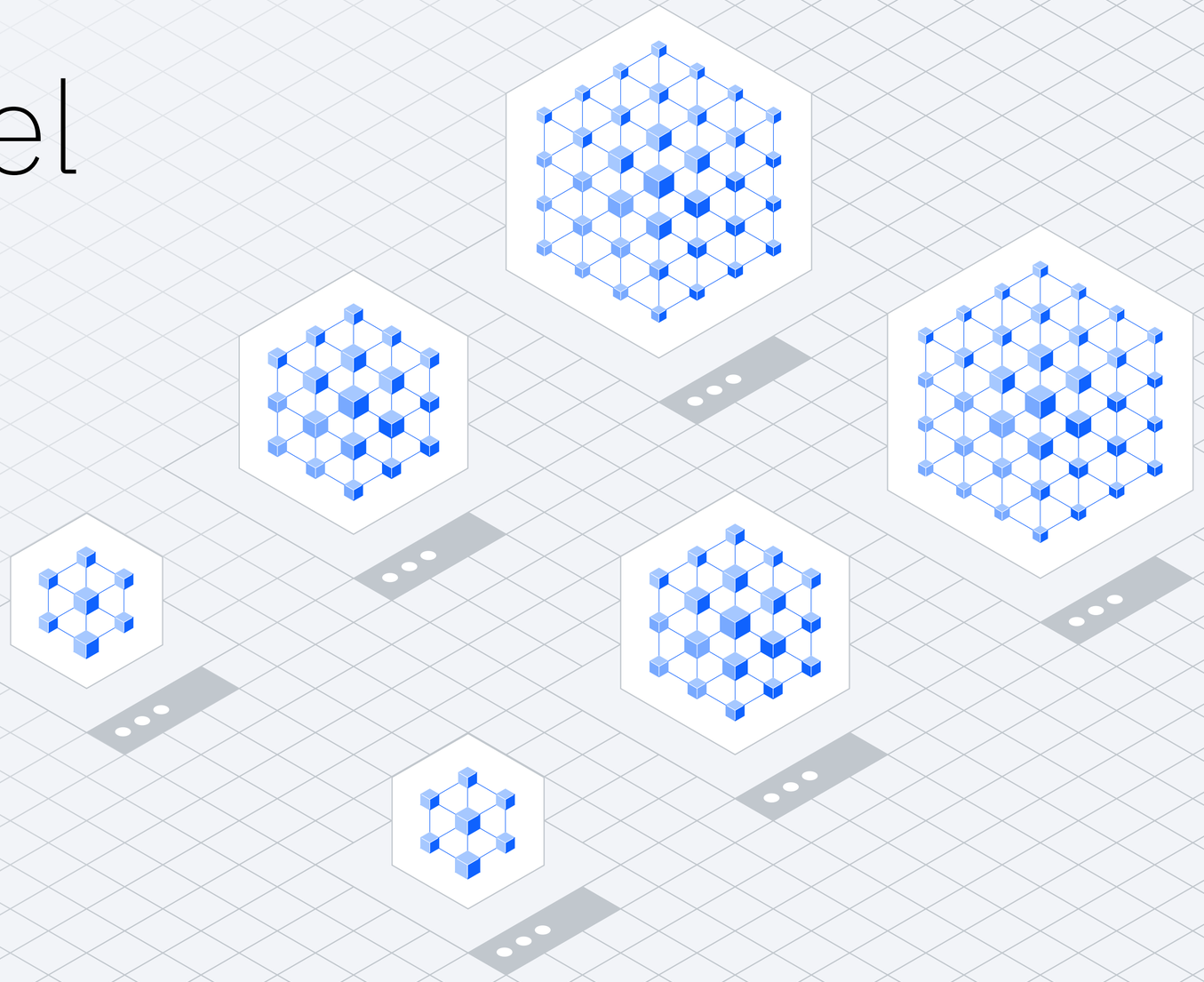
So wählen Sie das richtige Foundation Model für die KI



