

Kritische KI-Infrastruktur neu denken

Wie sich die On-Device-Infrastruktur als Plattform für die Entwicklung
und Betrieb von KI in Unternehmen etabliert

FORSCHUNGSSTUDIE VON IM AUFTRAG VON

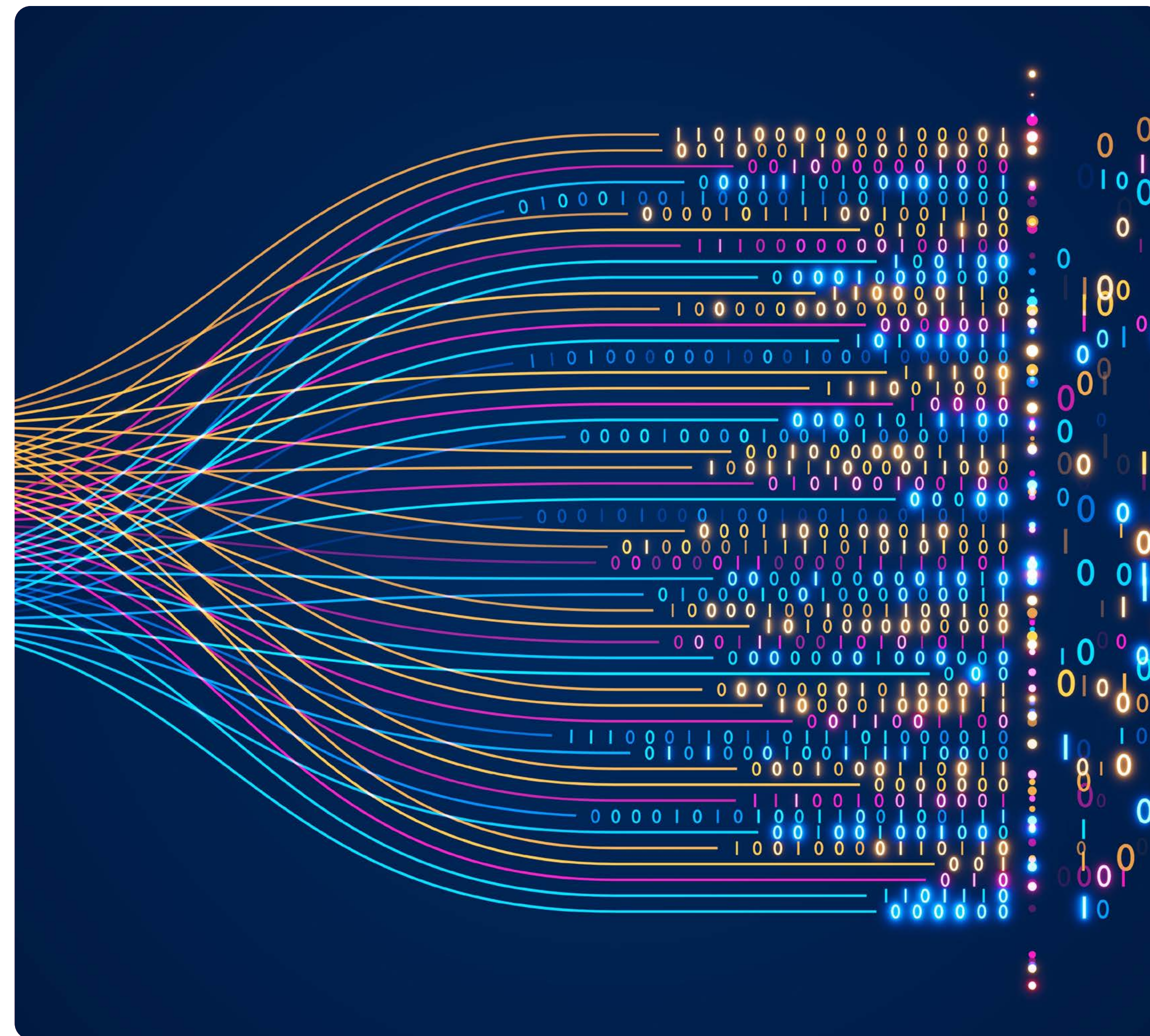


Kritische KI-Infrastruktur durch On-Device-Ansätze neu denken

In der sich rasant entwickelnden KI-Landschaft von heute stützen sich Unternehmen stark auf Cloud- und lokale Infrastrukturen, um die Entwicklung und den Einsatz von KI voranzutreiben. Viele Unternehmen entscheiden sich für einen hybriden Ansatz und nutzen die für die jeweilige Aufgabe am besten geeignete Infrastruktur. Viele haben jedoch noch nicht vollständig evaluiert, wo On-Device-KI messbare Vorteile bietet.

Erkenntnisse aus der exklusiven Studie von Omdia unter mehr als 1.500 Führungskräften und Fachleuten aus dem Bereich Unternehmenstechnologie – darunter CIOs, CTOs, Softwareentwickler:innen und KI-Spezialist:innen – zeigen, dass Unternehmen zunehmend hinterfragen, ob ihre aktuelle Infrastruktur drei entscheidende Anforderungen wirklich erfüllt: ihre Sicherheitsanforderungen, unvorhersehbare und laufende Kosten sowie die Notwendigkeit, die Infrastruktur an die Workload-Anforderungen anzupassen.

Diese Grundpfeiler definieren den Vorteil der On-Device-Verarbeitung und bieten eine architektonische Lösung für die Problemfelder in Bezug auf Sicherheit, Wirtschaftlichkeit und Arbeitslast, denen moderne Unternehmen gegenüberstehen.





Sicherheit

Datenschutz durch Architektur statt durch prozessbasierte Notlösungen

Drei Viertel der Unternehmen geben an, dass sie über mögliche Datenlecks durch Cloud-Dienste besorgt sind. On-Device-KI hat insofern einen klaren Vorteil, als Daten, die niemals übertragen werden, auch nicht während der Übertragung offengelegt werden können. Da 99 % der Unternehmen in ihren KI-Workflows proprietäre Daten verarbeiten, wird die Möglichkeit der lokalen Verarbeitung zu einer entscheidenden Sicherheitsanforderung – sie eliminiert Übertragungsrisiken und sichert KI-gestützte Workflows, ohne dass man sich auf prozedurale Kontrollen verlassen muss.



Wirtschaftlichkeit

Unbegrenzte Experimentiermöglichkeiten und vorhersehbare Produktionskosten

On-Device-Infrastruktur verursacht nach der Anfangsinvestition nahezu keine Grenzkosten mehr, sodass unbegrenzte Experimente ohne Budgetbeschränkungen möglich sind. Bei der Produktionsinferenz ersetzen feste Hardwarekosten die unvorhersehbaren Betriebskosten der Cloud, die mit dem Erfolg skalieren. Und im Gegensatz zu Cloud-Arbeitsplätzen, die bei geringer Akzeptanz durch die Nutzer:innen zu ungenutzten Ressourcen werden, behält die Hardware auf dem Gerät unabhängig von den Nutzungsmustern der KI ihren vollen Produktivitätswert.



Durchführbarkeit der Workload

Anpassung der Infrastruktur an Modelle

Unternehmen benötigen keine umfangreichen Foundation-Modelle, um Wert zu generieren – 57 % der Unternehmensmodelle haben weniger als 10 Milliarden Parameter und liegen damit deutlich innerhalb der Leistungsfähigkeit moderner Geräte wie dem MacBook Air oder das MacBook Pro der Einstiegsklasse. Überraschenderweise ist die Verlagerung hin zu On-Device-Lösungen im oberen Bereich sogar noch ausgeprägter: Unternehmen, die hauptsächlich On-Device-Infrastrukturen nutzen, setzen fast doppelt so häufig Modelle mit über 100 Milliarden Parametern ein.

Die aktuelle KI-Landschaft in Unternehmen

KI-reife Unternehmen setzen verstärkt auf hybride Ansätze durch mehr On-Device-Nutzung

Unternehmen, die sich in fortgeschrittenen Stadien der KI-Reife befinden, weisen eine höhere Akzeptanz von hybriden Bereitstellungsansätzen auf. Dies deutet darauf hin, dass jede Infrastrukturart unterschiedliche Zwecke für Unternehmen erfüllt, die KI in mehreren Geschäftsfunktionen einsetzen oder dies planen.

Die Infrastruktur auf dem Gerät übernimmt Entwicklungsworkflows, die Verarbeitung sensibler Daten, die verteilte Bereitstellung kleinerer Modelle über Teams hinweg und die Produktionsinferenz für Workloads, die eine lokale Verarbeitung erfordern. Während lokale GPU-Cluster für Batch-Training und gemeinsam genutzte Umgebungen mit mehreren Benutzern bevorzugt werden können, bietet die Cloud-Infrastruktur elastische Skalierung und Produktions-APIs, die eine hohe Verfügbarkeit erfordern.

Beachtenswert ist, dass Unternehmen sich nicht ausschließlich für eine Option entscheiden, sondern verschiedene Infrastrukturtypen strategisch kombinieren. Unter den Unternehmen, die Mac für KI-Workloads auf Geräten einsetzen, verwenden 56 % auch lokal NVIDIA-GPUs, was zeigt, dass Plattformen auf Geräten spezialisierte Rechenressourcen ergänzen und nicht ersetzen.

Das Muster spiegelt die Wirtschaftlichkeit der Arbeitslast wider: Sicherheitsrelevante Entwicklungen erfolgen auf kontrollierten Workstations, für groß angelegtes Training werden bei Bedarf dedizierte Cluster verwendet, und die Produktionsbereitstellung passt die Infrastruktur an die Skalierungsanforderungen an.

Unternehmen setzen Infrastruktur dort ein, wo sie die beste Kombination aus Leistung, Kosten und Kontrolle für bestimmte Aufgaben bietet. Ein Finanzdienstleistungsunternehmen könnte Prototypen von Betrugserkennungsmodellen auf Workstations erstellen, Produktionsversionen bei Bedarf auf GPU-Clustern trainieren, umfangreiche Inferenzprozesse in der Cloud bereitstellen und Modelle auf Zweigstellenebene auf lokalen Systemen ausführen, wobei jede Architektur dort eingesetzt wird, wo sie den optimalen Nutzen bietet.

Die Vielfalt der Plattformen ermöglicht strategische Flexibilität. Unternehmen können ihre Infrastruktur an die Anforderungen anpassen, anstatt alle Workloads auf eine einzige Architektur zu beschränken. Mit der Weiterentwicklung der KI-Fähigkeiten und der Veränderung der Workload-Muster lassen sich Infrastrukturportfolios leichter anpassen als Lösungen, die auf einer einzigen Plattform basieren.

