

Kleine spezialisierte Modelle als europäischer KI-Weg

Was die verbreitete These trägt, wo sie bricht — und welche Fassung von ihr standhält

Datum 26.08.2026 — Autor Alessandro Salomon, stellvertretender Vorstand — Status Veröffentlicht am 15.09.2026 — Version 1.0

Veröffentlichte Fassung. Der Text lag Mitgliedern und Beirat des Instituts zur Querlesung vor. Der Vorstand hat die Freigabe am 15. September 2026 erteilt. **Zwei Abschnitte stehen unter Vorbehalt** und sind im Text gekennzeichnet; sie werden nachgeführt, sobald die offenen Fragen geklärt sind. Rückmeldungen an hello@dizk.de.

Das Wichtigste in Kürze

- **Für eine feste, wiederkehrende Aufgabe ist ein kleines Modell die bessere Wahl** — rund zehn- bis dreißigfach günstiger im Betrieb und gut genug. **Für ein ganzes Fachgebiet gilt das nicht:** In einer medizinischen Vergleichsstudie schlugen führende Allzweckmodelle spezialisierte Fachsysteme in allen drei Prüfbereichen.
- **Souveränität entsteht nicht aus der Modellgröße und nicht aus dem Herkunftsland**, sondern aus dem Besitz der Gewichte und der Verfügungsgewalt über die Daten. Im Juni 2026 musste ein US-Anbieter auf Anordnung seiner Regierung allen Ausländern den Zugang zu zwei Spitzenmodellen sperren, über zwei Wochen lang.
- **Der eigentliche Engpass ist nicht die Rechenleistung, sondern sind die Daten.** Seit August 2026 gibt es dafür einen öffentlichen Preis: zehn Millionen Dollar für die Betriebsdaten einer insolventen Fluggesellschaft.
- **Und hier steht Europa besser da, als die Debatte nahelegt.** Das Urheberrecht lässt die Verfügungsgewalt bei den Urhebern, die KI-Verordnung macht die Herkunft der Trainingsdaten prüfbar. Genutzt wird beides bisher kaum.

Die Position: Europas Weg führt nicht über das größte Modell, sondern über die Schicht darüber: über Modelle, die **eine Aufgabe** können und nicht ein Fachgebiet, deren Gewichte man besitzt und die dort rechnen, wo die Daten liegen. Große Rechenzentren bleiben notwendig, um Grundlagenmodelle zu **erschaffen** — nicht als Ort, an dem Europas Arbeit stattfindet.

1. Zwei Bewegungen zur selben Zeit

Wer 2026 über Künstliche Intelligenz liest, sieht meist nur eine der beiden Bewegungen.

Die erste ist laut. Ein chinesischer Konzern trainiert nach Presseberichten ein Modell mit bis zu zehn Billionen Parametern. Das bislang größte chinesische Modell hat 2,8 Billionen, das stärkste offen verfügbare 2,4 Billionen; für das westliche Spitzenmodell nennen Branchenschätzungen rund acht Billionen. Die wichtigste Einschränkung steht in derselben Meldung: Aus der Parameterzahl allein lässt sich nicht auf die Leistungsfähigkeit schließen.

Parallel baut Europa die Infrastruktur, die solche Läufe ermöglicht. Am 30. Juli 2026 startete die Ausschreibung für die KI-Gigafabriken: bis zu zehn Milliarden Euro öffentliche Mittel, zwanzig Milliarden von privat, im ersten Los vier Rechenzentren mit der Leistung von 75.000 Beschleunigern einer heute vier Jahre alten Generation. Die Haltung dahinter fasste ein Verantwortlicher in einen Satz: **Man wolle nicht hundert kleine Rechenzentren.** Drei Details gehören in jede ehrliche Souveränitätsdebatte: Die Anlagen müssen auf Unionsgebiet stehen und unter europäischer Kontrolle, bei den Beschleunigern bleibt man von wenigen Herstellern abhängig, und Deutschland ist erst in der zweiten Runde dabei.

Die zweite Bewegung ist leise — und für die meisten Organisationen die wichtigere. Anfang August 2026 wurden die Gewichte eines Modells mit 27 Milliarden Parametern frei verfügbar. In voller Rechengenauigkeit bräuchte es rund 120 Gigabyte Speicher; komprimiert passt es in 22 Gigabyte, in einer optimierten Fassung in

16 — also in eine Grafikkarte ab etwa 420 bis 550 Euro. Ein Fachtest ließ dieses Modell ohne Rückfragen eine vollständige Programmierschnittstelle mit Benutzerverwaltung und Rechteverwaltung entwickeln. Nach zwölf Minuten übersetzte sich der Code beim ersten Versuch; das Urteil der Tester lautete, das Ergebnis sei von einem kommerziellen Spitzenmodell nicht zu unterscheiden.

Eine Landschaftsaufnahme derselben Woche bringt die Lage auf eine Formel: **Die chinesischen Sprachmodelle werden größer, die amerikanischen dagegen werden offen.** Europa kommt darin mit kleinen Modellen vor — aus der Schweiz in verbesserter Fassung, aus Deutschland erstmals überhaupt, dort mit Startschwierigkeiten. Diese Dreiteilung ist wichtiger, als sie aussieht: Offene Gewichte kommen inzwischen auch aus den USA. Souveränität und Herkunft sind zwei verschiedene Fragen.

2. Die These und ihre drei Behauptungen

Die geprüfte These lautet: **Kleine spezialisierte Modelle mit rund 27 Milliarden Parametern sind für eine festgelegte Aufgabe besser geeignet als große Alleskönner aus den USA. Damit macht man sich unabhängig und braucht keine großen Rechenzentren.** Sie enthält drei Behauptungen, die getrennt geprüft werden müssen.

Behauptung	Ergebnis der Prüfung
A — klein und spezialisiert schlägt groß und allgemein	nur eingeschränkt. Trägt für eine wiederkehrende Aufgabe. Für ein ganzes Fachgebiet widerlegt (Abschnitt 4).
B — das macht unabhängig	ja, aber anders begründet. Nicht die Größe macht souverän, sondern Besitz der Gewichte und Verfügungsgewalt über die Daten (Abschnitte 3 und 5).
C — dafür braucht es keine großen Rechenzentren	in dieser Form nicht haltbar. Kleine Modelle werden aus großen destilliert. Richtig wird der Satz erst getrennt: für den Betrieb nein, für die Erschaffung ja.

Zwei Begriffe entscheiden über das Ergebnis. **Klein** heißt in der Forschung nicht schwach, sondern: Das Modell passt auf ein Gerät, das eine einzelne Person besitzen kann. Diese Größenklasse ist die Grenze zwischen Rechenleistung mieten und Rechenleistung besitzen. **Spezialisiert** ist doppeldeutig: Es kann heißen, auf ein Fachgebiet ausgerichtet — oder auf eine Aufgabe zugeschnitten. Diese beiden Bedeutungen führen zu entgegengesetzten Ergebnissen. Das ist der Kern von Abschnitt 4.