Inhalte

01

Einführung

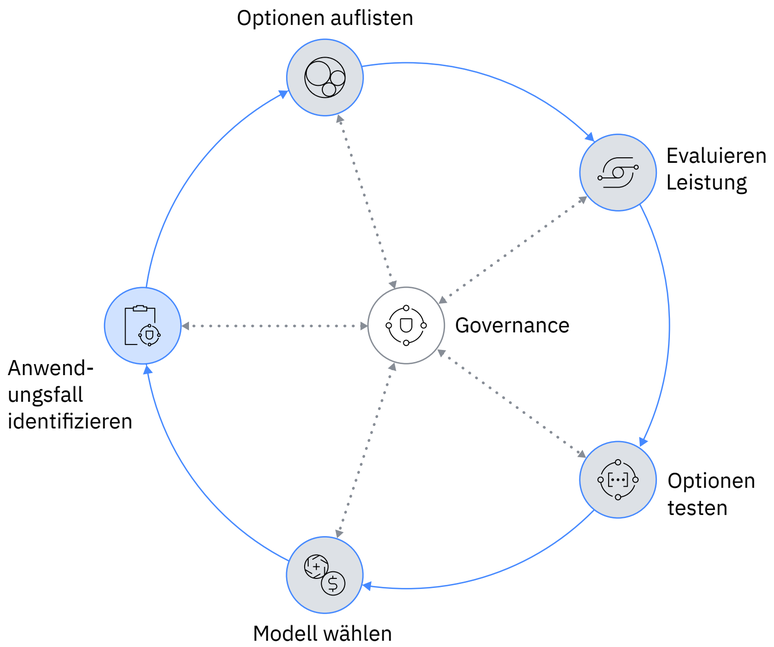
Die meisten Unternehmen sind sich darüber im Klaren, welche Ergebnisse sie von generativer KI erwarten. Weniger klar ist jedoch, wie man diese Ergebnisse erreichen kann. Unterschiedliche Ergebnisse erfordern unterschiedliche Ansätze – hinsichtlich der Datensätze, die Sie vorbereiten, und der KI-Modelle, die Sie verwenden.

Die Wahl des falschen Modells kann schwerwiegende Auswirkungen auf alle Aspekte des Unternehmens haben – von den Finanzen über die Strategie und die Rechtsabteilung bis hin zu Ihrer Belegschaft. Die Risiken reichen von Bias, der von den Trainingsdaten oder Algorithmen eines bestimmten Modells ausgeht, bis hin zu einem fehlerhaften Ergebnis eines vorgelagerten Modells, das sich zu einem großen Problem auswachsen kann, was möglicherweise zu Rechtsstreitigkeiten und Reputationsschäden führt.

Ein Framework zur Evaluierung hilft Ihnen, die unterschiedlichen Bedürfnisse und Fähigkeiten aller Arten von KI-Entscheidungsträgern und -Nutzern in Ihrem Unternehmen zu berücksichtigen. Die Endnutzer von KI-Modellen können von Data Scientists und Ingenieuren für maschinelles Lernen bis hin zu Business-Analysten, Rechts- und Compliance-Teams und Entscheidungsträgern reichen. Es ist wichtig, all ihre spezifischen Anforderungen zu berücksichtigen, damit Sie das richtige Modell für jeden Anwendungsfall identifizieren und die KI erfolgreich implementieren können, um den ROI zu steigern.

Framework für die Auswahl von KI-Modellen

Im Mittelpunkt dieses fünfstufigen zyklischen Prozesses zur Modellauswahl steht die Governance.



Beschreiben Sie klar und deutlich Ihren Anwendungsfall

Dies könnte beispielsweise Textgenerierung sein. Sagen wir, Sie möchten, dass die KI personalisierte E-Mails für eine Verkaufs- oder Marketingkampagne schreibt.

Listen Sie alle Modelloptionen auf und ermitteln Sie die Größe, Leistung und Risiken der einzelnen Modelle

Lassen Sie uns für dieses Beispiel zwei Modelle vergleichen, die für die Texterstellung entwickelt wurden: Ein 70B großes Allzweckmodell und ein 13B spezialisiertes Modell.

Bewerten Sie Modellgröße, Leistung und Risiken

Einer der Nachteile des 70B-Allzweckmodells ist seine geringere Geschwindigkeit; Andererseits verspricht es eine ganz besonders hohe Genauigkeit. Das 13B-Spezialmodell ist schneller als das 70B-Allzweckmodell, besitzt aber eine geringere Genauigkeit, da es mit einem kleineren Datensatz trainiert wurde. Das Risiko ist in dieser Phase für beide Modelle auch ein wichtiger Faktor.

Optionen testen

Wählen Sie das Modell aus, das voraussichtlich die gewünschte Leistung liefert. Prüfen Sie, ob es funktioniert. Beurteilen Sie die Leistung des Modells und die Qualität der Ausgabe anhand von Metriken wie Perplexität oder BLEU-Score. Versuchen Sie, die gleiche Leistung mit kleineren Modellen zu erreichen, indem Sie Techniken wie Prompt Engineering und Modell-Tuning einsetzen und Ihre Auswahl verfeinern, um die Kosten und Ihre Einsatzanforderungen zu optimieren. Fügen Sie Ihre Datensätze zum Stack hinzu und testen Sie die Modelle, um die Genauigkeit zu verbessern.

Entscheiden Sie sich für die Option mit dem größten Mehrwert

Ob Sie hohe Geschwindigkeit oder hohe Genauigkeit benötigen, hängt vom Anwendungsfall, Ihren Kostenvorgaben und den Bereitstellungsmethoden ab. Vielleicht entscheiden Sie sich am Ende für das spezialisierte Modell 13B, wenn der für Sie wichtigste Aspekt die schnelle Texterstellung ist.

