Agora Digitale Transformation

Whitepaper

Souverän in der Mitte?

Europas offene KI-Modelle auf Hugging Face

Autor:innen-Team:

Yannick Kalff

Danksagung

Vielen Dank an alle Gesprächspartner:innen, von deren Wissen dieser Text profitiert hat.

Agora Digitale Transformation gGmbH
Krausenstraße 8
10117 Berlin

T. +49(0)30 81 45 03 78 80
www.agoradigital.de
info@agoradigital.de

Ansprechperson:



Dr. Yannick Kalff
yannick.kalff@agoradigital.de

Gefördert von: STIFTUNG MERCATOR

Souverän in der Mitte? Europas offene KI-Modelle auf Hugging Face

Design: Jakub Stejskal, Make Studio
Satz: Agora Digitale Transformation
Grafiken: Nele Schacht, Parzelle34

Dieser Beitrag unterliegt einer CreativeCommons-Lizenz (CC BY-SA). Die Vervielfältigung, Verbreitung und Veröffentlichung, Veränderung oder Übersetzung von Inhalten der Agora Digitale Transformation, die mit der Lizenz „CC BY-SA“ gekennzeichnet sind, sowie die Erstellung daraus abgeleiteter Produkte sind unter den Bedingungen „Namensnennung“ und „Weiterverwendung unter gleicher Lizenz“ gestattet. Ausführliche Informationen zu den Lizenzbedingungen finden Sie hier: http://creativecommons.org/licenses/by-sa/4.0/

Februar 2026 | DOI 10.5281/zenodo.23808610

Executive Summary

Die europäische KI-Entwicklung ist dynamischer, als die öffentliche Debatte vermuten lässt. Unsere Auswertung der Plattform Hugging Face zeigt: Im laufenden Jahr 2026 wächst die Zahl neu veröffentlichter Modelle aus Europa um rund 70 Prozent und damit etwas schneller als im Vorjahr. Das ist eine höhere Wachstumsrate als in China oder den USA. Absolut bleibt Europa zurück: rund 40 neue Modelle gegenüber rund 140 aus China. Der Anteil kleiner und mittlerer Modelle (30 bis 130 Milliarden Parameter) liegt in Europa bei 31,2 Prozent. Viele dieser Modelle sind frei verfügbar und lassen sich auf eigener Recheninfrastruktur betreiben. An diese Dynamik gilt es anzuschließen.

Die öffentliche Debatte kreist jedoch um wenige große Sprachmodelle, die Frontier-Modelle. Ihre Leistungsfähigkeit bestimmt den Stand der Technik und damit den technischen Rückstand Europas. Von allen untersuchten Modellen stammen etwas mehr als 10 Prozent aus Europa; mit den Spitzenmodellen aus den USA oder China kann keines davon konkurrieren.

Wir sehen in offenen Modellen jenseits der Frontier eine politische Chance für Europa. Für viele Aufgaben in Industrie und Verwaltung genügen kleine und mittlere Modelle, die auf einen konkreten Zweck zugeschnitten sind. Sie benötigen weniger Rechenleistung für Training und Betrieb (Inferenz). Deshalb kann Europa die Entwicklung und den produktiven Einsatz solcher Modelle gezielt fördern und ausbauen. Die europäische KI-Entwicklung kann so zu einem tragfähigen Standbein für den souveränen Einsatz von KI werden – trotz fehlender Rechenzentren und verzögerter Gigafactory-Strategie. Denn Souveränität entscheidet sich nicht allein an der Leistungsspitze der Frontier.

Zugleich bieten kleine und mittlere Modelle einen Skalierungshebel für ein souveränes europäisches KI-Ökosystem, das langfristig auch Frontier-Modelle entwickeln und betreiben kann. Investitionen in offene Modelle nach Open-Source- und Open-Weights-Prinzipien bauen technisches und institutionelles Erfahrungswissen auf. Davon profitieren auch größere Vorhaben.

Wir empfehlen, das Mittelfeld der europäischen KI-Entwicklung stärker in den Blick zu nehmen und gezielt zu unterstützen. Konkret schlagen wir vor, öffentlich-private Konsortien für die Modellentwicklung zu fördern, Anwendungsfälle in öffentlich-privaten Partnerschaften zu erproben und dabei technische Beratung und Implementierungshilfe bereitzustellen sowie eine anwendungsorientierte, vernetzte Forschungsförderung zu finanzieren, die in konkreten Transfer rückgebunden wird. So lässt sich ein souveränes KI-Ökosystem stärken, das Entwicklung und Einsatz von KI in offenen Strukturen vorantreibt und vorhandenes Know-how vernetzt.