3. Was für die These spricht

3.1 Für eine feste Aufgabe reicht klein

Der meistzitierte Beleg stammt von einem Grafikkartenhersteller, dessen Geschäft eher am Gegenteil hängt. Eine Forschungsarbeit aus dem Jahr 2025 legt dar, dass kleine Modelle für den Betrieb von KI-Agenten hinreichend leistungsfähig, ihrer Natur nach besser geeignet und notwendigerweise wirtschaftlicher seien. Die Zahlen sind ungewöhnlich klar: Ein Modell mit sieben Milliarden Parametern zu betreiben ist **zehn- bis dreißigmal günstiger** als eines mit siebzig bis 175 Milliarden, gerechnet in Wartezeit, Energieverbrauch und Rechenoperationen. In drei untersuchten Agentensystemen ließen sich **40 bis 70 Prozent** der Anfragen an das große Modell durch ein kleines ersetzen.

Entscheidend ist die Bedingung, unter der das gilt, und sie steht in der Zusammenfassung: Es geht um Agenten, die **eine kleine Zahl spezialisierter Aufgaben wiederholt und mit wenig Abweichung** ausführen. Der Grund ist banal und wird selten ausgesprochen: Ein Vorgang, der immer dieselbe Entscheidung verlangt, braucht kein Universalwissen. Wo dagegen allgemeine Gesprächsfähigkeit gebraucht wird, empfehlen dieselben Autoren gemischte Systeme.

Bemerkenswert ist, was die Autoren als Hindernisse benennen — es sind keine technischen: die bereits getätigten Investitionen in zentrale Recheninfrastruktur, die Gewohnheit, auch Spezialmodelle an Allgemeinbildungstests zu messen, und fehlende Bekanntheit. Die Vorentscheidung für große Modelle ist damit vor allem eine wirtschaftliche und kulturelle, keine sachliche.

3.2 Die Rechnung im Betrieb — und warum sie sich verschiebt

Für den lokalen Betrieb ist die wichtigste Kennzahl der Arbeitsspeicher, nicht die Rechengeschwindigkeit. Als Faustregel für komprimierte Modelle gilt: rund ein halbes Gigabyte je Milliarde Parameter, plus etwa zwei Gigabyte für den laufenden Kontext.

Modellgröße	Speicherbedarf komprimiert	Läuft auf
7 bis 8 Milliarden	etwa 6 bis 8 GB	Gerät der 16-GB-Klasse, auch Notebook
13 bis 14 Milliarden	etwa 12 bis 16 GB	Gerät der 24- bis 32-GB-Klasse
rund 27 Milliarden	etwa 17 bis 22 GB	eine Grafikkarte ab etwa 550 Euro
rund 70 Milliarden	etwa 35 bis 48 GB	64 GB oder mehrere Karten

Dazu die laufenden Kosten. Ein kompakter Rechner mit gemeinsamem Speicher zieht unter Last etwa 30 bis 40 Watt, eine Hochleistungs-Grafikkarte 350 bis 575 Watt. Wer heute 150 bis 250 Euro monatlich für Nutzungsgebühren ausgibt, hat einen Einstiegsrechner der 600-Euro-Klasse in **unter sechs Monaten** eingespielt.

Das Inferenzparadoxon: Ein Marktforschungshaus erwartet, dass sich die Kosten je agentischem Arbeitsablauf bis 2028 **auf mehr als das Fünffache** erhöhen — und zwar **obwohl** die Preise je Modell sinken. Der Grund: Sinkende Preise machen aufwendigere Abläufe wirtschaftlich, der Verbrauch steigt schneller, als der Preis fällt. Dasselbe Haus erwartet, dass **mehr als 40 Prozent** der geplanten oder eingeführten KI-Agenten bis 2027 wieder eingestellt werden — wegen steigender Kosten, ausbleibenden Nutzens und unzureichender Risikokontrolle.

Für die Frage nach dem eigenen Betrieb heißt das: Der sinkende Modellpreis ist das schwächste Argument für die Miete. Er lässt die Rechnung langsamer wachsen, nicht kleiner. Wer den Verbrauch nicht in der Hand hat, hat die Kosten nicht in der Hand.

3.3 Was ein eigener Test zeigt

Die folgenden Beobachtungen stammen aus einem dokumentierten Auswahlverfahren für eine lokal betriebene Fachanwendung mit hinterlegten Rechtsquellen. Geprüft wurde mit einem festen Aufgabensatz und vorher festgelegten Bestehensschwellen; zwischen den Durchläufen wurde ausschließlich das Modell getauscht. Die Modelle sind namentlich genannt, damit die Beobachtung nachprüfbar bleibt; sie ist **keine Anbieterbewertung**, sondern gilt für genau eine Aufgabe.

Das aufwendigere Vorgehen war nicht das bessere. Ein mit eigenem Aufwand angepasstes Mistral-7B (zwei Adapter) wurde von **Qwen2.5-7B** geschlagen, einem frei verfügbaren Basismodell derselben Größenklasse — **ohne jede Anpassung**: 52 von 60 Punkten gegenüber 48, Trefferquote der Quellensuche 0,90 gegenüber 0,80. Die beiden eigenen Anpassungen lagen mit 40 und 33 Punkten **unter** dem unveränderten Ausgangsmodell.

Die Lehre daraus: Der Hebel ist selten das Training, sondern der Datensatz und die Quellensuche davor. Das ist die gute Nachricht für kleine Organisationen: Der teure Teil ist der, den man nicht braucht. Und die unbequeme: Der Teil, den man braucht, ist Ordnung in den eigenen Daten.

3.4 Das Recht steht auf dieser Seite

Ein Argument, das in der technischen Debatte fast nie vorkommt, in der politischen aber entscheidend ist: Die europäische KI-Verordnung behandelt große und kleine Modelle unterschiedlich. Ein Modell mit allgemeinem Verwendungszweck gilt nach Artikel 51 als Modell mit systemischem Risiko, sobald für sein Training mehr als **10 hoch 25** Rechenoperationen aufgewendet wurden. Dann greifen die Pflichten aus Artikel 55: Modellbewertungen, Prüfung auf Angriffsszenarien, Meldung schwerwiegender Vorfälle, Cybersicherheit und Berichte über den Energieverbrauch. Für eine Organisation, die ein Modell der 27-Milliarden-Klasse zuschneidet und im eigenen Haus betreibt, ist dieser Teil nicht einschlägig. **Der kleine Weg ist damit auch regulatorisch der einfachere.**

3.5 Der Vorfall, der die Souveränitätsfrage aus dem Konjunktiv holt

Bis hierher ist Souveränität ein Argument. Im Juni 2026 wurde sie ein Vorgang. Auf Anordnung der US-Regierung sperrte ein kalifornischer Anbieter den Zugang zu seinen neuesten Spitzenmodellen — nicht für einzelne Verdächtige, sondern **für alle Ausländer**. Begründung: nationale Sicherheit. Die Sperre wurde nach mehr als zwei Wochen wieder aufgehoben.