In den folgenden Kapiteln werden wir die wichtigsten Überlegungen zur Auswahl eines KI-Modells im Detail untersuchen: Anwendungsfall, Größe, Leistung, Risiko, Kosten und Bereitstellungsanforderungen.

03

Identifizierung eines klaren Anwendungsfalls

Bei der Beurteilung von KI-Modellen ist eines der wichtigsten Bewertungskriterien, wie gut sich die Funktionalitäten des Modells mit den Anforderungen Ihres Anwendungsfalls überschneiden.

Der einfachste Weg, das richtige Modell für Ihren Anwendungsfall zu finden, besteht darin, die Anfrage und die ideale Antwort zu formulieren und die Arbeit danach auszurichten um die Daten zu finden, die für die gewünschte Antwort erforderlich sind.

Arbeiten Sie eng mit den Produkt- und Entwicklungsteams und den Sponsoren aus dem Unternehmen zusammen, um die tatsächlichen Eingabeaufforderungen zu verstehen, die Sie zur Lösung der anstehenden Geschäftsprobleme benötigen. Berücksichtigen Sie die Besonderheiten Ihres Unternehmens (z. B. Branchenterminologie und Standarddefinitionen) als Teil der Eingabeaufforderung und machen Sie die Eingabeaufforderungen so spezifisch für die Anwendungsfälle wie möglich. Sind Sie in der Lage den Anwendungsfall in Eingabeaufforderungen aufzuschlüsseln? Können Sie die Besonderheiten Ihres Unternehmens an mehreren Stellen in die Eingabeaufforderungen integrieren?

All dies ist wichtig, da selbst subtile Nuancen in den Eingabeaufforderungen einen großen Unterschied machen können, wenn es um die Auswahl eines ganz bestimmten Modells geht. Um dies zu veranschaulichen, betrachten Sie die Aufforderung „Geben Sie mir die ID-Nummer von John Doe.“ In einer HR-Funktion wäre damit die Mitarbeiter-ID gemeint, im Kundenservice könnte es jedoch die Kundenbindungsnummer sein. Je spezifischer die Eingabeaufforderung ist, desto genauer ist die Antwort. Eine spezifischere Eingabeaufforderung könnte Sie dazu bringen, ein Modell auszuwählen, das bereits zwischen verschiedenen persönlichen ID-Nummern unterscheiden kann, ohne dass zusätzliches Training erforderlich ist.

Die Suche nach dem richtigen Partner
Möchten Sie beurteilen, ob ein Modell für Sie geeignet ist, überprüfen Sie die Modellkarte, um zu sehen, ob es ein Modell gibt, das speziell für Ihren Zweck mit Daten trainiert wurde. Vortrainierte Foundation Models sind auf bestimmte Anwendungsfälle wie Stimmungsanalyse oder Dokumentzusammenfassung abgestimmt und ermöglichen Ihrem Team die Verwendung von Zero-Shot-Prompting, um die gewünschten Ergebnisse zu erzielen. Dieser Approach ermöglicht auch schnelles Experimentieren mit gezielten, bereichsspezifischen Modellen. Da weniger interne Schulungen und Fachkenntnisse erforderlich sind, um die Modelle an Ihre Bedürfnisse anzupassen, können Sie eine schnellere Wertschöpfung erzielen und Wettbewerbsvorteile aufbauen.

Generative KI-Anwendungsfälle mit dem größten zu erwartenden Potenzial für Unternehmen, laut IDC.[1]

- Wissensmanagement (46 %)
- Konversationsanwendung (46 %)
- Designanwendungen (44 %)
- Codegenerierungsanwendungen (43 %)
- Marketinganwendungen (36 %)

Die beliebtesten Anwendungen von KI im aktuellen Szenario, laut IDC.[1]

- Softwareentwicklung (29,4 %)
- Produktentwicklung/-design (24,7 %)
- Kundenbindung (23,4 %)
- Lieferkette (20,5 %)
- Finanzen (18,2 %)
- Umsatz (18 %)
- Forschung und Entwicklung (15,4 %)
- Personalwesen (15,2 %)
- Fertigung (15 %)
- Marketing/PR (13,9 %)

Fragen Sie sich selbst: Benötige ich ein großes Modell, um meine Aufgabe zu erfüllen?

In den meisten Fällen lautet die Antwort darauf nein. Der bessere Ansatz besteht darin, die Größe des Modells an Ihren spezifischen Anwendungsfall anzupassen.

