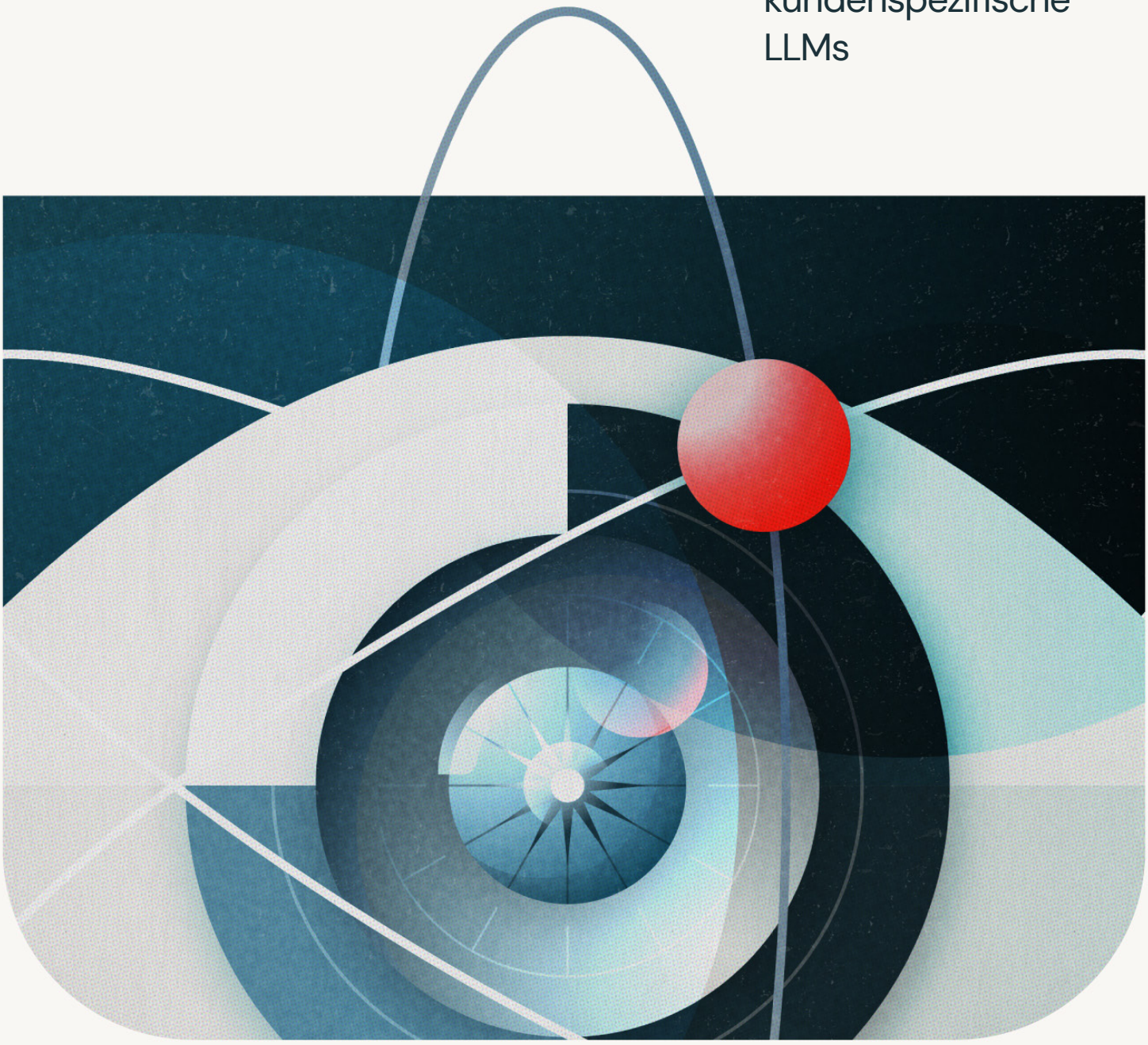
databricks

DATEN UND KI: DER AKTUELLE STAND

Data Intelligence und der Wettstreit um kundenspezifische LLMs



Unternehmen drängen darauf, Daten und KI zu demokratisieren

Einführung

Generative KI läutet eine neue Ära von Innovation, Kreativität und Produktivität ein. Gerade einmal 18 Monate nachdem das in die Mainstream-Diskussionen eingezogen ist, investieren Unternehmen überall in GenAI, um ihre Strukturen zu transformieren. Sie haben erkannt, dass ihre Daten entscheidend sind, um ihren Nutzern ein hochwertiges GenAI-Erlebnis bieten zu können. Die drängendste Frage unter Führungskräften lautet jetzt: *Wie können wir unsere Daten am schnellsten und besten integrieren?*

Isolierte Daten- und KI-Plattformen machen es Teams schwer, ihre GenAI-Projekte voranzutreiben – unabhängig davon, ob sie natürliche Sprache zur Datenabfrage nutzen oder intelligente Apps entwickeln. Wir sind davon überzeugt, dass Data-Intelligence-Plattformen eine radikale Datendemokratisierung in den Unternehmen bewirken werden. Diese neue Plattformkategorie setzt auf GenAI, um Daten einfacher zu erfassen und zu nutzen, und senkt die technischen Hürden bei der Wertschöpfung daraus. Bei unseren eigenen Kunden beobachten wir bereits eine deutliche Beschleunigung der KI-Nutzung.

Der Report *Daten und KI: Der aktuelle Stand* vermittelt einen Überblick über die Prioritäten, die Unternehmen bei Daten- und KI-Initiativen setzen. Die Erkenntnisse stammen von mehr als 10.000 Kunden weltweit – darunter über 300 der Fortune 500 –, die die Databricks Data Intelligence Platform nutzen. Finden Sie heraus, wie die innovativsten Unternehmen mit maschinellem Lernen auf Erfolgskurs gelangen, wie sie GenAI einsetzen und die sich wandelnden Governance-Anforderungen bewältigen.

Dieser Report soll Unternehmen bei der Entwicklung wirksamer Datenstrategien mit einer Enterprise-KI helfen, die sich im ständigen Wandel befindet.

Wesentliche Erkenntnisse

11× mehr KI-Modelle wurden in diesem Jahr in die Produktion überführt

Nachdem die Unternehmen mit der KI jahrelang nur experimentiert haben, implementieren sie jetzt wesentlich mehr Modelle in der Praxis als noch vor einem Jahr.

Im Schnitt sind die Unternehmen heute bei der Einführung von Modellen in die Produktion mehr als dreimal so effizient.

Dabei ist Natural Language Processing (NLP) die meistgenutzte und am schnellsten wachsende Anwendung für maschinelles Lernen.

70 % der Unternehmen, die GenAI nutzen, erweitern ihre Basismodelle mithilfe von Tools und Vektordatenbanken

Nach der Integration wurde LangChain innerhalb von noch nicht einmal einem Jahr zu einem der meistverwendeten Daten- und KI-Produkte.

Unternehmen arbeiten intensiv daran, LLMs unter Verwendung von von Retrieval Augmented Generation (RAG) mit ihren eigenen Daten anzupassen.

RAG erfordert Vektordatenbanken, deren Nutzung im Vergleich zum Vorjahr um 377 % gestiegen ist (Nutzung einschließlich Open-Source- und Closed-Source-LLMs).

76 % der Unternehmen, die LLMs nutzen, entscheiden sich für Open Source, oft parallel zu proprietären Modellen

Viele Unternehmen setzen auf kleinere Open-Source-Modelle als akzeptablen Kompromiss zwischen Kosten, Leistung und Latenz.

Nur 4 Wochen nach der Markteinführung macht Meta Llama 3 bereits 39 % der gesamten Nutzung von Open-Source-Modellen aus.

Überraschenderweise treten bei GenAI vor allem stark regulierte Branchen als Vorreiter auf. Der Finanzdienstleistungssektor – Spitzenreiter bei der GPU-Nutzung – entwickelt sich mit einem Wachstum von 88 % innerhalb von 6 Monaten diesbezüglich am schnellsten.

Methodik: Wie hat Databricks diesen Report erstellt?

Der Report *Daten und KI: der aktuelle Stand 2024* basiert auf vollständig aggregierten und anonymisierten Daten, die wir von unseren Kunden zur Nutzung der Databricks Data Intelligence Platform und ihres weitreichenden Integrationsökosystems erhoben haben.

Dieser Report befasst sich schwerpunktmäßig mit Trends im maschinellen Lernen, der Akzeptanz von GenAI, Integrationen und Anwendungsfällen. Die Kunden kommen aus allen wichtigen Branchen, und ihre Bandbreite reicht von Start-ups bis hin zu vielen internationalen Global Playern. Sofern nicht anders angegeben, werden in diesem Bericht Daten aus dem Zeitraum vom 1. Februar 2023 bis zum 31. März 2024 präsentiert und analysiert, und die Nutzung wird anhand der Anzahl der Kunden gemessen. Wenn möglich, stellen wir Vorjahresvergleiche an, um Entwicklungen im zeitlichen Verlauf aufzuzeigen.

Maschinelles Lernen

KI ist in der Produktion angekommen

UNTERNEHMEN WETTEIFERN UM DEN EINSATZ VON ML-MODELLEN IN DER PRODUKTION

In diesem Jahr haben wir bei der KI eine Verlagerung vom Experimentieren hin zu Produktionsanwendungen erlebt. Mit dem Aufschwung des maschinellen Lernens (ML) lernen die Unternehmen jetzt, mit den beiden separaten Lebenszyklushälften von ML-Modellen zurechtzukommen. Zunächst erstellen Unternehmen ihre ML-Modelle durch *experimentelles Testen*, um durch Einsatz verschiedener Algorithmen und Hyperparameter die besten Modelle zu finden. Erst dann werden diese Modelle in die *Produktion* überführt. In diesem Prozess haben die Teams zwei widerstreitende Ziele: Einerseits müssen sie sicherstellen, dass die Experimentierphase so zeiteffizient wie möglich abläuft, andererseits dürfen sie nur strikt ausgetestete Modelle in die Produktion geben.

Die Implementierung von Modellen in der Produktion war in der Vergangenheit mit Schwierigkeiten verbunden: uneinheitliche Daten- und KI-Plattformen, komplexe Implementierungs-Workflows, fehlende Zugriffskontrollen für die Governance, Mangel an Kontrollmöglichkeiten usw. Unsere Daten zeigen, wie Unternehmen diese Herausforderungen mit der Einführung von Data-Intelligence-Plattformen bewältigen.

Unternehmen setzen verstärkt auf ML in der Produktion

Daten von MLflow (einer von Databricks entwickelten Open-Source-MLOps-Plattform) zeigen, wie häufig unsere Kunden Modelle *protokollieren* (was für Experimente steht) und *registrieren* (wenn sie sie in die Produktion einführen).

Die Ergebnisse belegen, dass nicht nur mehr experimentiert wird, sondern dass die Unternehmen auch wesentlich mehr Effizienz beim Übergang in die Produktion an den Tag legen.

