DISKUSSIONSPAPIER

On-Premises vs. Cloud

Wo läuft die KI der Zukunft?

Eine unternehmerische Abwägung zwischen Kontrolle, Kosten und Geschwindigkeit

Pro On-Premises **Pro Cloud** **Pro Hybrid**

HERAUSGEBER

KI Bundesverband & KPMG

VERANSTALTUNG

KI-Stammtisch Nord, Hamburg – Debattenabend am 24. August 2026

MITWIRKENDE

Mitglieder und Teilnehmende des KI-Stammtisch Nord, darunter:

Dr. Christian Geier, prokube.ai GmbH

Peter C. Krell, AI.Mag Redaktion

Frederic Voigt, M.M.Warburg & CO

Victor Wiedermann, Scala Value

und Weitere.

VORWORT

Zu diesem Papier

Dieses Diskussionspapier fasst die Ergebnisse des Debattenabends des KI-Stammtisch Nord vom 24. August 2026 zusammen.

Das Format war bewusst kontrovers angelegt: Die Teilnehmenden wurden einer von drei Delegationen – On-Premises, Cloud oder Hybrid – zugeordnet, um die jeweilige Position argumentativ zu vertreten.

HINWEIS ZUR EINORDNUNG

Die hier festgehaltenen Argumente und Thesen geben ausschließlich den Argumentationsstrang des Abends wieder. Sie stellen ausdrücklich keine Positionen dar – weder die persönliche Überzeugung der einzelnen Teilnehmenden noch die offizielle Haltung, Strategie oder Geschäftsposition ihrer Arbeitgeber.

Dies gilt insbesondere und ausdrücklich auch für die Partnerorganisationen KPMG und den KI Bundesverband: Die im Papier wiedergegebenen Argumente spiegeln in keiner Weise deren strategische Ausrichtung, Produkt- oder Marktposition wider.

Ziel ist nicht eine abschließende Empfehlung, sondern eine nachvollziehbare, faktenbasierte Darstellung der zentralen Abwägungen.

EXECUTIVE SUMMARY

Ergebnisse des Debattenabends

24.08.2026
Debattenabend des KI-Stammtisch Nord in Hamburg

3
Delegationen: On-Premises, Cloud und Hybrid

4
Spannungsfelder: Kontrolle, Kosten, Geschwindigkeit, Skalierbarkeit

Die Wahl der KI-Infrastruktur – On-Premises oder Cloud – ist für Unternehmen längst keine reine IT-Entscheidung mehr, sondern eine strategische Weichenstellung mit Auswirkungen auf Kontrolle, Kosten, Geschwindigkeit und Wettbewerbsfähigkeit. Eine universell „richtige" Antwort gibt es nicht: Die passende Architektur hängt von Auslastung, Workload-Profil, Datensensibilität und organisatorischer Kompetenz eines Unternehmens ab.

Dieses Diskussionspapier dokumentiert den Debattenabend des KI-Stammtisch Nord vom 24. August 2026 in Hamburg, bei dem drei Delegationen – On-Premises, Cloud und Hybrid – ihre Position entlang von vier Spannungsfeldern vertreten haben: Kontrolle und Unabhängigkeit vom Anbieter, Kosten und Betrieb, Geschwindigkeit und Marktzugang sowie Skalierbarkeit und Flexibilität.

Der Abend hat die Frage verschoben. Diskutiert wurde weniger, ob Daten in der Cloud abfließen, als vielmehr, wer welches Risiko trägt, wie bewusst er es trägt und wie schnell er reagieren kann, wenn es eintritt. Die On-Premises-Delegation argumentierte, dass vertragliche Zusagen keine Kontrolle ersetzen: Ob Daten bereits abgeflossen sind, lässt sich nur eingeschränkt nachvollziehen, und die Einhaltung von Zusagen wird in der Praxis kaum kontrolliert. Die Cloud-Delegation hielt dem entgegen, dass es nicht nur um Daten geht: Ein professioneller Anbieter deckt Verlässlichkeit, Disaster Recovery und Patch-Zyklen besser ab als die meisten IT-Abteilungen, und wer Microsoft 365 einsetzt, hat die Grundsatzentscheidung längst getroffen. Die Hybrid-Delegation verstand ihre Position nicht als Kompromiss, sondern als Aufteilung nach Datenklasse und Anwendungsfall: sensible Daten intern, große Frontier-Modelle extern, Lastspitzen über die Cloud.

In der Kostenfrage blieb der Abend ohne Einigung. Dass es einen Break-even gibt, war unstrittig. Strittig war, wann er eintritt. Für On-Premises sprach, dass danach im Wesentlichen Energiekosten bleiben und nicht länger pro Token gezahlt wird – zugespitzt in der Frage an den Saal:

> **Geht ihr noch ins Internetcafé, oder habt ihr einen eigenen Laptop?**
>
> Frage aus der On-Premises-Delegation an den Saal

Für die Cloud sprachen explodierende Hardwarepreise, im Normalbetrieb ineffizient ausgelastete Eigensysteme und der Umstand, dass jede Preissenkung pro Token den Break-even nach hinten verschiebt. Beim Leistungsabstand zu den führenden Modellen zeichnete sich hingegen eine gemeinsame Linie ab: Er schrumpft, und für einen erheblichen Teil der Anwendungsfälle ist er ohne praktische Bedeutung.