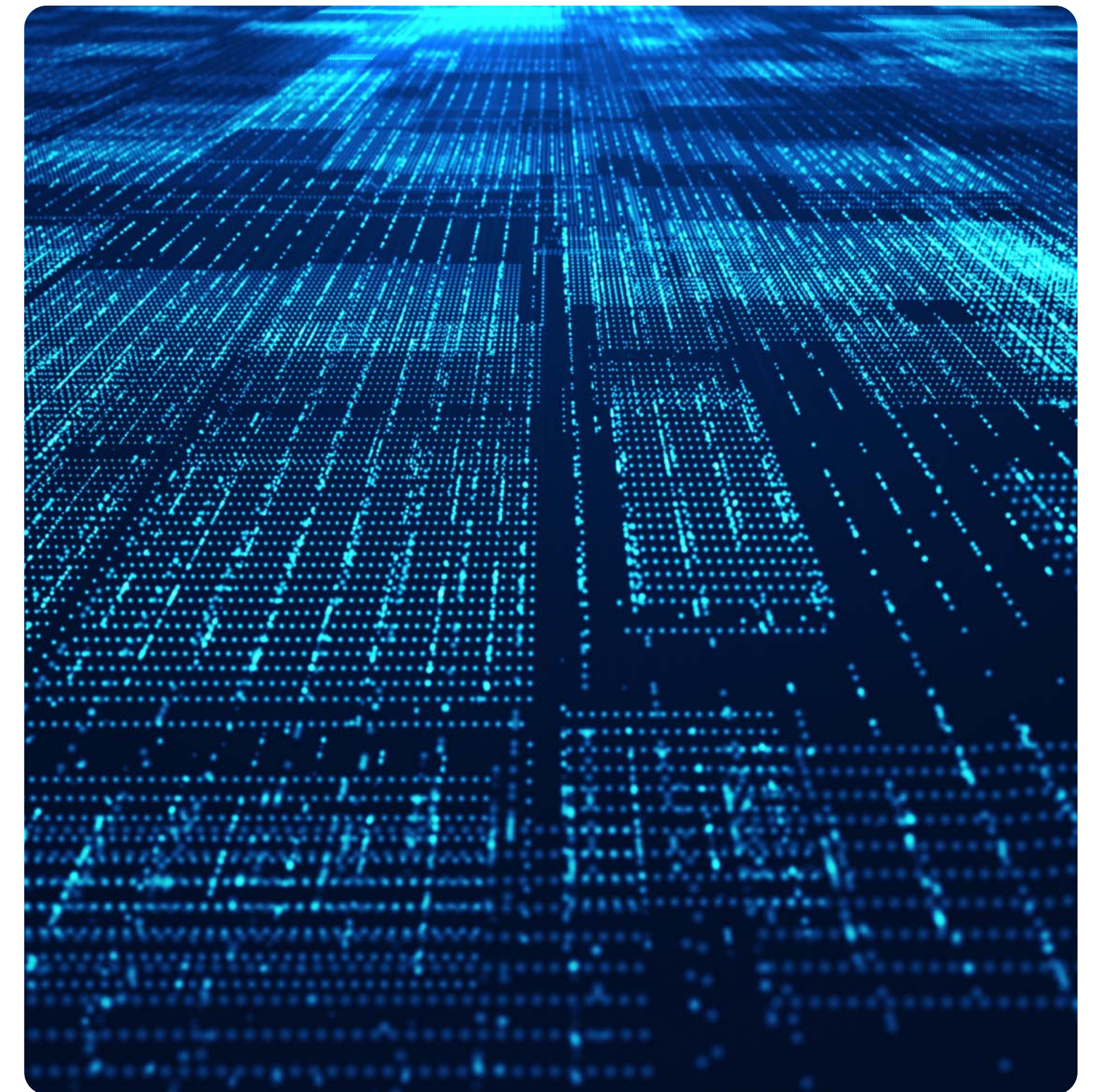
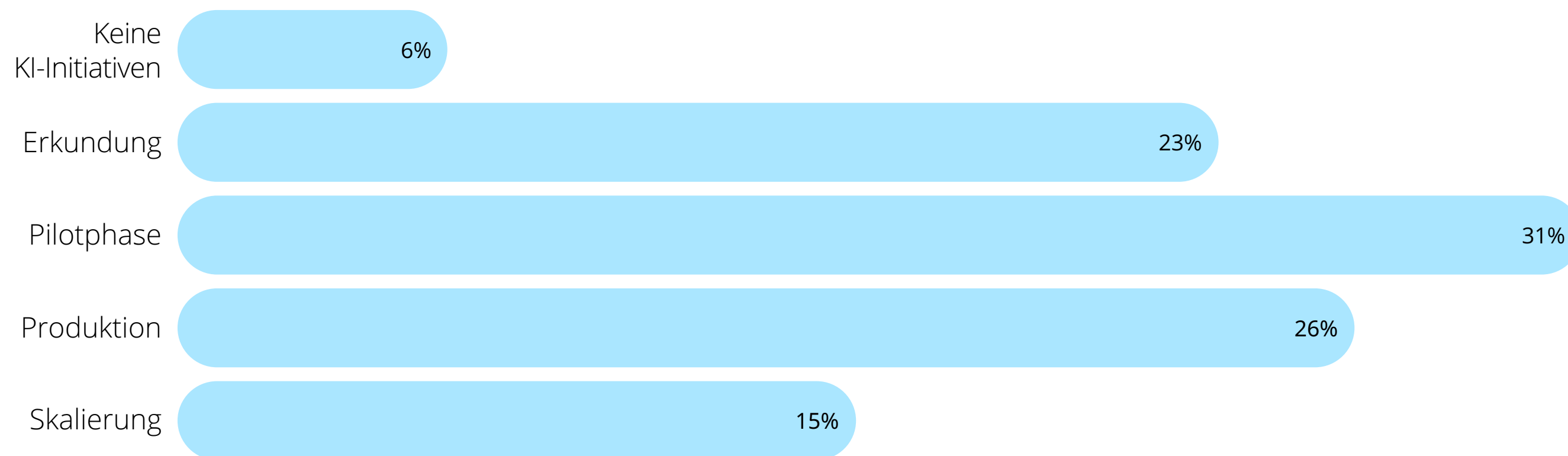
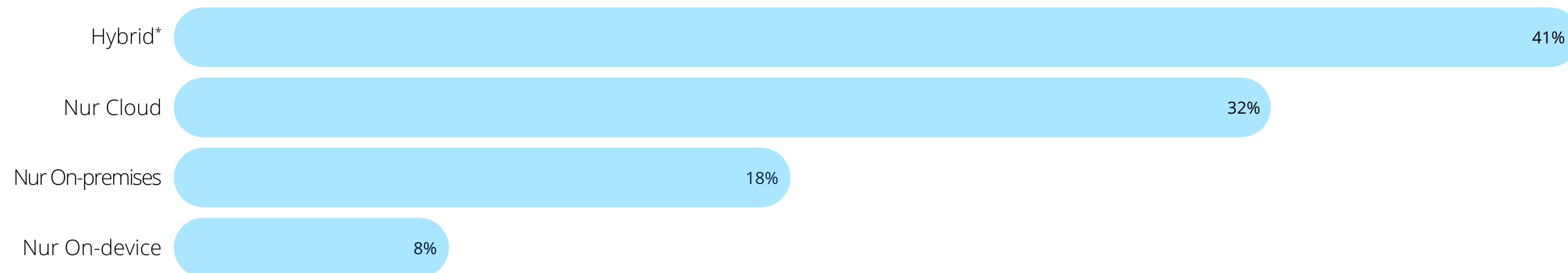


Abbildung 1: 41 % setzen KI aktiv in Geschäftsprozessen ein

Was ist der aktuelle Reifegrad Ihres Unternehmens im Bereich KI?

**Abbildung 2:** 73 % nutzen Cloud- oder Hybrid-Infrastrukturen für die Bereitstellung

Wo stellen Sie KI-Workloads bereit?



QUELLE: OMDIA

ANMERKUNGEN: N=1584, *HYBRIDSTRATEGIEN KOMMEN ZUM EINSATZ, WENN KI-IMPLEMENTIERUNGEN EINE KOMBINATION AUS CLOUD-, ON-PREMISES- UND ON-DEVICE-RECHENLEISTUNGEN UMFASSEN. ES HANDELT SICH VORWIEGEND UM EINE KOMBINATION AUS CLOUD- UND ON-PREMISES-SERVERN.

© 2025 OMDIA

Während die Hälfte der Unternehmen, die KI über mehrere Geschäftsfunktionen hinweg skaliert haben, auf hybride Bereitstellungsstrategien setzen, tut dies nur etwa ein Drittel der Unternehmen, die KI-Programme pilotieren.

Unternehmen verlagern einige KI-Workloads auf Geräte

Über ein Drittel der Unternehmen plant, in den nächsten 12 Monaten mehr KI-Workloads auf Geräte zu verlagern. Diese Unternehmen haben bestimmte Workloads identifiziert, bei denen die On-Device-Infrastruktur Lücken in ihren aktuellen Ansätzen schließt.

Bei sicherheitskritischen Workloads verursachen prozedurale Kontrollen einen kontinuierlichen Mehraufwand. Jedes Projekt, das geschützte Informationen, Kundendaten oder regulierte Inhalte umfasst, löst Dokumentationsanforderungen, Auditzyklen und Risikobewertungen für die Übertragung aus. Die Verarbeitung auf dem Gerät reduziert diesen Aufwand, indem die Daten lokal gespeichert werden. Was nicht übertragen wird, erfordert keine Überprüfung der Übertragungssicherheit oder Genehmigungen für den Zugriff durch Dritte.

Bei Entwicklungsabläufen bestimmt die Iterationsgeschwindigkeit, wie schnell Teams Modelle und Anwendungen optimieren können. Teams müssen Hypothesen testen, Parameter optimieren und Ansätze validieren, ohne auf die Zuweisung von Cloud-Ressourcen oder die Überwachung von Nutzungskontingenten warten zu müssen. Die On-Device-Infrastruktur beseitigt diese Einschränkungen: Entwicklungsteams arbeiten direkt mit lokalen Datensätzen, ohne Übertragungsverzögerungen, ohne Token-Beschränkungen und ohne Grenzkosten pro Iteration.