Inhalt

1 Hinter der Frontier – ruhiges Hinterland? 4

2 Open-Weights oder Open-Source? 7

3 Das europäische Open-Weights KI-Ökosystem auf Hugging Face 8

3.1 Verteilung nach Regionen 9

3.2 Verteilung der Modelle nach Parameterzahl 9

3.3 Entwicklung im Zeitverlauf 11

3.4 Lizenzmodelle 12

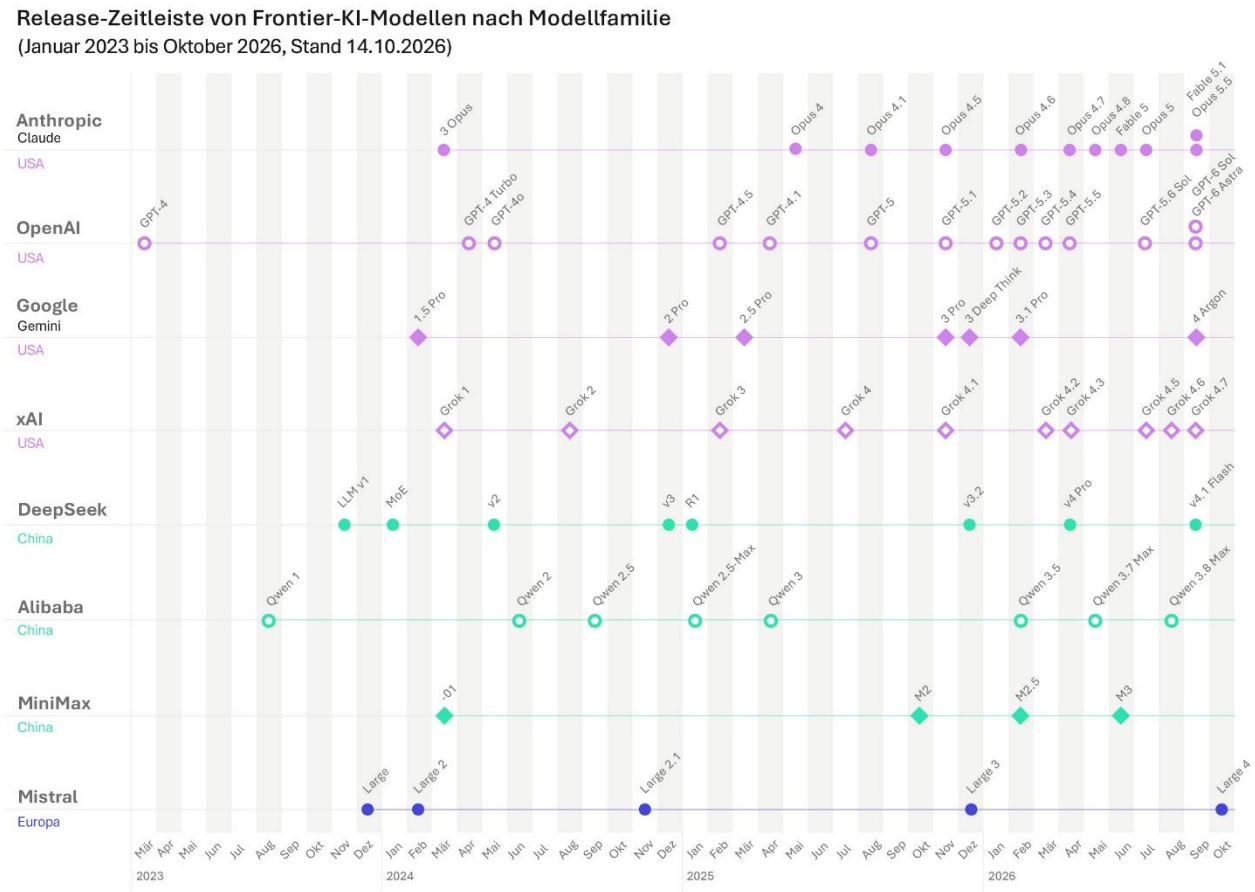
4 Souveräne KI-Entwicklung und Einsatz durch Open-Weights-Modelle stützen 14

Literaturverzeichnis 15

1 Hinter der Frontier – ruhiges Hinterland?

In der Debatte um Künstliche Intelligenz dominieren die großen Sprachmodelle sowohl medial als auch politisch. Sie markieren die Grenze des aktuellen Entwicklungsstands und werden daher als Frontier-Modelle bezeichnet. Anthropics Claude, OpenAIs GPT-Serie und die Modelle chinesischer Anbieter wie Alibabas Qwen oder DeepSeeks v4 verschieben diese Grenze in immer kürzeren Abständen (Abbildung 1). Mistral ist derzeit das einzige europäische Unternehmen, das in diese Reihe gestellt werden kann und in vergleichsweise langen Abständen neue Modelle veröffentlicht.

Abbildung 1 Modellhistorie



Quelle: Eigene Recherche

Hinter der Frontier beginnt das vermeintlich ruhige Hinterland. Kleine und mittelgroße Modelle, die für spezifische Anwendungsfälle entwickelt und spezialisiert nachtrainiert (Fine-Tuning) werden, erhalten in der Wahrnehmung und in der politischen Debatte kaum Aufmerksamkeit. Gerade hier liegt jedoch ein Betätigungsfeld für europäische Anbieter:innen und Anwender:innen sowie für eine tragfähige industriepolitische Initiative. Entschiedene Aufmerksamkeit für diesen Bereich könnte Akteure dabei unterstützen, Modelle zu entwickeln und in Anwendung zu bringen und dadurch das politische Ziel eines souveränen europäischen KI-Ökosystems zu tragen.

Denn europäische Anbieter haben bisher kaum ausreichende Potenziale aufgebaut, um im Spitzenfeld mitzuspielen. Lediglich Mistral gilt als bedeutender Entwickler verschiedener LLMs, die

an die Spitze heranrücken. Zudem setzt Mistral – wie chinesische Anbieter auch – auf einen Open-Weights-Ansatz und stellt Modelle und deren Gewichte bereit.[1] Ein genauer Blick ins Hinterland lohnt sich dennoch, denn europäische Entwickler:innen bieten hier durchaus Angebote an. Ihre Modelle erreichen zwar in Anwendungsdomänen wie der Programmierung oder der IT-Sicherheit nicht die Leistungsfähigkeit der Frontier-Modelle. Sie sind jedoch aufgrund ihrer geringeren Komplexität effizienter und benötigen weniger Rechenkapazität – sowohl für ihr Nachtraining (Fine-Tuning) als auch für ihren produktiven Einsatz (Inferenz).