04

Bewertung von Größe, Leistung und Risiken

Die richtige Dimensionierung eines Modells für Ihren Anwendungsfall

Große KI-Modelle sind so leistungsfähig, weil sie auf riesigen Datenmengen trainiert werden. Allein die schiere Größe der Trainingsdaten dieser Modelle ermöglicht es ihnen, die komplexen Muster und nuancierten Verbindungen in den Daten zu erfassen und qualitativ hochwertigen Output zu generieren. Sie benötigen jedoch mehr Rechenleistung, sind teurer im Betrieb und dabei nicht immer genauer. Die Frage, die Sie sich stellen müssen, lautet also: Benötige ich ein großes Modell, um meine Aufgabe zu erfüllen? In den meisten Fällen lautet die Antwort darauf nein.

Der beste Ansatz besteht hier darin, die Größe des Modells an Ihren spezifischen Anwendungsfall anzupassen. Beginnen Sie mit dem größten oder leistungsstärksten Modell und nutzen Sie eine grundlegende Eingabeaufforderung, um Ihre optimale Leistung zu erzielen. Verkleinern Sie nun Ihr Modell und verwenden Sie Techniken wie das Prompt-Tuning, um zu sehen, ob Sie die gleichen Ergebnisse erzielen können. Das Prompt-Tuning eines kleineren Modells ist weitaus kosteneffizienter als die Feinabstimmung desselben Modells, die beträchtliche Daten- und Rechenressourcen erfordert, und dennoch genaue und kontextbezogene Antworten liefert.

Wenn Sie sich für ein größeres Allzweck-LLM entscheiden, z. B. für die Überprüfung von Lebensläufen, können Sie HR-spezifische Eingabeaufforderungen und Modelltuning anwenden, um die gewünschten Ergebnisse zu erzielen. Zu den effektiven Techniken der Modellabstimmung gehört das Prompt-Tuning mit für den Anwendungsfall spezifischen Daten, um optimale Ergebnisse zu erzielen. Wenn Sie die Frage „Wie hoch waren die jährlichen Kosten für das Onboarding in den letzten 10 Jahren?“ in „Wie hoch waren die jährlichen Kosten für das Onboarding neuer Mitarbeiter in den letzten 10 Jahren?“ ändern, werden Sie womöglich eine genauere und relevantere Antwort erhalten. Die frühere Eingabeaufforderung hätte Informationen über das Onboarding neuer Kunden statt über das Onboarding neuer Mitarbeiter liefern können.

Bewertung der Modellleistung

Bei der Leistungsbewertung eines Modells sind Genauigkeit, Zuverlässigkeit und Geschwindigkeit die wichtigsten Kriterien. Die Gewichtung der einzelnen Kriterien ist von Anwendungsfall zu Anwendungsfall unterschiedlich. Abhängig von den spezifischen Anforderungen des Anwendungsfalls ist häufig ein Kompromiss zwischen diesen Faktoren erforderlich.

Genauigkeit
Genauigkeit kann objektiv und wiederholt gemessen werden und gibt an, wie nahe der generierte Output zu dem gewünschten Output liegt. Dies ist die primäre Methode zur Bewertung eines Modells – im Vergleich zu Branchen-Benchmarks. Wählen Sie Bewertungsmetriken aus, die für Ihren Anwendungsfall relevant sind. So kann beispielsweise die Leistung der Inhaltszusammenfassung oder der RAG durch ROUGE (Recall-Oriented Understudy for Gisting Evaluation) bewertet werden, während BLEU (Bilingual Evaluation Understudy) die Qualität der Textübersetzung anzeigt.

Reliabilität
Zuverlässigkeit ist ein Maß dafür, wie gut das Modell den gleichen Output erzeugt. Sie hängt von mehreren Faktoren ab, wie z. B. Konsistenz, Erklärbarkeit und Vertrauenswürdigkeit, sowie davon, wie gut ein Modell Toxizität (Hassreden, anstößige Sprache) und Bias vermeidet. Dies ist insbesondere bei nach außen gerichteten Anwendungen ein entscheidender Gesichtspunkt. Normalerweise ist ein Modell zuverlässiger, das Transparenz hinsichtlich seiner Trainingsmethodik und -daten bietet, da es wichtige Probleme wie Governance, Risiken und Datenschutz berücksichtigt.

Geschwindigkeit
Bei der Geschwindigkeit geht es darum, wie schnell ein Benutzer eine Antwort auf eine Frage erhält. Dies ist insbesondere bei Echtzeitanwendungen von entscheidender Bedeutung, da hier eine geringe Latenzzeit erforderlich ist. Bedenken Sie Anwendungsfälle, die von Chatbots bis zum Finanzhandel reichen, wo zeitnahe Antworten den entscheidenden Unterschied ausmachen können, im Vergleich zu Finanzprognosen, bei denen Genauigkeit wichtiger ist. Der Kompromiss zwischen Geschwindigkeit und Genauigkeit ist hier eine entscheidende Überlegung, und die Priorisierung eines Faktors gegenüber dem anderen ist eine Entscheidung, die ganz vom vorliegenden Anwendungsfall abhängt.