Das wirksamste Argument des Abends war der Kill Switch. Er wurde von allen drei Delegationen unterschiedlich gedeutet: als Begründung für den Eigenbetrieb, als beherrschbares Restrisiko und als Beleg dafür, dass eine zweite Säule notwendig ist, weil auf europäische Intervention nicht zu zählen ist. Entschärft wurde er durch die Fallback-Option, die in beide Richtungen gelesen werden kann: Wer in der Cloud betreibt, kann im Ernstfall auf eigene Kapazität wechseln, und wer selbst betreibt, kann Ausfälle über die Cloud abfedern. Welche Richtung die realistischere ist, blieb use-case-spezifisch.

GEMEINSAMES ERGEBNIS

Entscheidend ist weniger die gewählte Architektur als die Fähigkeit, sie zu wechseln. Diese Fähigkeit muss vorab aufgebaut werden, nicht erst im Ernstfall. Und die Entscheidung fällt nicht einmalig für das Unternehmen, sondern je Anwendungsfall nach Datenklasse, Auslastung und Kritikalität.

KAPITEL 1

Einleitung

1.1 Warum die Entscheidung zwischen On-Premises, Cloud und Hybrid unternehmerisch zählt

Die Entscheidung für eine KI-Infrastruktur bindet Unternehmen langfristig: kapitalintensiv bei On-Premises-Investitionen, vertraglich und technisch bei Cloud-Lock-in. Sie berührt zugleich mehrere unternehmerische Dimensionen, die selten gemeinsam betrachtet werden:

- Kontrolle & Unabhängigkeit vom Anbieter
- Kosten (Investition vs. laufender Betrieb, Total Cost of Ownership)
- Geschwindigkeit & Zugang zu den leistungsfähigsten Modellen
- Skalierbarkeit und Flexibilität bei schwankendem Bedarf
- Wettbewerbsfähigkeit gegenüber international agierenden, cloud-nativen Konkurrenten

Für den deutschen Mittelstand ist diese Entscheidung dabei selten reversibel. Ein Architekturwechsel nach getroffener Wahl ist mit erheblichem Migrationsaufwand und Kosten verbunden.

1.2 Der Anthropic-Fall als Ausgangslage der Debatte

Als Anthropic am 13. Juni kurzfristig zwei Claude-Modelle für einen Teil der Nutzer abstellte, standen Unternehmen, die diese Modelle produktiv im Einsatz hatten, von einem Moment auf den nächsten vor einem Ausfall. Ohne Einfluss, ohne Vorwarnung. Dieser Fall zeigt exemplarisch, worum es bei der Infrastrukturentscheidung im Kern geht: Wer eine KI-Fähigkeit über einen externen Anbieter bezieht, gibt damit auch ein Stück Kontrolle über deren Verfügbarkeit ab, unabhängig davon, wie gut der Vertrag oder das Service Level Agreement ausgestaltet ist.

Am Ende bedeutet digitale Souveränität nicht nur die Frage, ob man ein deutsches oder europäisches Modell nutzt, sondern auch, wie schnell man im Zweifel auf eine verlässliche Alternative umsteigen kann.

FÜR DIE DEBATTE

Der Fall eignet sich als Einstieg in die Kernfrage des Abends: Ist ein solches Ausfallrisiko ein Preis, den man für Geschwindigkeit und Zugang zu den besten Modellen bewusst in Kauf nimmt – oder ein Argument, das strukturell für mehr Kontrolle durch On-Premises-Betrieb spricht?

1.3 Format des Abends: Debatte statt Panel

Der KI-Stammtisch Nord folgte an diesem Abend bewusst einem anderen Format als sonst üblich: Statt Impulsvortrag und Panel stand eine offene Positionsdebatte im Zentrum.

Teilnehmende haben sich bereits bei der Anmeldung einer von drei Delegationen zugeordnet – On-Premises, Cloud oder Hybrid. Diese Zuteilung muss keine persönliche Überzeugung widerspiegeln: Wer eine Position verteidigt, schärft dadurch die eigenen Argumente unabhängig von der eigenen Meinung. Die Hybrid-Delegation vertritt dabei explizit die These, dass eine bewusste Kombination aus On-Premises und Cloud der eigentlich überzeugendste Ansatz ist – nicht ein unentschlossenes „bisschen von beidem".

1.4 Leitfrage & Zielsetzung

Wie viel Kontrolle ist einem Unternehmen, wie viel Geschwindigkeit, Kosten und Marktzugang wert – und wer trägt im Ernstfall die Konsequenzen einer Abhängigkeit vom Anbieter?

Dieses Papier hat nicht das Ziel, einen Konsens vorwegzunehmen. Es soll ein geschärftes Verständnis der zentralen unternehmerischen Abwägungen vermitteln und als Grundlage für die Breakout Sessions und die Debatte dienen. Aus der Debatte entsteht im Anschluss ein vollständiges Diskussionspapier, das der KI Bundesverband veröffentlicht.

KAPITEL 2

Optionsspektrum

Für die Debatte legen wir folgendes Spektrum an Infrastrukturoptionen zugrunde. Die Debatte selbst wird bewusst mit drei Delegationen geführt – On-Premises, Cloud und Hybrid –, da Hybrid-Architekturen in der unternehmerischen Praxis oft die tatsächlich gewählte Lösung sind und eine eigenständige, ernstzunehmende Position verdienen.