Zwei Wochen, ohne Vorwarnung, ohne Vertragsbruch, ohne dass ein europäischer Kunde etwas hätte tun können. Die Europäische Kommission hat darauf mit einem Aktionsplan reagiert und entwickelt Notfallmaßnahmen für den Fall, dass ein Drittstaat den Zugang zu KI mit bedrohlichen Fähigkeiten einschränkt. Die zuständige Kommissarin fasste die Lehre so zusammen: Bei Fähigkeiten, die für die eigene Sicherheit entscheidend sind, könne man sich nicht allein auf außereuropäische Lösungen verlassen; das sei kostspielig, aber die Kosten, führende Modelle **nicht** aufzubauen, seien höher.

Dem lässt sich ein Bild aus einem Praxistest gegenüberstellen. Ein Journalist traute dem Ergebnis seines lokalen 27-Milliarden-Modells nicht — es war zu gut. Um sicherzugehen, dass nicht heimlich ein Cloud-Modell mitrechnete, zog er das Netzkabel aus dem Router. Das System arbeitete weiter. **Ein abgeschalteter Zugang, gegen den kein Vertrag hilft. Und ein gezogenes Kabel, das nichts ändert.**

4. Was die These aushalten muss

4.1 Die stärkste Gegenevidenz: wenn Spezialisierung verliert

Ein Forschungsteam einer US-amerikanischen Universitätsklinik hat 2026 in einer medizinischen Fachzeitschrift untersucht, was die These im Kern behauptet, nur in der anderen Richtung: Sind spezialisierte klinische KI-Systeme bei medizinischen Fragen besser als allgemeine Spitzenmodelle? Geprüft wurden zwei am Markt etablierte Fachsysteme gegen die drei führenden Allzweckmodelle, in drei Durchgängen.

Prüfung	Allgemeine Spitzenmodelle	Spezialisierte Fachsysteme
500 Fragen im Stil der ärztlichen Zulassungsprüfung	97,4 % / 94,2 % / 90,2 %	89,6 % / 88,4 %
500 Aufgaben, Bewertung nach ärztlichen Kriterien (von 100)	88,0 / 79,3 / 77,0	62,6 / 61,3
100 echte Anfragen von Ärztinnen und Ärzten, von zwölf Medizinern verblindet bewertet (1 bis 4)	3,62 / 3,54 / 3,52	3,24 / 3,17

In allen drei Bereichen lagen die Allzweckmodelle vorn. Aus 1.800 Einzelbewertungen ergaben sich drei weitere Befunde: Bei der Sicherheit war kein statistisch signifikanter Unterschied messbar, doch in den Freitextanmerkungen fielen bei einem der Fachsysteme unvollständige klinische Inhalte und sicherheitsrelevante Auslassungen besonders häufig auf. Das zweite Fachsystem **verweigerte 19 Prozent** der realen Anfragen die Antwort, gegenüber einem bis drei Prozent bei den allgemeinen Modellen. Und der bitterste Wert: Die **gewöhnliche Suchmaschine mit KI-Antwort** erreichte 3,27 und lag damit auf dem Niveau beider Fachsysteme.

Was daraus folgt — und was nicht: Es folgt: Der Vorteil kleiner, zugeschnittener Modelle trägt in der Bedeutung **auf eine Aufgabe zugeschnitten**, nicht in der Bedeutung **auf ein Fachgebiet ausgerichtet**. Wer ein System kauft, weil das Etikett **für die Medizin** oder **für das Recht** darauf steht, verlässt sich auf etwas, das diese Studie nicht hergibt. — **Es folgt nicht**, dass lokale Modelle unterlegen wären: Die Studie hat keine getestet und nennt keine Parameterzahlen. Verglichen wurden **Produkte**, nicht Größenklassen.

Die Autorinnen und Autoren vermuten, dass die breiteren Trainingsdaten und schnelleren Entwicklungszyklen der führenden Modelle stärker wiegen als eine nachträgliche Spezialisierung. Sie schränken selbst ein, dass stark spezialisierte Teilgebiete, komplexe lokale Arbeitsabläufe oder eigene Modelle einer Einrichtung anders ausfallen können. Ihre Empfehlung deckt sich mit dem Befund aus Abschnitt 3.3: Entscheidend sei, wie gut ein System in **unabhängigen Tests auf realistischen Aufgaben** funktioniert, nicht ob es als Spezialprodukt vermarktet wird.

4.2 Das kleine Modell ist ein Kind des großen Laufs

Die guten kleinen Modelle entstehen nicht dadurch, dass jemand ein kleines Modell trainiert. Sie entstehen durch **Destillation aus einem großen**. Der technische Bericht einer führenden Modellreihe beschreibt das Verfahren offen: Das Wissen der Flaggschiffe wird auf die leichten Modelle der Reihe übertragen; für ein Modell der Acht-Milliarden-Klasse liefert das bessere Ergebnisse als bestärkendes Lernen, zu einem Bruchteil der Rechenkosten.

Im Klartext: **Ohne den Lauf mit Billionen Parametern gäbe es das Modell mit 27 Milliarden nicht**. Wer den kleinen Weg für den Verzicht auf große Rechenzentren hält, hat die Lieferkette nicht zu Ende gelesen. Die Behauptung wird erst in der Trennung richtig: **für den Betrieb nicht nötig, für die Erschaffung schon**. Das ist keine akademische Unterscheidung — sie entscheidet, ob die europäischen Milliarden für Rechenzentren eine Fehlinvestition sind oder die Voraussetzung dafür, dass Europa eigene Grundlagenmodelle hat, aus denen sich kleine ableiten lassen.

4.3 Herkunft ist nicht Souveränität

Das Modell, das derzeit als beste selbst betreibbare Alternative gilt, kommt aus China. Wer ein US-Abonnement dadurch ersetzt und das Souveränität nennt, hat den Begriff nicht verstanden. Der Unterschied liegt woanders: Ein Abonnement ist eine laufende Beziehung — der Anbieter sieht die Daten, ändert Preise und Regeln und kann abschalten. **Frei verfügbare Gewichte sind ein Gegenstand**. Sie liegen auf der eigenen Festplatte. Niemand kann sie verteuern, mitlesen oder abschalten, und keine fremde Regierung kann den Hersteller dazu anweisen.

Der Prüfsatz: Souveränität ist keine Frage der Flagge auf dem Modell, sondern eine Frage des Besitzes. Nicht: **Wo wurde es gebaut?** Sondern: **Kann es mir jemand wegnehmen, verteuern oder mitlesen?** — Die Herkunft bleibt trotzdem relevant, weil Lizenzen, eingebaute Beschränkungen und Wertvorstellungen mitkommen. Sie ist nur nicht das entscheidende Kriterium.