In der Praxis treiben vor allem ökonomische und Effizienzfragen den Einsatz von Open-Weights-Modellen voran. Aber auch Risiko- und Sicherheitsabwägungen werden zunehmend relevant:

1. Die meisten Alltagsanwendungen benötigen weder die neuesten Modellgenerationen noch die größten Modelle. Wo bei der Qualität Spielraum besteht, liefern kleinere Modelle schnellere Ergebnisse.
2. Frontier-Modelle können teuer sein, weil die Anbieter ihre Komplexität und Fähigkeiten einpreisen. Das gilt insbesondere für US-amerikanische Anbieter, die bis zu sechsmal so teuer sind wie chinesische Anbieter.[2] Auch die Preisgestaltung der KI-Unternehmen verändert sich: Statt einer Flatrate für alle KI-Modelle bieten Abo-Tarife nun ein begrenztes Kontingent, und für Frontier-Modelle zahlen Nutzende zusätzlich. Kleinere Modelle kosten pro Token (Abrechnungseinheit für Text) deutlich weniger. Laufen sie lokal, entfallen die Abo- und Token-Kosten ganz. Die Investitionskosten für Infrastruktur und Hardware bleiben bestehen.
3. Multi-Modell-Plattformen wie OpenRouter haben den Bedarf an Flexibilität zu ihrem Geschäftsmodell gemacht. Nutzende können zwischen Modellen verschiedener Anbieter wählen und so die Bindung an einen Anbieter (Lock-in) vermeiden. Die Plattformen führen jedoch eine Zwischenschicht zwischen dem Modellangebot und der Nachfrage ein. Sie vermitteln zudem den Zugang zu chinesischen Anbietern, betreiben deren Modelle jedoch in der Regel nicht selbst.
4. Lokale Modelle können in der Verwaltung und in sensiblen Unternehmensbereichen zielführender sein, weil ihre Nutzung weniger Risiken birgt: Informationen und Daten gehen nicht an Dritte. Auch die begrenzten Fähigkeiten kleiner und mittlerer Modelle

Ein gängiges Praxisproblem

Ein mittelständisches Unternehmen will KI im Kundenservice, in der Dokumentenverwaltung und in der Administration einsetzen. Die dafür erforderlichen Daten, wie Konstruktionsunterlagen, Verträge, interne Berichte und Kund:innenkommunikation, sind unternehmenskritisch und hochsensibel. Sie dürfen das Haus nicht verlassen. Zudem ist der Einsatz von KI für das Unternehmen auch eine Kostenfrage: sei es durch Abo-Gebühren pro Nutzer:in, eine verbrauchsbasierte Abrechnung oder eine Mischung aus beidem. Ein eigenes Modell in der eigenen IT kann eine alternative Lösung für beide Probleme sein. Sie setzt jedoch Rechenleistung sowie Betriebs- und Implementierungswissen voraus. Genau an diesem Punkt entscheidet sich die digitale Souveränität im Bereich der KI.

[1] Lang et al., 2026

[2] Mozilla Foundation, 2026, S. 51. Unternehmen zeigen zunehmend eine Wechselbereitschaft zu Open-Weights-Modellen, die vor allem kostengetrieben ist (Tan, 2026). Bitkom zufolge nutzen knapp 11 Prozent der befragten Unternehmen Open-Weights-Modelle, am häufigsten aus den USA (Bitkom e.V, 2026).

tragen dazu bei, Sicherheitsvorfälle zu vermeiden. Von ihnen gehen geringere Risiken aus als von Frontier-Modellen.

Für KI-Projekte, die Prozessdigitalisierung oder KI-Agenten in Verwaltung und Industrie umfassen, bieten kleinere Sprachmodelle daher eine gangbare Alternative abseits der Frontier-Modelle und der Bindung an Hyperscaler.[3] Mit Fokus auf europäische Modelle und die Betriebsinfrastruktur wird daraus zugleich ein Souveränitätsthema.

Die Debatte reduziert die Wahl oft auf „USA oder China“. Es gibt eine dritte Option. Auf der einen Seite stehen proprietäre US-Frontier-Modelle, deren Zugang aufgrund erratischer politischer Entscheidungen nicht jederzeit gewährleistet zu sein scheint.[4] Auf der anderen Seite stehen chinesische Open-Weights-Modelle mit möglicherweise zensierten Trainingsdaten und Antworten oder mit inhärenten Sicherheitsrisiken.[5] Dazwischen ermöglichen europäische Open-Weights-Modelle einen souveränen und verlässlichen Einsatz von KI in der Industrie und in der Verwaltung.

Wir empfehlen daher die hier verfügbaren Modellangebote aufzugreifen und in konkrete Anwendungsfälle zu überführen. Dies kann durch eine nachhaltige industriepolitische Förderung gelingen, die auch in die digitale Souveränität einzahlt.

[3] Bundesdruckerei, 2026

[4] Mozilla Foundation, 2026

[5] Stein, 2025

2 Open-Weights oder Open-Source?

Die Debatte um nicht-proprietäre Modelle wird häufig mit dem Begriff „Open Weights“ in Verbindung gebracht. Der Begriff ist vor allem bei chinesischen Modellen verbreitet. Große Anbieter wie DeepSeek oder Moonshot machen ihre Modelle nicht nur über eine Online-Schnittstelle (API) zugänglich, sondern veröffentlichen sie auch auf Plattformen wie Hugging Face. Ihre Nutzungsmöglichkeiten werden durch unterschiedliche Lizenzmodelle geregelt. Je nach Lizenzierungsansatz werden unterschiedliche Verwertungsmöglichkeiten eröffnet, die von restriktivem Ausschluss kommerzieller Interessen bis hin zu uneingeschränktem Einsatz reichen und unterschiedliche Rechte einräumen.

Open-Weights suggeriert eine Nähe zur Open-Source-Bewegung, was jedoch zu kurz greift. Offene und einsehbare Modellgewichte heben die Modelle zwar von proprietären Modellen wie jenen von Anthropic oder OpenAI ab; für tatsächlich offene Modelle nach der Open-Source-Definition sind jedoch zusätzlich transparente Auskünfte über die Art und Zusammensetzung der Trainingsdaten, den konkreten Trainingsdatensatz sowie den Trainingscode erforderlich.

Was bedeutet Open-Weights?