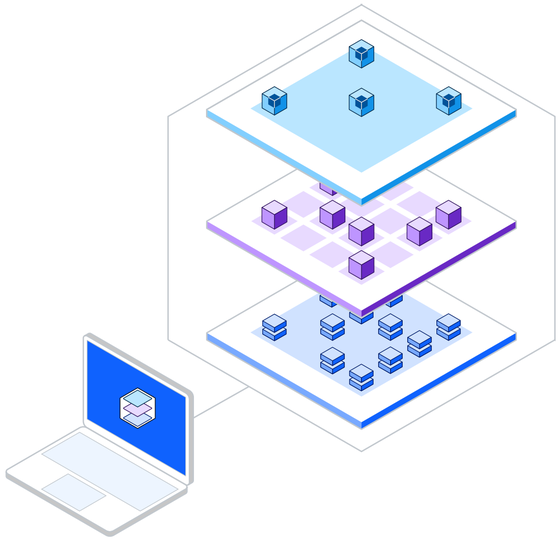
Bewertung von Risiken und Governance

Für jedes Unternehmen sind Datenschutz und Sicherheit wichtige Aspekte, unabhängig vom Anwendungsfall. Auch die Transparenz und Nachvollziehbarkeit der Trainingsdaten sowie die Genauigkeit und Zuverlässigkeit der Ergebnisse – die frei von Verzerrungen, Toxizität und Bias sein sollten – sind entscheidende Aspekte bei der Modellauswahl.

Daher gilt die KI-Governance für den gesamten Auswahlprozess – von der Leistungsbewertung und -optimierung über die schnelle Entwicklung bis hin zur Validierung und Kostenkontrolle. Das ist die ein wesentlicher Bestandteil, den Sie benötigen, um das richtige Modell für Ihren Anwendungsfall mit Attributen wie Risiko, Transparenz, Zuverlässigkeit und Vertrauenswürdigkeit auszuwählen, und es bleibt ein fortlaufender Prozess während das Modell einer kontinuierlichen Überwachung und Bewertung unterzogen wird.

In fast allen Anwendungsfällen ist eine vollständige Transparenz der Trainingsmethodik von entscheidender Bedeutung, um einen verantwortungsvollen Einsatz der Modelle zu ermöglichen und wichtige Fragen wie Governance, Risikobewertung, Datenschutz und Vermeidung von Bias zu klären. Ein hochtransparentes Modell ermöglicht es den Nutzern, eine größere Zuverlässigkeit und Vertrauenswürdigkeit zu erreichen, da sie die Leistung und die operativen Risiken leicht überwachen und optimieren können. Und dann gibt es noch Modelle, die einen Schritt weiter gehen und den Vorteil eines vertraglichen Schutzes für die Nutzung oder Entschädigung beinhalten.

Wenn Ihr Ökosystem Datenlücken aufweist oder der Datenschutz aufgrund Ihres Anwendungsfalls gefährdet ist, sollten Sie sich nach einem Modell umsehen, das synthetische Tabellendaten generieren kann, um die Lücken zu schließen und es den Benutzern zu ermöglichen, das Modell mit minimalem Risiko des Datenschutzes oder des Bias zu trainieren.

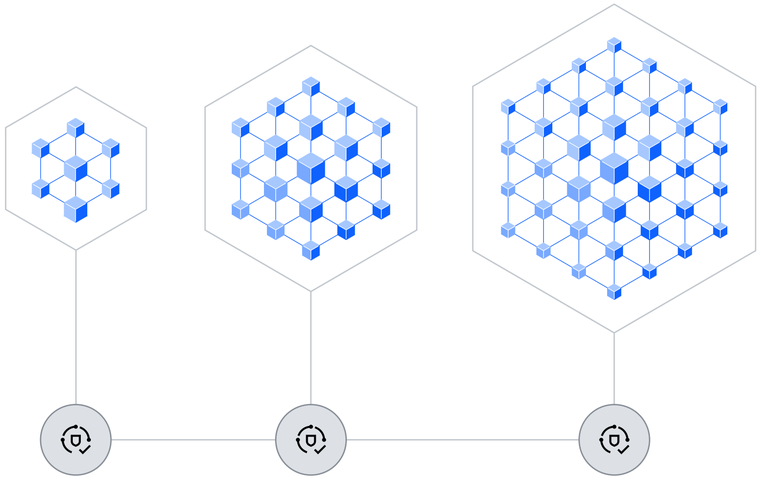


Boston Scientific wendete KI-Modelle aus dem Open-Source-Bereich auf seinen Inspektionsprozess an. Das Ergebnis waren direkte Einsparungen in Höhe von fünf Millionen US-Dollar bei einem bescheidenen Budget von etwa 50.000 US-Dollar sowie eine Genauigkeit, die den bestehenden Inspektionsprozess übertraf.[2]

05

Verfeinerung der Auswahl auf Basis von Kosten und Bereitstellungsbedarf

Bei der Modellauswahl sind die wichtigsten Kriterien der Anwendungsfall, die Größe und Leistung des Modells und die Art der Bereitstellung, aber auch die Kosten und der ROI.