4.4 Europas eigene Modelle sind noch nicht so weit

Die Landschaft existiert: ein europäisches Modell der 22-Milliarden-Klasse seit Ende 2025, ein Schweizer Modell, das sogar seine Trainingsdaten offenlegt, ein deutsches Forschungsmodell für die EU-Amtssprachen, ein Verbund aus rund zwanzig Forschungseinrichtungen. Die Einordnung in der Fachpresse ist zurückhaltend: nicht an der Weltspitze, Forschungsmodelle kleiner Größenklasse, Engpässe bei der Rechenkapazität, das erste genuin deutsche Modell mit Startschwierigkeiten.

Belegt ist das auch durch einen eigenen Fehlschlag. In dem in Abschnitt 3.3 beschriebenen Auswahlverfahren war **EuroLLM-9B-Instruct** aus Souveränitätsgründen der liebste Kandidat. Es erreichte **22 statt der geforderten 48 Punkte** in der einen und **2 von 40** in der anderen Kategorie; ein Kontrolllauf mit identischen Einstellungen schloss einen Konfigurationsfehler aus. Wichtig für die Einordnung: **SauerkrautLM-Nemo-12B**, ebenfalls europäisch und deutsch angepasst, **bestand dieselbe Prüfung** mit 47 Punkten.

Ein Nachtrag, der in beide Richtungen zeigt. Das Schweizer Team hat angekündigt, aus seinem Modell **Fachmodelle für Medizin, Recht und den Finanzsektor** abzuleiten. Das ist genau der Weg, dem die Studie aus Abschnitt 4.1 den Vorsprung abspricht — und zugleich die einzige Ausnahme, die deren Autorinnen und Autoren selbst einräumen: Modelle einer Einrichtung selbst könnten anders ausfallen als zugekaufte Fachprodukte. **Ob der Unterschied trägt, wird sich messen lassen.** Das ist eine der wenigen Stellen, an denen die Frage dieses Papiers in absehbarer Zeit empirisch entschieden wird.

Die Lehre: Ein Modell ist nicht gut oder schlecht, sondern **passend oder unpassend** — und das entscheidet die Aufgabe, nicht die Herkunft und nicht die Rangliste. Wer europäische Modelle aus Prinzip beschafft, ohne sie an der eigenen Aufgabe zu messen, bekommt kein souveränes System, sondern ein schlechtes. **Souveränität, die man nicht prüft, ist eine Behauptung.** Es ist derselbe Fehler wie bei den Fachprodukten in Abschnitt 4.1, nur mit einem anderen Etikett.

4.5 Klein heißt auch anfälliger

Die ehrlichste Beschreibung der Grenzen stammt von jemandem, der von der Technik begeistert war. Nach einem ausführlichen Praxistest mit einem lokalen 27-Milliarden-Modell benannte der Tester drei Schwächen: **Halluzinationen** — auf die Bitte um zwanzig wenig bekannte Sehenswürdigkeiten seiner Stadt waren alle zwanzig frei erfunden, jede mit überzeugenden Details. **Angreifbarkeit** — kleine Modelle sind anfälliger für

versteckte Anweisungen in Webseiten und Dokumenten. Und **Sprachgefühl** — im Umgang mit Sprache lag das Modell erkennbar hinter kommerziellen Angeboten.

Dazu zwei praktische Hürden, die in Werbeaussagen nie vorkommen: Für agentische Arbeit muss der Kurzzeitspeicher des Modells mindestens 65.000 Zeichenbausteine fassen, besser 200.000 — und dieser Speicher kostet zusätzlichen Arbeitsspeicher. Und die Einrichtung ist Facharbeit. Das Fazit des Testers blieb trotz aller Begeisterung: brauchbar, aber fehleranfällig, und noch nicht auf dem Niveau eines Spitzenmodells.

Die Betriebsregel, die daraus folgt: Kleine Modelle gehören dorthin, wo eine falsche Antwort auffällt — in Abläufe mit hinterlegten Quellen, mit Prüfschritt, mit menschlicher Freigabe. Nicht dorthin, wo frei aus dem Gedächtnis geantwortet wird. Das ist zugleich ein Argument für die enge Fassung der These: Der aufgabenbezogene Einsatz ist der einzige, in dem die Schwäche kleiner Modelle beherrschbar bleibt.

4.6 Offene Gewichte sind nicht dasselbe wie quelloffen

Wer Unabhängigkeit über frei verfügbare Modelle sucht, muss das Kleingedruckte lesen. Eine Untersuchung von Ende 2025 kam zu dem Ergebnis, dass viele Modelle unter falscher Flagge segeln: als quelloffen beworben, aber mit Einschränkungen der kommerziellen Nutzung; Lizenzen, die nur die kleineren Varianten freigeben; Bedingungen, die bestimmte Einsatzzwecke ausschließen. Nach den strengen Kriterien der einschlägigen Organisation erfüllten zum Erhebungszeitpunkt nur **zwei** Modelle weltweit die Anforderungen: **Apertus** aus der Schweiz und **Olmo 2** aus den USA. Auch bei europäischen Modellen gilt das im Übrigen: Eines der bekanntesten ist in der verbreiteten Fassung **nicht kommerziell nutzbar**. Wer beschafft, prüft den Einsatzzweck gegen die Lizenz, bevor er baut.

Das ist die Zahl, an der sich die Debatte messen lassen muss: **Von allen bekannten Sprachmodellen sind zwei tatsächlich offen. Eines davon ist europäisch.**

5. Die Frage, die alles vorentscheidet: Wem gehören die Trainingsdaten?

Bis hierher ging es um Modelle, Rechenleistung und Betrieb. Der eigentliche Engpass liegt woanders. **Ein Modell ist im Kern eine Funktion seiner Trainingsdaten**. Wer über Daten verfügt, kann Modelle bauen; wer nicht verfügt, kann sie nur mieten. Diese Frage entscheidet über den europäischen Weg mehr als jede Parameterzahl — und sie ist in der öffentlichen Debatte unterbelichtet.

5.1 Betriebsdaten haben einen öffentlichen Marktpreis

Im **August 2026** ersteigerte ein US-Konzern im Insolvenzverfahren einer Fluggesellschaft deren anonymisierte Unternehmensdaten für **zehn Millionen Dollar** und überbot damit einen Mitbieter mit 7,5 Millionen. Marktbeobachter werten den Vorgang als erstes öffentliches Preissignal für betriebliche Unternehmensdaten als eigenständiges KI-Gut.

Dahinter steht ein Markt, der in vier Jahren aus dem Nichts entstanden ist: **null** öffentlich bekannte Lizenzabschlüsse 2022, zwölf 2023, achtundzwanzig 2024, rund sechshunddreißig für 2026 erwartet; das Volumen wird für 2026 auf etwa 4,4 Milliarden Dollar geschätzt.