Mit Open-Weights-Modellen sind Modelle gemeint, deren trainierte Gewichte öffentlich heruntergeladen und selbst betrieben werden können. Das ist mehr als nur ein API-Zugang, aber weniger als vollständige Offenheit: Trainingsdaten, Trainingscode und die Qualität der Datengrundlage bleiben in der Regel unbekannt. Die Open-Source-Initiative unterscheidet deshalb ausdrücklich zwischen Open-Weights und Open-Source.[1]

Open-Weights-Modelle stehen in der Regel frei zum Download bereit. Jede:r mit geeigneter Hardware kann sie selbst betreiben – lokal oder in einem Rechenzentrum nach Wahl. Hinzu kommt die Möglichkeit, die Modelle durch Fine-Tuning an eigene Anwendungsfälle anzupassen.[6] Eins zu eins nachbauen lassen sie sich damit allerdings nicht. Auch die Qualität der Datengrundlage bleibt intransparent. Mögliche datenseitige Verzerrungen lassen sich daher nur durch systematische Tests und regelmäßige Audits statistisch annähern.

Auch wenn Open-Weights-Modelle nicht vollständig offen sind, liegt ihr besonderer Vorteil in ihrer freien Verfügbarkeit und im Einsatz in eigener Infrastruktur. Neben den bereits genannten Kostenvorteilen bietet der Eigenbetrieb auch Sicherheitsvorteile, da Betriebsgeheimnisse nicht abfließen und DSGVO-Vorschriften nicht verletzt werden, wenn die fraglichen Daten in der eigenen Infrastruktur verarbeitet werden. Lokale KI-Inferenz mit Open-Weights-Modellen ist damit ein relevanter Faktor für die Einhaltung von Compliance-Vorschriften.

[6] Stapel, 2026

3 Das europäische Open-Weights KI-Ökosystem auf Hugging Face

Open-Weights-Modelle werden über Distributionsplattformen wie Hugging Face bereitgestellt. Diese ähneln dem Aufbau von Softwarequellcode-Repositorien wie GitHub und bieten Organisationen und Einzelpersonen eine Anlaufstelle, um eigene entwickelte Modelle bereitzustellen oder auf fremde Modelle unkompliziert zuzugreifen. Hugging Face gilt als Quasi-Standard für die Veröffentlichung von Modellen und ist eine ideale Quelle, um den Stand des europäischen KI-Ökosystems auch im Vergleich zu China und den USA genauer zu bestimmen. Stand Oktober 2026 bietet Hugging Face über 3 Millionen Modelle für verschiedene Anwendungsarten an.[7]

Was sagen Parameterwerte aus?

Parameter sind die trainierten Zahlenwerte eines neuronalen Netzes; die Anzahl der Parameter – beispielsweise 7 Milliarden (7B) – gibt Aufschluss über die Größe und Kapazität eines Modells sowie über dessen Ressourcenbedarf. Die Anzahl der Parameter sagt jedoch wenig über die Qualität eines Modells aus. Ebenso wichtig sind Trainingsdaten, Modellarchitektur und Trainingsmethoden. Für Qualitäts- und Leistungsvergleiche sind vergleichende Benchmarks und die Messung der konkreten Aufgabenerfüllung aussagekräftiger.

Wir haben nur Modelle zur Textgenerierung betrachtet, die auf Hugging Face verfügbar sind, und ihre Metadaten – Anzahl der Downloads, Veröffentlichungsdatum, Anzahl der Parameter, Lizenzierung – über die API-Schnittstelle ausgelesen. Die Daten wurden weiter eingegrenzt:

1. Nur Basismodelle werden gezählt, das heißt, Modelle, die von den ursprünglichen Entwickler:innen bereitgestellt werden. Fine-Tunings und erneut hochgeladene Modelle Dritter fließen nicht in die Auswertung ein.
2. Die Modelle müssen mindestens 7 Milliarden (7B) Parameter aufweisen. Diese kleinen Modelle sind bereits für einfache Aufgaben leistungsfähig und besonders attraktiv für kostengünstige Entwicklung und schnelles Fine-Tuning.[8]
3. Der Veröffentlichungszeitraum auf Hugging Face erstreckt sich vom 1. Januar 2023 bis zum 30. September 2026. Dadurch werden alle Modelle, die wenige Wochen nach der öffentlichen Vorstellung von ChatGPT im November 2022 bereitgestellt wurden, bis hin zu den aktuellen Entwicklungen gezählt.
4. Von allen Einträgen auf Hugging Face zählen wir nur Organisationsprofile und Profile, die zwar als Individualprofile eingetragen sind, aber Organisationen repräsentieren. Diese Profile ordnen wir anhand der Profilinformationen und automatisierter Websuchen einem oder mehreren Herkunftsländern zu. Ist keine Zuordnung möglich oder handelt es sich um Individual-Accounts, beziehen wir die Fälle nicht ein.
5. Europa verstehen wir geografisch und zählen alle Mitgliedstaaten der Europäischen Union (EU27), Norwegen, die Schweiz und das Vereinigte Königreich dazu.

Mit diesen Einschränkungen sind auf Hugging Face insgesamt 1.037 Modelle verfügbar.

[7] Ghosh et al., 2026

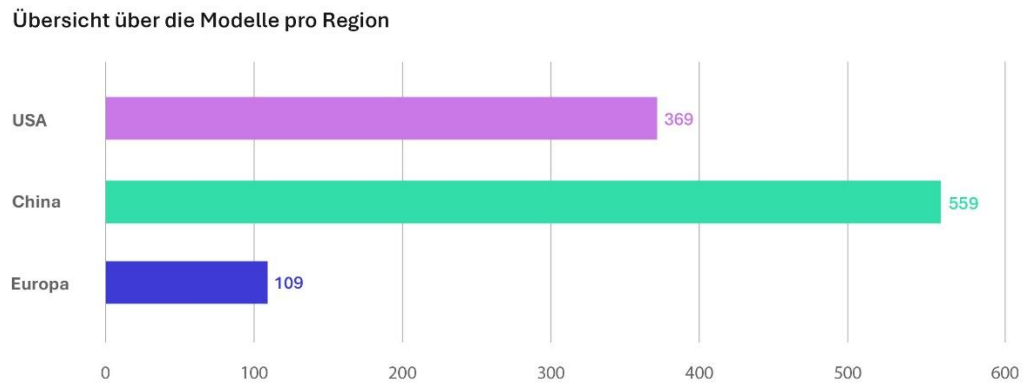
[8] Stapel, 2026, S. 366–367

3.1 Verteilung nach Regionen

Abbildung 2 zeigt, wie sich die Modelle auf die drei Regionen verteilen. China führt mit deutlichem Vorsprung bei den absoluten Modellzahlen und macht über die Hälfte aller Veröffentlichungen aus. Die USA folgen mit etwas Abstand und stellen mehr als ein Drittel der Modelle auf Hugging Face bereit. China und die USA stellen zusammen etwa 90 Prozent aller ausgewerteten Modelle. Der Anteil europäischer Modelle beträgt nur etwas mehr als 10 Prozent.