VERHÄLTNIS VON PROTOKOLLIERTEN EXPERIMENTEN ZU REGISTRIERTEN MODELLEN

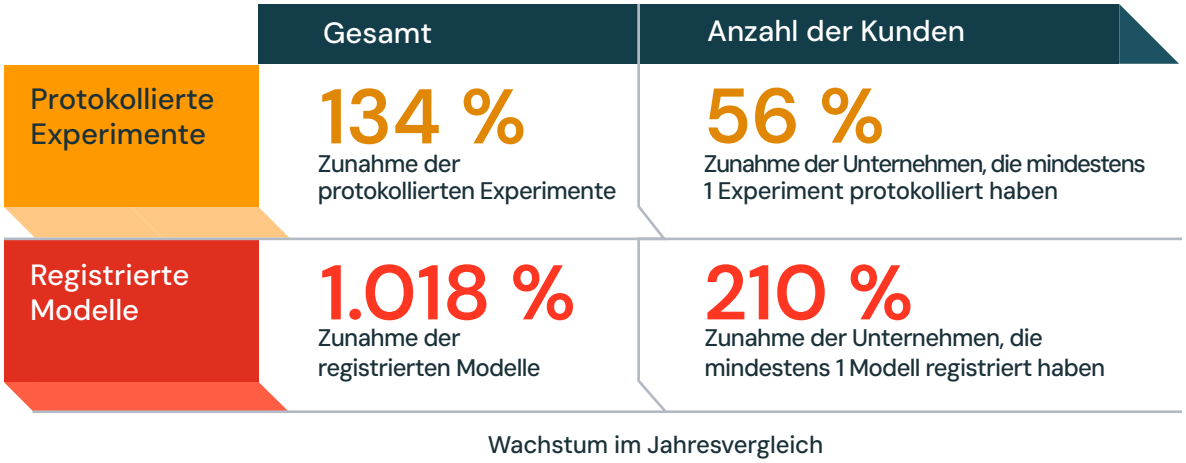
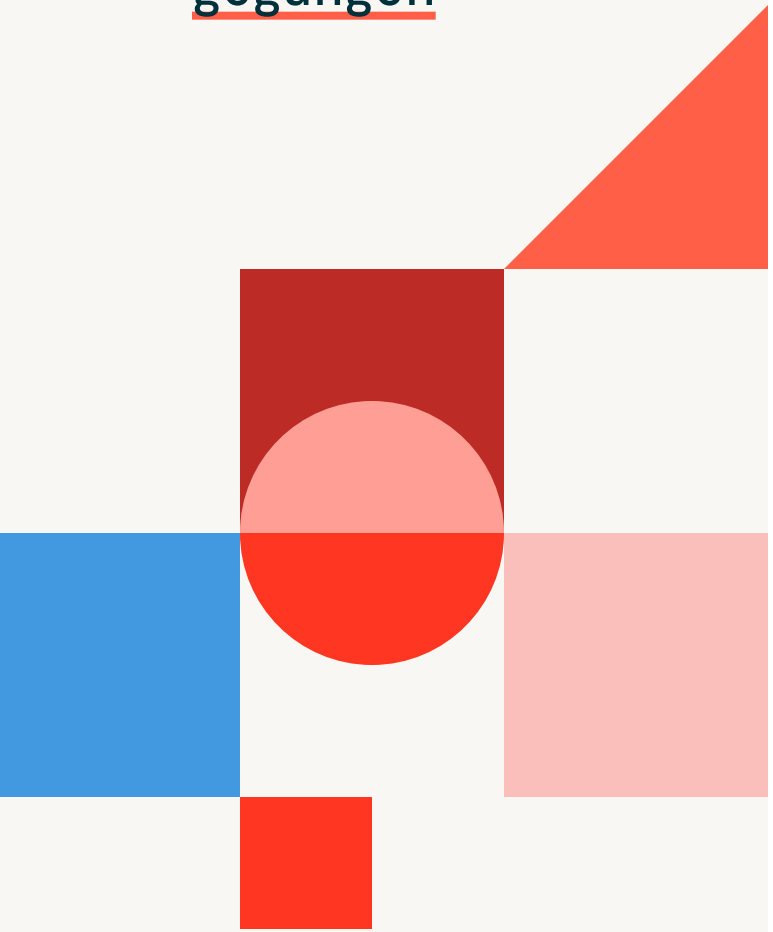


Abbildung 1: Im Jahresvergleich hat das Wachstum der registrierten Modelle das Wachstum der protokollierten Experimente deutlich übertroffen. Das deutet darauf hin, dass die Unternehmen verstärkt von der Experimentier- in die Produktionsphase wechseln.

Ein Riesenschritt: 11 Mal mehr Modelle sind in Produktion gegangen



Die Anzahl der Modelle hat spür- und messbar zugenommen.

DIE ZAHL DER UNTERNEHMEN, DIE IN ML INVESTIEREN, IST SPRUNGHAFT ANGESTIEGEN

Unsere Daten zeigen, dass 56 % mehr Unternehmen experimentelle Modelle protokollieren als noch vor einem Jahr, gleichzeitig aber 210 % mehr Modelle registrieren. Das deutet darauf hin, dass viele Unternehmen, die sich im letzten Jahr vorrangig mit Experimenten befassten, nun zur Produktion übergegangen sind.

DIE ZAHL DER ML-MODELLE IST BEI ALLEN UNTERNEHMEN GESTIEGEN

Nach Jahren intensiver Beschäftigung mit Experimenten steigen die Unternehmen jetzt in die Produktion ein. Die Registrierung von Modellen stieg in diesem Jahr um 1.018 % – deutlich mehr als der Zuwachs von 134 % bei den erfassten Experimenten. Diese Tendenz erkennen wir auch auf Unternehmensebene. Im Durchschnitt hat ein Unternehmen in diesem Jahr 261 % mehr Modelle registriert und 50 % mehr Experimente protokolliert.

DIE QUINTESSENZ

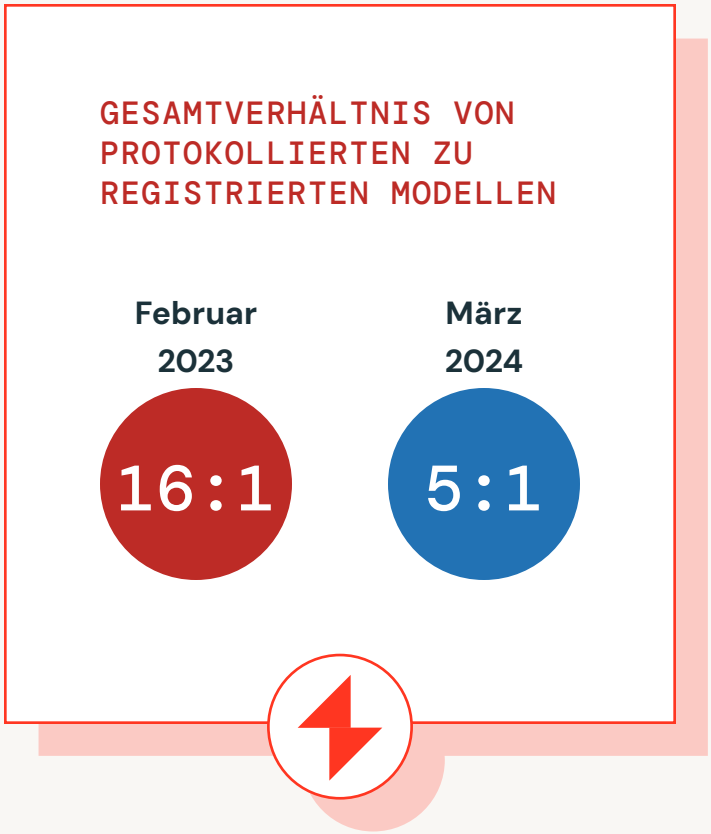
ML ist entscheidend dafür, wie innovativ Unternehmen sind und wie sie sich von der Konkurrenz abheben. Da die Kompetenz der Unternehmen in diesem Bereich weiter fortschreitet, ist davon auszugehen, dass sich dieser Trend in den kommenden Jahren weiter fortsetzen wird. Der neuere Zweig der GenAI befindet sich noch in der Testphase, aber die Unternehmen fangen an, Fuß zu fassen.

Die Unternehmen sind heute dreimal so effizient bei der Einführung von Modellen in die Produktion.

Die ML-Effizienz hat einen echten Nutzwert, der in Zeit, Geld und Ressourcen gemessen werden kann. Modellentwicklung und -erprobung sind zwar entscheidend, doch müssen sich diese Modelle letztlich in realen Anwendungsfällen bewähren, um den geschäftlichen Mehrwert zu erhöhen.

Wir haben das Verhältnis von protokollierten zu registrierten Modellen bei allen Kunden ausgewertet, um den Fortschritt zu beurteilen. Im Februar 2023 lag das Verhältnis von protokollierten zu registrierten Modellen bei 16:1. Das bedeutet, dass für jeweils 16 experimentelle Modelle genau ein Modell in die Produktion überführt wurde. Zum Ende des Beobachtungszeitraums sank das Verhältnis von protokollierten zu registrierten Modellen drastisch auf 5:1 – eine Verbesserung um das Dreifache.

Die wichtigste Erkenntnis lautete, dass die Effizienz der Unternehmen beim Implementierung von Modellen in Produktionsumgebungen deutlich zugenommen hat; gleichzeitig fließen immer weniger Ressourcen in Experimente, die keinen realen Mehrwert bieten.



Effizienz auf Branchenebene

Branchen haben sehr unterschiedliche Datasets, strategische Ziele und Risikoprofile. Daher erwarten wir, dass sich auch bei den ML-Konzepten Unterschiede abzeichnen werden – auch und gerade beim Verhältnis von experimentellem ML und ML in der Produktion.

Wir haben sechs Schlüsselsektoren ausgewertet, um diese Trends besser nachzuvollziehen.

VERHÄLTNIS VON PROTOKOLLIERTEN EXPERIMENTEN ZU REGISTRIERTEN MODELLEN, NACH BRANCHE

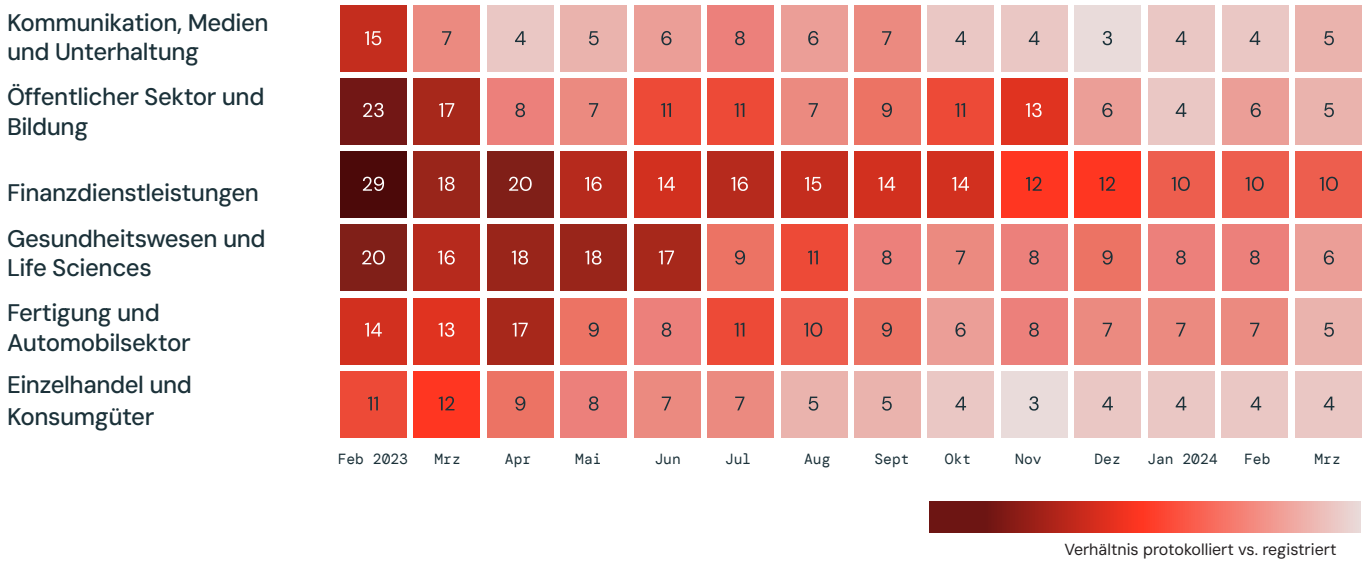


Abbildung 2: Das Verhältnis von protokollierten zu registrierten Modellen sank zwischen dem 1. Februar 2023 und dem 31. März 2024 stetig. Das deutet darauf hin, dass Unternehmen experimentelle Modelle verstärkt in die Produktion überführt haben.

HINWEIS: Aufgrund Änderungen in der Modellregistrierungs-API und beim Tracking sind die diesjährigen Daten nicht direkt mit den Vorjahresdaten vergleichbar.

DIE EFFIZIENTESTE BRANCHE – DER HANDEL – BRINGT 25 % DER MODELLE IN DIE PRODUKTION.

Einzelhandel und Konsumgüter erreichten mit einem Verhältnis von 1:4 zwischen Produktions- und Experimentiermodellen die höchste Effizienz der untersuchten Branchen. Wie im Report „MIT Technical Review Insights“ beschrieben, sind Einzelhandel und Konsumgüterbranche aufgrund des Wettbewerbsdrucks und der Verbrauchererwartungen schon lange ein Motor für die KI.

Effizienzgewinn: Financial Services hat die Effizienz bei der Überführung von Modellen in die Produktion fast verdreifachen können

FINANZDIENSTLEISTER VERZEICHNEN DEN STÄRKSTEN EFFIZIENZGEWINN

Der Finanzsektor ist die testlastigste Branche. Anfang 2023 wurden dort im Schnitt 29 Experimente pro registriertes Modell verzeichnet. Die Effizienz konnte aber fast verdreifacht werden: Bis März 2024 verbesserte sich das Verhältnis auf 10:1. In regulierten Branchen sind die Risiken für die Überführung von ML in die Produktion höher, was langwierige Testzyklen erforderlich macht.