In der Regel geht es nicht nur um die Genauigkeit der Aufgaben, sondern auch um die praktische Umsetzung von ROI und Kosteneffizienz. Bei der Auswahl der Modelle spielt der Kostenfaktor eine wichtige Rolle. Ein teureres, größeres Modell kann etwas genauer sein als ein deutlich kleineres, billigeres. Die Frage ist jedoch: Bietet Ihnen das teure Modell den ROI, der seinen Einsatz für diesen speziellen Anwendungsfall rechtfertigt? Die Antwort ist nicht immer ja.

Es kommt wirklich darauf an, den Sweet Spot zwischen Leistung, Geschwindigkeit und Kosten zu finden. Ein kleineres, preiswerteres Modell bietet vielleicht nicht die gleiche Leistung oder Genauigkeit wie ein teureres Modell, ist aber dennoch dem letzteren vorzuziehen, wenn Sie die zusätzlichen Vorteile berücksichtigen, die das Modell bieten könnte, wie z. B. eine geringere Latenz und eine größere Transparenz der Modelleingaben und -ausgaben. Das kleinere Modell könnte beispielsweise in Ihrem Unternehmen für mehrere Anwendungsfälle skaliert werden, was seinen Gesamtwert erhöht. Auf der anderen Seite könnten Sie sich für das größere Modell entscheiden, wenn Ihr Anwendungsfall ein hohes Maß an Genauigkeit und Präzision bei den Ergebnissen erfordert – allerdings gibt es einen Punkt, an dem der ROI abnimmt, sodass es am Ende eher eine Frage der Abwägung ist.

Eine andere Möglichkeit, sich dem Kompromiss zwischen Größe und Leistung zu nähern, besteht darin, Prompt-Tuning-Techniken auf ein kleineres Modell anzuwenden, um die gleichen oder bessere Ergebnisse zu erzielen als beim Prompt-Engineering eines größeren, teuren Modells. Prompt-Tuning ist eine effiziente, kostengünstige Methode zur Anpassung eines Modells an einen bestimmten Anwendungsfall, da das Modell nicht neu trainiert werden muss und mit Ihren vorhandenen Ressourcen durchgeführt werden kann.

Verfeinerung der Auswahl auf Basis von Kosten und Bereitstellungsbedarf

Ein weiterer entscheidender Faktor, der sich auf die Gesamtkosten und den Energieverbrauch auswirkt, ist die Methode der Modellbereitstellung und die entsprechenden Anforderungen an die Infrastruktur und den Grafikprozessor.

LLMs sind zweifellos ressourcenintensiv. Wie können Sie diese Modelle in Ihr Unternehmen integrieren, ohne Ihre ESG-Ziele wesentlich zu beeinträchtigen? Es ist wichtig zu überlegen, wie die Modelle nachhaltig bereitgestellt werden können, um die Gesamtbetriebskosten niedrig zu halten.

Die Bereitstellungsentscheidung

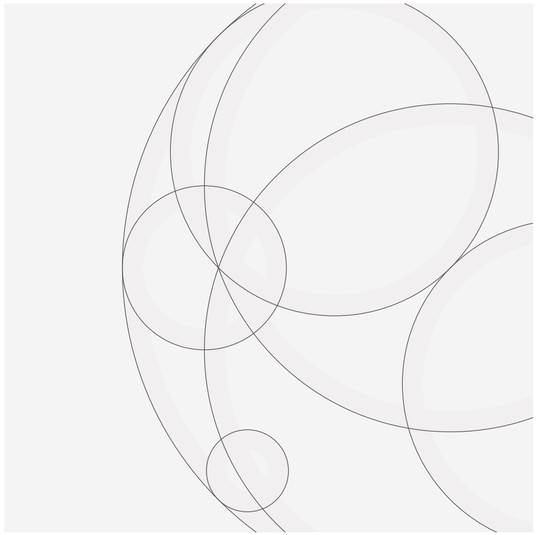
Eine weitere wichtige Überlegung bei der Bewertung der Modellauswahlkosten ist, wo und wie Sie das Modell und die Daten bereitstellen möchten. Die KI-Modelle, die Sie auswählen, sollten mit Ihren Anwendungen, Ihren bestehenden Anbietern und Partnern sowie Ihrer gesamten KI- und Datenplattform kompatibel sein – unabhängig davon, ob diese vor Ort, in der Cloud oder in einer hybriden Umgebung betrieben wird.