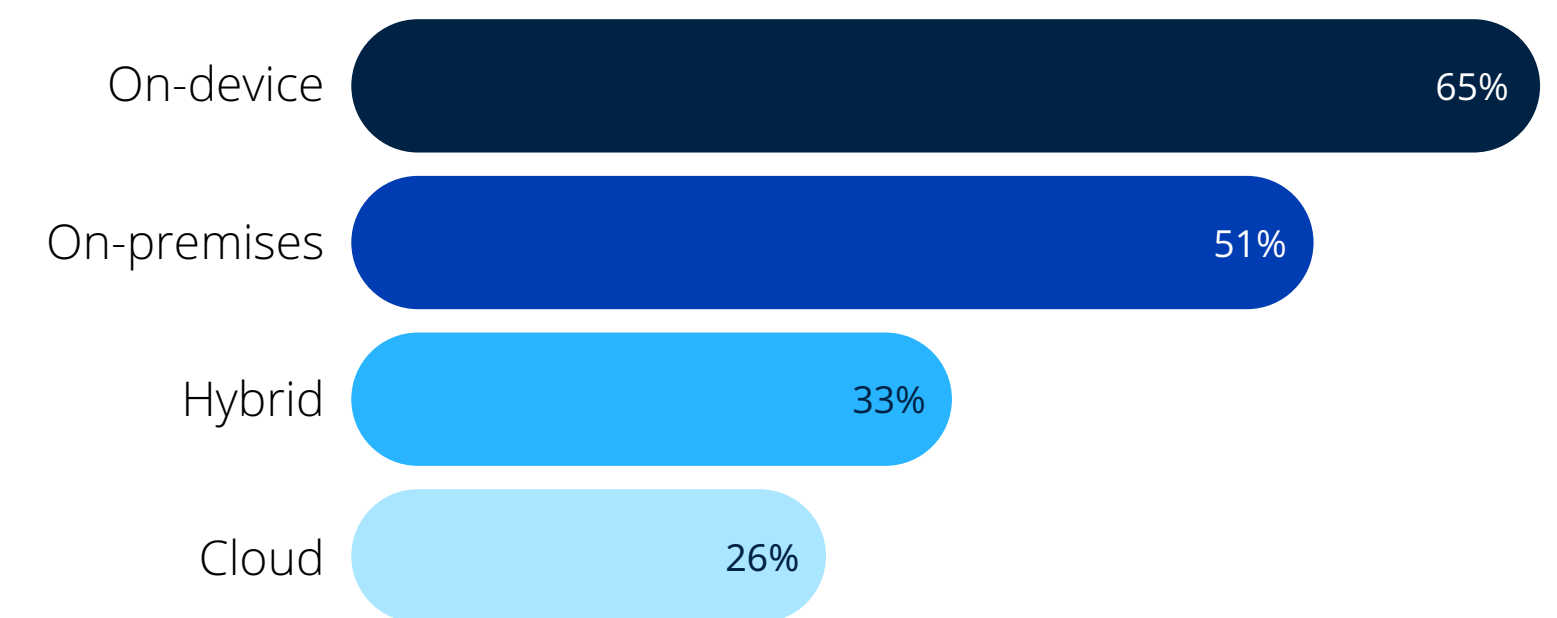
Unternehmen verzichten nicht auf Cloud-Infrastrukturen, sondern folgen demselben Impuls, der viele von ihnen ursprünglich dazu veranlasst hat, hybride Strategien einzuführen: Sie benötigen das richtige Werkzeug für die jeweilige Aufgabe. Jede Architektur erfüllt unterschiedliche Zwecke. Das Muster unter KI-erfahrenen Unternehmen ist eher eine strategische Diversifizierung als die Abhängigkeit von einer einzigen Plattform.



Unter den Unternehmen, die eine lokale Infrastruktur nutzen, planen 51%, KI-Workloads stärker auf den Geräten auszuführen. Von den Unternehmen, die bereits eine On-Device-Infrastruktur einsetzen, planen 65%, weitere KI-Workloads auf die Geräte zu verlagern.

Abbildung 3: Anteil der Unternehmen, die planen, mehr KI-Workloads auf Geräte zu verlagern, aufgeschlüsselt nach aktueller Bereitstellungsstrategie

Wie wird sich Ihrer Meinung nach die Verteilung zwischen On-Device-, On-Premises- und Cloud-KI in den nächsten 12 Monaten verändern?*



QUELLE: OMDIA
ANMERKUNGEN: N=1568, **ES WERDEN NUR „MEHR ON-DEVICE“-ANTWORTEN PRO AKTUELLER PRIMÄRER INFRASTRUKTUR GEZEIGT.
© 2025 OMDIA

On-Device ergänzt KI-Anbieter, indem es deren Lücken schließt

Nur 9 % berichten über keine signifikanten Lücken bei ihrem strategischen KI-Partner

Strategische KI-Partnerschaften bieten Funktionen, die viele Unternehmen allein nicht aufbauen könnten: vortrainierte Modelle, Bereitstellungsinfrastruktur und spezialisierte Tools. Doch Partnerschaften bringen auch Einschränkungen mit sich. Sie erfordern Datenzugriff, führen zu komplexen Preisstrukturen und erzeugen Integrationsabhängigkeiten. Unternehmen entdecken diese Reibungspunkte in unterschiedlichen Phasen: Sicherheitsbewusste Unternehmen stoßen darauf, wenn sie sensible Daten bereitstellen, während kostenbewusste Unternehmen sie entdecken, wenn die Rechnungen eintreffen.

Die Studie von Omdia zeigt eine weit verbreitete Unzufriedenheit. Unzureichende Datenschutzkontrollen, unvorhersehbare Kosten und eine mangelhafte Integration in bestehende Systeme stellen sich als durchgängige Herausforderungen für Unternehmen dar, die mit großen Cloud-Anbietern, traditionellen Anbietern von Unternehmenssoftware oder KI-nativen Unternehmen zusammenarbeiten. Nur 9 % der Unternehmen geben an, dass ihre Partnerschaften ihren Anforderungen vollständig entsprechen.

Diese Unzufriedenheit schafft Raum für einen hybriden Ansatz. Unternehmen lehnen Partnerschaften im Bereich KI nicht ab, sondern suchen nach Flexibilität, um spezifische operative

Herausforderungen zu bewältigen, die mit reinen Cloud-Strategien nicht vollständig gelöst werden können. Durch die lokale Entwicklung auf einer kontrollierten Infrastruktur erhalten Unternehmen eine Hochgeschwindigkeitsumgebung für sensible Daten und Echtzeitaufgaben, die sowohl vorhersehbare Kosten als auch direkten Datenzugriff bietet.

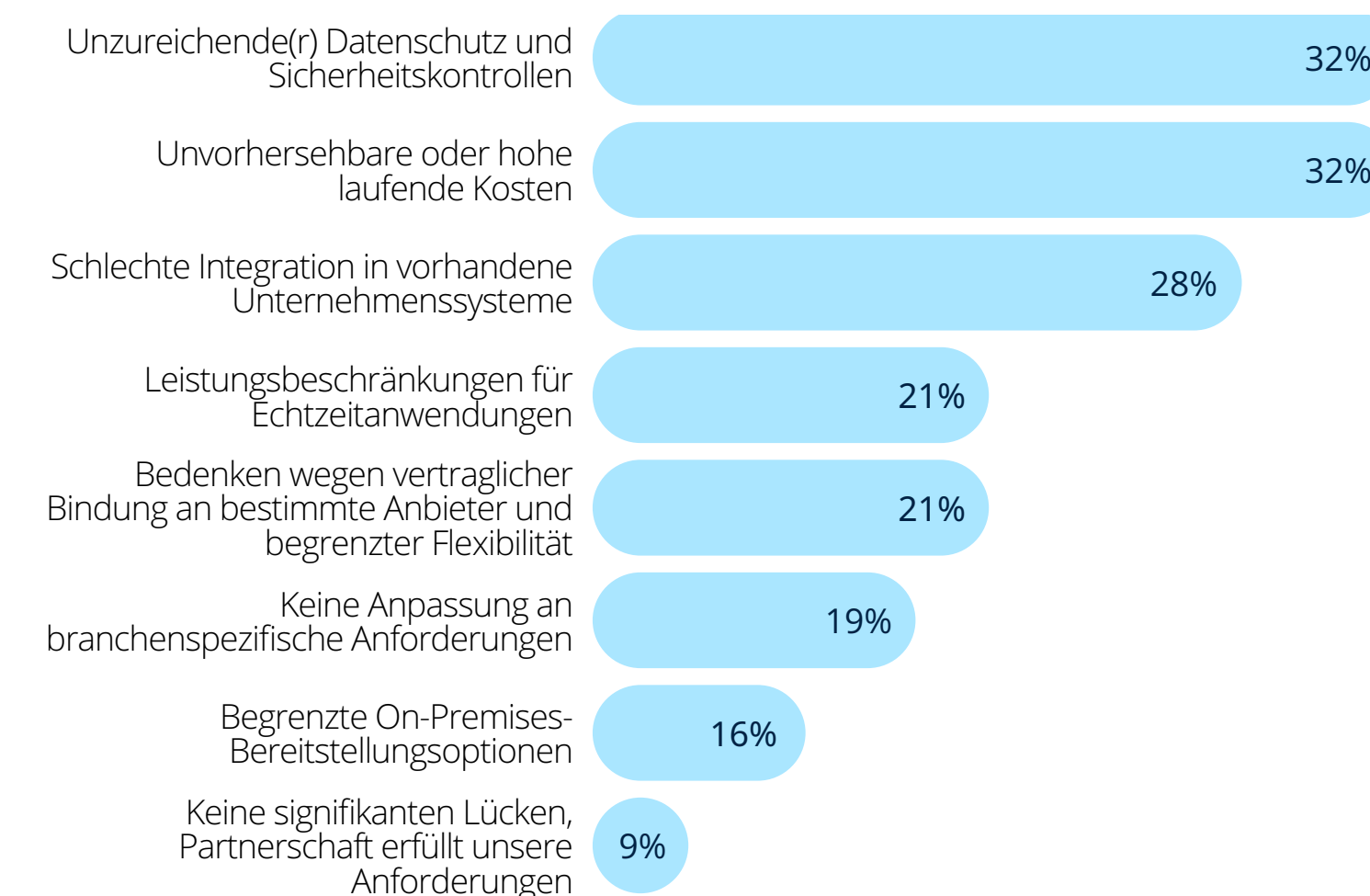
Die Ergänzung der Infrastrukturkapazität durch On-Device-Computing schließt die funktionalen Lücken, die bei reinen Cloud-Partnerschaften bestehen. Es optimiert die Abläufe, indem es KI in jeder Phase eines Workflows ermöglicht – von der Annotation bis zur Bereinigung sensibler Daten –, wo zuvor Datenschutzkontrollen oder Ausgabekosten ein Hindernis darstellten. Die Nutzung der Infrastruktur auf dem Gerät als Grundlage gewährleistet, dass strategische Cloud-Ressourcen für die groß angelegte Verteilung und zentralisierte Koordination priorisiert werden, anstatt für Aufgaben verwendet zu werden, die besser für das Gerät geeignet sind.



Die On-Device-Infrastruktur bietet Ihnen eine sichere Basis, von der aus Sie Ihr gesamtes KI-Ökosystem verwalten können.

Abbildung 4: Sicherheit, Kosten und unzureichende Integration sind anhaltende Herausforderungen in Partnerschaften

Was sind die zwei größten Einschränkungen der aktuellen KI-Partnerschaft? Wählen Sie bitte bis zu zwei Antworten aus.



QUELLE: OMDIA
ANMERKUNGEN: N=1352
© 2025 OMDIA

Unternehmen verfolgen sichtbare Kosten und übersehen dabei die wesentlichen Kostentreiber

Über zwei Drittel erfassen Cloud-Kosten, jedoch weniger als die Hälfte erfassen sicherheitsbezogene Kosten

Investitionen in Cloud-Computing, Sicherheit und Softwarelizenzen stellen die wichtigsten Bedenken für die Gesamtbetriebskosten (TCO) der KI-Infrastruktur dar. Allerdings bedeuten diese Bedenken nicht automatisch eine genaue Nachverfolgung und Zuordnung.