Aber waren in diesem Jahr gerade dort mehr Unternehmen in der Lage, ihre Modelle verstärkt in die Produktion zu überführen? Ein möglicher Grund ist die Verfügbarkeit von Data-Intelligence-Plattformen, die Nutzern eine standardisierte, offene Umgebung für den gesamten ML-Lebenszyklus bereitstellen. Unternehmen können alle Phasen – von der Datenaufbereitung über das Modelltraining bis hin zu Echtzeitbereitstellung und Monitoring – auf derselben Plattform durchführen und dabei Governance, Datenschutz und Sicherheit gewährleisten. Das trägt zu einer höheren Qualität der Produktionsmodelle bei und unterstützt die Bereitschaft dafür.

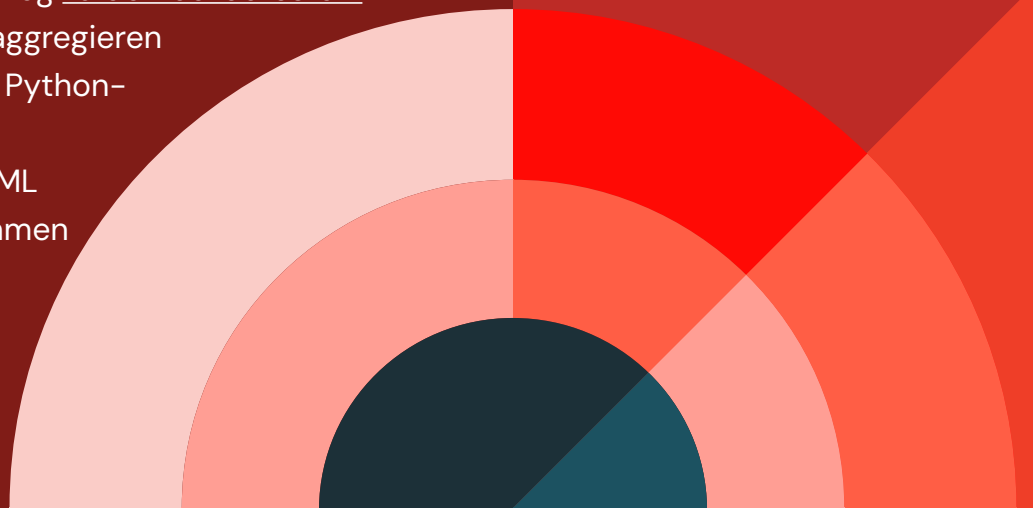
Machine-Learning-Anwendungen

NLP explodiert förmlich

NLP IST DAS ZWEITE JAHR IN FOLGE DIE FÜHRENDE ANWENDUNG FÜR DATA SCIENCE UND ML

Unstrukturierte Daten sind branchen- und regionsübergreifend vorzufinden. Daher sind Verfahren zur Verarbeitung natürlicher Sprache (NLP) unerlässlich, um den Sinn dieser Daten zu verstehen. GenAI ist ein wesentlicher Anwendungsfall für NLP.

Die folgenden Diagramme beziehen sich vorrangig auf Python-Bibliotheken, da diese bei der Weiterentwicklung von ML und KI eine Vorreiterrolle spielen und durchweg zu den beliebtesten Programmiersprachen gehören. Wir aggregieren die Daten zur Nutzung spezialisierter Python-Bibliotheken, um die fünf wichtigsten Anwendungen für Data Science und ML (DS/ML) zu ermitteln, die in Unternehmen eingesetzt werden.



WICHTIGSTE DS/ML-ANWENDUNGEN, NACH BRANCHE

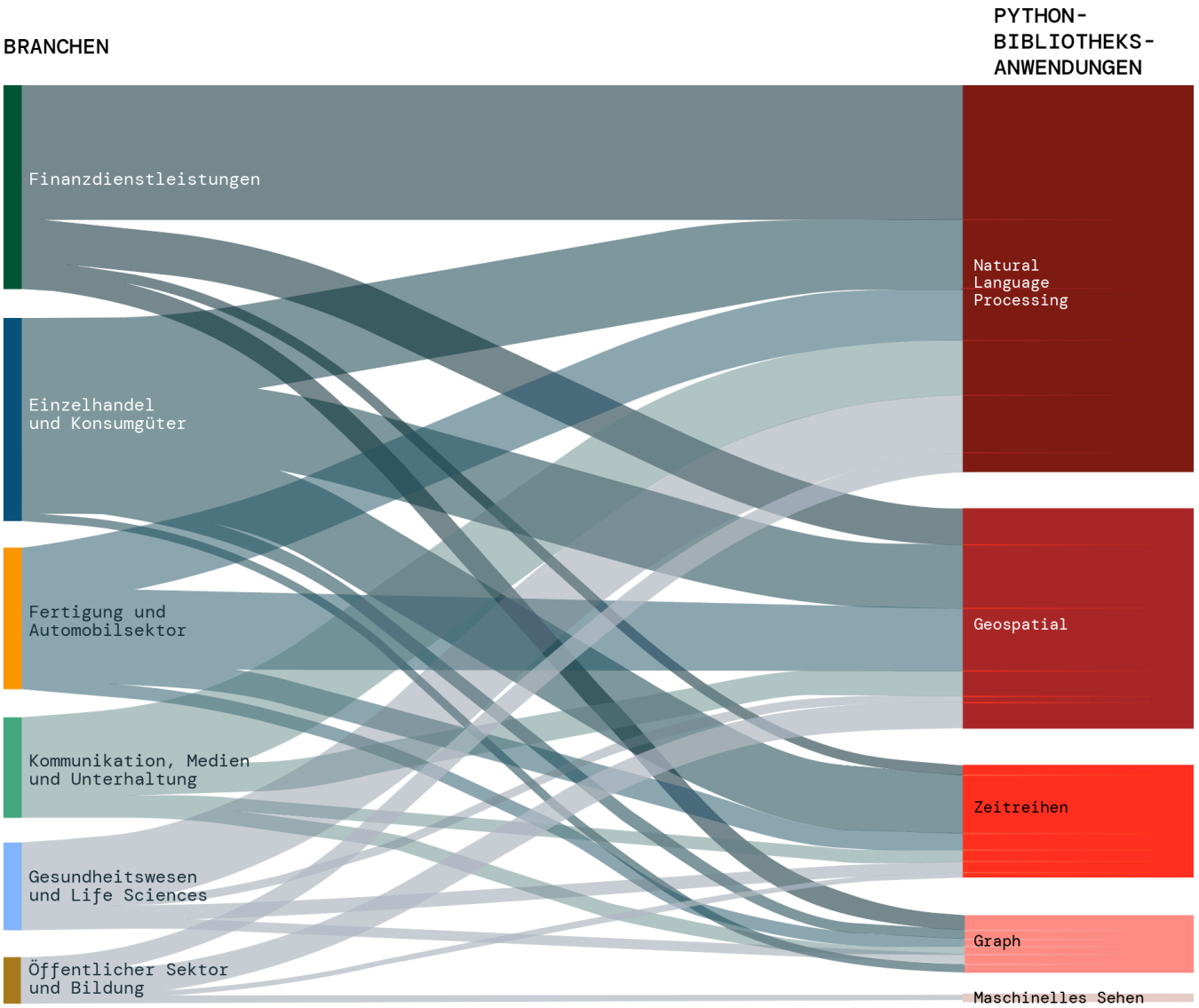


Abbildung 3: NLP ist die meistgenutzte Python-Bibliotheksanwendung und wird in allen von uns untersuchten Branchen intensiv genutzt.

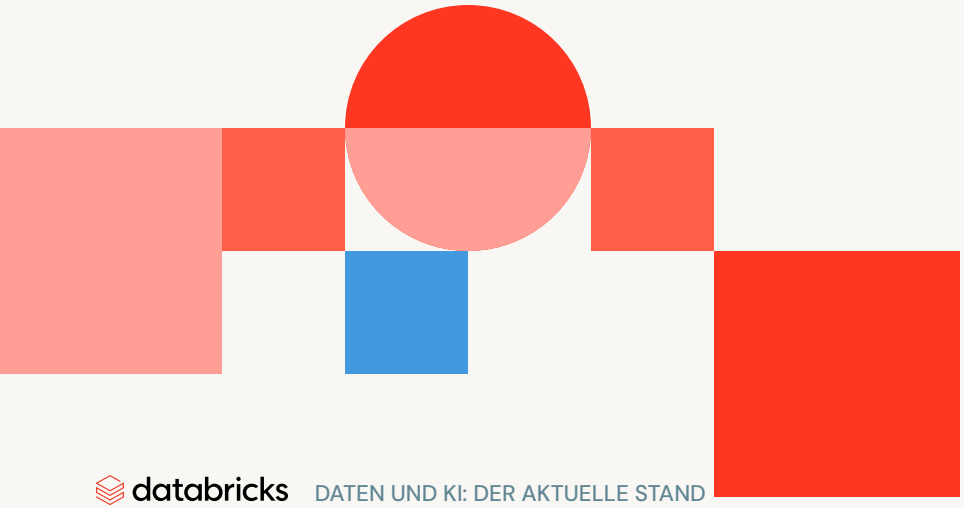
HINWEIS: Dieses Diagramm zeigt die Zahl der von ML-Bibliotheken verwendeten Notebooks je Kategorie. Nicht enthalten sind Bibliotheken, die in Tools zur Datenaufbereitung und -modellierung genutzt werden.

Im zweiten Jahr in Folge zeigen unsere Daten, dass NLP die wichtigste DS/ML-Anwendung ist: 50 % der genutzten spezialisierten Python-Bibliotheken betreffen NLP.

Datenteams begeistern sich auch für die Nutzung von Geospatial- und Zeitreihenanwendungen. Geospatial-Bibliotheken, oft für standortbezogene Analysen zur Optimierung von Benutzererlebnissen verwendet, machen mit 30 % der Nutzung von Python-Bibliotheken den zweithäufigsten Anwendungsfall aus.

HÖCHSTE AKZEPTANZ VON NLP IM BEREICH GESUNDHEITSWESEN UND LIFE SCIENCES

Im Bereich Gesundheitswesen und Life Sciences liegt der Anteil der Python-Bibliotheksnutzung für NLP bei 69 % – der höchste unter den vorgestellten Branchen. Laut einer von Arcadia in Zusammenarbeit mit der Healthcare Information and Management Systems Society durchgeführten Studie erzeugt das Gesundheitswesen 30 % des weltweiten Datenvolumens und wächst schneller als jeder andere Sektor. NLP unterstützt die Analyse in der klinischen Forschung, beschleunigt die Markteinführung neuer Medikamente und steigert die kommerzielle Effizienz von Vertrieb und Marketing.

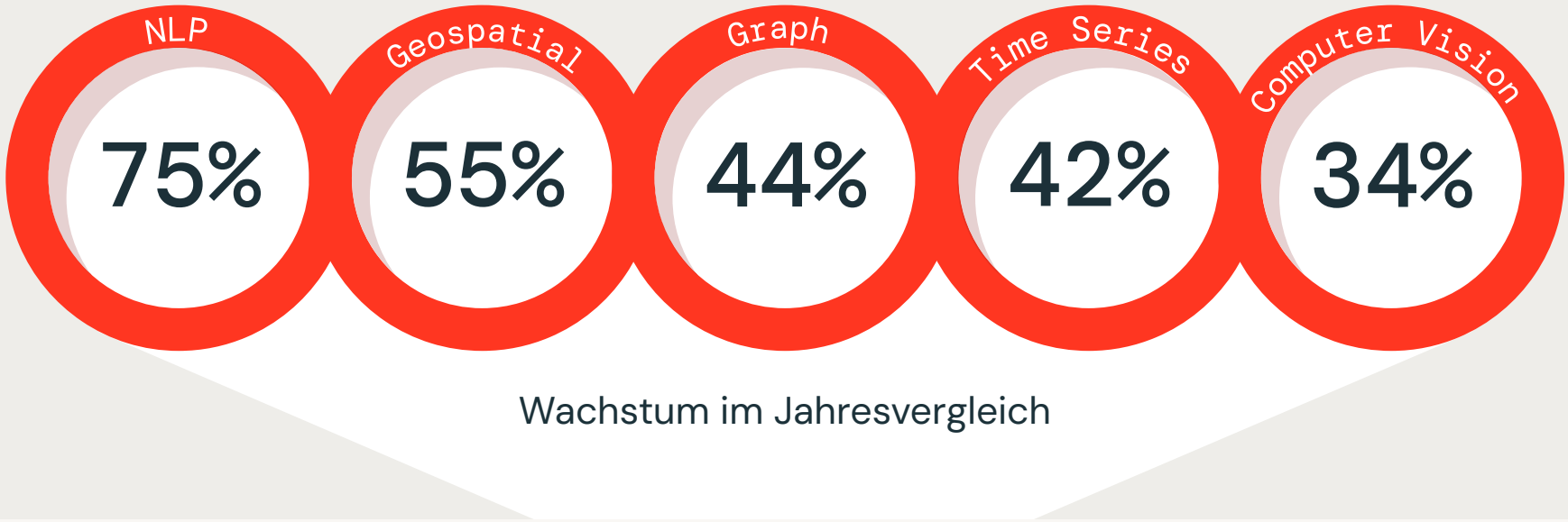


50 % der Nutzung spezialisierter Python-Bibliotheken stehen mit NLP in Verbindung.