Auf den ersten Blick erscheint das europäische Ökosystem damit weit abgeschlagen. Die absoluten Zahlen sind jedoch mit Vorsicht zu lesen. Sie umfassen auf allen drei Seiten möglicherweise Forschungsmodelle, die nicht für den produktiven Einsatz geeignet sind. Gelöschte Modelle tauchen ebenfalls nicht auf.

Abbildung 2 Verteilung der auf Hugging Face verfügbaren Modelle auf drei Regionen USA, China, EU27 + CH + GB + NO



Quelle: Eigene Erhebung

3.2 Verteilung der Modelle nach Parameterzahl

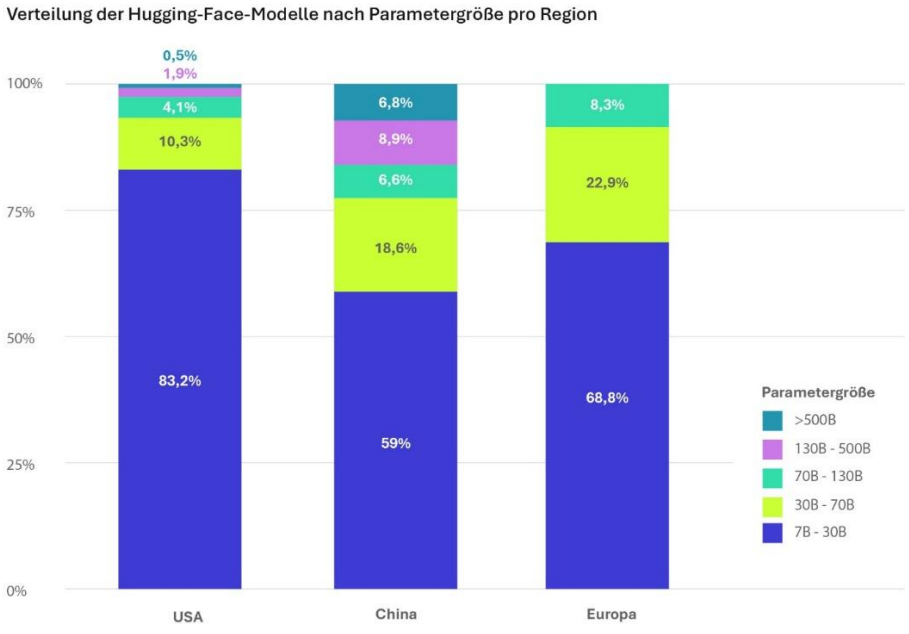
Neben den absoluten Zahlen lohnt sich ein Blick auf die Modellkomplexität aller Modelle, gemessen an der Anzahl der Parameter. Die Anzahl der Parameter gibt die Kapazität eines Sprachmodells an. Bei der Wiedergabe von Fakten helfen viele Parameter, da auch kleineres Detailwissen besser im Modell abgebildet wird. Gleichwohl ist die Anzahl der Parameter nur eingeschränkt als Qualitätsmaß geeignet. Außerdem sind proprietäre Frontier-Modelle aus den USA nicht auf Hugging Face veröffentlicht und fließen daher nicht in die Auswertung ein.

Abbildung 3 zeigt, dass sich die Spitze europäischer Modelle auf Hugging Face im mittleren Größensegment zwischen 30B und 130B Parametern verteilt. Europa hat in dieser Klasse den höchsten relativen Anteil unter den drei Regionen. Oberhalb von 130 Milliarden Parametern fehlt Europa jedoch vollständig,[9] während China mit 15,7 Prozent deutlich vor den USA liegt. Europa ist damit vor allem auf das Mittelfeld fokussiert und entsprechend aufgestellt. Die Zahlen sind jedoch vor dem Hintergrund der insgesamt sehr geringen Zahl an europäischen Modellen kritisch zu betrachten.

[9] Mistrals im Oktober 2026 veröffentlichtes Modell mistral/mistral-large-4 soll Ende Oktober 2026 auf Hugging Face verfügbar gemacht werden und eine Billionen Parameter aufweisen.

In den Daten zeigt sich ein strategischer Unterschied: Chinesische Anbieter veröffentlichen auch ihre komplexesten Modelle auf Hugging Face.[10] US-Anbieter stellen ihre leistungsfähigsten Modelle stattdessen nicht zum Download bereit, sondern über eigene API-Schnittstellen.

Abbildung 3 Verteilung der auf Hugging Face verfügbaren Modelle je Region nach Parameterbereichen.



Quelle: Eigene Erhebung

Tabelle 1 listet die jeweils größten Modelle der Regionen auf. Deutlich erkennbar hebt sich China bei den Modellgrößen, gemessen an der Anzahl der Parameter, ab. Anders als in den USA oder in der europäischen Region werden hier Parameterzahlen jenseits der Billionen-Marke (1T) mit Open-Weights-Modellen erreicht.

Tabelle 1: Überblick über die jeweils fünf größten Large Language Modelle je Region

Modell	Parameter[11]	Erstellt	Downloads abs./Alter
China			
moonshotai/Kimi-K3	2.8T	06/2026	1.757.174
Qwen/Qwen3.8-2.4T-A95B	2.4T	08/2026	84.899
meituan-longcat/LongCat-2.0	1.8T	07/2026	3.518
deepseek-ai/DeepSeek-V4-Pro-Base	1.6T	04/2026	28.132
nex-agi/Nex-N2.5-Max	1.6T	09/2026	12.870
Europa			
mistralai/Mistral-Medium-3.5-128B	128B	03/2026	181.428
mistralai/Devstral-2-123B-Instruct-2512	123B	12/2025	35.010

[10] Die vom chinesischen Handelsministerium angekündigten Exportkontrollen für KI-Modelle werden voraussichtlich diese Politik ändern. Die Regelungen beschränken die Verfügbarkeit von Modellen nach Trainingsaufwand, Modellgröße oder Leistung in gängigen KI-Benchmarks. Das schließt auch die Veröffentlichung von Gewichten und Modellen auf Hugging Face ein (Wu, 2026).