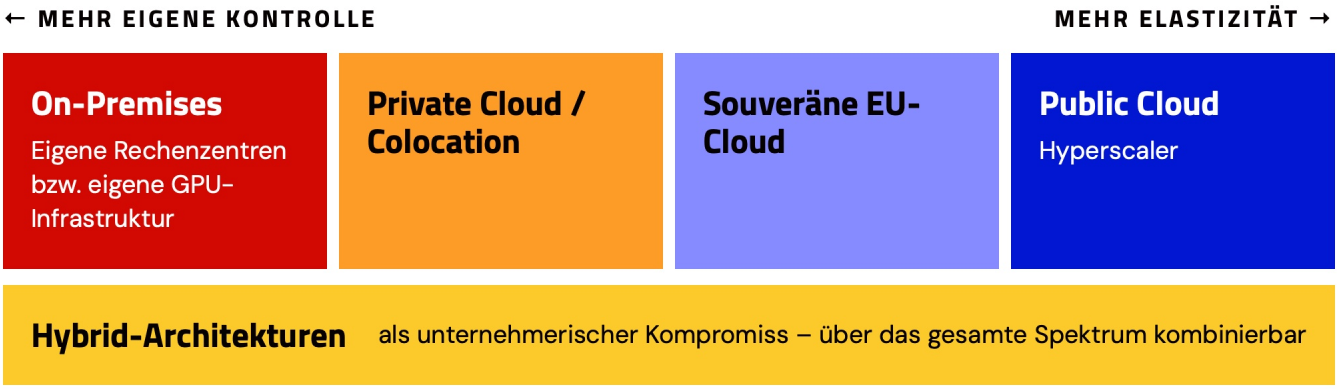


Abb. 1: Spektrum der Infrastrukturoptionen, die der Debatte zugrunde liegen.

WICHTIGE DIFFERENZIERUNG FÜR DIE DEBATTE

Kontrolle über die Infrastruktur ist nicht gleichzusetzen mit Kontrolle über das Modell. Ein Unternehmen kann die physische Infrastruktur vollständig besitzen (On-Premises) und dennoch die Kontrolle über ein per API genutztes Frontier-Modell verlieren – während On-Premises-Betrieb mit offenen Modellen echte Modellkontrolle ermöglicht, aber ggf. Leistungseinbußen gegenüber Frontier-Modellen bedeutet. Diese zweite Achse sollte in der Debatte nicht mit der ersten vermischt werden.

KAPITEL 3

Die Debatte: Pro On-Premises vs. Pro Cloud vs. Pro Hybrid

Die folgenden vier Abschnitte strukturieren die zentralen Spannungsfelder aus unternehmerischer Sicht – jeweils aus der Perspektive aller drei Delegationen. Die Dimensionen sind auf Basis der Recherche des KI Bundesverbands vorbereitet; die eigentlichen Argumente werden in den Breakout Sessions erarbeitet und von den Delegierten in der Debatte vorgetragen.

Leitfragen

01 Ist On-Premises für die meisten Unternehmen noch eine ernsthafte Option, oder angesichts des Innovationstempos der Hyperscaler ein Auslaufmodell?

02 Wie viel Leistungsabstand zu den führenden Cloud-Modellen ist ein Unternehmen bereit zu akzeptieren, um mehr Kontrolle über die eigene Infrastruktur zu gewinnen?

03 Wer trägt im Ernstfall die Konsequenzen einer Abschaltung oder Leistungsänderung – der Anbieter, oder implizit der Kunde selbst durch die Wahl des Anbieters?

04 Verschiebt sich die Verhandlungsmacht zwischen Kunde und Cloud-Anbieter strukturell zugunsten weniger großer Hyperscaler – und was bedeutet das für den deutschen Mittelstand?

05 Ist eine vermeintlich „sichere" Cloud-Wahl mit Mehrkosten für höhere Kontrolle (z. B. souveräne EU-Cloud-Angebote) für den Mittelstand wirtschaftlich tragbar, oder untergräbt sie die Wettbewerbsfähigkeit gegenüber global agierenden, cloud-nativen Konkurrenten?

06 Sollte die Infrastrukturentscheidung zentral von der Unternehmensführung getroffen werden, oder dezentral je Anwendungsfall und Fachbereich?

07 Ist Hybrid eine eigenständige, überzeugende dritte Position – oder in Wahrheit nur ein Ausweg vor der eigentlichen Grundsatzentscheidung, der am Ende die Nachteile beider Welten (Komplexität und Kontrollverlust) statt ihrer Vorteile vereint?

CHATHAM HOUSE RULE

Die im Rahmen der Debatte genannten Argumente laufen standardmäßig unter der Chatham House Rule und werden im finalen Diskussionspapier ohne Namensnennung verarbeitet. Einzelne Diskutantinnen und Diskutanten können mit Namen, ggf. Foto und Zitat aufgenommen werden, sofern sie dem vorab oder vor Ort ausdrücklich zustimmen.

3.1 SPANNUNGSFELD 1
Kontrolle & Unabhängigkeit vom Anbieter

LEITFRAGEN

Wie real ist das Risiko einer plötzlichen Abschaltung oder Leistungsänderung durch den Anbieter (vgl. Anthropic-Fall)? Wie viel Verhandlungsmacht hat ein einzelnes Unternehmen gegenüber einem Hyperscaler? Was bedeutet Business Continuity jenseits klassischer Service-Level-Agreements?