NLP als meistgenutzte DS/ML-Anwendung lässt nicht nach

Mit der Zunahme KI-gestützter Anwendungen steigt branchenübergreifend auch die Nachfrage nach NLP-Lösungen. NLP dominiert nicht nur die Nutzung von Python-Bibliotheken, sondern weist mit 75 % auch das höchste Wachstum aller Anwendungen gegenüber dem Vorjahr auf.

Wachstumsstärkste DS/ML-Anwendungen



WACHSTUMSSTÄRKSTE DS/ML-ANWENDUNGEN, NACH BRANCHE

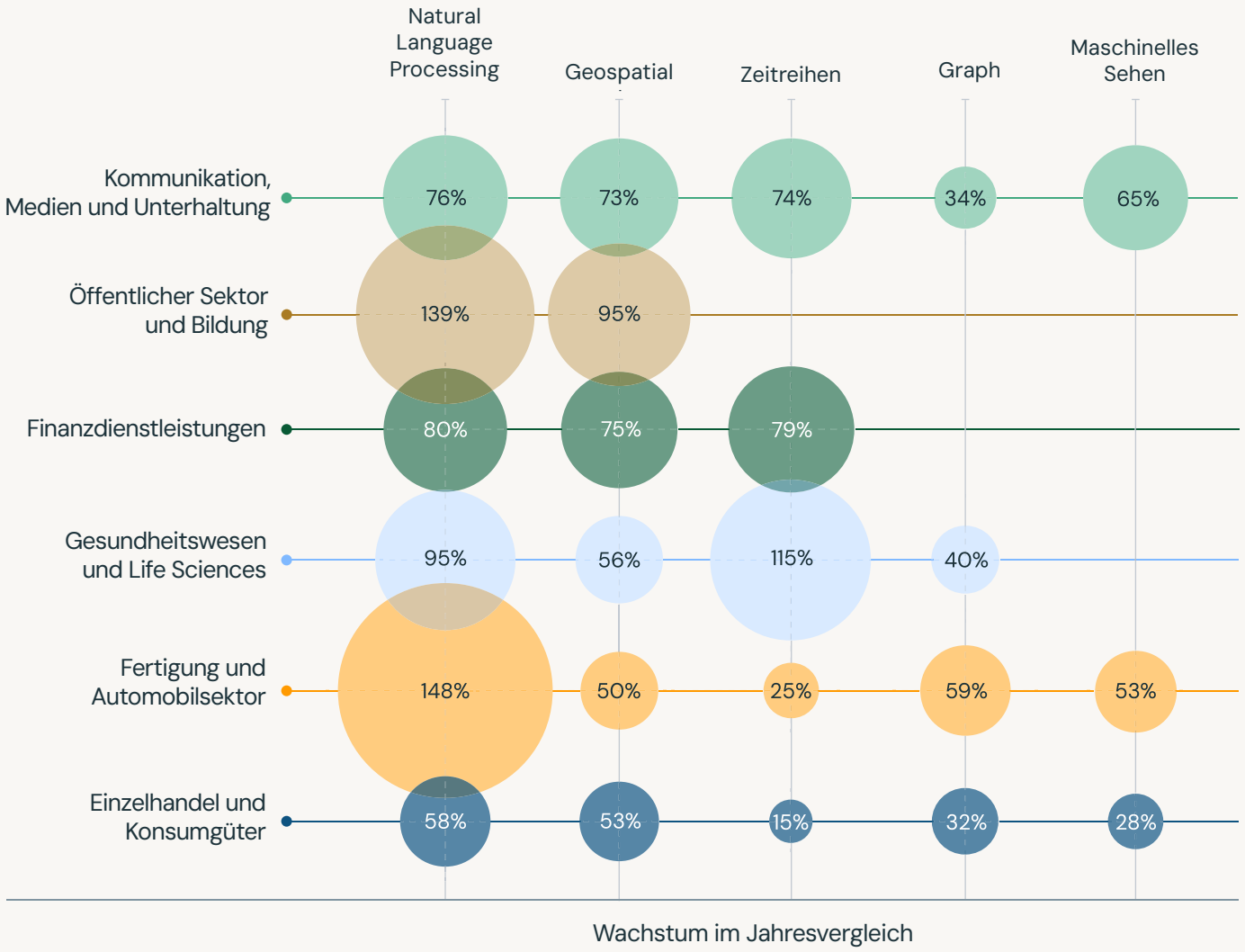
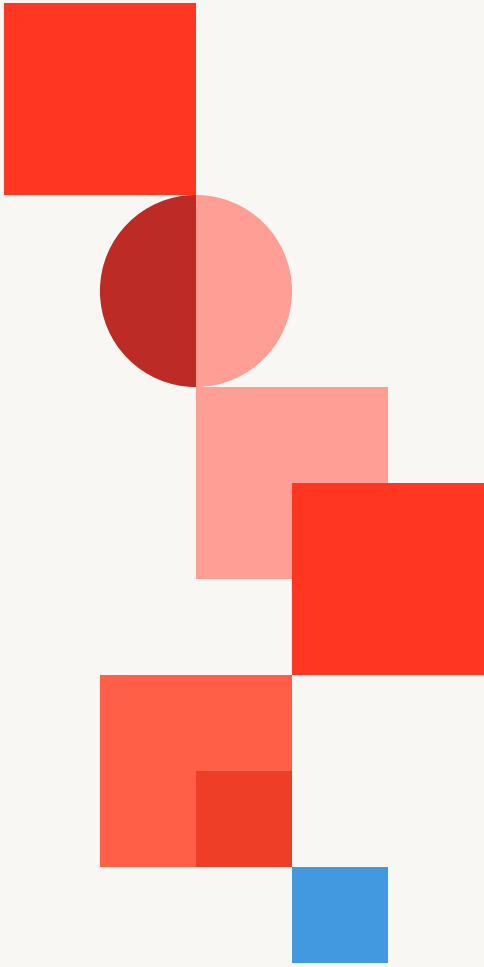


Abbildung 4: NLP erfährt das größte Wachstum bei den Anwendungen. Mit 148 % weist die NLP-Nutzung in der Fertigungs- und Automobilbranche das stärkste Wachstum im Jahresvergleich auf.

ALLE BRANCHEN INVESTIEREN KRÄFTIG IN NLP

Unter den von uns betrachteten Branchen verzeichnete die Fertigungs- und Automobilbranche den größten Zuwachs bei der NLP-Nutzung mit einem Anstieg von 148 % im Vergleich zum Vorjahr. NLP hilft dem Sektor bei allem Möglichen – von der Analyse des Kundenfeedbacks über Qualitätskontrolle bis hin zum Einsatz von Chatbots – und versetzt Unternehmen so in die Lage, ihre betriebliche Effizienz zu verbessern. Dicht dahinter folgt das NLP-Wachstum im öffentlichen Sektor und im Bildungswesen mit 139 % gegenüber dem Vorjahr.



OB WALDBRÄNDE ODER VOGELGRIPPE: AKTUELLE EREIGNISSE DECKEN SICH MIT DEM WACHSTUM DES MASCHINELLEN LERNENS.

Die zweite Anwendung, die in allen sechs Branchen deutlich gewachsen ist, ist Geospatial. Unternehmen suchen verstärkt nach Mustern, Trends und Korrelationen in Positionsdaten. Das starke Wachstum bei Geospatial im öffentlichen Sektor und im Bildungswesen könnte mit den Bereichen Katastrophenmanagement und Notfallplanung zusammenhängen.

Die dritthöchste Wachstumsrate weist anwendungs- und branchenübergreifend die Nutzung von Zeitreihenbibliotheken im Bereich Gesundheitswesen und Life Sciences auf: 115 % gegenüber dem Vorjahr. Zeitreihen unterstützen die Risikovorhersage für Patienten, Versorgungsprognosen und die Medikamentenerforschung. In einem Gutachten der National Institutes of Health von 2023 heißt es: „Die Zeitreihenanalyse ermöglicht es uns, durch Schätzungen direkt aus den Daten schnell und einfach präzise kurzfristige Vorhersagen für neue Pandemien zu treffen.“[1]

1 Applications of Time Series Analysis in Epidemiology: Literature Review and Our Experience During COVID-19 Pandemic, 16. Oktober 2023.

DER MODERNE DATEN- UND KI-STACK

Entwicklung zur GenAI

SPITZENPRODUKTE FÜR DATEN UND KI BELEGEN DIE NÄCHSTE GENAI-ENTWICK-LUNGSSTUFE

Führungskräfte aus dem Datenbereich suchen stets nach den optimalen Tools zur KI-Strategieumsetzung. Unsere Top 10 der Daten- und KI-Produkte zeigt die meistverwendeten Integrationen auf der Databricks Data Intelligence Platform. Die Kategorien sind DS/ML, Data Governance und Sicherheit, Orchestrierung, Datenintegration und Datenquellenprodukte.

9 der 10 besten Produkte sind Open-Source. Die Unternehmen setzen auf Flexibilität und umgehen die Beschränkungen proprietärer Lösungen. Wie wir weiter unten noch erörtern werden, erfreuen sich auch quelloffene LLMs zunehmender Beliebtheit.

DATEN- UND KI-PRODUKTE: DIE TOP 10

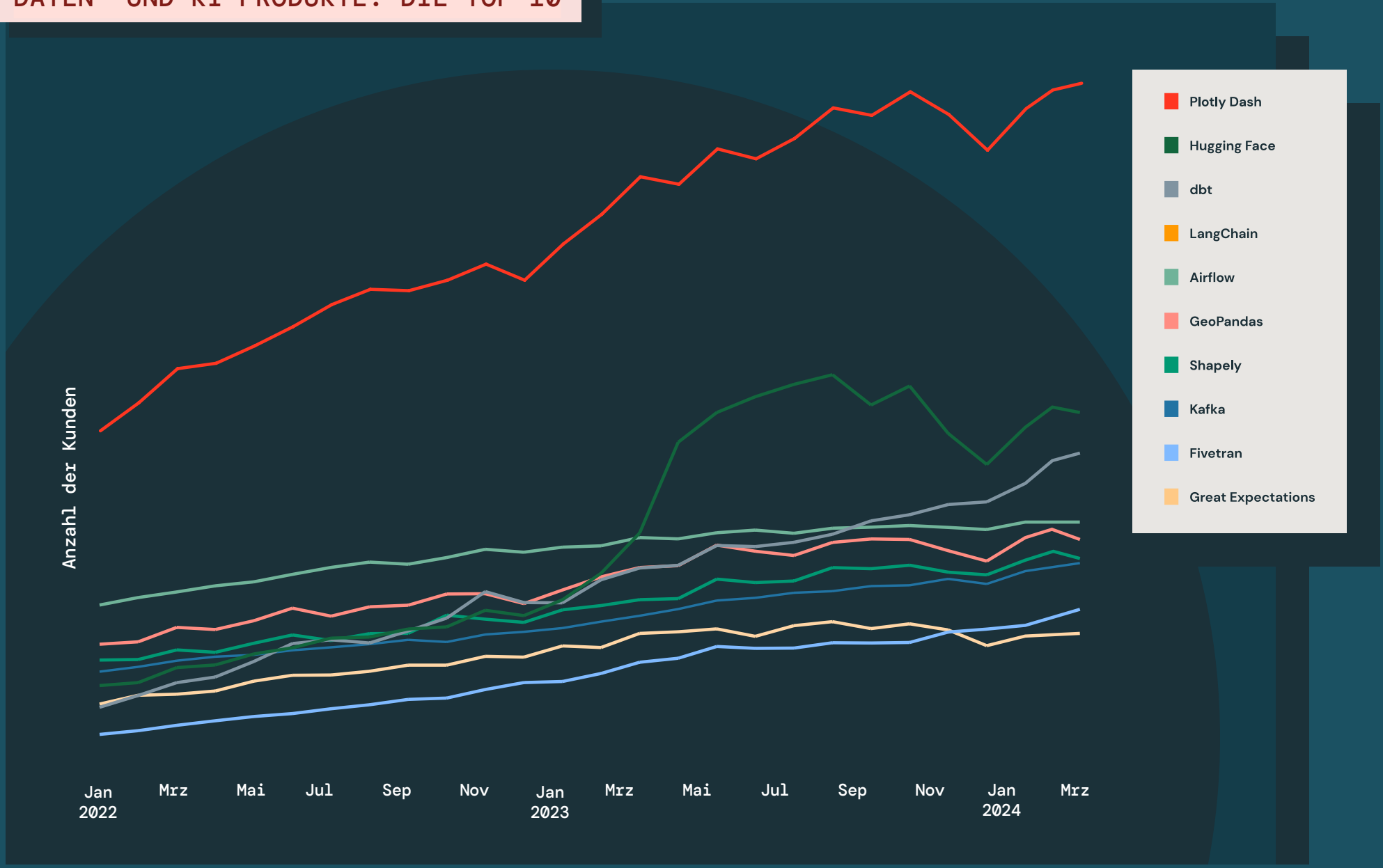


Abbildung 5: Unsere Top 10 der Daten- und KI-Produkte umfassen die Kategorien DS/ML, Data Governance und Sicherheit, Orchestrierung, Datenintegration und Datenquellenprodukte.