Von denjenigen, die Cloud Computing als TCO-Problem einstufen, geben 72 % an, dass sie dies systematisch verfolgen und als separaten Posten in ihrem KI-Infrastrukturbudget erfassen. Dies ist sinnvoll, da Cloud-Anbieter regelmäßig Rechnungen mit einer übersichtlichen Aufschlüsselung der einzelnen Dienste ausstellen.

Bei der Sicherheit sieht es anders aus. Obwohl dies das zweitwichtigste TCO-Anliegen ist, erfasst weniger als die Hälfte die Sicherheitskosten als separaten Bestandteil der TCO für die KI-Infrastruktur. Sicherheitskosten sind real – Compliance-Aufwand, Auditzyklen, Reaktion auf Vorfälle, spezialisierte Tools –, jedoch werden sie auf bestehende Budgets und organisatorische Silos verteilt, was die Zuordnung erheblich erschwert.

Diese Sichtbarkeitslücke führt zu strategischen Schwachstellen.

Unternehmen optimieren das, was sie messen können. Wenn Cloud-Kosten in einer detaillierten Rechnung aufgeführt werden, während Sicherheitsinvestitionen weiterhin auf Beschaffungs-, Personal- und Betriebsbudgets verteilt sind, konzentrieren sich Entscheidungsträger verständlicherweise auf die sichtbaren Zahlen. Die unsichtbaren Kosten könnten jedoch größer und für den langfristigen Erfolg entscheidender sein.

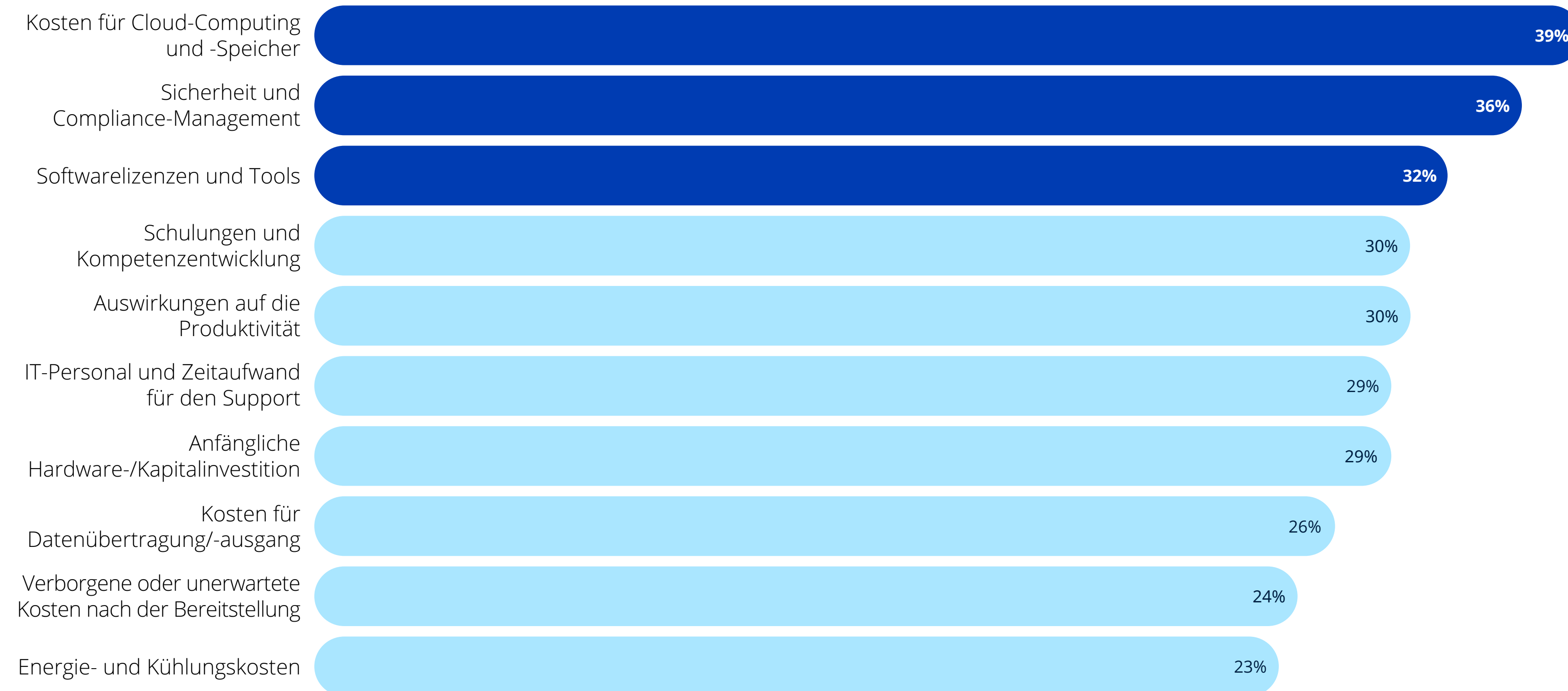
Die Lücke verdeckt auch mögliche Kostenvorteile für Bereitstellungsstrategien, die über die reine Cloud hinausgehen. Unternehmen neigen dazu, hybride Ansätze systematisch zu unterschätzen, da die potenziellen Kosteneinsparungen nicht unmittelbar sichtbar sind, während die Vorabinvestitionen die Kostenbewertung dominieren und den langfristigen Wert überschatten können. Diese Asymmetrie in der Sichtbarkeit kann die Entscheidungsfindung zugunsten cloudzentrierter Strategien beeinflussen, selbst wenn die Gesamtbetriebskosten einen ausgewogeneren Ansatz begünstigen.



Attributionslücken
und Budgetstrukturen
können die TCO-
Vorteile hybrider
Bereitstellungsstrategien
verschleiern.

Abbildung 5: Cloud, Sicherheit und Softwarelizenzen sind die wichtigsten Kostenbedenken für die Gesamtbetriebskosten der KI-Infrastruktur

Was sind Ihre wichtigsten Bedenken hinsichtlich der Gesamtbetriebskosten der KI-Infrastruktur?
Wählen Sie drei Antworten aus.



QUELLE: OMDIA
ANMERKUNGEN: N=1584
© 2025 OMDIA

Das Sicherheitsmodell stimmt

Sicherheit als Reibungspunkt vs. Sicherheit als Architektur

Minimierung der Datenübertragung durch Architektur

Aktuelle cloudzentrierte Ansätze führen zu einer inhärenten Spannung: Um die KI-Fähigkeiten der Cloud nutzen zu können, müssen Daten das Unternehmen verlassen und an Infrastrukturen von Drittanbietern übertragen werden. Diese Übertragung führt zu einer Gefährdung, die von 76 % der Unternehmen als problematisch angesehen wird. Die Besorgnis betrifft nicht unbedingt absichtliche Angriffe, sondern vielmehr unbeabsichtigte Übertragungs- oder Konfigurationsfehler, die weitaus häufiger auftreten.

Diese Besorgnis verstärkt sich, wenn Unternehmen Modelle unter Verwendung ihrer wertvollsten Informationen optimieren oder trainieren: Kundendaten, interne Wissensdatenbanken und firmeneigene Forschungsergebnisse. Obwohl 87 % der Unternehmen erkennen, dass die Speicherung sensibler Daten vor Ort oder auf Geräten kritisch oder wichtig ist, können Cloud-First-Architekturen nur eine Risikominderung, jedoch keine Risikobeseitigung bieten.

Die grundlegende Herausforderung: Innovation erfordert Datenzugriff, während Datenschutz die Einschränkung der Datenübertragung erfordert. Entwicklungsteams benötigen sensible, geschützte Daten, um effektive Modelle zu erstellen.

Jedes neue Projekt löst Compliance-Prüfungszyklen aus, was zu Planungsaufwand führt, der die Experimentiergeschwindigkeit bei neuen Anwendungsfällen einschränkt.

In regulierten Branchen ist diese Spannung besonders ausgeprägt. Unternehmen im Gesundheitswesen, die Modelle auf der Grundlage von Patientengesundheitsdaten trainieren, unterliegen vorgeschriebenen Kontrollen. Finanzdienstleistungen, die Transaktionshistorien und Finanzdaten von Kunden verwenden, müssen die Beschränkungen hinsichtlich der Datenresidenz beachten. Und Unternehmen, die Fachdienstleistungen erbringen und mit vertraulichen Kundendaten arbeiten, unterliegen vertraglichen Verboten hinsichtlich des Zugriffs auf Daten durch Dritte.



Sicherheitsanforderungen sind einer der Hauptgründe, warum Unternehmen angeben, dass sie eine On-Device-Infrastruktur verwenden. Unternehmen bevorzugen zunehmend architektonische Lösungen gegenüber prozeduralen Sicherheitskontrollen.

⚠ Die Sorge

76%

der Unternehmen sind besorgt oder sehr besorgt über mögliche Datenlecks durch Cloud-Dienste

✅ Die Präferenz

87%

der Unternehmen geben an, dass es wichtig ist, sensible Daten vor Ort oder auf Geräten zu speichern

KI-Entwicklung auf dem Gerät stärkt die Sicherheit

Starke Verringerung der Risiken durch Datenschutz durch Architektur statt durch prozedurale Workarounds

On-Device-Entwicklung löst das Spannungsfeld zwischen Innovation und Sicherheit auf architektonischer Ebene. Die Daten verlassen niemals die sichere Umgebung, wodurch ganze Kategorien von Sicherheitsüberprüfungen und Compliance-Aufwand entfallen. Dies wird durch Anforderungen an die Datenhoheit noch verstärkt – wenn Unternehmen in verschiedenen Rechtsräumen tätig sind, sehen sie sich mit länderspezifischen Vorschriften konfrontiert, die vorschreiben, dass Daten innerhalb der geografischen Grenzen bleiben müssen. Cloud-Architekturen können die Einhaltung von Vorschriften erschweren; die Verarbeitung auf dem Gerät erleichtert die Einhaltung von Vorschriften.

Der Sicherheitsvorteil des On-Device-Ansatzes ist klar: Daten, die niemals übertragen werden, können auch nicht während der Übertragung exponiert werden. Entwicklungsteams können direkt mit lokalen Datensätzen arbeiten, ohne auf die Infrastruktur von Drittanbietern zugreifen zu müssen. Modelle werden dort trainiert, wo sich die Daten befinden, sodass keine Kosten für die Datenübertragung anfallen,

keine Verschlüsselung der Übertragung überprüft werden muss, keine Kontrollen des Cloud-Anbieters geprüft werden müssen und keine Dokumentation für grenzüberschreitende Datenübertragungen erforderlich ist.