Pro On-Premises | **Pro Cloud** | **Pro Hybrid**

DIMENSION **Abhängigkeit vom Anbieter**

- Daten fließen nicht ab – darum geht es vor allem
- Vertragliche Zusagen ersetzen keine Kontrolle: Ob Daten bereits abgeflossen sind, ist nur eingeschränkt nachvollziehbar; Zusagen werden zu wenig kontrolliert (z. B. öffentliche Verwaltung)
- US CLOUD Act: behördlicher Zugriff bleibt möglich, unabhängig vom Serverstandort
- Politische Abhängigkeit: Die EU hat im Handelsstreit mit den USA wenig Einfluss, so einige Teilnehmende (vgl. Reaktion auf den 13. Juni)
- In Branchen wie dem Finanzsektor, bei Versicherern und Krankenkassen sind praktisch alle Daten hochsensibel – hier kann kein Risiko eingegangen werden
- Nicht jeder Geschäftsprozess braucht das leistungsfähigste Modell

- Es geht nicht nur um Daten – und einen Datenabfluss gibt es bei professionellen Anbietern nicht
- Konsistenzfrage: Gilt die Kritik nur für KI oder auch für Microsoft 365 und die übrige Cloud-Nutzung? Die meisten Unternehmen sind längst dort
- Cloud ist eine Form der Risikokontrolle, nicht deren Aufgabe: Sicherheit hat viele Aspekte, die ein Anbieter professioneller abdeckt
- Abhängigkeiten sind real und geopolitisch sensibel – aber vertraglich und architektonisch adressierbar

- Hybrid heißt bewusste Aufteilung nach Use Case und Datenklasse – nicht „ein bisschen von beidem“
- Zwei Lösungen bedeuten keine einseitige Abhängigkeit: Sensible Daten bleiben intern, große Frontier-Modelle kommen aus der Cloud
- Kill Switch: Als Fable am 13. Juni abgeschaltet wurde, hat die EU nicht eingegriffen – deshalb nie nur eine Säule

3.1 KONTROLLE & UNABHÄNGIGKEIT VOM ANBIETER (FORTSETZUNG)

Pro On-Premises | **Pro Cloud** | **Pro Hybrid**

DIMENSION Reaktionsfähigkeit im Ernstfall

- Kill Switch: Wird ein Modell abgeschaltet, läuft der eigene Betrieb weiter
- Szenario: Ein Anbieter schafft die nächste Finanzierungsrunde nicht – wer selbst betreibt, ist davon unabhängig
- Zuverlässigkeit: Auch Cloud-Dienste können überlastet sein („später erneut versuchen")
- Ausfallsicherheit ist gestaltbar: Auch ein Cloud-Rechenzentrum kann brennen, eigene Systeme lassen sich mehrfach absichern
- Wie fehleranfällig die Cloud tatsächlich ist, ist von außen nicht einsehbar

- Verlässlichkeit ist höher als im Eigenbetrieb: mehrere Sicherheitsebenen, redundante Standorte
- Disaster Recovery läuft schneller und ist erprobt
- Ein vollständiger Cloud-Ausfall ist unwahrscheinlich – und dafür gibt es Fallback-Architekturen, use-case-spezifisch
- Kill-Switch-Szenario: Falls es eintritt, bleibt der Wechsel zu On-Premises möglich

- Bei Abschaltung oder Ausfall trägt die zweite Säule – Fallback in beide Richtungen
- Kritische Prozesse laufen intern weiter, während extern nachgesteuert wird
- Der Ernstfall ist eingeübt, weil beide Wege im Regelbetrieb genutzt werden

DIMENSION Verhandlungsmacht & Vertragsgestaltung

- Vertrauen als Entscheidungsgrundlage ist obsolet – Verträge regeln Haftung, nicht Kontrolle
- Regulatorische Zusicherungen für neue Modelle sind kaum zu bekommen
- Compliance-Vorgaben sind auch mit EU-Servern nicht zu 100 % einzuhalten
- Kein Vendor-Lock-in: verhandeln ohne Migrationsdruck
- Bei hochsensiblen Daten gibt es nichts zu verhandeln

- Risiko-Shift: Betriebsrisiko und Sicherheitsverantwortung liegen beim Anbieter, der sie besser tragen kann
- Viele Unternehmen gehen dieses Risiko bewusst und kalkuliert ein – Risiko null ist kein Geschäftsmodell
- Verhandelt wird über Preis und Flexibilität; im Eigenbetrieb verhandelt man mit Hardwarelieferanten in einem angespannten Markt
- Exit-Strategie und Multi-Cloud sind Vertragsgegenstand

- Eine glaubwürdige Exit-Option ist das stärkste Verhandlungsargument
- Gegenfrage an die Cloud-Delegation: Wie souverän sind Entscheidungen dort tatsächlich?
- Anbieterwechsel ist real möglich, nicht nur vertraglich zugesichert

3.2 SPANNUNGSFELD 2
Kosten & Betrieb

LEITFRAGEN

Wie ist die Auslastungsabhängigkeit der Kosten zu bewerten (Break-even bei ca. 40–50 % vs. ca. 65 % Dauerauslastung)? Welche Rolle spielen Total Cost of Ownership, Datenexport-Kosten und Hardware-Wertverlust? Welche Kompetenzen (MLOps vs. FinOps) benötigt welches Modell im Betrieb?