Der Satz, der mit Technik nichts zu tun hat: Betriebsdaten sind ein Vermögenswert — und Vermögenswerte werden in der Insolvenz verwertet. Wer heute Daten in fremde Systeme gibt, entscheidet damit auch über einen Fall, den er später nicht mehr steuert.

5.2 Für die meisten ist der Preis trotzdem null

Ein Markt zahlt für Knappheit und Verhandlungsmacht. Wer beides nicht hat, bekommt nichts. Das strukturelle Problem lässt sich so beschreiben: Solange die Teilnahme am Index einer Suchmaschine faktisch auch die Teilnahme an deren KI-Trainingspipeline bedeutet, ist der Preis für diese Daten **effektiv null** — die Alternative wäre der vollständige Ausschluss aus der Sichtbarkeit, und diese Rechnung kann sich kaum jemand leisten. Deshalb ist Lizenzieren keine Antwort auf die Frage nach Datensouveränität. Es ist eine Antwort für die wenigen, die einen großen, exklusiven Bestand kontrollieren.

5.3 Die Rechtslage in Europa: erlaubt, solange niemand maschinenlesbar widerspricht

Der europäische Rahmen ist präziser, als die Debatte vermuten lässt, und er steht seit 2021 im Gesetz. Nach **Paragraf 44b des Urheberrechtsgesetzes** ist die Vervielfältigung rechtmäßig zugänglicher Werke zum Zweck des Text- und Datenauswertens erlaubt — **aber nur, wenn der Rechtsinhaber sich die Nutzung nicht**

vorbehalten hat. Und dieser Vorbehalt ist bei online zugänglichen Werken **nur wirksam, wenn er in maschinenlesbarer Form erfolgt.** Für die wissenschaftliche Forschung gilt eine weitere Schranke, die keinen solchen Vorbehalt kennt.

An diesem einen Wort — **maschinenlesbar** — hängt in Europa die Verfügungsgewalt über Trainingsdaten. Wie schwer es wiegt, zeigt eine Entscheidung des Oberlandesgerichts Hamburg vom 10. Dezember 2025. Eine Bildagentur hatte automatisierte Downloads in ihren Nutzungsbedingungen ausdrücklich untersagt — in natürlicher Sprache und zusätzlich lesbar im Quellcode der Seite. Der Senat ließ **ausdrücklich offen**, ob ein in natürlicher Sprache formulierter Vorbehalt den Anforderungen des § 44b Abs. 3 Satz 2 UrhG genügt; die Frage sei umstritten. Entschieden hat er über die **Darlegungslast**: Wer sich auf einen Vorbehalt beruft, muss darlegen, dass dieser **zum Zeitpunkt der Nutzung** maschinenlesbar war — nach dem damaligen Stand der Technik. Diesen Nachweis sah der Senat nicht als geführt an (OLG Hamburg, Az. 5 U 104/24; Vorinstanz LG Hamburg, Az. 310 O 227/23; die Revision ist beim Bundesgerichtshof anhängig, Az. I ZR 281/25).

Die praktische Lehre für alle, die Inhalte veröffentlichen: Ob ein Vorbehalt in natürlicher Sprache genügt, ist ungeklärt — wer sich darauf verlässt, trägt das Risiko. Vor Gericht muss der Rechteinhaber **belegen**, dass sein Vorbehalt zum Zeitpunkt der Nutzung maschinenlesbar war. Wer nicht möchte, dass seine Inhalte in Trainingsdaten landen, setzt den Vorbehalt deshalb technisch und dokumentiert, seit wann — und zwar, bevor gesammelt wird. Ein Vorbehalt wirkt nur nach vorn.

Für die Ausgabeseite hat das Landgericht München I am 11. November 2025 eine Gegenlinie gezogen: Die **Speicherung geschützter Texte im Modell** stellt eine urheberrechtlich relevante Vervielfältigung dar; die Schranke für unwesentliches Beiwerk greife nicht, und der Anbieter sei keine privilegierte Forschungseinrichtung (Az. 42 O 14139/24). In einem zweiten Verfahren bestätigte dasselbe Gericht am 31. Juli 2026, dass die Verantwortung für urheberrechtlich geschützte Trainingsdaten **beim Anbieter der KI** liegt (Az. 42 O 763/25). **Beide Urteile sind nicht rechtskräftig.**

Vorbehalt zu diesem Abschnitt: Die Rechtsfrage ist in Deutschland **höchstrichterlich weiterhin nicht geklärt.** Der I. Zivilsenat des Bundesgerichtshofs hat die Revision im oben genannten Verfahren am **3. September 2026** verhandelt (Aktenzeichen **I ZR 281/25**). Eine Entscheidung war zum Redaktionsschluss dieser Fassung am **15. September 2026** weder in der Entscheidungsdatenbank noch in den Pressemitteilungen des Gerichts veröffentlicht. Zur Klärung stehen unter anderem: Ist der Aufbau eines Trainingsdatensatzes noch Text- und Datenauswertung? Bleibt eine Forschungseinrichtung privilegiert, wenn ihr Datensatz anschließend kommerziellen Anbietern offensteht? Und welche Anforderungen gelten genau an die Maschinenlesbarkeit? **Dieser Abschnitt wird nachgeführt, sobald die Entscheidung vorliegt.**

5.4 In den USA gilt ein anderer Maßstab, mit einer bemerkenswerten Trennlinie

Ein US-Bundesgericht hat entschieden, dass das **Training** auf rechtmäßig erworbenen Büchern im Kern umgestaltend und damit vom dortigen Schrankenrecht gedeckt ist. **Nicht** gedeckt war, wie sich der Anbieter die Bücher beschafft hatte: Das Herunterladen und Vorhalten rechtswidriger Kopien blieb unzulässig. Der daraus folgende Vergleich wurde im Juli 2026 endgültig genehmigt — **1,5 Milliarden Dollar** für rund 500.000 Werke, etwa 3.000 Dollar je Werk, verbunden mit der Vernichtung des rechtswidrig beschafften Datensatzes. Er betrifft ausdrücklich nur die Vergangenheit, nicht die Ausgaben des Modells.

Die Trennlinie ist die eigentliche Nachricht: **Nicht das Lernen war das Problem, sondern die Herkunft der Daten.** Genau diese Unterscheidung fehlt in der europäischen Debatte oft — hier wird meist über das Training an sich gestritten, dort über die Beschaffung.

5.5 Die KI-Verordnung macht die Herkunft der Daten zur Beschaffungsfrage

Seit dem **2. August 2025** gilt für Anbieter von Modellen mit allgemeinem Verwendungszweck Artikel 53 der KI-Verordnung: Sie müssen eine Strategie zur Einhaltung des Urheberrechts vorhalten, einschließlich der Beachtung maschinenlesbarer Vorbehalte, und eine **öffentlich zugängliche, hinreichend detaillierte Zusammenfassung der Trainingsinhalte** veröffentlichen, nach einer Vorlage des europäischen KI-Büros.