On-Device-Infrastruktur schafft bisher unmögliche Workflow-Vorteile: Nutzung von KI zur Vorbereitung von Daten für KI. Unternehmen, die Cloud-Umgebungen einsetzen, müssen sensible Daten vor der Übertragung bereinigen und vorbereiten, indem sie personenbezogene Daten entfernen oder vertrauliche Angaben schwärzen. Ohne On-Device-KI-Funktionen erfolgt diese Datenaufbereitung mit herkömmlichen Tools, was zu erheblichen Engpässen führt.

Mit On-Device-KI können Entwicklungsteams KI-Tools lokal einsetzen, um den Bereinigungs- und Vorbereitungsprozess zu beschleunigen – mithilfe von Modellen zur Identifizierung sensibler Felder, zur Automatisierung der Schwärzung, zur Validierung der Datenqualität und zur Transformation von Datensätzen –, und zwar bevor Daten die sichere Umgebung verlassen.

Unternehmen können Innovationen basierend auf sensiblen Daten ohne Kompromisse mit voller Geschwindigkeit vorantreiben. Entwicklungszyklen, die zuvor mehrere Tage für Sicherheitsüberprüfungen erforderten, werden erheblich verkürzt. Entwicklungsteams gewinnen an Geschwindigkeit zurück, und Sicherheitsteams gewinnen Vertrauen durch architektonische Garantien anstelle von prozeduralen Kontrollen.

Dies bedeutet nicht, dass alle Sicherheitsanforderungen entfallen. Die physische Gerätesicherheit, Zugriffskontrollen und Audit-Protokollierung bleiben weiterhin von entscheidender Bedeutung. Es schützt jedoch vor den mit der Übertragung verbundenen Sicherheitsbedenken, denen Cloud-Architekturen aufgrund ihrer Konzeption ausgesetzt sind. Eine On-Device-Infrastruktur beseitigt das architektonische Risiko und auch die prozeduralen Reibungspunkte.



99%

der Unternehmen verwenden proprietäre Daten, wenn sie KI-Modelle trainieren oder feinabstimmen. Daher stellt die On-Device-KI einen entscheidenden Workflow-Vorteil dar, da sie Übertragungsrisiken eliminiert, den Verfahrensaufwand reduziert und KI-gestützte Workflows zur Datenaufbereitung ermöglicht.

Die Wirtschaftlichkeit stimmt

On-Device-Entwicklung eliminiert die Experimentierkosten

Die KI-Entwicklung profitiert von Iteration. Die besten Lösungen entstehen, wenn Teams frei experimentieren können. Dies ermöglicht es ihnen, Hypothesen zu überprüfen, Modelle zu optimieren und kreative Ansätze ohne Einschränkungen zu erforschen. Für die 54 % der Unternehmen, die sich noch in der Phase der Untersuchung oder Erprobung von KI-Lösungen befinden, führt die frühzeitige Schaffung einer ausgewogenen Infrastruktur direkt zu Kosteneinsparungen, da die Experimentierphase intensiviert wird und die Workloads auf die Produktion skaliert werden.

Eine On-Device-Infrastruktur verändert die Wirtschaftlichkeit von Experimenten. Eine feste Investition in Hardware ermöglicht unbegrenzte Iterationen. Sobald die Hardware bereitgestellt ist, entstehen für jede Iteration, jeden Testlauf und jeden Validierungszyklus keine zusätzlichen Kosten.

Im Gegensatz dazu verursachen cloudzentrierte Strategien Kosten für jede einzelne Erkundung. Bei jedem Versuchslauf fallen Kosten für Rechenzeit, Datenübertragung und Speicherung an. Wenn ein vielversprechender Ansatz nach 20 Iterationen scheitert, werden diese 20 Iterationen dennoch auf der Rechnung ausgewiesen. Es kostet genauso viel, aus erfolglosen wie aus erfolgreichen Ansätzen zu lernen.

Bei länger laufenden Projekten sehen sich Unternehmen mit erheblich unterschiedlichen Gesamtkosten zwischen Cloud und Gerät konfrontiert. Die Cloud-Kosten steigen mit jeder Version, während Hardware, die KI vollständig auf dem Gerät ausführen kann, unbegrenzte Experimentiermöglichkeiten ohne zusätzliche Kosten bietet. Die gleiche Hardware-Investition kann während ihrer gesamten Lebensdauer für mehrere Projekte genutzt werden.

Der strategische Vorteil verstärkt sich mit der Zeit. Entwicklungsteams können gründlich validieren, Sackgassen erkunden und schnell iterieren, ohne sich um das Budget sorgen zu müssen. Die Einschränkung liegt nun eher in der menschlichen Aufmerksamkeit und der Entwicklungszeit als in der Wirtschaftlichkeit der Infrastruktur. Dadurch werden finanzielle Hindernisse beseitigt, die Teams dazu zwingen, Experimente einzuschränken, Validierungsläufe zu überspringen oder alternative Ansätze nicht zu prüfen.

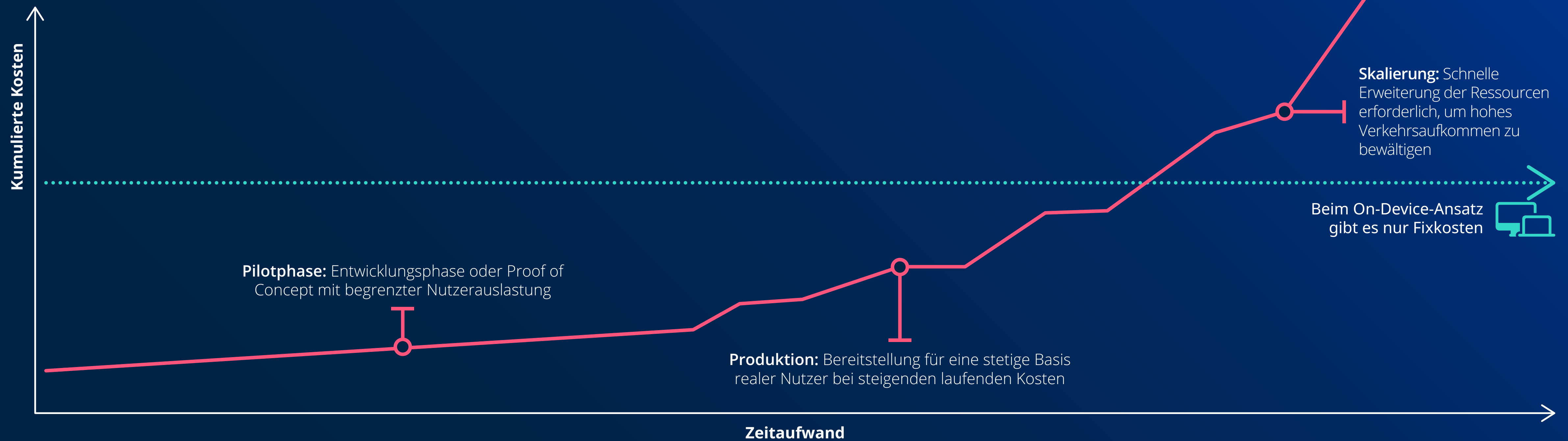
Eine On-Device-Infrastruktur senkt nicht nur die Kosten, sondern ermöglicht auch schnelle Iterationszyklen, die zu einer besseren KI führen. Teams liefern bessere Produkte und Lösungen, da sie es sich leisten können, mehr zu testen und aus mehr Fehlern zu lernen.



Nach der anfänglichen Investition bedeutet On-Device-KI, dass jedes Experiment und jede Iteration im Wesentlichen keine Grenzkosten verursacht.

Die Bereitstellung auf dem Gerät macht unvorhersehbare Kosten vorhersehbar

Abbildung 6: Cloud-Kosten steigen mit der Zeit und skalieren mit der Nutzung, während die Kosten für On-Device-Lösungen konstant bleiben



Die Bereitstellung von KI auf Endgeräten schützt Unternehmen vor ausufernden Inferenzkosten und ungenutzten Cloud-Lizenzen

Über die Entwicklung hinaus hat die Wahl der Infrastruktur einen erheblichen Einfluss auf die Kosten der Produktionsinferenz. Für cloudzentrierte KI gibt es zwei primäre Preismodelle: Gebühren pro API-Aufruf, die sich mit jeder Anfrage summieren, oder arbeitsplatzbasierte Abonnements, die unabhängig von der Nutzungsintensität fest bleiben. Beide stellen wirtschaftliche Herausforderungen dar, die im Laufe der Zeit zu erhöhten Kosten führen.

Eine On-Device-Infrastruktur verändert die Wirtschaftlichkeit der KI-Inferenz grundlegend. Nach der anfänglichen Investition in die Hardware fallen für jede Inferenzanforderung keine zusätzlichen Kosten an. Die Generierung von 1.000 oder 100.000 Token auf dem Gerät führt zu denselben Gesamtbetriebskosten. Diese Vorhersagbarkeit erweist sich als besonders wertvoll für Unternehmen, die KI in ihrer gesamten Belegschaft einsetzen, wo es aufgrund von Abteilungsgrenzen oft schwierig ist, die Cloud-Kosten unternehmensweit genau und in Echtzeit zu verfolgen.

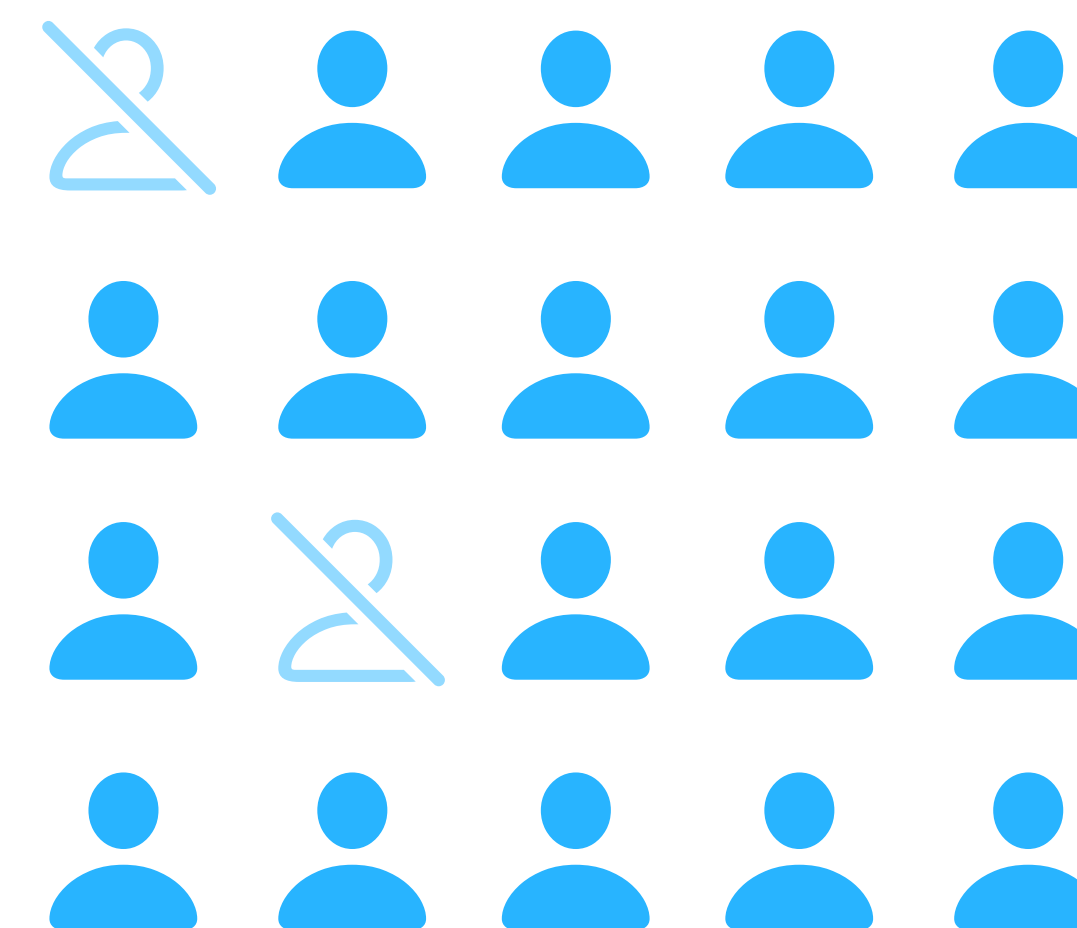
Das arbeitsplatzbasierte Lizenzmodell stellt eine andere Herausforderung dar. Unternehmen, die Cloud-Abonnements abschließen und eine breite Akzeptanz erwarten, stellen häufig fest, dass die tatsächliche Nutzung hinter den Erwartungen zurückbleibt. 32 % der Unternehmen gaben an, dass sie mit der Akzeptanz von KI-Tools durch die Endnutzer:innen nicht zufrieden sind, weitere 30 % waren nur eingeschränkt zufrieden. Diese ungenutzten Lizenzen verursachen laufende Kosten ohne Produktivitätsgewinn.