Pro On-Premises | **Pro Cloud** | **Pro Hybrid**

DIMENSION Kapitalbindung vs. laufende Kosten

- Der Break-even wird erreicht – danach fallen außer Energie keine Zusatzkosten an, es wird nicht mehr pro Token gezahlt
- Ist On-Premises wirklich unwirtschaftlicher? Der Abstand ist kleiner als angenommen; kleinere Pakete und Ausbaustufen sind buchbar
- Vergleichsmaßstab sind die Personalkosten, die der KI-Einsatz einspart
- „Kaufe ich einen eigenen Rechner oder gehe ich ins Internetcafé?“ Nicht jeder Use Case erfordert den gleichen Aufwand
- Die Eintrittshürde ist real – aber eine Frage der Prioritäten: Bei sensiblen Daten hat deren Schutz Vorrang

- Geringe Einstiegshürde: keine Kapitalbindung, sofort startfähig
- Hardwarekosten explodieren – eigene GPU-Investitionen sind ein Preisrisiko
- Der Break-even ist umstritten: Die Rechnung geht hin und her und hängt an Annahmen, die sich ständig verschieben
- Betriebskosten bleiben planbar und nutzungsabhängig

- Es entstehen Kosten für beides, aber nicht doppelte Kosten – teurere Token nur dort, wo sie nötig sind
- Kleinere Basisinvestition, weil nicht die Spitzenlast dimensioniert werden muss
- Den Break-even gibt es, er ist aber nicht schnell erreichbar – Hybrid überbrückt die Zeit bis dahin

3.2 KOSTEN & BETRIEB (FORTSETZUNG)

Pro On-Premises	Pro Cloud	Pro Hybrid

DIMENSION Kosten bei niedriger / schwankender Auslastung

- Freie Kapazität ist nicht verloren: Interne Cloud und weitere Dienste können sie auffüllen (Vorbild STACKIT: eigener Bedarf plus Angebot an Dritte)
- Leitfrage: Welche weiteren Cases ließen sich auf derselben Hardware abdecken?
- Keine Preisspirale pro Token bei wachsender Nutzung

- Eigene Hardware läuft im Normalbetrieb ineffizient – bezahlt wird sie trotzdem
- Nachskalieren ist On-Premises nicht kurzfristig möglich; Überdimensionierung ist der Regelfall
- Bezahlt wird nur, was genutzt wird – Schwankungen kosten nichts extra

- Basislast intern, Spitzen extern – die eigene Hardware läuft hoch ausgelastet
- Keine Überdimensionierung für Ausnahmefälle
- Kostenflexibilität: Man kann jederzeit umschichten

DIMENSION Kosten bei hoher Dauerauslastung

- Bei hoher Auslastung ist On-Premises teilweise günstiger – hier kippt die Rechnung
- Nach dem Break-even bleiben im Wesentlichen die Energiekosten
- Kostenniveau vergleichbar mit zusätzlichem Personal, nicht mit einer Lizenz

- Auch bei hoher Last: Hardware veraltet und muss ersetzt werden – der Wertverlust trifft den Eigenbetrieb
- Energie, Rechenzentrum und Personal werden in On-Premises-Rechnungen systematisch unterschätzt
- Der Break-even verschiebt sich mit jeder Preissenkung pro Token nach hinten

- Hohe Grundlast rechtfertigt eigene Kapazität – genau dort läuft sie
- Teure Token nur für die Fälle, die das beste Modell wirklich brauchen

DIMENSION Benötigte Betriebskompetenz

- Aufwand beherrschbar: monatliche Update-Zyklen statt Dauerbaustelle
- Der Aufwand skaliert mit dem Anspruch – nicht jeder Fall braucht ein eigenes Rechenzentrum
- Aufgebaute Kompetenz bleibt im Unternehmen statt beim Anbieter
- Datenqualität als eigentlicher Erfolgsfaktor wird ohnehin intern erarbeitet

- Welche Fähigkeiten braucht man im Team? Patch-Ketten, Updates und Sicherheit sind im Eigenbetrieb eine Dauerpflicht
- Knappe Fachkräfte arbeiten an Use Cases statt an Infrastruktur
- Betriebskosten inklusive Personal sind in der Cloud transparent

- Offener Nachteil: Wartung und zwei Kompetenzprofile im Haus
- Antwort darauf: klare Zuordnung pro Use Case und Standardisierung der Schnittstellen
- Der Aufwand ist der Preis für Wahlfreiheit – und geringer als der Preis eines Fehlgriffs

3.3 SPANNUNGSFELD 3
Geschwindigkeit, Innovation & Marktzugang

LEITFRAGEN

Wie ist der aktuelle Leistungsabstand zwischen den größten Cloud-Modellen und offen verfügbaren Modellen zu bewerten (aktuell ca. vier Monate Rückstand, bei Alltagsaufgaben nahezu geschlossen)? Wie schnell verändert sich dieser Abstand? Welche Rolle spielt der Zugang zu den jeweils neuesten Modellen für die eigene Wettbewerbsfähigkeit?

Pro On-Premises | **Pro Cloud** | **Pro Hybrid**

DIMENSION Zugang zu den leistungsfähigsten Modellen

- Hängt man wirklich hinterher? Der Abstand zu den Frontier-Modellen schrumpft
- Braucht es überhaupt das neueste Modell? Für Aufgaben wie E-Mails oder Zusammenfassungen nicht
- Der Leistungsunterschied ist kaum quantifizierbar und gerade für KMU praktisch irrelevant
- Hochspezialisierte Cases: Eigene Modelle lassen sich genauer trainieren – entscheidend ist die Datenqualität

- Zugang zu den leistungsfähigsten Modellen ohne eigene Investition
- Die Schwelle für Innovation sinkt deutlich – ausprobieren kostet fast nichts
- Neue Modelle stehen am Tag der Veröffentlichung bereit

- Frontier-Modelle über die Cloud, wo sie den Unterschied machen
- Datenschutz über On-Premises, wo die Daten es verlangen
- Man braucht nicht für jeden Prozess das beste Modell

