seine **3B-/7B-Modellfamilie RedPajama-INCITE** eine eigene Datenbasis aufgebaut, die sich an der von Llama orientiert. Sie steht allerdings im Verdacht, geschützte Werke zu enthalten, die aus illegalen Quellen stammen; konkret handelt es sich um die vom KI-Forscher Shawn Presser erstellte „Books3"-Sammlung, die mittlerweile abgeschaltet und nicht mehr zugänglich ist. Technisch stützt sich das Instruction-Modell auf GPT-NeoX von EleutherAI.

Fazit

Die Open-Source-KI-Landschaft ist fragmentiert und undurchsichtig, kommerzielle Akteure ringen mit den gemeinnützigen Initiativen um die Deutungshoheit. Wichtig ist, dass der Begriff nicht verwässert und Open-Washing verhindert wird, wie unser Überblick zeigt. Die meisten als Open Source angepriesenen Sprachgeneratoren entpuppen sich bei näherem Hinsehen als Open-Weights-Modelle. Die haben durchaus ihre Berechtigung, weil sie sich unkompliziert lokal installieren und in Betrieb nehmen lassen und für die meisten Zwecke kostenlos sind. Dennoch handelt es sich um proprietäre Software.

Demgegenüber verfolgt echte Open-Source-KI das Ziel, Firmen und Wissenschaftlern volle Transparenz und uneingeschränkte Nutzung zu garantieren – von Modifizierungen der Modellarchitektur über das Training mit eigenen Daten bis hin zum Alignment und Finetuning. Der zur Grundlagenforschung befähigte Personenkreis ist zwar eingeschränkt – wer hat schon einen Supercomputer? Aber vor allem größere Firmen, die ein möglichst hohes Maß an Rechtssicherheit, Datenschutz und Transparenz anstreben oder dazu verpflichtet sind, können solche Modelle mit vergleichsweise geringem Aufwand mit ihren eigenen Daten optimieren (Finetuning).

Den konsequentesten Weg haben das gemeinnützige US-Forschungsinstitut Allen AI mit OLMo und die schweizerischen Forscher mit Apertus eingeschlagen: Beide haben ihre Datensammlungen sorgfältig kuratiert und versuchen, geschütztes Material herauszufiltern. Außerdem stehen ihre LLMs sowohl in kleinen als auch in einer mittleren Größe zur Verfügung – was sie auch für anspruchsvollere Aufgaben qualifiziert. Der Abstand zu einem ChatGPT oder Gemini mag anfangs groß erscheinen, derzeit hinken sie den Frontier-Modellen etwa zwei bis drei Jahre hinterher. Mit größerer Unterstützung und mehr finanziellen Mitteln könnten sie den Abstand aber verringern und kämen keinen privatisierten Firmen, sondern der Allgemeinheit zugute. *(atr@ct.de)* c't

Alle im Artikel erwähnten Links:
ct.de/yzne

Open-Source-Sprachmodelle

Modell	Apertus	OLMo 2	OpenELM	Pythia	RedPajama INCITE
Anbieter, URL	ETH Zürich/EPFL, swiss-ai.org	Allen Institute for AI (AI2), allenai.org	Apple ML Research, machinelearning.apple.com	EleutherAI, eleuther.ai	Together.ai, together.ai/blog/redpajama
Art	Text	Text	Text	Text	Text
verfügbare Größen (Parameterzahl)	8B, 70B	7B, 13B, 32B	270M, 450M, 1.1B, 3B	16 Modelle, 70M bis 12B	3B, 7B (Base, Instruct)
Größe des Trainingskorpus (in Token)	15 Billionen (über 1000 Sprachen)	6 Billionen	1,8 Billionen	300 Milliarden	1,2 Billionen
max. Kontextfenster	65.000 Token	4100 Token	2048 Token	2048 Token	k. A.
Lizenz	Apache 2.0	Apache 2.0	Apple Sample Code License (ggf. Einschränkungen bezüglich Patenten)	Apache 2.0	Apache 2.0
fürs Training verwendete Hardware	trainiert auf Alps-Supercomputer mit über 4000 GH200-GPUs	OLMo 2 32B wurde auf 8 H100-GPUs trainiert.	k. A.	z. B. Pythia-1B: 4 × A100 in 18 Tagen oder 8 × RTX 3090 in 30 Tagen	RedPajama-INCITE-7B-Base wurde mit 3072 V100-GPUs trainiert.
Hardware-Anforderungen für lokalen Betrieb (Inferenz)[1]	8B-Version: High-End-Consumer-GPU / 70B: A100- oder H100-Cluster	7B: leistungsfähige Einzel-GPU; 32B: sehr große GPUs oder Quantisierung nötig	3B-Version: High-End-Consumer-GPU, z. B. RTX 4090	High-End-Consumer-GPUs mit 16–24 GB VRAM)	RedPajama-INCITE-3B: 8-GB-GPU bei Standard oder 6 GB bei Int8 quantisiert

k. A. keine Angabe [1] für alle Modelle stehen auch quantisierte Versionen für gängige ML-Frameworks zum Download