Eine On-Device-Infrastruktur vermeidet dieses Risiko vollständig. Unabhängig davon, ob die Hardware für KI-Workloads oder für Standard-Computing-Aufgaben eingesetzt wird, bietet sie während ihres gesamten Lebenszyklus einen hohen Mehrwert. Unternehmen können die mit ungenutzten Abonnements verbundenen Kosten vermeiden und gleichzeitig flexibel auf sich ändernde Anforderungen reagieren.

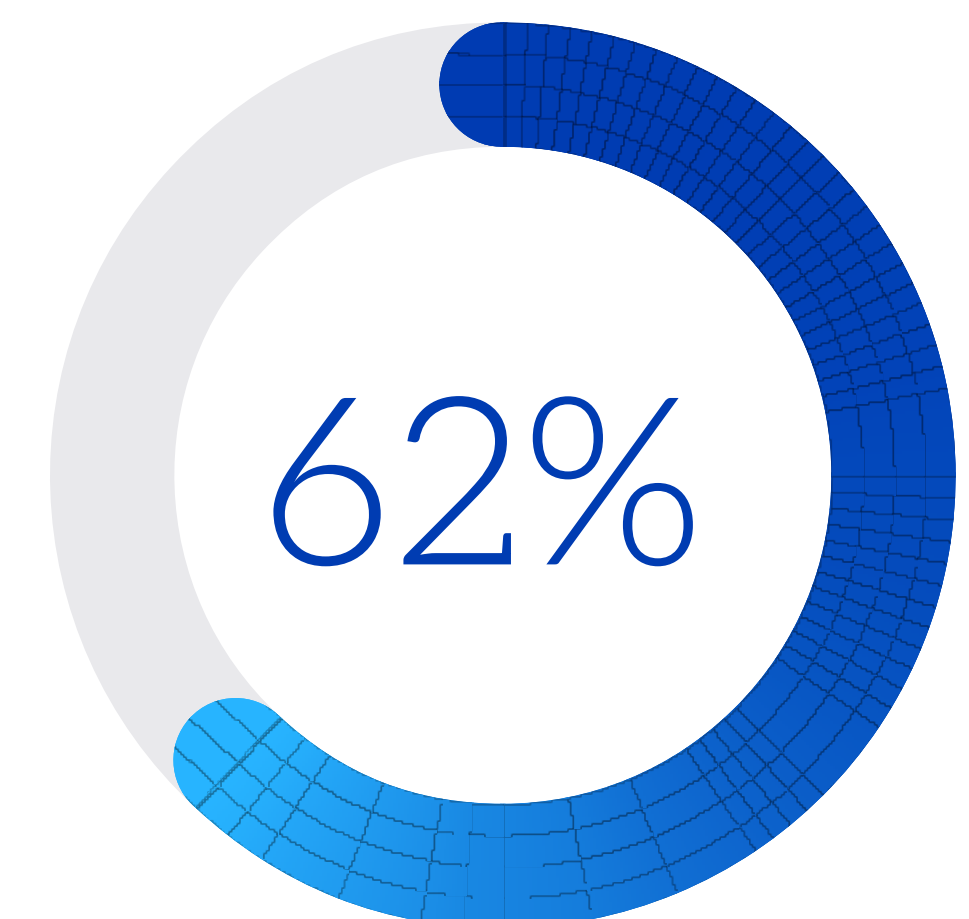


Der Einsatz von KI auf Geräten schützt Unternehmen vor ausufernden Inferenzkosten und ungenutzten Cloud-Lizenzen.

Abbildung 7: Pro Arbeitsplatz berechnete Cloud-Lizenzen kosten unabhängig von der Nutzung gleich viel



Angenommen, ein Unternehmen kauft 1.000 Lizenzen für cloudzentrierte KI zu einem Preis von 30 US-Dollar pro Benutzer und Monat, und nur 10 % davon werden nicht genutzt, **dann werden monatlich 3.000 US-Dollar verschwendet.**



der Unternehmen geben an, dass sie mit der Nutzung von KI-Tools durch Endnutzer:innen nur eingeschränkt zufrieden oder unzufrieden sind.

Die Workloads stimmen

Unternehmens-KI-Modelle sind zweckoptimiert

Die Modelle, die mit Billionen von Parametern im Vortraining Schlagzeilen machen, haben den Eindruck erweckt, dass alle bedeutenden KI-Arbeiten Hyperscale-Rechenzentren erfordern. Dies spiegelt nicht die Realität der Bereitstellung in Unternehmen wider.

95% der Unternehmen erstellen Modelle nicht von Grund auf neu, sondern optimieren bestehende Modelle und führen Inferenzen durch. Diese Aufgaben sind deutlich weniger rechenintensiv.

Darüber hinaus verwenden 57% der Modelle, die Unternehmen in der Regel einsetzen, weniger als 10 Milliarden Parameter. Selbst für die umfangreichsten Modelle mit mehreren hundert Milliarden Parametern stehen Optionen zur Verfügung, die eine lokale Entwicklung und Bereitstellung auf Desktop-Geräten in Clustern ermöglichen. Mit anderen Worten: Die Infrastrukturanforderungen für die Entwicklung von KI in Unternehmen entsprechen häufig den Fähigkeiten moderner PCs.



Unternehmen, die bereits On-Device-Lösungen einsetzen, verwenden doppelt so häufig Modelle mit über 100 Milliarden Parametern

Abbildung 8: 95% der Unternehmen optimieren und führen Inferenz durch

Was ist die typische Größe der KI-Modelle, die Ihr Unternehmen einsetzt? Führen Sie Inferenz, Optimierung oder Training Ihrer eigenen Modelle durch?



QUELLE: OMDIA
ANMERKUNGEN: N=1529
© 2025 OMDIA

Die Mehrzahl der Unternehmensmodelle hat weniger als 10 Milliarden Parameter – das liegt problemlos im Leistungsbereich moderner Geräte wie dem MacBook Air oder das MacBook Pro der Einstiegsklasse



Ein Modell mit 10 Milliarden Parametern benötigt bei Quantisierung etwa 12 GB Speicherplatz. Dies wird von Geräten mit 24 GB Speicher, die ausreichend Spielraum für gleichzeitige Anwendungen und Arbeitsabläufe bieten, problemlos bewältigt. Selbst Basiskonfigurationen mit 16 GB können kleinere Modelle mit einer Milliarde Parametern unterstützen.

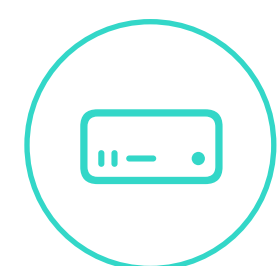
Unternehmen, die vorrangig On-Device-Lösungen einsetzen, implementieren doppelt so häufig Modelle mit über 100 Milliarden Parametern

Eine weit verbreitete Annahme bezüglich der KI-Infrastruktur ist, dass größere Modelle automatisch eine Cloud-Bereitstellung erfordern. Die Analyse von Omdia ergab keinen signifikanten Zusammenhang zwischen der Größe der von einem Unternehmen eingesetzten Modelle und dessen Präferenzen hinsichtlich der Infrastrukturbereitstellung.

Unternehmen, die Modelle mit mehr als 100 Milliarden Parametern betreiben, bevorzugten Cloud-Infrastrukturen nicht mehr als solche, die Modelle mit 10 Milliarden Parametern betreiben. Die Faktoren, die die Wahl der Bereitstellungsumgebung tatsächlich beeinflussen, sind Sicherheitsanforderungen, Kostenüberlegungen und Anforderungen an den Entwicklungsworkflow.

Dieses Muster spiegelt eine technische Realität wider: Die Einschränkung liegt nicht in der Modellgröße an sich, sondern vielmehr in der verfügbaren Speicherkapazität und Architektur. Der entscheidende Faktor ist, ob die Infrastruktur ausreichend Speicher für die Arbeitslast bereitstellen kann, und nicht, ob das Modell einen willkürlichen Größenschwellenwert überschreitet.

Für die 20 % der Unternehmensimplementierungen, die Modelle mit mehr als 100 Milliarden Parametern verwenden, wird die Verfügbarkeit von Speicherplatz zu einem entscheidenden Faktor. Eine Infrastruktur, die Speicheranforderungen von über 100 GB bewältigen kann, sei es durch Einzelkonfigurationen mit großen einheitlichen Speicherpools oder durch geclusterte Multi-Device-Bereitstellungen, ermöglicht die lokale Verarbeitung dieser größeren Modelle.



Die Speicherarchitektur bestimmt die Mächtigkeit: Einheitliche Speichersysteme ermöglichen Konfigurationen mit mehr als 100 GB über ein einzelnes Gerät anstelle einer Multi-GPU-Infrastruktur.