PLOTLY DASH BEHAUPTET DIE SPITZENPOSITION

Plotly Dash ist eine Low-Code-Plattform, mit der Data Scientists Datenanwendungen bequem entwickeln, skalieren und bereitstellen können. Produkte wie Dash helfen Unternehmen, Anwendungen schneller und einfacher zu liefern, um mit dynamischen Geschäftsanforderungen mitzuhalten. Seit über 2 Jahren hält Dash seine Position als Nummer eins, was für den wachsenden Druck auf Data Scientists spricht, produktionsreife Daten- und KI-Anwendungen zu entwickeln.

HUGGING FACE TRANSFORMERS SPRINGT AUF PLATZ 2

Vor einem Jahr noch auf Platz 4, ist Hugging Face Transformers heute das bei unseren Kunden zweitbeliebteste Produkt. Oft werden die vortrainierten Transformationsmodelle der Open-Source-Plattform zusammen mit den eigenen Unternehmensdaten für die Erstellung und Optimierung von Basismodellen genutzt. Das unterstützt einen zunehmenden Trend, den wir bei RAG-Anwendungen beobachten.

LANGCHAIN WIRD NUR WENIGE MONATE NACH DER INTEGRATION ZU EINEM SPITZENPRODUKT

LangChain – eine Open-Source-Toolchain zum Erstellen von und Arbeiten mit proprietären LLMs – hat im Frühjahr den Sprung in die Spitzengruppe geschafft und steht jetzt, noch nicht einmal ein Jahr nach der Integration, auf Platz 4. Unternehmen, die eigene moderne LLM-Anwendungen entwickeln und mit auf Transformationen spezialisierten Python-Bibliotheken arbeiten, um Modelle zu trainieren, können mit LangChain Prompt-Oberflächen und Integrationen für andere Systeme entwickeln.

UNTERNEHMEN INVESTIEREN IN PRODUKTE ZUR ERSTELLUNG HOCHWERTIGER DATASETS

Die Präsenz von drei Datenintegrationsprodukten in unseren Top 10 belegt, dass sich die Unternehmen vorrangig dem Aufbau vertrauenswürdiger Datasets widmen: dbt (Datentransformation), Fivetran (Automatisierung von Datenpipelines) und Great Expectations (Datenqualität) verzeichnen alle ein stetiges Wachstum. Namentlich dbt ist im letzten Jahr um zwei Plätze nach oben geklettert.

Aufsteiger

John Snow LABS

John Snow Labs ist ein KI- und NLP-Anbieter, der Unternehmen aus Gesundheitswesen und Life Sciences bei Aufbau, Implementierung und Betrieb von KI-Projekten unterstützt. Mithilfe moderner NLP, ML-Modelle und GenAI trägt John Snow Labs zur Verbesserung von Diagnosen, Medikamentenerforschung und Patientenversorgung bei.

John Snow Labs verdient besondere Erwähnung: Trotz des vorwiegenden Einsatzes im Gesundheitswesen belegt es Platz 15 unserer Daten- und KI-Produkte. Die beliebte Spark NLP-Bibliothek unterstützt eine Vielzahl von NLP-Aufgaben wie Textklassifizierung, Entity Recognition und Stimmungsanalyse und ist damit auch in anderen Branchen wie etwa Finanzdienstleistungen nützlich.

Vektordatenbanken

Unternehmen brauchen individualisierte LLMs

LLMs unterstützen mit ihrem Sprachverstehen und ihren Generierungsfunktionen eine Vielzahl von Geschäftsanwendungen. Allerdings sind LLMs allein – insbesondere in großen Unternehmen – Grenzen gesetzt. Sie können als Informationsquellen unzuverlässig sein und neigen dazu, Falschinformationen – so genannte Halluzinationen – auszugeben. Im Kern sind eigenständige LLMs nicht auf das Branchenwissen und die Bedürfnisse einer konkreten Organisation zugeschnitten.

Unsere Daten bestätigen, dass immer mehr Unternehmen auf RAG setzen, statt sich nur auf eigenständige LLMs zu verlassen. RAG befähigt Unternehmen dazu, ihre eigenen Daten einzusetzen, um LLMs besser anzupassen und hochwertige GenAI-Apps zu entwickeln. Wenn LLMs zusätzliche relevante Informationen erhalten, können die Modelle zutreffendere Antworten geben und halluzinieren seltener.

RAG: Wegbereiter für GenAI in großen Unternehmen

Im Vorjahr zeigte unsere Grafik zu LLM-Python-Bibliotheken, dass SaaS-LLMs in gerade einmal gut 5 Monaten um 1.310 % zugenommen hatten. SaaS-LLMs wie GPT-4 werden mit enormen Text-Datasets trainiert und haben sich vor noch nicht einmal zwei Jahren etabliert.

Dieses Jahr dagegen sind es die Vektordatenbanken, die unsere Tabelle aufmischen. Diese Kategorie insgesamt ist gegenüber dem Vorjahr um 377 % gewachsen, davon allein 186 % seit der öffentlichen Preview von Databricks Vector Search. [2]

WAS IST RAG?

Retrieval Augmented Generation (RAG) ist ein GenAI-Anwendungsmuster, das Daten und Dokumente findet, die für eine Frage oder Aufgabe relevant sind, und sie als Kontext für das LLM bereitstellt, um präzisere Antworten zu geben.

WIE ARBEITEN VEKTORDATENBANKEN UND RAG ZUSAMMEN?

Vektordatenbanken erzeugen Darstellungen vorwiegend unstrukturierter Daten. Diese dienen der Informationsgewinnung in RAG-Anwendungen, um Dokumente oder Unterlagen auf der Grundlage ihrer Ähnlichkeit mit Schlüsselwörtern in einer Abfrage zu finden.

RAG-Anwendungen bieten gegenüber Lösungen von der Stange eine Menge Vorteile. RAG hat sich schnell zu einem beliebten Verfahren zur Einbindung eigener Echtzeitdaten in LLMs entwickelt, weil Kosten und Zeitaufwand für die Feinabstimmung oder das Modellvortraining entfallen.

Das exponentielle Wachstum bei den Vektordatenbanken legt nahe, dass Unternehmen mehr RAG-Anwendungen entwickeln, um ihre Unternehmensdaten in ihre LLMs zu integrieren.

2 Databricks Vector Search ist am 7. Dezember 2023 in die öffentliche Preview gegangen.

NUTZUNG VON LLM-PYTHON-BIBLIOTHEKEN

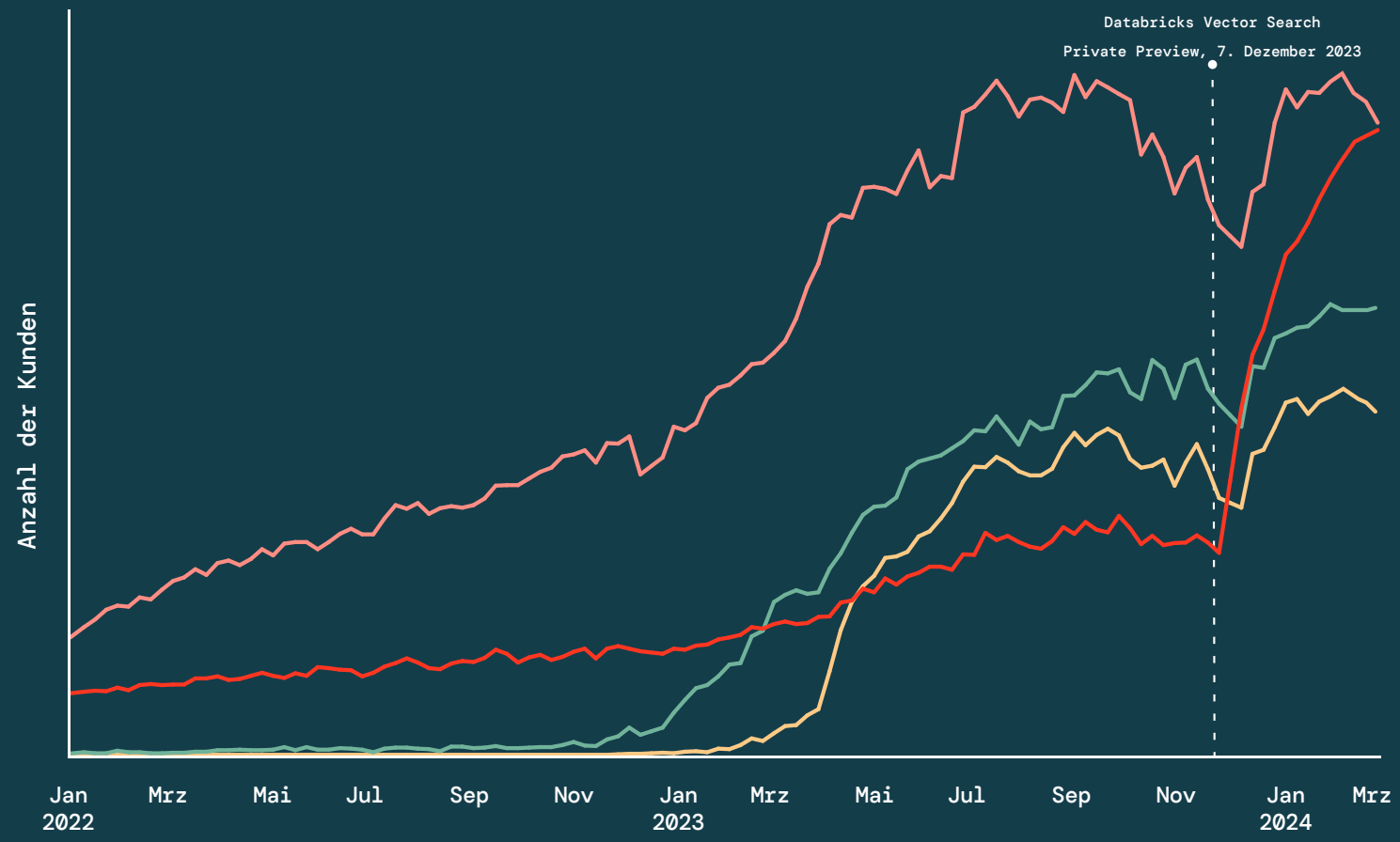


Abbildung 6: Seit der öffentlichen Preview von Databricks Vector Search ist die Vektordatenbankkategorie insgesamt um 186 % gewachsen – deutlich mehr als jede andere LLM-Python-Bibliothek.

HINWEISE: Da Kunden oft mehrere Tools aus einer Kategorie verwenden, werden diese ggf. mehrfach gezählt. Die Nutzung wird durch die Verwendung von Paketen, die sich mit externen Vektordatenbankservices verbinden, und durch API-Aufrufe auf unserer Plattform gemessen. Die Trendlinien zwischen dem 18. Dezember und dem 1. Januar wurden gemittelt, um saisonale Schwankungen zu berücksichtigen.