Das ist der Hebel, der in Beschaffungsverfahren bisher kaum genutzt wird. Denn aus der Transparenzpflicht des Anbieters wird eine **Prüfpflicht des Beschaffers**: Wer ein Modell einsetzt, kann und sollte nach dieser Zusammenfassung fragen. Ob sie viel hergibt, ist eine andere Frage — Beobachter halten die bislang veröffentlichten Zusammenfassungen für dünn. Aber sie ist eine **prüfbare Anforderung**, und prüfbare Anforderungen sind selten in diesem Feld.

Wo Offenlegung an ihre Grenze kommt: Alles bisher Gesagte gilt für **allgemein verfügbare Modelle im Markt**. Für Systeme im hoheitlichen Kernbereich — Sicherheitsbehörden, Nachrichtendienste, Verteidigung — ist Offenlegung weder daten- noch geheimchutzrechtlich durchsetzbar. **Das Institut hat dazu bereits Stellung genommen:** Ein Positionspapier vom Juni 2026 zur automatisierten Datenanalyse in Sicherheitsbehörden plädiert dort ausdrücklich **nicht** für vollständige Offenlegung, sondern für ein **abgestuftes Auditmodell** — unabhängige Sachverständige erhalten unter Geheimchutz Zugriff auf Quellcode, Modelllogik, Software-Stückliste und Validierungsdaten. Geprüft wird, veröffentlicht wird nicht. **Der Grundsatz bleibt derselbe, nur der Adressat wechselt:** nicht „jeder muss es sehen können“, sondern „jemand Unabhängiges muss es prüfen können“.

5.6 Es geht auch sauber — und der Preis ist bekannt

Dass rechtstreues Training keine Theorie ist, zeigt ein europäisches Beispiel. Ein Schweizer Forschungsmodell mit 70 Milliarden Parametern und über 1.000 Sprachen legt nicht nur die Gewichte offen, sondern auch die **Trainingsdaten, die Filterpipeline, die Zwischenstände und die Nachbearbeitungsverfahren**. Obwohl für ein Modell dieser Größe die strengsten Anforderungen der KI-Verordnung gar nicht gelten, haben die Forschenden sich ihnen **freiwillig unterworfen**. Es heißt **Apertus** und ist damit eines der beiden Modelle aus Abschnitt 4.6, die alles offenlegen.

Der bemerkenswerteste Teil ist ein **nachträglicher Widerspruch**: Die Forschenden fragen die aktuellen Einwilligungen ab und bereinigen damit rückwirkend die Datenbestände der Jahre 2013 bis 2024. Damit lösen sie das Problem, an dem der Rechtevorberehalt sonst scheitert — dass ein Widerspruch von heute die Daten von gestern nicht mehr erreicht.

Die Zahl, die man sich merken sollte: Durch den Verzicht auf urheberrechtlich bedenkliches Material gehen **rund 8 Prozent der englischen und 4 Prozent der mehrsprachigen Daten** verloren. Das Ergebnis liegt in Standardprüfungen zum Allgemeinwissen dennoch etwa auf dem Stand vergleichbarer offener Modelle seiner Generation. **Rechtstreue kostet rund 8 Prozent der Trainingsdaten — und ist damit machbar.** Das ist die beste verfügbare Antwort auf den Einwand, europäische Regeln machten wettbewerbsfähige Modelle unmöglich. Sie machen sie etwas teurer. Nicht unmöglich.

Hinweis zur Belegtiefe: Die Zahlen zum Schweizer Modell sind am Originalartikel geprüft. Die zitierten Gerichtsentscheidungen wurden bislang über Fachbesprechungen ausgewertet, nicht im Volltext.

5.7 Sauber beschaffen ist nur die halbe Aufgabe

Ein Filter vor dem Training löst das Beschaffungsproblem. Er löst nicht das Problem, das Abschnitt 5.3 juristisch entschieden hat: **Ein Modell, das Trainingsdaten wörtlich wiedergibt, vervielfältigt sie** — unabhängig davon, wie sauber sie beschafft wurden. Die Münchner Urteile knüpfen die Rechtsverletzung genau daran.

Die üblichen Gegenmaßnahmen setzen spät an. Anbieter bringen ihren Modellen in der **Nachbearbeitung** bei, bestimmte Ausgaben zu verweigern. Das ist aufwendig, es braucht mühsam zusammengestellte Datensätze, und es hält nicht zuverlässig: Forschende haben an mehreren Spitzenmodellen gezeigt, dass sich solche Sperren mit geschicktem Fragen oder gezielten Nachtrainings-Angriffen umgehen lassen.

Das Schweizer Team geht einen anderen Weg und setzt **schon im Vortraining** an, also in der ersten Stufe. Ein Verfahren namens **Goldfish Loss** verändert über eine gezielte Maskierung die Fehlerfunktion so, dass das Modell das wörtliche Nachsprechen gar nicht erst als optimal lernt. Es soll dadurch grundsätzlich vom Original abweichend formulieren, statt das Nachsprechen später zu verbieten.

Warum das für die Rechtsdebatte zählt: Wenn sich das Nachsprechen bereits im **Bauverfahren** des Modells unterdrücken lässt, ist das Auswendiglernen keine unvermeidliche Eigenschaft großer Modelle mehr, sondern eine **Konstruktionsentscheidung**. Damit verschiebt sich die Frage: nicht mehr, ob ein Modell memorisiert, sondern **ob der Hersteller etwas dagegen getan hat**. Rechtstreue ist danach nicht nur bezahlbar, sondern technisch gestaltbar — und wer sie nicht gestaltet, hat sich dagegen entschieden. Das ist ein anderer Satz als: Das ließ sich nicht vermeiden.

Belegstand: Das Verfahren ist in der Quelle beschrieben, seine Wirksamkeit dort aber nicht beziffert. Dieser Abschnitt behauptet deshalb keinen Erfolg, sondern einen Ansatzpunkt.

5.8 Was daraus für die These folgt

Die Datenfrage stützt die zweite Behauptung — **und verschiebt ihre Begründung ein zweites Mal**. Bisher lautete die Antwort: Souveränität ist Besitz der Gewichte plus Ort der Verarbeitung. Das bleibt richtig, ist aber unvollständig.

Vollständig lautet sie: Souverän ist, wer über die Daten verfügt, aus denen ein Modell entsteht — und wer die Verfügungsgewalt behält, wenn sich Recht, Preise oder Eigentümer ändern.