Warum Speicherarchitektur wichtig ist

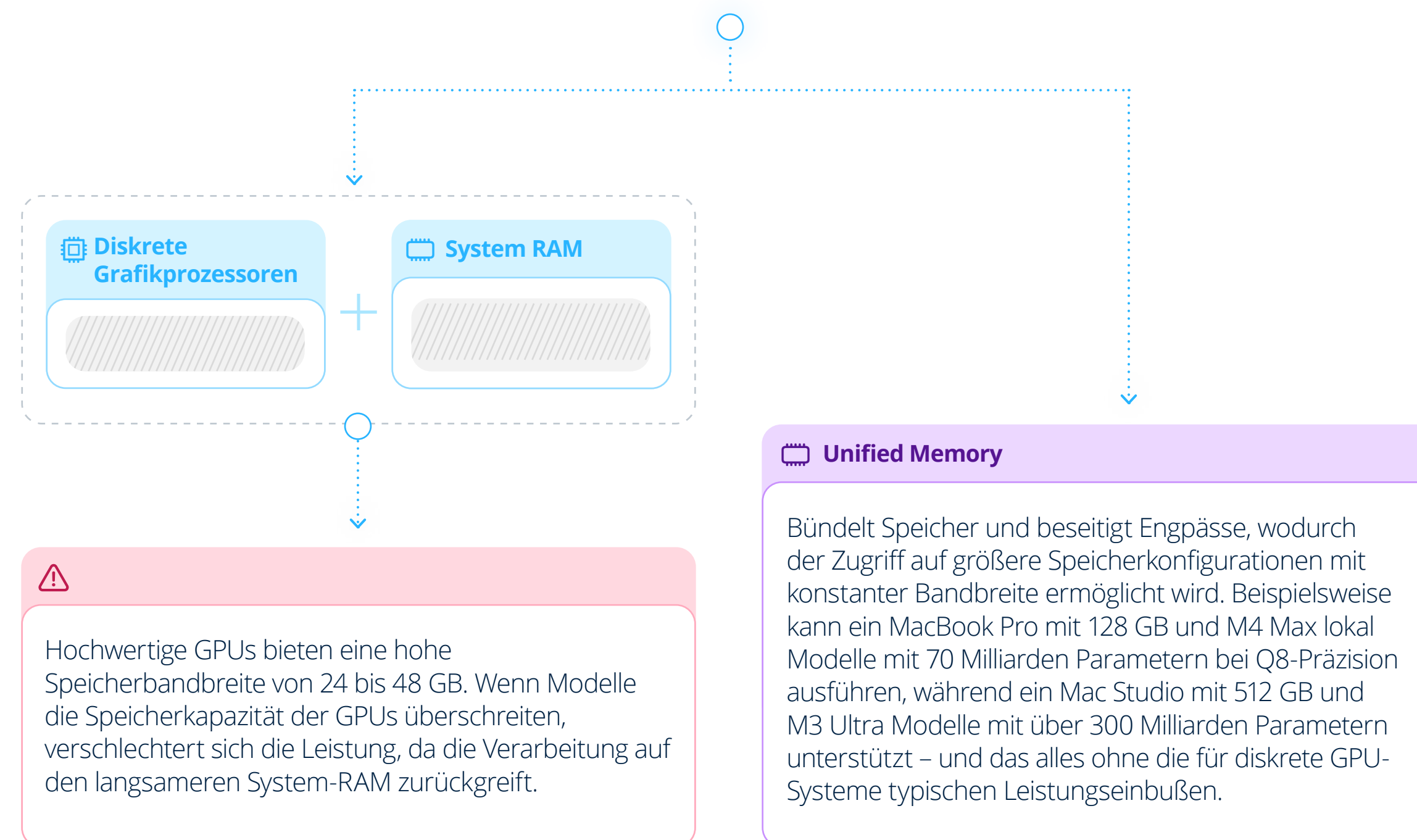
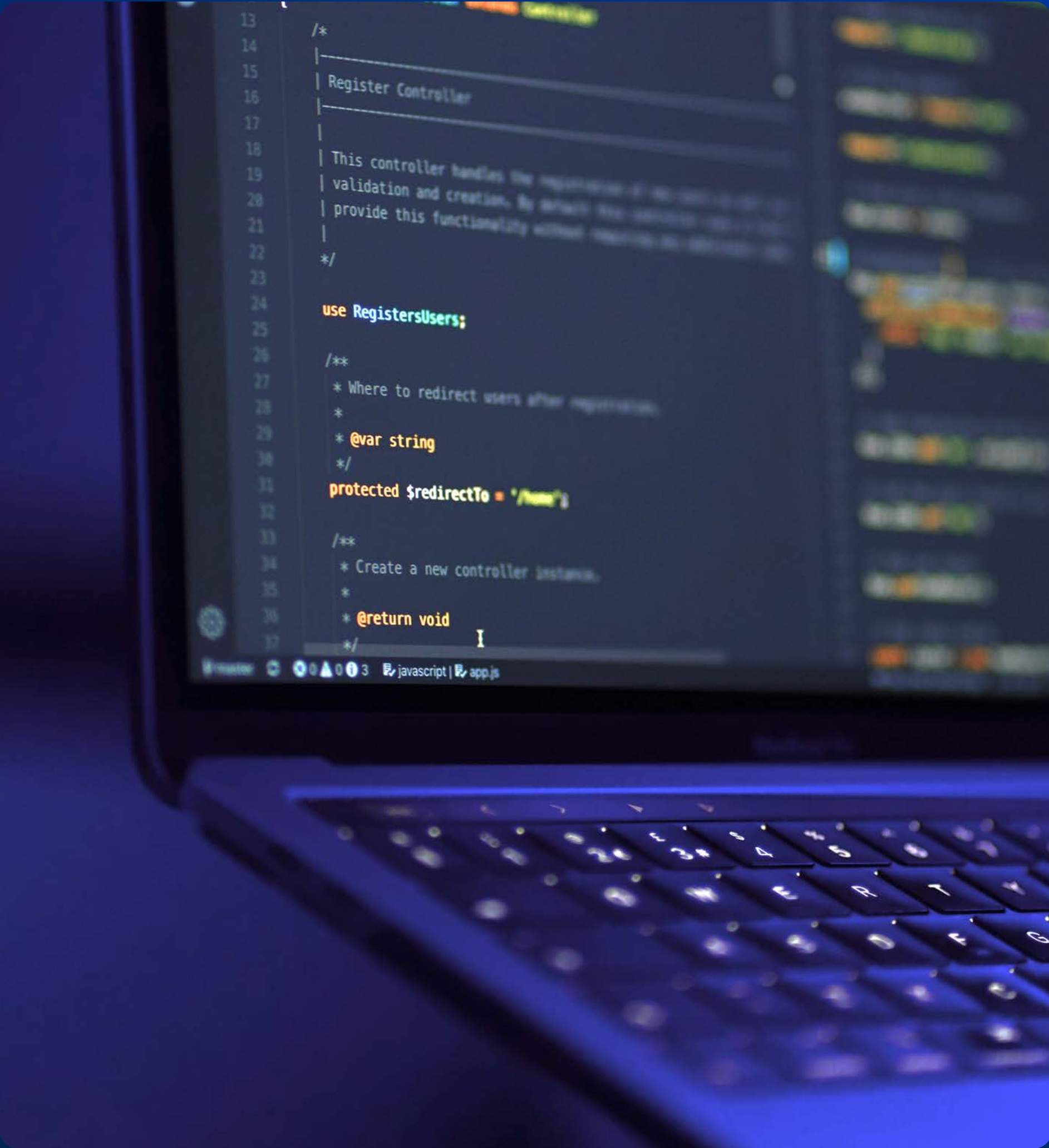


Abbildung 9: Modellspeicheranforderungen für die Modellinferenz, nach Präzision
Beinhaltet Modellgewichte + 20% zusätzliche Inferenzkosten

ANZAHL DER PARAMETER UND INDIKATIVE FÄHIGKEITEN	Q8	FP16	FP32
1 milliarde: Textklassifizierung, Sentiment-Analyse, einfache Extraktion	1.2GB	2.4GB	4.8GB
10 milliarde: Konversations-KI, Fragen & Antworten, Textzusammenfassung, grundlegende Code-Unterstützung	12GB	24GB	48GB
15 milliarde: Dokumentenanalyse, Übersetzung, strukturierte Inhaltserstellung	18GB	36GB	72GB
70 milliarde: Erweiterte Code-Generierung, mehrstufige Aufgaben	84GB	168GB	336GB
100 milliarde: Multimodale Analyse, Forschungssynthese	120GB	240GB	480GB
300 milliarde: Frontier-Klasse, komplexes Schlussfolgern, autonome Agenten-Workflows	360GB	720GB	1440GB

INDIKATIVE FÄHIGKEITEN NACH MODELLGRÖSSE (TATSÄCHLICHE LEISTUNG VARIERT JE NACH ARCHITEKTUR UND OPTIMIERUNG).
Q8 (8-BIT-QUANTISIERUNG), FP16 (16-BIT-GLEITKOMMA) UND FP32 (32-BIT-GLEITKOMMA) BEZIEHEN SICH AUF DIE NUMERISCHE PRÄZISION ZUR SPEICHERUNG VON MODELLGEWICHTEN. GERINGERE PRÄZISION REDUZIERT DEN SPEICHERBEDARF BEI WEITGEHENDEM ERHALT DER AUSGABEQUALITÄT – 8-BIT-QUANTISIERUNG GILT FÜR DIE MEISTEN INFERENZAUFGABEN ALS NAHEZU VERLUSTFREI. Q8 IST DIE SPEICHEREFFIZIENTESTE HIER GEZEIGTE OPTION; IN DER PRAXIS WIRD AUCH 4-BIT-QUANTISIERUNG HÄUFIG FÜR ON-DEVICE-BEREITSTELLUNGEN EINGESETZT, WO DER SPEICHER STÄRKER BEGRENZT IST.
QUELLE: OMDIA



Unternehmen, die stark in KI-Workloads investieren, entscheiden sich für Mac

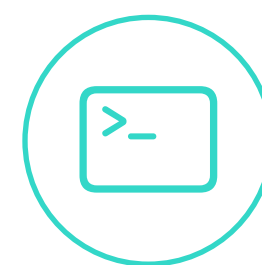
Die Beziehung zwischen KI-Strategie und Plattformauswahl zeigt ein signifikantes Muster. Unternehmen, die KI-Lösungen intern entwickeln, setzen Mac für KI-Workloads fast doppelt so häufig ein wie Unternehmen, die kommerzielle Lösungen kaufen.

Dies ist ein eindeutiger Fall von offener Präferenz. Interne Entwicklungsteams haben die höchsten technischen Anforderungen. Sie benötigen Flexibilität bei der Iteration ohne Grenzkosten, absolute Datensicherheit für proprietäre Informationen und eine Leistung, die sie direkt beobachten, steuern und optimieren können.

Diese Anforderungen entsprechen den Stärken der On-Device-Infrastruktur, und Entwickler:innen, die intensiv lokal entwickeln, entscheiden sich am häufigsten für Mac.

Unternehmen, die es für unerlässlich erachten, dass ihre Infrastrukturplattformen Entwicklungsworkflows unterstützen, benötigen schnelle Iterationen, lokalen Datenzugriff und zuverlässige Leistung – sie werden am stärksten von Unternehmen mit hochentwickelten KI-Fähigkeiten akzeptiert. Die Teams, die wissen, was sie benötigen, entscheiden sich für Mac.