DIMENSION Time-to-Market für neue KI-Anwendungsfälle

- Modelle sind auch On-Premises wechselbar – offene Gewichte lassen sich austauschen
- Entscheidung pro Use Case möglich, nicht als Grundsatzfestlegung
- Auch als kleiner Player lohnt es sich, früh eigene Kompetenz aufzubauen
- In regulierten Bereichen dauert ein Modellwechsel ohnehin – der Geschwindigkeitsvorteil der Cloud verpufft

- Niedrigschwelliger Zugang: Start in Tagen statt in Beschaffungszyklen
- Kein Infrastrukturaufbau, bevor der erste Use Case Wert liefert
- Scheitern ist billig – das erhöht die Zahl der Versuche

- Schnell in der Cloud starten, bewährte Cases nach innen holen
- Keine Blockade durch Beschaffung, keine Sackgasse durch Anbieterbindung
- Zwei Lösungen im Haus: Nimm die, die zum Fall passt

3.3 GESCHWINDIGKEIT, INNOVATION & MARKTZUGANG (FORTSETZUNG)

Pro On-Premises | **Pro Cloud** | **Pro Hybrid**

DIMENSION

Zukunftssicherheit der Infrastrukturentscheidung

- Kein Vendor-Lock-in: Die Wechselfähigkeit bleibt im eigenen Haus
- On-Premises smart bauen: Architektur so anlegen, dass Modelle schnell austauschbar sind
- Regulatorische Zusicherungen für neue Modelle bleiben schwierig – eigene Kontrolle ist die robustere Wette
- Das Kill-Switch-Risiko entfällt
- Keine Investition, die technologisch veraltet, bevor sie sich rechnet
- Modernisierung ist Aufgabe des Anbieters – Risiko-Shift nach außen
- Wer sich nicht auf Hardware festlegt, bleibt beim nächsten Technologiesprung anschlussfähig
- Keine Wette auf eine einzige Architektur
- Lehre aus dem Fable-Fall (Anthropic): Wer beide Wege betreibt, übersteht Anbieterentscheidungen
- Das Gewicht zwischen beiden Säulen lässt sich mit dem Markt verschieben

3.4 SPANNUNGSFELD 4

Skalierbarkeit & Flexibilität

LEITFRAGEN

Wie gut lässt sich unvorhersehbare Nachfrage (z. B. saisonale Spitzen, neue Anwendungsfälle) abfangen? Welchen Wert hat die Elastizität der Cloud gegenüber der Planungssicherheit von On-Premises?

Pro On-Premises | **Pro Cloud** | **Pro Hybrid**

DIMENSION

Umgang mit Nachfragespitzen

- Lastspitzen sind eine Frage der Dimensionierung, kein Argument gegen eigene Kapazität
- Cloud ist bei Spitzen nicht automatisch verfügbar – auch dort gibt es Überlast
- Hybrid als Ventil klingt naheliegend, kombiniert aber die Nachteile beider Welten
- Elastizität ist das Kernargument: Spitzen werden abgefangen, ohne ganzjährig dafür zu zahlen
- On-Premises kann nicht nachskalieren – die Spitze definiert die Investitionshöhe
- Skalierbarkeit ohne Planungsvorlauf
- Kernargument: Spitzen über die Cloud abfangen, Grundlast intern halten
- Keine Investition, die nur für Ausnahmelast existiert
- Nachteil: Die Umschaltung muss technisch vorbereitet sein

3.4 SKALIERBARKEIT & FLEXIBILITÄT (FORTSETZUNG)

Pro On-Premises	Pro Cloud	Pro Hybrid

DIMENSION Vorlaufzeit bei Kapazitätserweiterung

- Erweiterung in kleineren Paketen statt großer Sprünge verkürzt die Vorlaufzeit
- Die Vorlaufzeit ist planbar – anders als eine kurzfristige Anbieterentscheidung
- Zwischenzeitlich freie Kapazität lässt sich mit internen Diensten auslasten

- Kapazität ist sofort verfügbar, keine Lieferzeiten für Hardware
- Keine Abhängigkeit von einem angespannten Hardwaremarkt
- Fehleinschätzungen sind korrigierbar, nicht abgeschrieben

- Die Cloud überbrückt, während eigene Kapazität aufgebaut wird
- Wachstum wird nicht durch Lieferzeiten gebremst
- Ausbau erfolgt datenbasiert, nicht auf Verdacht

DIMENSION Flexibilität bei sich ändernden Anforderungen

- Pro Use Case entscheiden: sensible Fälle intern, unkritische ggf. extern
- Modelltausch bleibt möglich, wenn die Architektur darauf ausgelegt ist
- Hybrid bedeutet doppelte Investition und Mitarbeitende, die für beide Systeme ausgebildet werden müssen – viel Komplexität, die man beherrschen muss

- Use-case-spezifische Wahl von Modell und Kapazität, jederzeit änderbar
- Fallback-Architekturen sind Standard – die Cloud kann selbst der Fallback sein
- Keine Bindung an eine einmal getroffene Dimensionierung

- Einfach switchen: Workloads lassen sich nach Kosten, Datenlage und Modellqualität verschieben
- Nachteil klar benannt: komplexere Infrastruktur und höherer Wartungsaufwand
- Flexibilität ist kein Selbstzweck – sie muss pro Use Case begründet sein

KAPITEL 4

Empfehlungen: „Es kommt darauf an“ – Perspektiven nach Unternehmenstyp

Die Bewertung hängt maßgeblich von Unternehmensgröße, Auslastung und Workload-Typ ab. Für die folgenden Konstellationen stellen wir Kontext und Fragestellung dar; vorgestellt werden jeweils die naheliegende Empfehlung sowie das stärkste Gegenargument.