Möchten Sie beispielsweise vielleicht mit einem Open-Source-Modell wie Flan-UL2 arbeiten. Wenn Sie es jedoch mit Ihren eigenen Unternehmensdaten trainiert haben, müssen Sie es möglicherweise lokal bereitstellen. Lokal haben Sie im Vergleich zu einer Public-Cloud-Umgebung mehr Kontrolle und Sicherheit. Dies ist jedoch ein teures Unterfangen – vor allem, wenn man die Größe des Modells und die Rechenleistung berücksichtigt, einschließlich der Anzahl der GPUs, die für die Ausführung eines einzigen großen Sprachmodells erforderlich sind. Bei der Entscheidung für das richtige Modell und die richtige Bereitstellungsmethode kommt es also darauf an, die Kosten für das Modell-Hosting, die Leistung, die Sicherheit und die Governance-Aspekte abzuwägen.

06

So hilft eine KI- und Datenplattform

Um das wahre Potenzial von KI zu nutzen, sollten Unternehmen aufhören, KI als Zusatz zu bestehenden Systemen zu betrachten und sie stattdessen in den Kern ihres Geschäfts einbetten.



Dies ist jedoch leichter gesagt als getan. Bei der Implementierung von KI gibt es verschiedene Herausforderungen: Diese reichen von der Datenqualität und -verfügbarkeit über die Datensicherheit und den Datenschutz bis hin zur Interoperabilität mit bestehenden Systemen, der Skalierbarkeit für eine unternehmensweite Einführung sowie den Kosten. Wie überwachen Sie bei der Auswahl eines Modells dessen Genauigkeit und Vertrauenswürdigkeit – vor allem, wenn es sich um ein geschlossenes, proprietäres Modell handelt? Was können Sie tun, um die Kosten und den Energieverbrauch von groß angelegten Modellen zu minimieren? Wie machen Sie Ihre KI verantwortungsvoll, transparent und erklärbar?

Sie benötigen die richtige KI- und Datenplattform, um diese Bedenken auszuräumen und die Einführung von KI im gesamten Unternehmen voranzutreiben. IBM® watsonx ist eine vollumfängliche KI- und Datenplattform, die es Ihnen ermöglicht, die Wirkung von KI mit vertrauenswürdigen Daten zu skalieren und zu beschleunigen.

Mit watsonx können Sie KI in Ihrem gesamten Unternehmen trainieren, abstimmen und bereitstellen und dabei kritische Daten verwenden, wo auch immer sie sich befinden. Zu den wichtigsten Komponenten der KI- und Datenplattform watsonx gehören:

- IBM® watsonx.ai, ein Studio für Foundation Models, generative KI und maschinelles Lernen. Das watsonx.ai Studio bietet eine Vielzahl von Foundation Models– darunter proprietäre, Open-Source- und Drittanbieter-Modelle.
- IBM® watsonx.data, ein zweckmäßiger Datenspeicher, der auf einer offenen Data Lakehouse Architektur basiert.
- IBM® watsonx.governance, eine Lösung, die Unternehmen das nötige Toolkit an die Hand gibt, um Risiken zu managen, Transparenz zu schaffen und die Einhaltung von künftigen KI-spezifischen Vorschriften zu gewährleisten.

Das watsonx.ai Studio ist für einen Multimodell-Ansatz konzipiert. Um die Auswahl und Skalierung von Modellen zu vereinfachen, bietet das Studio einen hybriden Full-Stack Approach an, der Folgendes umfasst:

- Bibliothek mit von IBM entwickelten Foundation Models und ausgewählten Drittanbieter- und Open-Source-Modellen von Hugging Face.
- Prompt Lab zum Experimentieren mit Foundation Models und für die Entwicklung von Prompts für verschiedene Anwendungsfälle und Aufgaben.
- Tuning-Studio, das Ihnen dabei hilft, Ihre Foundation Models mit gekennzeichneten Daten für bessere Leistung und Genauigkeit abzustimmen.
- Data Science- und MLOps-Toolkit zur automatischen Erstellung von Modellen für maschinelles Lernen (ML) mit Modelltraining, Entwicklung, visueller Modellierung und Generierung synthetischer Daten.

Das watsonx.ai Studio bietet eine Vielzahl von Modellgrößen, die für verschiedene Anwendungsfälle trainiert wurden. Diese Modelle können vor Ort, in der Cloud oder in einer hybriden Umgebung eingesetzt werden und bieten somit große Flexibilität.
Sie können sich über die Vorteile der Zusammenarbeit mit dem Studio, die derzeit verfügbaren Foundation Models und ihre spezifischen Anwendungsfälle informieren.

Erkunden Sie Foundation Models innerhalb der watsonx-Plattform →

07

Zusammenfassung

Nicht alle KI-Modelle sind identisch, und auch Ihre Anwendungsfälle sind es nicht.