LLM-DEFINITIONEN

Transformer-Training: Bibliotheken zum Trainieren von Transformer-Modellen (z. B. Hugging Face Transformers)

SaaS-LLMs: Bibliotheken für den Zugriff auf API-basierte LLMs (z. B. OpenAI)

LLM-Tools: Toolchains für das Arbeiten mit und das Erstellen von proprietären LLMs (z. B. LangChain)

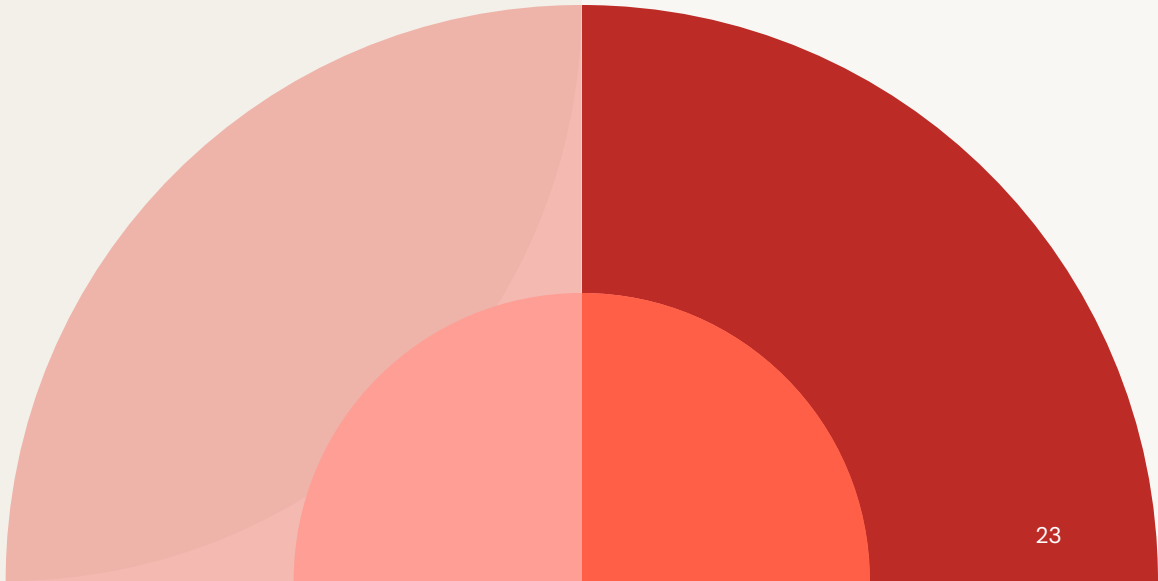
Vektordatenbanken: Vektor/KNN-Indizes (z. B. Pinecone und Databricks Vector Search)

UNTERNEHMEN WERDEN BEIM AUFBAU VON LLMS IMMER VERSIERTER

Im letzten Jahr waren LLMs von der Stange der letzte Schrei. Und auch jetzt ist die Zahl der Kunden, die SaaS-LLMs nutzen, im Vergleich zum Vorjahr noch einmal um 178 % gestiegen. Aber die Unternehmen übernehmen allmählich mehr Kontrolle über ihre LLMs und entwickeln Tools, die speziell auf ihre Bedürfnisse zugeschnitten sind.

Das stetige Wachstum bei Vektordatenbanken, LLM-Tools und Transformer-bezogenen Bibliotheken zeigt, dass viele Datenteams jetzt lieber selbst entwickeln, statt zu kaufen. Unternehmen investieren verstärkt in LLM-Tools wie LangChain, um unternehmenseigene LLMs zu nutzen und zu erstellen. Transformer-bezogene Bibliotheken wie Hugging Face werden für das LLM-Training verwendet und haben gemessen an der Kundenzahl immer noch die höchste Akzeptanz. Die Nutzung dieser Bibliotheken stieg im Jahresvergleich um 36 %. Gemeinsam deuten diese Trends darauf hin, dass Open-Source-LLMs verstärkt eingesetzt werden.

Die Zahl der Kunden, die Vektordatenbanken nutzen, ist im Jahresvergleich um 377 % gestiegen.



Unternehmen bevorzugen kleinere Open-Source-Modelle

NUTZUNG VON OPEN-SOURCE-LLMs

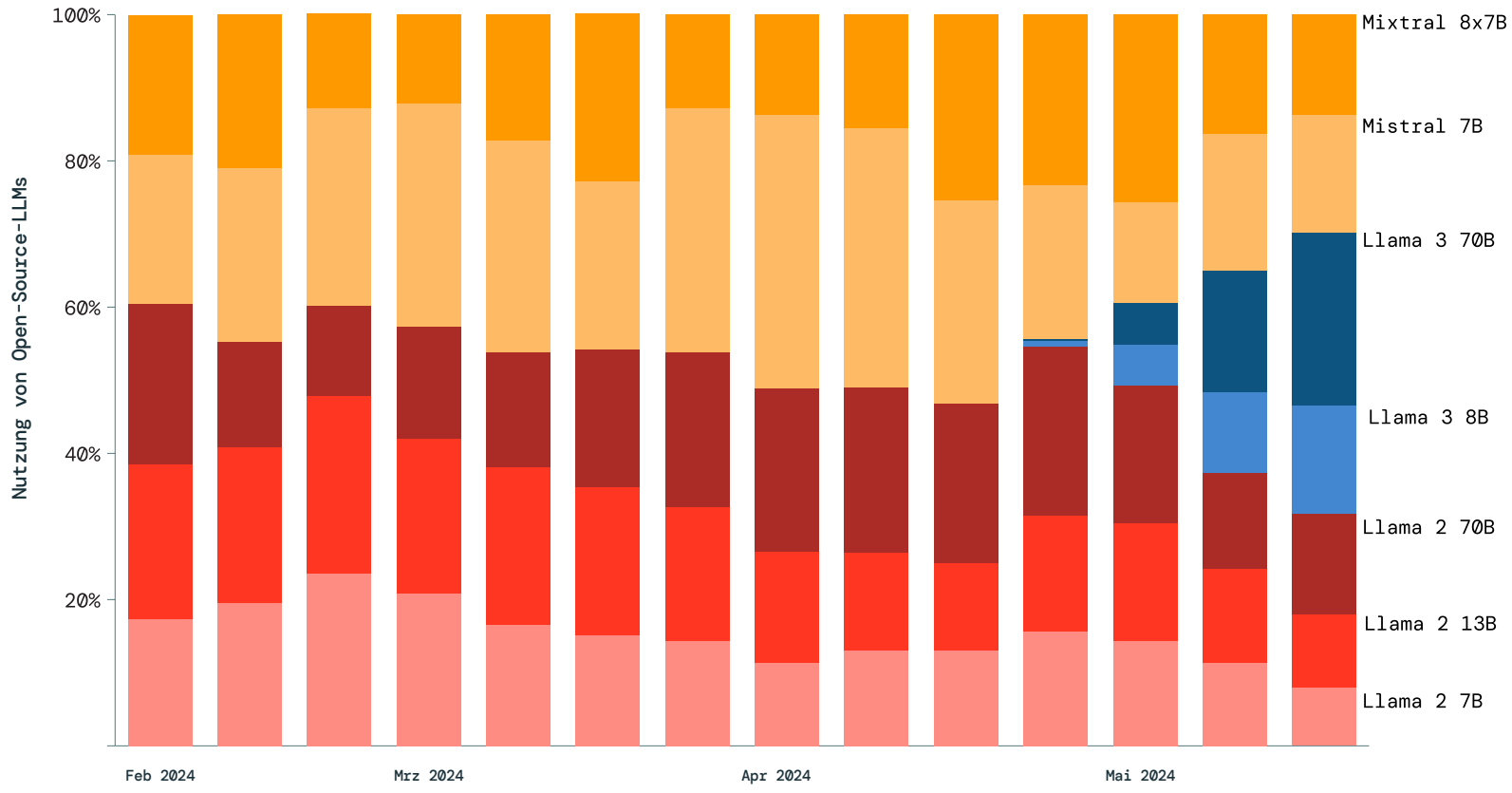


Abbildung 7: Relative Übernahme der Open-Source-Modelle Mistral und Meta Llama in den Basismodell-APIs von Databricks.

HINWEIS: Das Diagramm erstreckt sich auf den Zeitraum bis zum 19. Mai 2024, um den Launch von Meta Llama 3 zu berücksichtigen.

Einer der größten Vorteile von Open-Source-LLMs ist die Möglichkeit, sie für bestimmte Anwendungsfälle anzupassen – vor allem im Unternehmensumfeld. Folgende Frage wird oft gestellt: *Welches ist das beliebteste Open-Source-Modell?* In der Praxis probieren Kunden oft viele Modelle und Modellfamilien aus. Wir haben die Nutzung von Meta Llama und Mistral, den beiden größten Open-Source-Modellen, analysiert. Unsere Daten zeigen, dass der Bereich der quelloffenen LLMs in Bewegung ist und neue leistungsfähige Modelle gut angenommen werden.

Bei jedem Modell muss zwischen Kosten, Latenz und Leistung abgewogen werden. Die Nutzung der beiden kleinsten Meta Llama 2-Modelle (7 bzw. 13 Mrd. Parameter) ist deutlich höher als beim größten Modell, Meta Llama 2 70B. Insgesamt entscheiden sich von allen Benutzern von Meta Llama 2, Llama 3 und Mistral 77 % für Modelle mit maximal 13 Mrd. Parametern. Das legt nahe, dass Kosten und Latenz für die Unternehmen eine wichtige Rolle spielen.

UNTERNEHMEN ZEIGEN SICH EXPERIMENTIERFREUDIG

Meta Llama 3 kam am 18. April 2024 auf den Markt. Bereits in der ersten Woche begannen Unternehmen, das System gegenüber anderen Modellen und Anbietern vorzuziehen. Gerade einmal vier Wochen nach der Markteinführung machte Llama 3 bereits 39 % der gesamten Nutzung von Open-Source-LLMs aus.

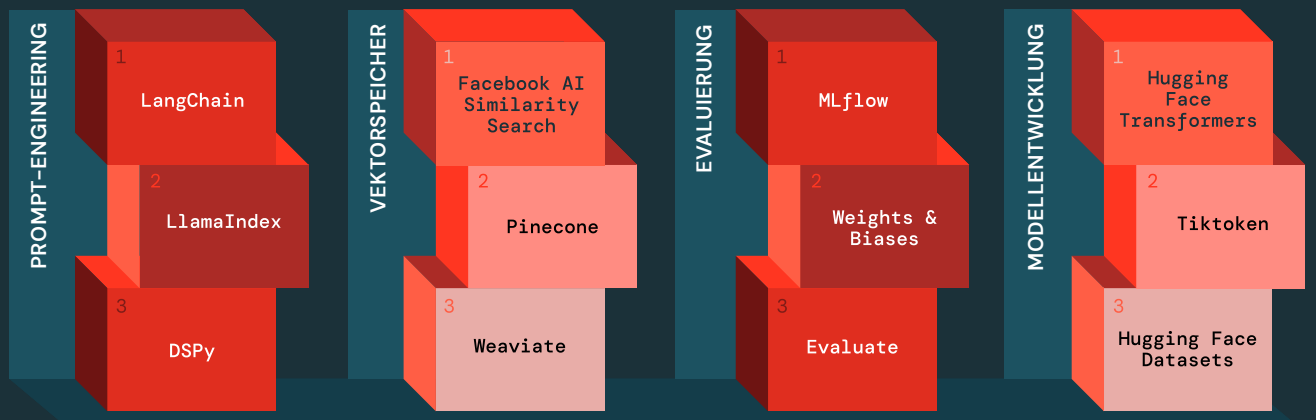
76 %

der Unternehmen, die LLMs nutzen, entscheiden sich für Open Source, wobei oft noch proprietäre Modelle parallel eingesetzt werden.

70 %

der Unternehmen, die GenAI nutzen, passen ihre Basismodelle mithilfe von Tools, Wiederverwendung und Vektordatenbanken an.

Wichtigste GenAI-Python-Pakete



Vor allem stark regulierte Branchen treten bei GenAI als Vorreiter auf.

Eigentlich haben stark regulierte Branchen den Ruf, risikoscheu zu sein und neue Technologien nur zögerlich einzuführen. Dafür gibt es viele Gründe, z. B. strenge Compliance-Anforderungen, eingefahrene Altsysteme, deren Austausch kostspielig wäre, und die Notwendigkeit einer behördlichen Genehmigung vor der Implementierung.

Zwar setzen alle Branchen auf neue KI-Innovationen, doch es sind vor allen zwei stark regulierte Bereiche – nämlich Finanzdienstleister sowie Gesundheitswesen und Life Sciences –, die mit den übrigen Sektoren Schritt halten und sie teils sogar übertreffen.

Im Dezember 2023 veröffentlichte Databricks Basismodell-APIs, die Sofortzugriff auf beliebte Open-Source-LLMs wie Meta Llama und MPT-Modelle bieten. Wir erwarten, dass das Interesse an Open Source deutlich zunehmen wird, da die Modelle weiterhin schnell besser werden, wie die jüngste Einführung von Llama 3 zeigt.