Für Europa ist das eine **bessere** Ausgangslage, als die Debatte über Rechenzentren nahelegt. Das europäische Recht lässt die Verfügungsgewalt grundsätzlich bei den Urhebern, die KI-Verordnung macht die Datenherkunft zu einer prüfbaren Eigenschaft, und das Schweizer Beispiel zeigt, dass ein rechtstreu trainiertes Modell konkurrenzfähig sein kann. Was fehlt, ist nicht der Rahmen. Es fehlt die Nutzung des Rahmens: **maschinenlesbare Vorbehalte, die tatsächlich gesetzt werden, und Beschaffungsverfahren, die tatsächlich nach der Datenherkunft fragen.**

6. Empfehlungen

Für Betriebe und Organisationen

- **Die Frage lautet nicht, welches Modell das beste ist, sondern welche Aufgabe es lösen soll.** Und Aufgabe heißt: ein Vorgang, der sich wiederholt — nicht ein Fachgebiet.
- **Vor der Beschaffung prüfen, nicht danach.** Zwanzig echte Fälle aus dem eigenen Haus, vorher festgelegte Bestehensschwellen, alle Kandidaten unter gleichen Bedingungen. Aufwand: ein bis zwei Tage, danach beliebig wiederholbar.
- **Misstrauen gegen Domänen-Etiketten.** Ein als Fachlösung vermarktetes System muss dieselbe Prüfung bestehen wie ein allgemeines.
- **In die Daten investieren, nicht ins Training.** Eine saubere Quellensuche über die eigenen Unterlagen bringt mehr als jeder Anpassungslauf und ist billiger.
- **Klein einsteigen.** Die Sieben- bis Acht-Milliarden-Klasse auf einem Gerät der 16-GB-Klasse deckt Zusammenfassen, Klassifizieren, Recherche in eigenen Unterlagen und Formulierhilfe ab. Die 27er-Klasse ist der nächste Schritt, nicht der erste.
- **Kleine Modelle nur mit Netz und doppeltem Boden** — hinterlegte Quellen, Prüfschritt, menschliche Freigabe, kein Vollzugriff auf Produktivsysteme.
- **Den Nutzungsvorbehalt maschinenlesbar setzen, heute.** Ein Verbot in den Nutzungsbedingungen genügt nach der Rechtsprechung nicht, und ein Vorbehalt wirkt nur nach vorn.

Für die öffentliche Hand

- **Souveränität als prüfbares Vergabekriterium formulieren:** Liegen die Gewichte vor? Wo wird verarbeitet? Was passiert, wenn der Zugang von außen gesperrt wird? Nicht als Herkunftsangabe.
- **Die Zusammenfassung der Trainingsdaten nach Artikel 53 zum Vergabekriterium machen.** Sie ist Pflicht, öffentlich und prüfbar — drei Eigenschaften, die sonst kaum ein Kriterium in diesem Feld hat.
- **Keine Vergabeentscheidung allein auf veröffentlichte Bestwerte stützen.** Sie sind ein Vorfilter und beantworten eine andere Frage als die des Beschaffers.
- **Die eigene Prüfung ist Verwaltungsaufgabe, nicht Lieferantenaufgabe.**
- **Verwaltungsdaten als Vermögenswert behandeln.** Sie sind nicht ersetzbar und nicht wiederbeschaffbar; ihre Verarbeitung außer Haus ist keine reine IT-Entscheidung.

Für die europäische Politik

- **Die großen Rechenzentren sind kein Widerspruch zum kleinen Weg, sondern seine Voraussetzung** — sofern das, was dort entsteht, als offene Gewichte herauskommt. Ein europäisches Grundlagenmodell hinter einer Schnittstelle wäre die teuerste Art, dieselbe Abhängigkeit noch einmal zu bauen.
- **Die Schwelle in Artikel 51 ist ein Standortvorteil, kein Hemmnis.**
- **Der Rechtevorbereitung ist der stärkste europäische Hebel und praktisch unbenutzt.** Wer ihn stärken will, muss bei der technischen Durchsetzbarkeit ansetzen: einheitliche Formate, Prüfbarkeit, und die Frage, was mit bereits gesammeltem Material geschieht. Das Schweizer Beispiel zeigt, dass ein nachträglicher Widerspruch technisch möglich ist — er ist bislang nur nirgends Pflicht.

- **Rechtstreue kostet rund 8 Prozent der Trainingsdaten, nicht die Wettbewerbsfähigkeit.** Diese Zahl gehört in jede Debatte über europäische Regeln als angeblichen Standortnachteil.
- **Rechenzeit, die nicht bei denen ankommt, die Modelle bauen, ist kein Beitrag zur Souveränität.** Während über Milliarden für neue Anlagen entschieden wird, berichten bestehende europäische Modellvorhaben von Kapazitätsengpässen.

7. Fazit

Die Ausgangsthese ist richtig — aber an einer anderen Stelle, als sie es selbst vermutet.

Der Teil, der am sichersten wirkte, ist der schwächste. Spezialisiert schlägt allgemein gilt nicht für ein ganzes Fachgebiet. Wo es geprüft wurde, verloren die Fachsysteme gegen die Allzweckmodelle, und zwar in allen drei Durchgängen. Was bleibt, ist der engere, aber solide Satz: Für **eine Aufgabe, die sich wiederholt**, ist das kleine Modell die bessere Wahl.

Der Teil, der nach Wunschdenken klang, hat den härtesten Beleg. Im Juni 2026 hat eine fremde Regierung angeordnet, allen Europäern den Zugang zu zwei Spitzenmodellen zu sperren. Dagegen half kein Vertrag, kein Datenschutzargument und kein europäischer Rechenzentrumsstandort. Es half nur, die Gewichte selbst zu haben.

Der Teil, der falsch war, ist der nützlichste. Kleine Modelle werden aus großen destilliert. Wer das weiß, versteht, warum Europa **beides** braucht — große Rechenzentren, um Grundlagenmodelle zu erschaffen, und die Grafikkarte für 550 Euro, um damit zu arbeiten. Das ist kein Widerspruch, sondern eine Arbeitsteilung.

Und der wichtigste Teil stand gar nicht in der These. Die Debatte dreht sich um Rechenzentren, weil Rechenzentren sichtbar sind. Der Engpass sind die Daten. Ausgerechnet hier steht Europa besser da, als jede Debatte über Gigawatt und Milliarden nahelegt: Das Recht lässt die Verfügungsgewalt bei den Urhebern, die KI-Verordnung macht die Datenherkunft prüfbar, und ein Schweizer Forschungsmodell hat vorgerechnet, dass Rechtstreue rund acht Prozent der Trainingsdaten kostet — nicht die Wettbewerbsfähigkeit. **Was fehlt, ist nicht der Rahmen. Es fehlt, dass ihn jemand benutzt.**

Die Leitlinie: Nicht das größte Modell. Nicht das mit dem passenden Etikett. Das geprüfte — im eigenen Haus, an den eigenen Aufgaben, mit Daten, über die man verfügt.