Etabliertes Industrieunternehmen mit sensiblen Konstruktionsdaten

NAHELIEGENDE EMPFEHLUNG

On-Premises für den Kern – Konstruktions-, Prozess- und Qualitätsdaten bleiben im eigenen Betrieb. Die Grundlast ist meist hoch genug für den Break-even, Betriebskompetenz ist im Haus vorhanden. Cloud für unkritische Assistenzfälle.

STÄRKSTES GEGENARGUMENT

Die Grenze zwischen kritisch und unkritisch ist im Alltag unscharf und wird von den Fachbereichen verschoben. Wer die stärksten Modelle nicht nutzt, verliert Tempo in der Engineering-Assistenz – und die eigene IT ist selten nachweislich sicherer als ein professionelles Rechenzentrum.

Wachstumsstarkes Scale-up / Tech-Start-up

NAHELIEGENDE EMPFEHLUNG

Cloud. Die Flexibilität der Cloud-Nutzung passt zu unvorhersehbaren Wachstumsfaktoren. Die Einstiegshürde ist niedrig, neue Modelle stehen sofort bereit, andere Modelle können schneller eingesetzt und genutzt werden.

STÄRKSTES GEGENARGUMENT

Kill-Switch- und Preisrisiko treffen genau die, die keine Alternative aufgebaut haben. Wer früh keine eigene Kompetenz entwickelt, kann später nicht wechseln. Die Transformation zu einem souveränen Ansatz kann letztlich hohe Kosten verursachen. Ebenso besteht das Risiko einer Kostenexplosion durch hohe Token-Nutzung.

4 PERSPEKTIVEN NACH UNTERNEHMENSTYP (FORTSETZUNG)

Mittelständischer Hidden Champion (Maschinenbau, Produktion)

NAHELIEGENDE EMPFEHLUNG

Portfolio statt Grundsatzentscheidung: Cloud für Standardaufgaben wie Texte, Übersetzung und Recherche, eigene Kapazität dort, wo Fertigungs- und Kundendaten hineinlaufen – in kleinen Ausbaustufen statt als Rechenzentrum.

STÄRKSTES GEGENARGUMENT

Zwei Wege bedeuten zwei Kompetenzprofile, und dafür fehlen im Mittelstand die Leute. Da der Leistungsunterschied der Modelle hier ohnehin kaum spürbar ist, spricht viel für die einfachere Lösung – also für einen Weg statt für zwei.

Konzern mit hoher, planbarer Dauerauslastung

NAHELIEGENDE EMPFEHLUNG

Eigene Kapazität für die planbare Dauerlast – hier kippt die Kostenrechnung, und nach dem Break-even bleiben im Wesentlichen Energiekosten. Cloud für Experimente und den Zugriff auf neue Modelle.

STÄRKSTES GEGENARGUMENT

Planbarkeit ist eine Momentaufnahme: Jede Modellgeneration verändert den Hardwarebedarf, und der Wertverlust trifft den Eigenbetrieb. Zudem verhandelt ein Konzern Cloud-Konditionen, die die Rechnung wieder drehen können.

KAPITEL 5

Entscheidungsraster für Unternehmen

Wir empfehlen ein sequenzielles Vorgehen zur Herleitung der individuellen Infrastrukturentscheidung:

1

Datensensibilität ehrlich einschätzen

Je sensibler Daten und Anwendungsfälle (z. B. Konstruktionsdaten, Geschäftsgeheimnisse), desto eher lohnt sich der Blick auf mehr Kontrolle durch On-Premises.

2

Reale Auslastung messen, nicht schätzen

Compute-/GPU-Auslastung der letzten 90 Tage erheben, statt auf Basis hypothetischer Volllast zu planen.

3

Workload-Profil klassifizieren

Kontinuierliche Inferenz, Batch-Training und experimentelle Last erfordern unterschiedliche optimale Architekturen.

4

Organisatorische Kompetenz ehrlich bewerten

Fehlende Betriebskompetenz entwertet den On-Premises-Vorteil, fehlende Kostensteuerungs-Kompetenz den Cloud-Vorteil.

Das Ergebnis dieses Rasters ist in aller Regel kein binäres „On-Premises oder Cloud“, sondern ein Portfolio-Ansatz: On-Premises für sensible, dauerhaft hoch ausgelastete Anwendungen; Cloud-Kapazität für planbares Training und Experimente; Hybrid als pragmatischer Zwischenschritt für Unternehmen, die (noch) nicht vollständig migrieren können oder wollen.

Das Fallback-Prinzip kommt hier besonders zum Tragen, da bei technischen Systemen immer die Möglichkeit eines Ausfalls besteht.

KAPITEL 6

Marktkontext: Wohin bewegt sich die Industrie?

Unabhängig vom Ausgang der Debatte lässt sich ein klarer Trend beobachten: Immer mehr Unternehmen in Europa überprüfen aktiv ihre Abhängigkeit von einzelnen Cloud-Anbietern.

Ein erheblicher Teil der westeuropäischen Unternehmen prüft derzeit, wie stark die eigene digitale Infrastruktur von einzelnen Anbietern außerhalb der EU abhängt, und verschiebt Investitionen in Richtung größerer eigener Kontrolle (Gartner, 2025). Vorfälle wie der Anthropic-Fall haben diese Prüfung beschleunigt, ausgelöst haben sie sie nicht.