OPEN-SOURCE-LLMs FÜR BRANCHENSPEZIFISCHE BEDÜRFNISSE NUTZEN

Die Bereiche Fertigung und Automobilbau sowie Gesundheitswesen und Life Sciences sind Spitzenreiter beim Einführen von Basismodell-APIs mit der höchsten Durchschnittsnutzung pro Kunde. In der Fertigung gelten Lieferkettenoptimierung, Qualitätskontrolle und Effizienz als vielversprechendste Anwendungsfälle.

Laut einem aktuellen Report von MIT Tech Review Insights zeigen sich die befragten CIOs aus Gesundheitswesen und Life Sciences davon überzeugt, dass GenAI für ihre Organisationen von Nutzen sein wird. Open-Source-LLMs ermöglichen es diesen stark regulierten Branchen, GenAI zu integrieren und dabei die größtmögliche Kontrolle über ihre Daten zu behalten.

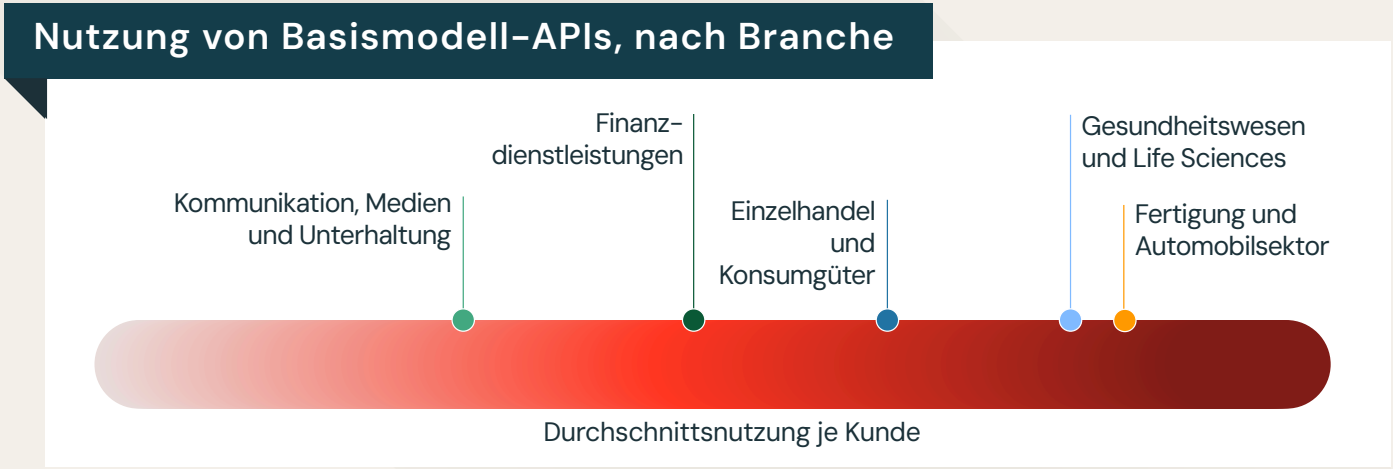


Abbildung 8: Fertigung und Automobilbau sowie Gesundheitswesen und Life Sciences sind Spitzenreiter beim Einführen von Grundmodell-APIs mit der höchsten Durchschnittsnutzung pro Kunde.

HINWEIS: Zeitraum: Januar bis März 2024.

CPUs und GPUs: Bei den Finanzdienstleistern nimmt der LLM-Einsatz in 6 Monaten um 88 % zu

CPUs sind Allzweckprozessoren, die eine Vielzahl von Aufgaben schnell, aber nur eine begrenzte Anzahl von Aufgaben parallel erledigen können. CPUs werden für klassisches ML eingesetzt. GPUs sind spezialisierte Prozessoren, die Tausende oder Millionen separater Tasks parallel verarbeiten können. Sie sind notwendig, um LLMs zu trainieren und zu bedienen.

Wir haben uns die CPU- und GPU-Nutzung und das Wachstum bei unseren Model Serving-Kunden angesehen, um die KI-Strategien zu verstehen. Die GPUs in unseren Daten stehen überwiegend in Verbindung mit LLMs.

BEI DEN FINANZDIENSTLEISTERN DOMINIERT DIE GPU-NUTZUNG

Finanzdienstleister – eine der am stärksten regulierten Branchen –, schrecken nicht vor GenAI zurück. Hier findet sich bei weitem die höchste durchschnittliche GPU-Nutzung pro Unternehmen – und mit 88 % in den letzten 6 Monaten auch das höchste Wachstum bei den GPUs. LLMs unterstützen geschäftskritische Anwendungen wie Betrugserkennung, Vermögensverwaltung und Anwendungen für Investoren und Analysten.

MODEL SERVING; KLASSISCHES ML UND LLM

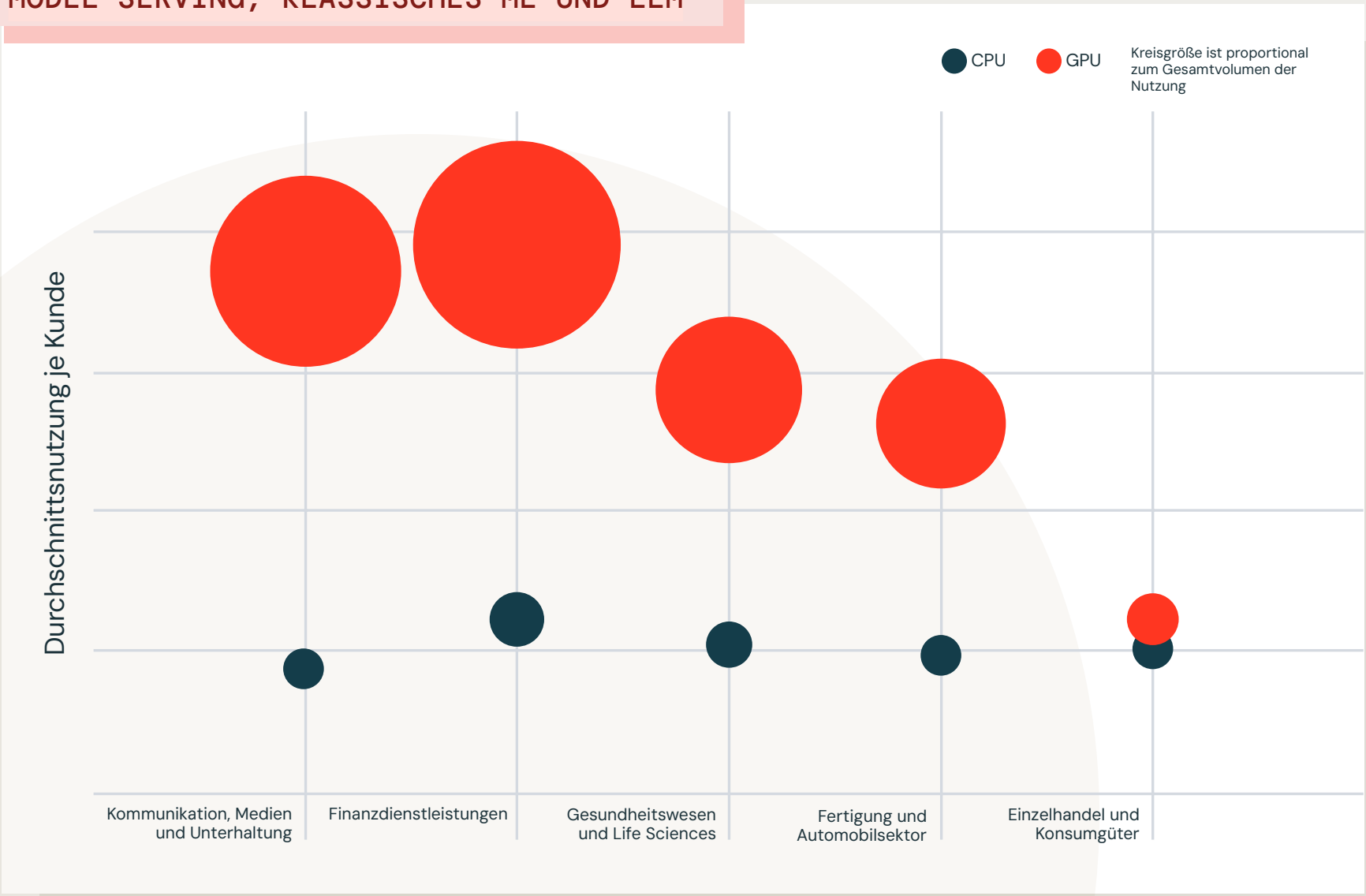


Abbildung 9: Finanzdienstleister haben die höchste Durchschnittsnutzung bei CPUs und GPUs.

HINWEIS: Zeitraum: Januar bis März 2024.

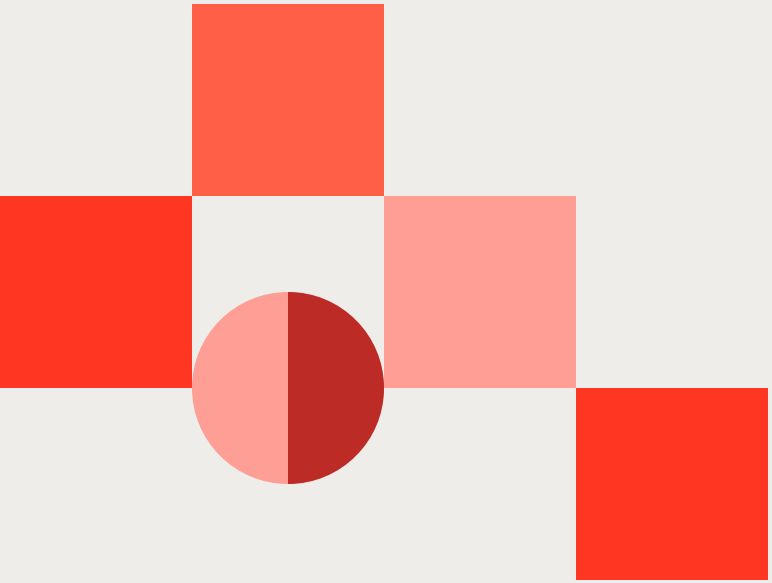
Stark regulierte Branchen führend bei der Einführung von Unified Governance

KI-Sicherheit und Governance sind für das Vertrauen in die KI-Initiativen eines Unternehmens von entscheidender Bedeutung. Sie helfen Datenfachleuten beim Entwickeln und Warten von Produkten unter Einhaltung präziser Richtlinien und Standards. Unified-Governance-Lösungen wie Databricks Unity Catalog umfassen alle Daten und KI-Assets und erleichtern Unternehmen das Trainieren und Nutzen von GenAI-Modellen mit den eigenen privaten Daten.

Laut Gartner sind Vertrauen, Risiken und Sicherheitsmanagement für KI die Top-Trends 2024, die bei Geschäfts- und Technologieentscheidungen eine Rolle spielen. Mehr denn je wollen Führungskräfte Daten und KI nutzen, um ihre Organisationen zu transformieren. Das spiegelt sich in der Akzeptanz von Unified Governance bei unseren Kunden wider.

FINANZDIENSTLEISTER: FÜHREND BEI DATEN- UND KI-GOVERNANCE

Die Einhaltung von Vorschriften und Sicherheitsbestimmungen gehört bei Finanzdienstleistern zur Unternehmenskultur. Nach den Umfragedaten des Reports CIO vision 2025 geht MIT Technology Review Insights davon aus, dass Finanzunternehmen den höchsten Investitionszuwachs bei Datenmanagement und -infrastruktur verzeichnen werden: „Die Befragten aus der Finanzbranche gaben an, dass ihre Investitionen bis 2025 um 74 % steigen werden; bei der Gesamtheit der Befragten sind es nur 52 %."