[11] Wir geben die Gesamtzahl der Parameter in Milliarden (B) oder Billionen (T) an. Mixture-of-Experts-Modelle (MoE) wie Qwen/Qwen3.8-2.4T-A95B haben neben Gesamtparametern eine niedrigere Zahl an aktiven Parametern, die bei Anfragen an das Model selektiv genutzt werden.

mistralai/Mistral-Large-Instruct-2407	122B	07/2024	193.601
mistralai/Mistral-Small-4-119B-2603	119B	03/2026	83.761
Infinimol/miiqu-f16	90B	03/2024	285
USA			
thinkingmachines/Inkling	952B	07/2026	616.361
nvidia/NVIDIA-Nemotron-3-Ultra.550B-A55B-Base-BF16	550B	06/2026	2.139
Snowflake/snowflake-arctic-base	482B	04/2024	560
meta-llama/Llama-3.1-405B	405B	07/2024	820.575
meta-llama/Llama-4-Maverick-17B-128E	400B	04/2025	4.754

Quelle: Eigene Erhebung

Europäische Modelle fallen bei der Obergrenze der Parameterzahl deutlich geringer aus als chinesische oder US-amerikanische Modelle. Vier der fünf größten Modelle aus Europa werden von Mistral bereitgestellt.

Tabelle 2 veranschaulicht die fünf Modelle mit den höchsten Downloadzahlen je Region, gewichtet nach dem Alter der Modelle. Drei der fünf am häufigsten heruntergeladenen Modelle aus Europa stammen aus dem Jahr 2023 und sind damit relativ alt. Auch nach den Downloadzahlen gemessen dominiert Mistral als europäischer Anbieter. Lediglich Swiss-AI hat mit einem sehr aktuellen Modell eine hohe Nachfrage erzielt. Die Zahlen verdeutlichen, dass in allen drei Regionen vor allem mittelgroße Modelle eine hohe Downloadnachfrage verzeichnen.

Tabelle 2: Überblick über die jeweils fünf häufigsten heruntergeladenen Large Language Modelle je Region

Modell	Parameter	Erstellt	Downloads abs. / Alter[12]
China			
Qwen/Qwen3.8-27B	27B	08/2026	12.979.013
Qwen/Qwen3-VL-8B-Instruct	8B	10/2025	6.538.505
zai-org/GLM-5.3-Flash	321B	08/2026	6.510.636
Qwen/Qwen3.6-35B-A3B	35B	04/2026	6.024.345
Qwen/Qwen3.6-27B	27B	04/2026	5.614.631
Europa			
mistralai/Mistral-7B-Instruct-v0.2	7B	12/2023	1.965.750
mistralai/Mistral-7B-v0.1	7B	09/2023	767.123
mistralai/Ministral-8B-Instruct-2410	8B	10/2024	370.876
swiss-ai/Apertus-v1.5-8B	8B	07/2026	351.278
mistralai/Mixtral-8x7B-v0.1	47B	12/2023	258.164
USA			
openai/gpt-oss-20b	20B	08/2025	6.323.616
openai/gpt-oss-120b	120B	08/2025	4.406.535
llava-hf/llava-1.5-7b-hf	7B	12/2023	1.471.892
thinkingmachines/Inkling	952B	07/2026	871.023
thinkingmachines/Inkling-Small	276B	07/2026	684.270

Quelle: Eigene Erhebung

3.3 Entwicklung im Zeitverlauf

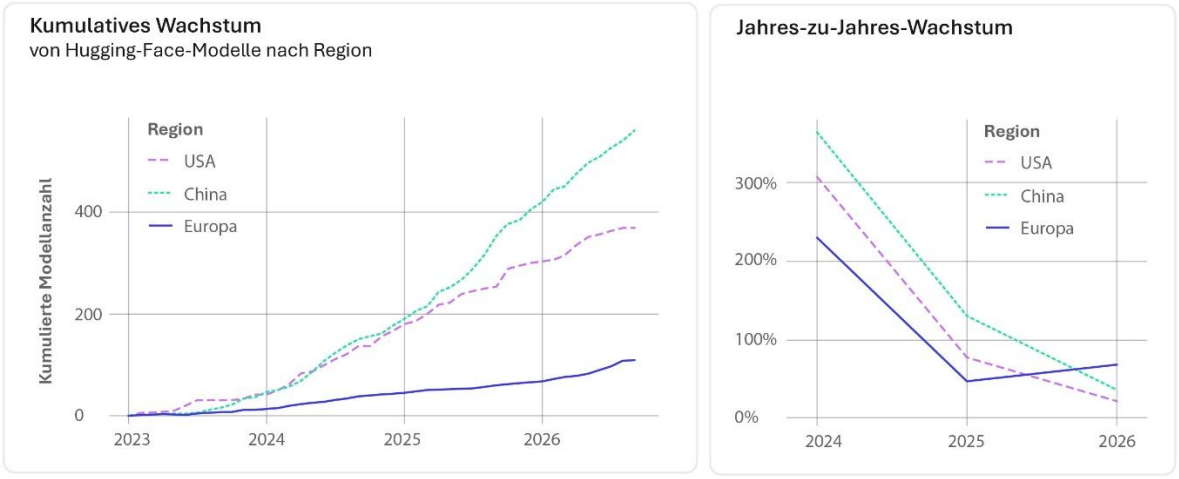
Die Zeitreihen in Abbildung 3 zeigen, wie sich die drei KI-Ökosysteme in den vergangenen Jahren entwickelt haben. Die Wachstumsquoten zeigen, dass Europa im laufenden Jahr 2026 (Januar bis September) vor den USA und China lag und erkennbar an Fahrt gewinnt. Die Dynamik bei US-