Gleichzeitig entsteht neue europäische Alternativ-Infrastruktur: Beispiele wie das Rechenzentrum von BearingPoint in Graz (Mai 2026) oder das neue Rechenzentrum von Mistral AI bei Paris (ca. 830 Mio. EUR Fremdkapital, ca. 13.800 NVIDIA-Chips, Produktivstart Q2 2026) zeigen, dass auch europäische Anbieter in Kontrolle und Leistungsfähigkeit investieren. Für Unternehmen bedeutet dies: Das Optionsspektrum jenseits der etablierten Hyperscaler wächst – und wird für die eigene Entscheidung zunehmend relevant.

ca. 830 Mio. €

Fremdkapital für das neue Rechenzentrum von Mistral AI bei Paris

ca. 13.800

NVIDIA-Chips im Mistral-AI-Rechenzentrum

Q2 2026

Produktivstart des Mistral-AI-Rechenzentrums

KAPITEL 7

Schluss, Fazit und Ausblick

KERNERGEBNISSE DES DEBATTENABENDS

01 Die Leitfrage verschob sich von der Datensicherheit zur Risikoverteilung: nicht ob ein Risiko besteht, sondern wer es trägt und wie schnell das Unternehmen reagieren kann.

02 Der Break-even ist unstrittig vorhanden, sein Zeitpunkt nicht: Er hängt an Annahmen, die sich mit jeder Preissenkung pro Token und jedem Hardwarezyklus verschieben.

03 Der Leistungsabstand zu den führenden Modellen gilt als schrumpfend und für viele Alltagsfälle als irrelevant – offen blieb, ob das auch für wettbewerbsentscheidende Kernprozesse zutrifft.

04 Der Kill Switch war das stärkste, aber kein eindeutiges Argument: Er wurde als Grund für den Eigenbetrieb, als beherrschbares Restrisiko und als Begründung für eine zweite Säule gelesen.

05 Wichtiger als die Architektur ist die Exit-Fähigkeit: Wer wechseln kann, verhandelt besser und übersteht Anbieterentscheidungen.

Fazit

Die Debatte hat die verbreitete Zuspitzung „On-Premises oder Cloud“ nicht bestätigt. Beide Seiten setzten ihre stärksten Argumente an unterschiedlichen Stellen an: On-Premises dort, wo Daten und Verfügbarkeit über den Geschäftsbetrieb entscheiden, Cloud dort, wo Zugang, Elastizität und die Verlagerung des Betriebsrisikos im Vordergrund stehen. Bei den Kosten hingegen fehlt bislang eine belastbare Datengrundlage dafür, welche Variante tatsächlich wirtschaftlicher ist – zumal der Break-even je nach Umfang des Rechenzentrums unterschiedlich ausfällt. Für die Zukunft gilt es, hier belastbare Vergleichsrechnungen auf Basis realer Auslastungsdaten zu erarbeiten.

Für die unternehmerische Praxis folgt daraus weniger eine Empfehlung als eine Reihenfolge: erst Datenklasse und Kritikalität eines Anwendungsfalls bestimmen, dann die reale Auslastung messen, dann die eigene Betriebskompetenz ehrlich bewerten. Vor allem gilt es, die Wechselfähigkeit in jedem Fall als Anforderung mitzuentwickeln – nicht als späteren Notausgang.

Der Fall vom 13. Juni hat gezeigt, wie schnell eine Fähigkeit entfallen kann, die zuvor als selbstverständlich verfügbar galt; er hat ebenso gezeigt, dass handlungsfähig bleibt, wer die Alternative kennt.

QUELLEN

Literaturverzeichnis

1 Sovereign AI in Europa 2026: Was die neuen Souveränitäts-Stacks bedeuten – BearingPoint, Mistral & Co.

2 Building the Business Case for On-Premise vs. Cloud GPU Infrastructure.

3 Cloud vs On-Premise Economics for AI – Kostenanalyse und Auslastungsmodelle.

4 Anthropic – Statement zur Aussetzung und Wiederherstellung des Zugangs zu Fable 5 / Mythos 5 (Juni/Juli 2026), anthropic.com/news/fable-mythos-access.

5 Open-Weight vs Closed Frontier Snapshot, Juni 2026 – Presenc AI.

6 Open models lag state-of-the-art closed models by 4 months – Epoch AI.

7 The Six-Lab Open-Weight Frontier Race.

8 Sovereign Cloud Stack and Gaia-X Federation Services.

9 Gartner, Inc.: Gartner Survey Reveals Geopolitics Will Drive 61% of CIOs and IT Leaders in Western Europe to Increase Reliance on Local Cloud Providers. Pressemitteilung, Barcelona, 12. November 2025. URL: https://www.gartner.com/en/newsroom/press-releases/2025-11-12-gartner-survey-reveals-geopolitics-will-drive-61-percent-of-cios-and-information-technology-leaders-in-western-europe-to-increase-reliance-on-local-cloud-providers

ÜBER DIESES PAPIER

Entscheidend ist weniger die gewählte Architektur als die Fähigkeit, sie zu wechseln.

Dieses Diskussionspapier dokumentiert den Debattenabend des KI-Stammtisch Nord am 24. August 2026 in Hamburg. Drei Delegationen – On-Premises, Cloud und Hybrid – haben ihre Positionen entlang von vier Spannungsfeldern vertreten. Die Argumente wurden unter der Chatham House Rule erhoben und ohne Namensnennung verarbeitet.

HERAUSGEBER

KI Bundesverband & KPMG

KONTAKT

Laura Bauer
laura.bauer@ki-verband.de