AKZEPTANZ VON UNITY CATALOG, NACH BRANCHE

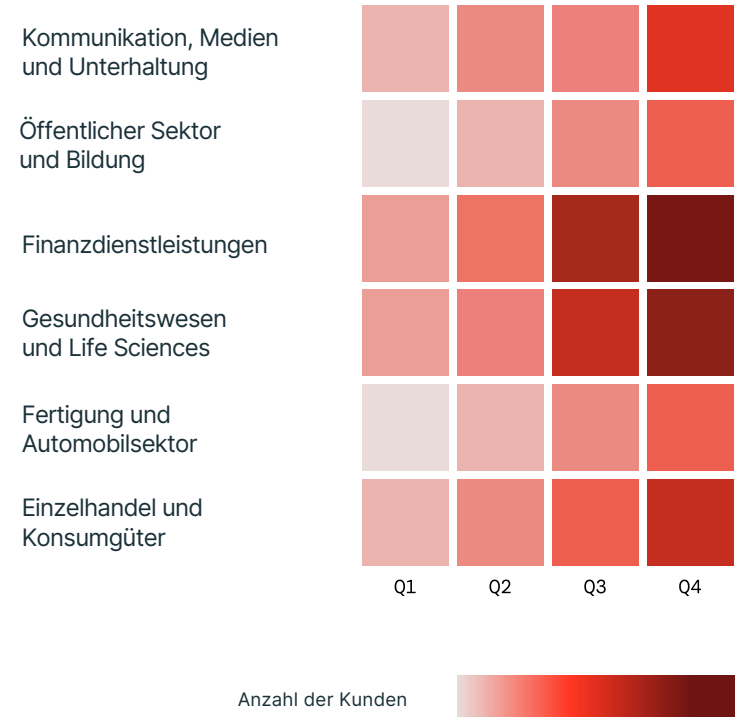


Abbildung 10: Finanzdienstleister sind führend bei der Einführung von Unity Catalog für Unified Governance für Daten und KI.

HINWEIS: Zeitraum: 1. Februar 2023 bis 31. Januar 2024.

AKZEPTANZ VON SERVERLESS MODEL SERVING, NACH BRANCHE

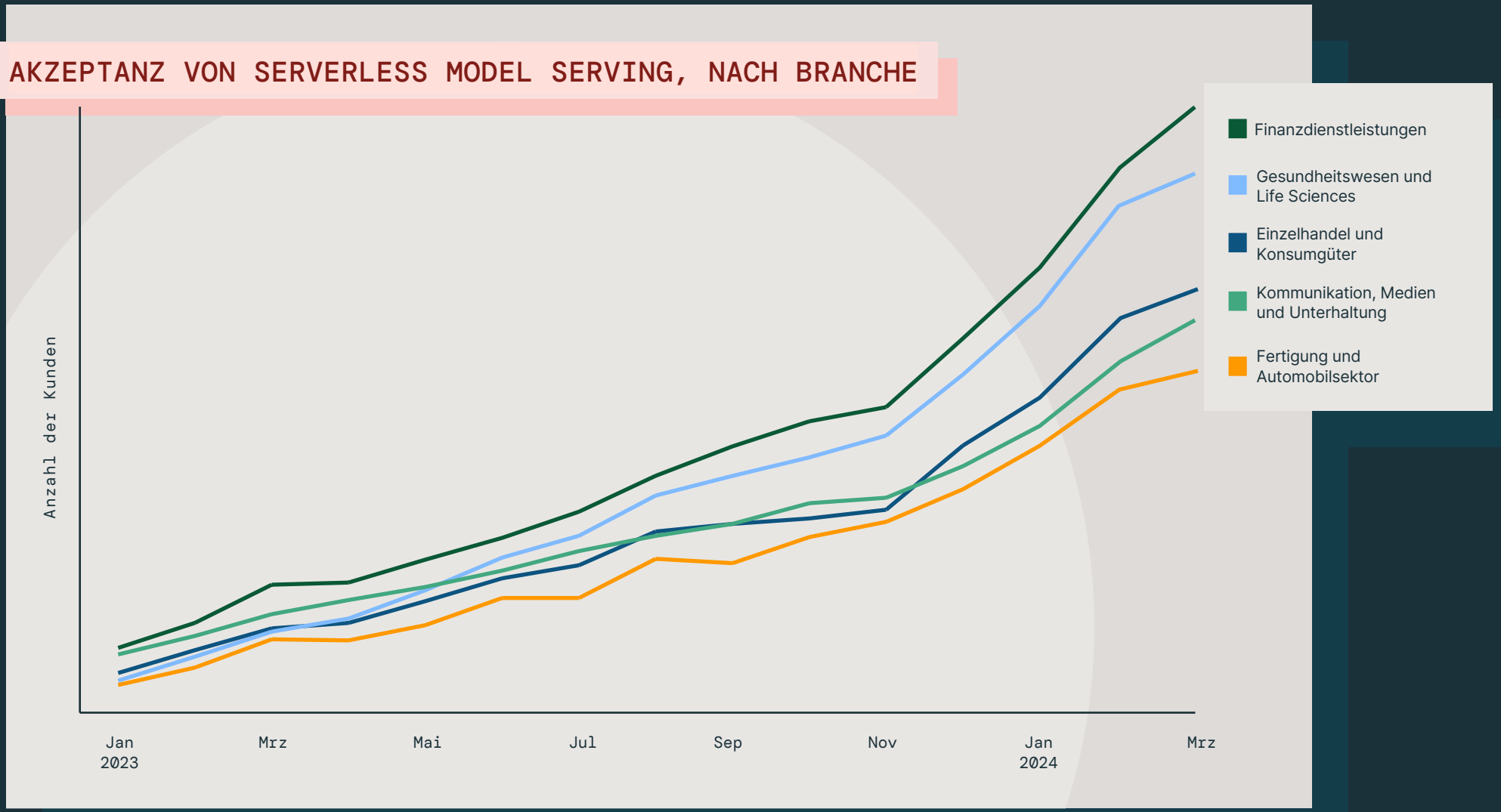


Abbildung 11: Finanzdienstleister sind führend bei der Nutzung von Serverless-Produkten, gefolgt von Gesundheitswesen und Life Sciences.

HINWEISE: Das umfasst Model Serving auf Serverless-Endpunkten, Databricks SQL, Lakehouse Monitoring und Serverless-Jobs. Im November 2023 wurde Serverless für weitere regionale Cloud-Plattformen verfügbar.

Unternehmen setzen bei Entwicklung von ML-Echtzeitanwendungen auf Serverless

ML-Echtzeitsysteme revolutionieren das Arbeiten in den Unternehmen, denn sie bieten die Möglichkeit, auf Basis eingehender Daten sofortige Vorhersagen oder Maßnahmen zu treffen. Sie brauchen jedoch eine schnelle und skalierbare Serverinfrastruktur, für deren Aufbau und Wartung Fachkenntnisse erforderlich sind.

Bei Serverless Model Serving erfolgt das Hoch- oder Herunterskalieren automatisch nach Bedarf, sodass die Unternehmen nur für den tatsächlichen Verbrauch zahlen. Unternehmen können ML-Echtzeitanwendungen entwickeln, deren Bandbreite von personalisierten Empfehlungen bis zur Betrugserkennung reicht. Model Serving unterstützt auch LLM-Anwendungen für Benutzerinteraktionen.

Wir haben stetiges Wachstum bei der Akzeptanz von Serverless Data Warehousing und Monitoring beobachtet, die ebenfalls nachfragegesteuert skalieren.

Bei den Finanzdienstleistern – dem größten Anwenderbereich von Serverless-Produkten – stieg die Nutzung innerhalb von 6 Monaten um 131 %. Hier sind Marktprognosen unverzichtbar, und Echtzeitprognosen erlauben bessere Marktanalysen.

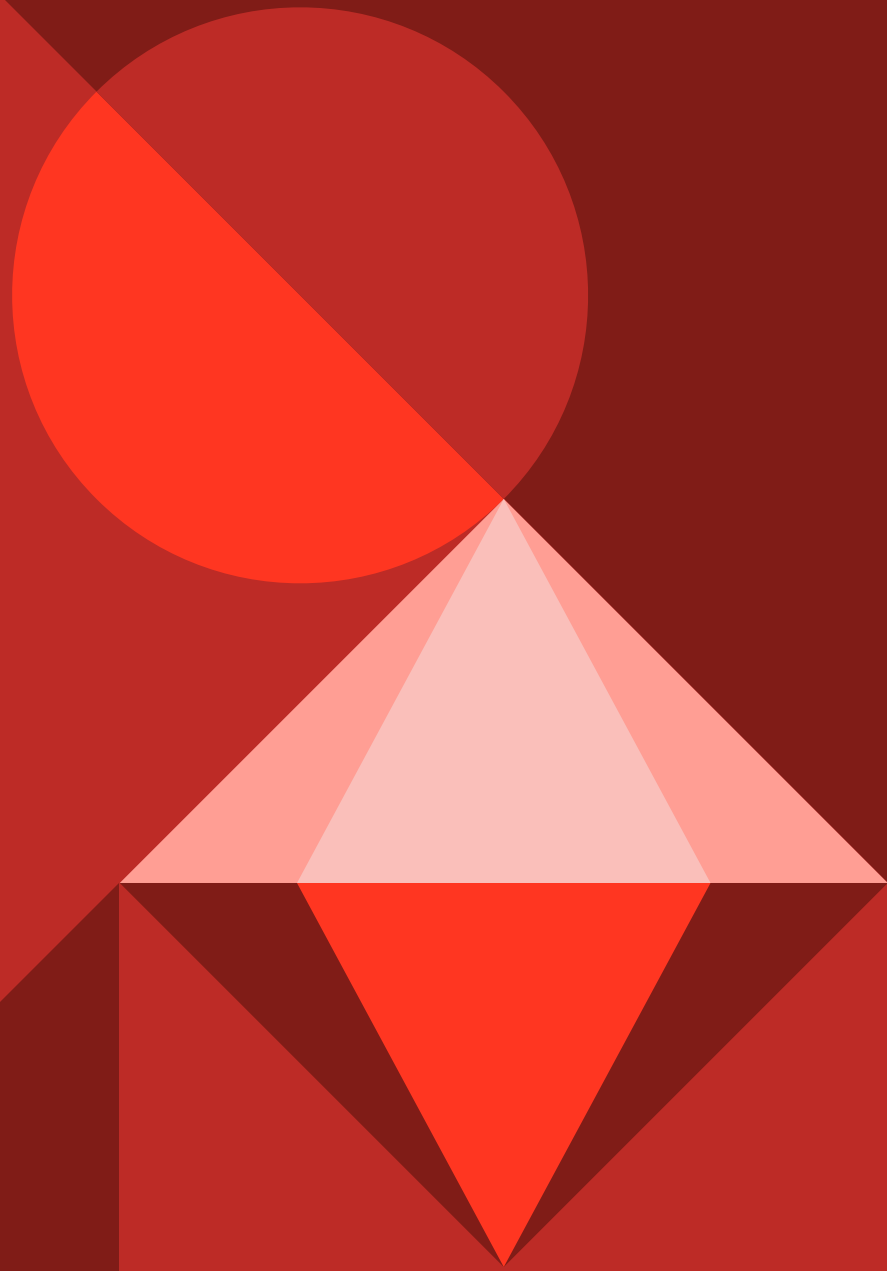
Im Bereich Gesundheitswesen und Life Sciences nahm die Nutzung in einem halben Jahr um 132 % zu. Die Branche ist im letzten Jahr von Platz 4 auf Platz 2 vorgerückt. Hier treten erhebliche Schwankungen bei den Anforderungen an die Datenverarbeitung auf – insbesondere in Spitzenzeiten oder bei großen Datasets wie Genomdaten oder medizinischer Bildgebung.

Fazit

Data Science und KI verhelfen Unternehmen zu mehr Effizienz, und GenAI eröffnet eine Vielzahl neuer Möglichkeiten. Mit Data-Intelligence-Plattformen erhalten Unternehmen einen zentralen und verwalteten Ort für die Nutzung von Daten und KI. Unsere Daten zeigen, dass Unternehmen aus allen Branchen diese Tools nutzen – und dass Early Adopters oft aus eher unerwarteten Branchen stammen.

Unternehmen erzielen durch den Einsatz von ML-Modellen in der Produktion messbare Gewinne. Sie setzen zunehmend NLP ein, um aus Daten Informationen zu generieren. Mithilfe von Vektordatenbanken und RAG-Anwendungen integrieren sie ihre eigenen Unternehmensdaten in ihre LLMs. Open-Source-Tools gehört die Zukunft, denn sie sind nach wie vor unsere beliebtesten Produkte. Unternehmen verfolgen Strategien für Unified Governance für Daten und KI.

Die Quintessenz: Wer Daten und KI am effektivsten nutzt, wird am Ende als Sieger aus dem Wettbewerb hervorgehen.



Über Databricks

Databricks ist das Unternehmen für Daten und KI. Mehr als 10.000 Unternehmen weltweit – darunter Block, Comcast, Condé Nast, Rivian, Shell und mehr als 60 Prozent der Fortune 500 – setzen auf die Databricks Data Intelligence Platform, um ihre Daten zu steuern und sie mithilfe von KI zu verwerten. Databricks wurde von den Erfindern von Lakehouse, Apache Spark™, Delta Lake und MLflow gegründet und hat seinen Hauptsitz in San Francisco mit Niederlassungen auf der ganzen Welt.

Wenn Sie mehr erfahren möchten, folgen Sie Databricks auf LinkedIn, X und Facebook.

databricks