Mittelstand-Digital
Zentrum
Chemnitz

NACHGELESEN

Die Qual der Wahl – Ein Leitfaden zur Auswahl von Large Language Models (LLM) für KMU

ALEXANDER RAUSCHER

Gefördert durch:

Bundesministerium
für Wirtschaft
und Energie

aufgrund eines Beschlusses
des Deutschen Bundestages

Mittelstand-
Digital

Künstliche Intelligenz hat das Potenzial, die Effizienz im Mittelstand massiv zu steigern. Doch die Entscheidung für ein bestimmtes Large Language Model (LLM) ist für kleine und mittlere Unternehmen (KMU) weit mehr als eine technische Spielerei.

Es geht um Haftung, Datensicherheit und die langfristige Wirtschaftlichkeit. Angesichts strenger EU-Regularien und einer unübersichtlichen Anbieterlandschaft stehen Unternehmen vor der Frage: Welches Modell passt zu meinem Sicherheitsbedürfnis und meinem Geschäftsmodell?

Dieses Nachgelesen strukturiert die Auswahlkriterien in

- rechtliche,
- sicherheitsrelevante,
- technische und
- ökonomische Dimensionen,

um KMU eine fundierte Entscheidungsgrundlage zu bieten.

Impressum

HERAUSGEBER
Mittelstand-Digital Zentrum Chemnitz
c/o TU Chemnitz
Erfenschlager Str. 73, 09125 Chemnitz
Tel: 0371 531 19935 Fax: 0371 531 819935
info@digitalzentrum-chemnitz.de
www.digitalzentrum-chemnitz.de

REDAKTION Bianca Eichler

GESTALTUNG UND PRODUKTION
PUNKT191 – Marketing und Design
www.punkt191.de

BILDNACHWEIS TITEL generiert mi KI
VERÖFFENTLICHUNG Mai 2026

Die Qual der Wahl – Ein Leitfaden zur Auswahl von Large Language Models (LLM) für KMU

Rechtliche Compliance: AI Act & Datenschutz als Wegweiser

Die rechtliche Sicherheit ist das Fundament jeder KI-Strategie. Besonders im deutschen Mittelstand, der eng mit sensiblen Kunden- und Prozessdaten arbeitet, sind drei Aspekte entscheidend:

- **Risikoklassen des AI Act:** Der EU AI Act klassifiziert KI-Systeme nach ihrem Risikopotenzial. KMU müssen prüfen, ob ihre Anwendung als „hochriskant" eingestuft wird (z. B. bei der automatisierten Vorauswahl von Bewerbern). In solchen Fällen gelten strenge Transparenz- und Qualitätsvorgaben für das gewählte Modell.

- **DSGVO & Datenort:** Die Verarbeitung personenbezogener Daten (z.B. der Mitarbeitenden oder der Kundschaft) erfordert höchste Sorgfalt, insbesondere mit Blick darauf, wo die Daten verarbeitet werden: in der EU oder in einem Drittland wie den USA. Das Data Privacy Framework (DPF) ist ein seit 2023 geltendes Abkommen zwischen der EU und den USA. Es ermöglicht zertifizierten US-Anbietern, Daten nach europäischen Schutzstandards zu verarbeiten und macht deren Nutzung für KMU grundsätzlich rechtlich zulässig. Allerdings besteht ein Restrisiko: Wie seine Vorgänger könnte auch das DPF durch den Europäischen Gerichtshof gekippt werden. Europäische Anbieter wie Mistral oder Ionos unterliegen direkt der DSGVO und bieten damit eine unmittelbar rechtssichere Alternative.

- **Modelltraining und Speicherung:** Ein ausschlaggebendes Kriterium ist, ob der Anbieter Konversationen und hochgeladene Dokumente dauerhaft speichert, um das Modell weiter zu trainieren. Für Unternehmen ist es essenziell, Verträge zu wählen, bei denen das Modelltraining mit Unternehmensdaten explizit ausgeschlossen ist, um den Schutz von Betriebsgeheimnissen zu gewährleisten.