Zur Methodik und Herkunft

Die externen Befunde stammen aus Fachpublikationen und Primärquellen des Jahres 2026: einer begutachteten medizinischen Vergleichsstudie, einer Forschungsarbeit zu kleinen Modellen in Agentensystemen, dem technischen Bericht einer Modellreihe, Fachberichterstattung zu Modellveröffentlichungen und Praxistests, den Rechtstexten der europäischen KI-Verordnung und des Urheberrechtsgesetzes sowie Fachbesprechungen deutscher und US-amerikanischer Gerichtsentscheidungen. Marktzahlen zum Datenmarkt sind Schätzungen und als Größenordnung zu lesen. Die Gerichtsentscheidungen wurden über Fachbesprechungen ausgewertet, nicht im Volltext; die beiden deutschen Urteile sind nicht rechtskräftig; die höchstrichterliche Klärung steht unmittelbar bevor. Dieser Punkt ist im Text als Vorbehalt gekennzeichnet. Die Zahlen zum Schweizer Modell sind am Originalartikel geprüft.

Die Beobachtungen in den Abschnitten 3.3 und 4.4 stammen aus einem dokumentierten Auswahlverfahren aus der Praxis eines Vorstandsmitglieds. Die geprüften Modelle sind dort namentlich genannt, das anwendende Unternehmen und die Fachanwendung nicht. Für die in Abschnitt 4.1 genannten Fachsysteme wurde bewusst auf Produktnamen verzichtet; die Studie ist öffentlich zugänglich.

Quellen

Aufgeführt sind die Belege, auf denen die Aussagen dieses Papiers ruhen. Eigene Beobachtungen aus der Praxis (Abschnitte 3.3 und 4.4) haben keine externe Quelle und sind im Text als solche gekennzeichnet. Stand aller Abrufe: 15. September 2026.

Zu Abschnitt 3 — Was für die These spricht

Belcak et al. (NVIDIA): „Small Language Models are the Future of Agentic AI“, arXiv:2506.02153 — <https://arxiv.org/abs/2506.02153>

Gartner, Pressemitteilung vom 17.08.2026, zu den Inferenzkosten je agentischem Arbeitsschritt; zugänglich über heise online, 18.08.2026 — <https://www.heise.de/news/Prognose-Inferenzkosten-fuer-KI-Agenten-verfuehnen-sich-bis-2028-11417470.html>

Google: „Introducing Gemma 3“ — <https://blog.google/technology/developers/gemma-3/>

heise online, 16.08.2026, Jan Mahn: „Lokale KI ausprobiert“ — <https://www.heise.de/-11415293>

heise online, 08.07.2026: „Europa stärkt KI-Sicherheit nach US-Modell-Blockade“ — <https://www.heise.de/-11358285>

Zu Abschnitt 4 — Was die These aushalten muss

Originalstudie zur Gegenevidenz, Nature Medicine: „General-purpose large language models outperform specialized clinical AI tools on medical queries“ — <https://doi.org/10.1038/s41591-026-04431-5>

Apertus Tech Report, Swiss AI Initiative (Primärquelle) — <https://github.com/swiss-ai/apertus-tech-report>

ETH Zürich, Medienmitteilung vom 02.09.2025 zu Apertus — <https://ethz.ch/de/news-und-veranstaltungen/eth-news.html>

Modellkarte Apertus-70B — <https://huggingface.co/swiss-ai/Apertus-70B-2509>

c't/heise, 01.12.2025, Andrea Trinkwalder: „Open-Source-KI: Was es gibt und wie offen die Modelle wirklich sind“ —

<https://www.heise.de/hintergrund/Open-Source-KI-Was-es-gibt-und-wie-offen-die-Modelle-wirklich-sind-10903968.html>

Zu Abschnitt 5 — Wem gehören die Trainingsdaten?

§ 44b UrhG, Text und Data Mining — https://www.gesetze-im-internet.de/urhg/_44b.html

Verordnung (EU) 2024/1689 (KI-Verordnung), Artikel 53 — <https://ai-act-service-desk.ec.europa.eu/de/ai-act/article-53>

KI-Verordnung, Erwägungsgrund 107 — <https://ai-act-law.eu/de/erwg/107/>

OLG Hamburg, Urteil vom 10.12.2025, Az. 5 U 104/24; Vorinstanz LG Hamburg, Urteil vom 27.09.2024, Az. 310 O 227/23; Revision beim BGH anhängig, Az. I ZR 281/25, Verhandlung am 03.09.2026

LG München I, Urteil vom 11.11.2025, Az. 42 O 14139/24 — nicht rechtskräftig

LG München I, Urteil vom 31.07.2026, Az. 42 O 763/25 — nicht rechtskräftig

Kritische Einordnung der Münchner Entscheidung, Verfassungsblog — <https://verfassungsblog.de/lg-muenchen-gema-open-ai/>

Zum US-amerikanischen Maßstab: Bartz v. Anthropic PBC, United States District Court for the Northern District of California, Az. 4:24-cv-05417; Berufung: Ninth Circuit, Az. 25-4843 — <https://www.courtlistener.com/docket/69058235/bartz-v-anthropic-pbc/>

heise/WiWo, 18.08.2026, zum Verkauf der Betriebsdaten einer insolventen Fluggesellschaft — <https://www.heise.de/-11416749>

Hinweis zu den Verweisen

Alle Netzverweise wurden am 15. September 2026 aufgerufen. Gerichtsentscheidungen sind zusätzlich über ihr Aktenzeichen zitiert, damit sie unabhängig vom Fortbestand einer Adresse auffindbar bleiben.

Über den Autor



Alessandro Salomon ist stellvertretender Vorsitzender des DIZK Deutsches Institut für zukunftsfähige KI e. V. und CTO der TrendTec UG (haftungsbeschränkt). Ausgebildeter Fachinformatiker für Anwendungsentwicklung, interner Auditor nach ISO 19011:2018.

Seine Arbeit liegt dort, wo KI-Systeme nicht entstehen, sondern laufen: Auslieferung, Schnittstellen, Datenwege — und die Frage, wer im Nachhinein noch nachvollziehen kann, warum ein System etwas getan hat. Die Beobachtungen in den Abschnitten 3.3 und 4.4 stammen aus dieser Praxis.

Interessenlage: Der Autor ist technischer Geschäftsführer eines Unternehmens, das KI-Lösungen für den Mittelstand entwickelt und betreibt. Er hat damit ein wirtschaftliches Interesse an dem Feld, über das dieses Papier urteilt. Das Institut hält die Offenlegung für den ehrlicheren Weg als den Verzicht auf Autoren mit Praxiserfahrung.

Kontakt zum Autor: stellv-vorsitz@dizk.de · Rückmeldungen zum Papier: hello@dizk.de

Dieses Positionspapier entstand mit Unterstützung eines KI-Assistenten; Recherche, Prüfung und Freigabe liegen beim Vorstand und den Mitgliedern des DIZK.

DIZK Deutsches Institut für zukunftsfähige KI e. V. hello@dizk.de dizk.de/publikationen

Satz in Bahnschrift.