[12] Wir berechnen den Zahlenwert als Gesamtzahl der Downloads, geteilt durch das Modellalter in Monaten. So berücksichtigen wir, dass ältere Modelle über einen längeren Zeitraum mehr Downloads ansammeln konnten. Datenstand: Oktober 2026.

amerikanischen und chinesischen Anbieter:innen verzeichnet einen anhaltenden Rückgang. Die Geschwindigkeit, mit der beide neuen Modelle auf Hugging Face veröffentlicht werden, nimmt ab. Für die USA fällt die Quote im Jahr 2025 unter die Verdopplungsschwelle. Im laufenden Jahr 2026 unterschreitet auch China diese Schwelle.

Die USA und China dominieren die absoluten Veröffentlichungszahlen auf Hugging Face. Seit 2025 laufen beide Regionen jedoch erkennbar auseinander: China schließt den Rückstand aus den Jahren 2023 und 2024 auf und überholt die USA Ende 2024. Die USA wachsen seitdem langsamer.

Tabelle 3: Modellveröffentlichungen auf Hugging Face im Zeitverlauf (links) und relative Jahr-zu-Jahr-Wachstumsquote (rechts)



Quelle: Eigene Erhebung

3.4 Lizenzmodelle

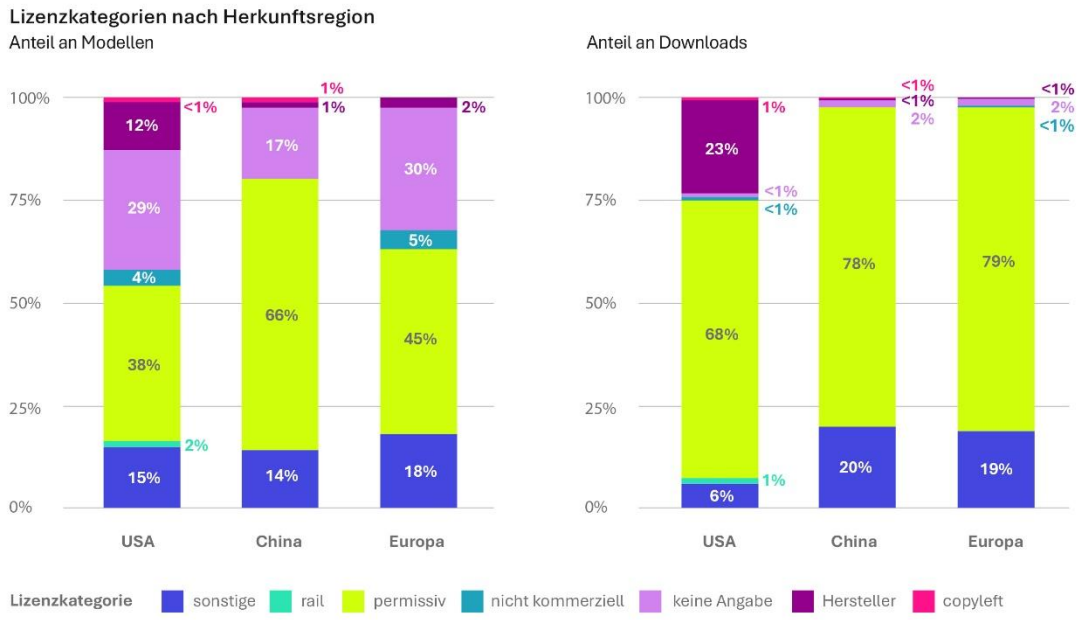
Verschiedene Lizenzmodelle sind das Aushängeschild eines offenen KI-Ökosystems, in dem Modelle auf Hugging Face eher permissiv oder restriktiv weiterverwendet werden können. Hier wird der Geist der Open-Source-Bewegung besonders deutlich, denn im Kern liegt die Möglichkeit, veröffentlichte Modelle je nach Lizenz zu modifizieren und kommerziell einzusetzen – wie Abbildung 4 zeigt, ist diese Möglichkeit bei 45 Prozent der europäischen Modelle tatsächlich gegeben. Ein Wert, der ausbaufähig ist und die Attraktivität der Modelle steigern kann.

Für Europa bietet sich die Chance, europäische Modelle insbesondere durch freizügige Lizenzen zu fördern und über die eigenen Grenzen hinweg zu verbreiten. Das heißt, die aktuell dominierende Veröffentlichungspraxis offener Gewichte könnte in einem weiteren Schritt in Richtung der Open-Source-Prinzipien weiterentwickelt werden, einschließlich der Offenlegung von Trainingsdaten und -methoden. So könnten Offenheit und Transparenz zu einem europäischen Qualitätsmerkmal werden.

Bislang sind europäische Modelle ähnlich strikt oder freizügig lizenziert wie Modelle aus China oder den USA. Zu den nicht kommerziell nutzbaren Modellen gehören vor allem solche, die unter Creative-Commons-Lizenzen veröffentlicht werden, die Derivate oder die explizite kommerzielle Nutzung untersagen. Verschiedene Entwicklungsfirmen wie Meta versehen ihre Modelle auch

mit eigenen Lizenzmodellen. Veröffentlichungen nach der RAIL (Responsible AI License) sind selten. Diese Lizenz soll den verantwortungsvollen Einsatz von Modellen und Source-Code regeln und ethisch bedenkliche Verwendungen, beispielsweise für Deepfakes oder Falschinformationen, ausschließen.

Abbildung 4 Lizenzkategorien nach Herkunftsregion



Quelle: Eigene Erhebung

4 Souveräne KI-Entwicklung und Einsatz durch Open-Weights-Modelle stützen

Die europäische Region ist bei der Entwicklung von Open-Weights-Modellen mittlerer Komplexität noch nicht verloren. Dafür sprechen zwei strukturelle Vorteile: Für die Entwicklung und das Training dieser Modelle werden weniger Ressourcen benötigt, und auch der Einsatz (Inferenz) erfordert weniger Rechenleistung. Daraus ergibt sich eine politische Chance, die europäische KI-Entwicklung in diesem Bereich als tragfähiges Standbein für den souveränen Einsatz von KI zu etablieren.