Jeder spezifische Anwendungsfall erfordert ein passendes KI-Modell. Das erklärt, warum ein Multi-Modell-Ansatz für den Erfolg mit generativer KI von zentraler Bedeutung ist. Letztendlich benötigen Sie zuverlässige, leistungsstarke und kosteneffiziente Foundation Models, die es Ihnen ermöglichen, verschiedene Parameter wie Kosten, Leistung und Risiko auf der Grundlage Ihrer Anwendungsfälle zu optimieren.

Mit der Flexibilität, den richtigen Modellmix zu kuratieren, können Sie viel erreichen:

- **Reduzieren Sie die Gesamtbetriebskosten** für Modelltraining, Inferenz, Tuning, Hosting, Berechnung und Produktion.
- **Optimieren Sie Rechenleistung und Kosten** sowie die Skalierbarkeit von Modellen über verschiedene Anwendungsfälle und Bereiche hinweg für einen optimalen ROI.
- **Nutzen Sie intuitive Benutzeroberflächen**, die das Human-in-the-Loop-Lernen erleichtern, um die Relevanz- und Genauigkeitswerte sowie die Leistung der Modelle nach Ihren Anforderungen zu verbessern.
- **Wählen Sie Modelle, die Transparenz** hinsichtlich der Schulungsmethodik bieten und vertraglichen IP-Schutz bieten, um eine verantwortungsvolle Bereitstellung und Nutzung zu ermöglichen.
- **Erstellen Sie Modelle mit eingebauten Verhaltensregeln** und etablieren Sie Best Practices, um wichtige Fragen wie Governance, Risikobewertung, Datenschutz und Vermeidung von Bias zu klären. So erhalten Sie ein Ergebnis, dem Sie hinsichtlich optimaler Leistung, Genauigkeit, Sicherheit und Zuverlässigkeit vertrauen können.

Der Einsatz von generativer KI kann die Geschäftsstrategie, die Produktpläne, das Talentmanagement, die Customer Experience und verschiedene andere Geschäftsbereiche beeinflussen. Allerdings sind Modellauswahl und -optimierung keineswegs ein einmaliger Prozess. Es ist wichtig, dass Sie jeden KI-Anwendungsfall hinsichtlich Relevanz, Modellgröße und Leistung sowie Bereitstellungsmethoden kontinuierlich überprüfen, um einen optimalen ROI und optimale Geschäftsergebnisse zu erzielen.

Nächste Schritte

Die KI- und Automatisierungsexperten von IBM können mit Ihnen zusammenarbeiten, um einen KI-Modellmix zusammenzustellen – und die Modelle zu operationalisieren –, um Ihre spezifischen Anwendungsfälle und Geschäftsanforderungen zu erfüllen. Erfahren Sie zunächst mehr über die im watsonx.ai-Studio angebotenen Modelle und die Vorteile von IBM watsonx, der vollumfänglichen KI- und Datenplattform für Unternehmen.

Demo buchen →
Mehr über watsonx.ai erfahren →

1. Future Enterprise Resiliency & Spending Survey Wave 2, IDC, March 2023

2. How to create business value with AI: 12 stories from the field, IBM IBV 2022-08-16

IBM Deutschland GmbH
IBM-Allee 1
71139 Ehningen
ibm.com/de
IBM Corporation
New Orchard Road
Armonk, NY 10504, USA

Hergestellt in den Vereinigten Staaten von Amerika
Januar 2024

Der Inhalt dieses Dokuments (einschließlich der Währungs- oder Preisangaben, die keine Steuern enthalten) ist zum Zeitpunkt der Erstveröffentlichung aktuell und kann von IBM jederzeit geändert werden. Nicht alle Angebote sind in allen Ländern verfügbar, in denen IBM tätig ist.

Es liegt in der Verantwortung der Anwender, die Nutzbarkeit anderer Produkte oder Programme neben den Produkten und Programmen von IBM zu evaluieren und verifizieren.

Erklärung zu bewährten Sicherheitsverfahren: Kein IT-System oder -Produkt sollte als vollkommen sicher angesehen werden, und kein einzelnes Produkt, kein Service und keine Sicherheitsmaßnahme kann eine missbräuchliche Nutzung oder einen missbräuchlichen Zugriff vollständig verhindern. IBM übernimmt keine Gewähr dafür, dass Systeme, Produkte oder Services vor böswilligem oder rechtswidrigem Verhalten von Dritten geschützt sind oder Ihr Unternehmen davor schützen.

Die Einhaltung sämtlicher geltender Gesetze und Vorschriften liegt in der Verantwortung des Kunden. IBM bietet keine Rechtsberatung an und gewährleistet nicht, dass die Services oder Produkte von IBM die Konformität von Gesetzen oder Verordnungen durch den Kunden sicherstellen.