Wenn technische Fachleute die Entscheidungen bezüglich der Plattform treffen, optimieren sie diese für die Tools, die ihre Produktivität steigern. Der Vorsprung von Mac bei der Akzeptanz unter internen Entwickler:innen deutet darauf hin, dass die Plattform etwas bietet, was andere Alternativen nicht bieten und was ihren Arbeitsabläufen entgegenkommt.



Unternehmen, die intern entwickeln, entscheiden sich vor allem aufgrund der Leistung und der Vertrautheit der Entwickler für Mac.

...beinahe doppelt so häufig wie bei Käufern kommerzieller Lösungen. Wenn technische Fachkräfte über Plattform-Entscheidungen bestimmen, wählen sie bevorzugt Tools, die ihre Produktivität steigern – wobei der Mac bei der internen KI-Entwicklung 1,8-mal häufiger eingesetzt wird.

Die wichtigsten Erkenntnisse



Aufbau einer ausgewogenen KI-Infrastruktur

Diese Studie zeigt eine Diskrepanz zwischen aktuellen Infrastrukturansätzen und den Anforderungen von Unternehmen an KI auf. Unternehmen sehen sich mit Sicherheitsbarrieren konfrontiert, die zu Mehraufwand führen, mit Kostenstrukturen, die Betriebsausgaben verschleiern, und mit Workflow-Einschränkungen, die die Entwicklungsgeschwindigkeit beeinträchtigen.

Der Weg in die Zukunft besteht darin, Infrastrukturportfolios mit On-Device-Fähigkeiten zu vervollständigen, damit Unternehmen die Workload-Anforderungen an die Infrastruktur anpassen können, anstatt alle KI-Initiativen über cloudzentrierte Ansätze zu leiten. Unternehmen, die diese Flexibilität aufbauen, verschaffen sich Vorteile, die sich mit der Weiterentwicklung der KI-Fähigkeiten und der Veränderung der Workload-Muster verstärken.



Anpassung der KI-Infrastruktur an die Workloads

Schreiben Sie die Abstimmung von Workload und Infrastruktur auf Grundlage von Modellgröße, Datensensibilität, Nutzungsmustern und Benutzeranforderungen vor. Die Mehrheit der Unternehmen, die Modelle mit weniger als 10 Milliarden Parametern betreiben, verfügt über eine andere optimale Infrastruktur als die KI-Forschenden, die bahnbrechende Modelle entwickeln. Entwicklungs-Workloads haben andere Anforderungen als Produktionsbetrieb. Vermeiden Sie standardisierte Infrastrukturen, die zu Diskrepanzen führen können.



Sicherheit architektonisch lösen

Bei Projekten, die geschützte oder sensible Informationen beinhalten, sollten Sie Infrastrukturoptionen prüfen, die Übertragungsrisiken durch die Architektur eliminieren, anstatt sie durch Prozesse zu verwalten. Die Verarbeitung auf dem Gerät ist nicht der einzige Ansatz, jedoch werden dadurch für Entwicklungsabläufe und Inferenz bei kleineren Modellen ganze Kategorien von Compliance-Aufwand und Verzögerungen durch Sicherheitsüberprüfungen vermieden.



Unvorhersehbare Kosten vorhersehbar machen

Weiten Sie die TCO-Messung über leicht nachverfolgbare Posten wie Cloud-Computing-Rechnungen und Hardware-Anschaffungen hinaus aus. Quantifizieren Sie Verzögerungen bei Sicherheitsüberprüfungen, Compliance-Kosten, Auswirkungen von Ressourcenengpässen auf die Produktivität und Grenzkosten von Experimenten bei nutzungsabhängiger Preisgestaltung. Unternehmen optimieren das, was sie messen. Unvollständige Kostentransparenz führt zu suboptimalen Infrastrukturentscheidungen.



Unterstützung interner technischer Teams bei der Entwicklung hybrider Lösungen

Befähigen Sie interne Entwicklungsteams zur Integration verschiedener Infrastrukturtypen, wobei jeder einzelne spezifische Arbeitsabläufe optimal unterstützt. Die Teams, die die anspruchsvollsten KI-Arbeiten durchführen, erstellen bereits hybride Portfolios, welche die Infrastruktur an die Workloads anpassen. Entwickeln Sie Beschaffungsrahmenbedingungen und Technologie-Richtlinien, die diese strategische Diversifizierung fördern, anstatt sie einzuschränken.

Der KI-Vorteil von Mac ist beabsichtigt, nicht zufällig

Die Präferenz für Mac in Unternehmen ist nicht willkürlich. Sie basiert auf einer Architektur, die direkt auf die Probleme und Lücken von cloudzentrierten oder lokalzentrierten Ansätzen eingeht.

Die einheitliche Speicherarchitektur von Mac bietet ausreichend Kapazität für typische KI-Workloads in Unternehmen und gleichzeitig eine beispiellose Skalierbarkeit der Speicherkapazität zu einem bestimmten Preis, sodass größere Modelle auf dem Gerät ausgeführt werden können. Dies behebt das Problem der Diskrepanz zwischen Workload und Infrastruktur und bietet eine hohe Leistung, ohne dass laufende Cloud-Kosten oder kostspielige lokale Serveroptionen erforderlich sind.



Datensicherheit

Die Verarbeitung auf dem Gerät eliminiert Übertragungsrisiken von vornherein. Für die überwiegende Mehrheit der Unternehmen, die sensible Daten lokal speichern müssen, bietet Mac architektonische Datenschutzfunktionen anstelle von prozeduralen Workarounds.



Unified Memory für KI-Inferenz

20 % der Unternehmensimplementierungen verwenden Modelle mit über 100 Milliarden Parametern – zu umfangreich oder kostspielig für diskreten GPU-VRAM, jedoch gut geeignet für den einheitlichen Speicher von Apple Silicon, der bis zu 128 GB auf dem MacBook Pro und 512 GB auf dem Mac Studio bietet.



Vorhersehbares Kostenmodell

Feste Hardware-Investitionen ersetzen nutzungsabhängige Cloud-Kosten, die sich im Laufe der Zeit skalieren. Entwicklungsteams profitieren von unbegrenzten Experimentiermöglichkeiten ohne Grenzkosten, während Produktionsbereitstellungen Kosten vermeiden, die mit der Nutzung skalieren.



Entwicklerpräferenz

Unternehmen, die KI intern entwickeln, entscheiden sich 1,8-mal häufiger für Mac als solche, die kommerzielle KI-Lösungen kaufen. Wenn technische Teams die Entscheidungen bezüglich der Plattform treffen, optimieren sie diese hinsichtlich Leistung, Sicherheit und Vertrautheit der Entwickler:innen.

Anhang



Über Omdia

Omdia

Kompetenz, die zuhört. Beratung, die weiterbringt.

Omdia ist ein weltweit führendes Technologieforschungsunternehmen, das aus der Fusion der Forschungsabteilung von Informa TechTarget (Ovum, Heavy Reading, Tractica und Canalys) und dem übernommenen Technologieforschungsportfolio von IHS Markit hervorgegangen ist. Wir vereinen das Fachwissen von mehr als 300 Analyst:innen aus dem gesamten Technologiespektrum und decken über 200 Märkte ab. Wir veröffentlichen jährlich über 3.000 Forschungsberichte und decken Tausende von Unternehmen aus den Bereichen Technologie, Medien und Telekommunikation ab. Dank unserer umfassenden Informationen und fundierten Technologiekompetenz sind wir in der Lage, umsetzbare Erkenntnisse zu gewinnen, die unseren Kunden helfen, in der sich ständig weiterentwickelnden Technologielandschaft von heute Zusammenhänge zu erkennen und ihre Geschäftsabläufe zu optimieren – heute und in Zukunft.



Über diese Studie – Methodik und Angaben zur Stichprobe

Die Umfrage zur kritischen KI-Infrastruktur wurde im vierten Quartal 2025 von Omdia durchgeführt und ausgewertet. Sie richtete sich an IT-Entscheidungsträger:innen in Unternehmen aus fünf globalen Märkten: Vereinigte Staaten, Vereinigtes Königreich, Deutschland, China und Indien.

Insgesamt wurden 1.584 Antworten von Unternehmen mit mehr als 500 Mitarbeitenden aus den Bereichen Gesundheitswesen, Finanz- und Fachdienstleistungen gesammelt. Alle Teilnehmenden verfügten entweder über direkte Budgethoheit für Technologiekäufe oder hatten erheblichen Einfluss auf die KI-Strategie ihres Unternehmens. Zu den Befragten gehörten unter anderem CIOs, CTOs, Softwareentwickler:innen und KI-Spezialist:innen.

Die Studie untersuchte die Einstellung von Unternehmen zu Entscheidungen hinsichtlich der KI-Infrastruktur, Sicherheits- und Datenschutzanforderungen, Überlegungen zu den Gesamtbetriebskosten und Plattformpräferenzen für KI-Workloads.

In welchem Land haben Sie Ihren Sitz?

LAND	
USA	20%
Vereinigtes Königreich	20%
Deutschland	20%
China	20%
Indien	20%

Was ist Ihre primäre Branche?

INDUSTRIE	
Gesundheitswesen	33%
Finanz- dienstleistungen	34%
Fach- dienstleistungen	33%

Wie viele Mitarbeitende weltweit?

500 – 999	19%
1.000 – 4.999	23%
5.000 – 9.999	22%
10.000 – 19.999	16%
20.000 – 49.999	12%
>50.000	10%

Wie hoch ist der Jahresumsatz Ihres Unternehmens?

< 100 Mio. USD	8%
100 bis 499 Mio. USD	16%
500 bis 999 Mio. USD	24%
1 bis 4,9 Mrd. USD	23%
5 bis 10 Mrd. USD	19%
> 10 Mrd. USD	9%
Gemeinnützige Organisation	1%

HINWEIS: DIE PROZENTUALEN ANGABEN ERGEBEN AUFGRUND VON RUNDUNGEN MÖGLICHERWEISE NICHT 100%.



The Omdia team of 400+ analysts and consultants are located across the globe

Americas

Argentina
Brazil
Canada
United States

Asia-Pacific

Australia
China
India
Japan
Malaysia
Singapore
South Korea
Taiwan

Denmark
France
Germany
Italy
Kenya
Netherlands
South Africa
Spain

United Kingdom

Europe, Middle East, Africa

Sweden
United Arab Emirates

Omdia

E insights@omdia.com
E consulting@omdia.com
W omdia.tech.informa.com

✉ OmdiaHQ
in Omdia

Citation Policy

Request external citation and usage of Omdia research and data via citations@omdia.com

COPYRIGHT NOTICE AND DISCLAIMER

© 2026 TechTarget, Inc. d/b/a Informa TechTarget. All rights reserved. The Informa TechTarget name and logo are subject to license. All other logos are trademarks of their respective owners. Informa TechTarget reserves the right to make changes in specifications and other information contained in this document without prior notice.