Wir sehen hierfür sechs notwendige Handlungsimpulse:

- Open-Weights-Initiativen mit europäischen Werten: Europäische Open-Weights-Modelle müssen dort punkten, wo US-amerikanische und chinesische Modelle intransparent bleiben oder die kommerzielle Nutzung ausschließen: bei Sicherheitsfragen und der Trainingsdatenkomposition einerseits – und bei Lizenzierungsmodellen andererseits. Ein konsequenter Fokus auf die zusätzliche Offenlegung von Trainingsdaten und Code neben den Gewichten sollte als Qualitätsmerkmal europäischer Modelle unbedingt gefördert werden.
- Europäische Modelle auf Hugging Face wie EuroLLM oder OpenGPT stammen aus Forschungsprojekten: In der Grundlagenforschung kann Europa auf belastbare Strukturen zurückgreifen. In öffentlich-privaten Konsortien können Synergien entstehen und der Zugang zu Rechenkapazität für das Training erleichtert werden. Mittelgroße Modelle sind ein geeignetes Format für kollaborative Entwicklung, in der öffentliche Hand, Forschung und Industrie ihre jeweiligen Stärken einbringen können.
- Forschung und Entwicklung braucht darüber hinaus für einen gelingenden Transfer eine aktualisierte Förder- und Nachnutzungsstrategie. Diese muss sicherstellen, dass die entstandenen Prototypen und Forschungsmodelle nach der Projektlaufzeit weiterentwickelt und gewartet werden, um eine fortlaufende Nutzung zu ermöglichen. Hierfür müssen über die Projektförderlaufzeiten hinaus Ressourcen bereitgestellt werden, um die Finanzierungslücke zwischen Projekt und Vermarktung zu schließen.
- Use Cases für den effizienten und produktiven Einsatz erproben. Erfahrungen der Verwaltung sowie der Industrie aus Leuchttürmen und Pilotprojekten müssen dokumentiert und vermittelt werden – insbesondere dort, wo Anwendungen ressourceneffiziente und produktive Ergebnisse liefern.
- Technische Beratungsangebote für Implementierung stärken. Viele Organisationen verfügen nicht über das Know-how, um Open-Weights-Modelle produktiv zu betreiben.[13] Daher braucht es tragfähige Geschäftsmodelle, die Beratung zur technischen und organisatorischen Einbettung und Anpassung von KI in betriebliche oder behördliche Arbeitsprozesse mit mittleren Open-Weights-Modellen attraktiv machen.
- Schließlich dürfen Offenheit und Souveränität nicht auf der Ebene der Modelle stehen bleiben. Ein souveränes, offenes Ökosystem braucht Verteilungsstrukturen für Modelle. Europäische Alternativen zu Hugging Face, wie die Initiative DeployAI, können kuratierte Open-Weights-Modelle verteilen, mit besonderem Augenmerk auf Lizenzierung oder auf einsehbare Trainingsdaten. Zusätzlich kann eine direkte Verbindung zu Rechenzentren für den produktiven Einsatz bereitgestellt werden.

[13] Mozilla Foundation, 2026

Literaturverzeichnis

Bitkom e.V. (2026, September 9). KI aus China? Für die deutsche Wirtschaft praktisch kein Thema. https://www.bitkom.org/Presse/Presseinformation/KI-aus-China-fuer-deutsche-Wirtschaft-kein-Thema

Bundesdruckerei (Hrsg.). (2026). Forschungsprojekt ADLER – effiziente und souveräne Sprachmodelle für die öffentliche Verwaltung [Whitepaper]. Bundesdruckerei Gruppe. https://www.bundesdruckerei.de/files/dokumente/pdf/Whitepaper_Forschungsprojekt_Adler.pdf

Ghosh, A., Kaffee, L.-A., Jernite, Y., & Solaiman, I. (2026, April 14). State of Open Source on Hugging Face: Spring 2026. huggingface.co. https://huggingface.co/blog/huggingface/state-of-os-hf-spring-2026

Lang, A.-S., Serve, V., & Lindern, J. von. (2026, Juni 5). KI-Unternehmen in Europa: Hat Europa das KI-Wettrennen wirklich schon verloren? Die ZEIT. https://www.zeit.de/wirtschaft/2026-06/ki-unternehmen-europa-mistral-aleph-alpha-anthropic

Mozilla Foundation. (2026). Open Source AI 2026 (Nr. V1.1). https://stateofopensource.ai/state-of-open-source-ai-v1-1.pdf

Open Source Initiative. (o.J.). What are Open Weights? https://opensource.org/ai/open-weights

Stapel, I. (2026). Souveräne KI: Strategie und Aufbau der eigenen Open-Source-Infrastruktur (1. Auflage). Rheinwerk Verlag.

Stein, S. (2025, November 20). CrowdStrike Researchers Identify Hidden Vulnerabilities in AI-Coded Software. CrowdStrike.Com. https://www.crowdstrike.com/en-us/blog/crowdstrike-researchers-identify-hidden-vulnerabilities-ai-coded-software/

Tan, E. (2026, September 4). Corporate America Is Getting Hooked on Open-Source A.I. The New York Times. https://www.nytimes.com/2026/09/04/technology/open-source-ai-anthropic-openai.html

Wu, Z. (2026, Oktober 16). China unveils draft export controls on AI technologies and chips. Financial Times. https://www.ft.com/content/e0a1295c-4492-4fdd-aaf5-55678a9ef9bf?syn-25a6b1a6=1

Ansprechperson bei der Agora Digitale Transformation

Yannick Kalff
Senior Policy Fellow KI und Public Interest Tech
yannick.kalff@agoradigital.de

Agora Digitale Transformation

www.agoradigital.de