Allgemeine Sicherheit: Bereitstellungsmodelle im Vergleich

Die Wahl des Speicher- und Rechenortes bestimmt maßgeblich das Schutzniveau unternehmens-kritischer Daten. In der Praxis lassen sich dabei grundlegende Bereitstellungsszenarien für den Einsatz von LLM unterscheiden:

PUBLIC CLOUD (Z. B. CHATGPT, GEMINI)

Cloud-basierte KI-Modelle etwa von Anbietern wie OpenAI oder Google bieten derzeit die höchste Leistungsfähigkeit und sind ohne technischen Aufwand sofort nutzbar. Sie verarbeiten Text, Bilder und weitere Datenformate und eignen sich damit für ein breites Anwendungsspektrum.

Die Daten werden auf Servern des Anbieters verarbeitet. Bei den meisten Anbietern lässt sich jedoch vertraglich ausschließen, dass eingegebene Daten für das Training neuer Modelle verwendet werden. Das grundsätzliche Risiko einer externen Datenübertragung bleibt dennoch bestehen und sollte bei vertraulichen Inhalten berücksichtigt werden.

PRIVATE / ENTERPRISE CLOUD (Z. B. MISTRAL ÜBER IONOS, AZURE OPENAI, LANGDOCK)

Bei dieser Variante werden KI-Modelle innerhalb einer abgeschlossenen, vertraglich definierten Cloud-Umgebung betrieben. Die Daten verlassen dabei nicht den festgelegten Bereich und bleiben vollständig DSGVO-konform.

Anbieter wie Microsoft über Azure OpenAI ermöglichen es, leistungsstarke Modelle in einer privaten Umgebung zu nutzen mit klar geregelten Datenschutzbedingungen und ohne, dass Daten für das Modelltraining verwendet werden. Die Leistung ist im Vergleich zur öffentlichen Cloud gelegentlich etwas geringer, das rechtliche und sicherheitstechnische Risiko jedoch deutlich überschaubarer – insbesondere für KMU, die mit personenbezogenen oder vertraulichen Daten arbeiten.

LOKALER BETRIEB (ON-PREMISE / EDGE KI)

Bei dieser Variante wird das KI-Modell direkt auf den eigenen Servern oder Rechnern im Unternehmen installiert - ohne Cloud-Anbindung. Genutzt werden sogenannte Open-Weight-Modelle, also frei verfügbare Modelle, die kostenlos heruntergeladen und lokal betrieben werden können. Beispiele hierfür sind Modelle westlicher Herkunft wie Llama (Meta), Mistral (Frankreich), Gemma (Google) oder GPT-OSS (OpenAI), aber auch leistungsstarke Modelle aus China wie DeepSeek oder Qwen.

Der entscheidende Vorteil ist, dass keine Daten das Unternehmen verlassen. Alle Eingaben, ob Kundendaten, interne Dokumente oder vertrauliche Informationen, bleiben vollständig unter eigener Kontrolle.

Der Nachteil liegt in der Hardware, denn meist sind leistungsstarke GPUs erforderlich. Einfachere Aufgaben wie Textzusammenfassungen funktionieren gut. Bei komplexen Analysen stoßen lokal betriebene Modelle jedoch schneller an ihre Grenzen als Cloud-basierte Alternativen. On-Premise-KI ist die datenschutzsicherste Lösung, erfordert jedoch eine realistische Einschätzung der eigenen IT-Infrastruktur.

HYBRIDE KOLLABORATION (LOCAL-CLOUD-COLLABORATION)

Als vierte Möglichkeit bietet sich ein hybrider Ansatz an, der gezielt die Vorteile unterschiedlicher Betriebsmodelle kombiniert. Dabei verbleiben sensible Unternehmensdaten etwa aus ERP-Systemen, CAD-Daten oder internen Dokumenten innerhalb der eigenen IT-Infrastruktur, während leistungsstarke Sprachmodelle aus der Cloud für die eigentliche Verarbeitung und Generierung genutzt werden.

Der Markt entwickelt sich weg vom Prinzip „ein Modell, ein Anbieter“ hin zu flexiblen Architekturen, die je nach Aufgabe das passende Modell einsetzen: leistungsstarke Cloud-KI für komplexe Analysen, lokale Modelle für einfachere oder datenschutzkritische Aufgaben. Vertrauliche Informationen werden dabei ausschließlich auf eigener Hardware verarbeitet und verlassen das Unternehmen nicht.

Werkzeuge wie n8n oder Langdock übernehmen die Koordination und entscheiden automatisiert, welche Aufgabe wohin geleitet wird. So müssen KMU sich nicht dauerhaft an einen Anbieter binden und müssen nicht zwischen Leistung und Datenschutz wählen.

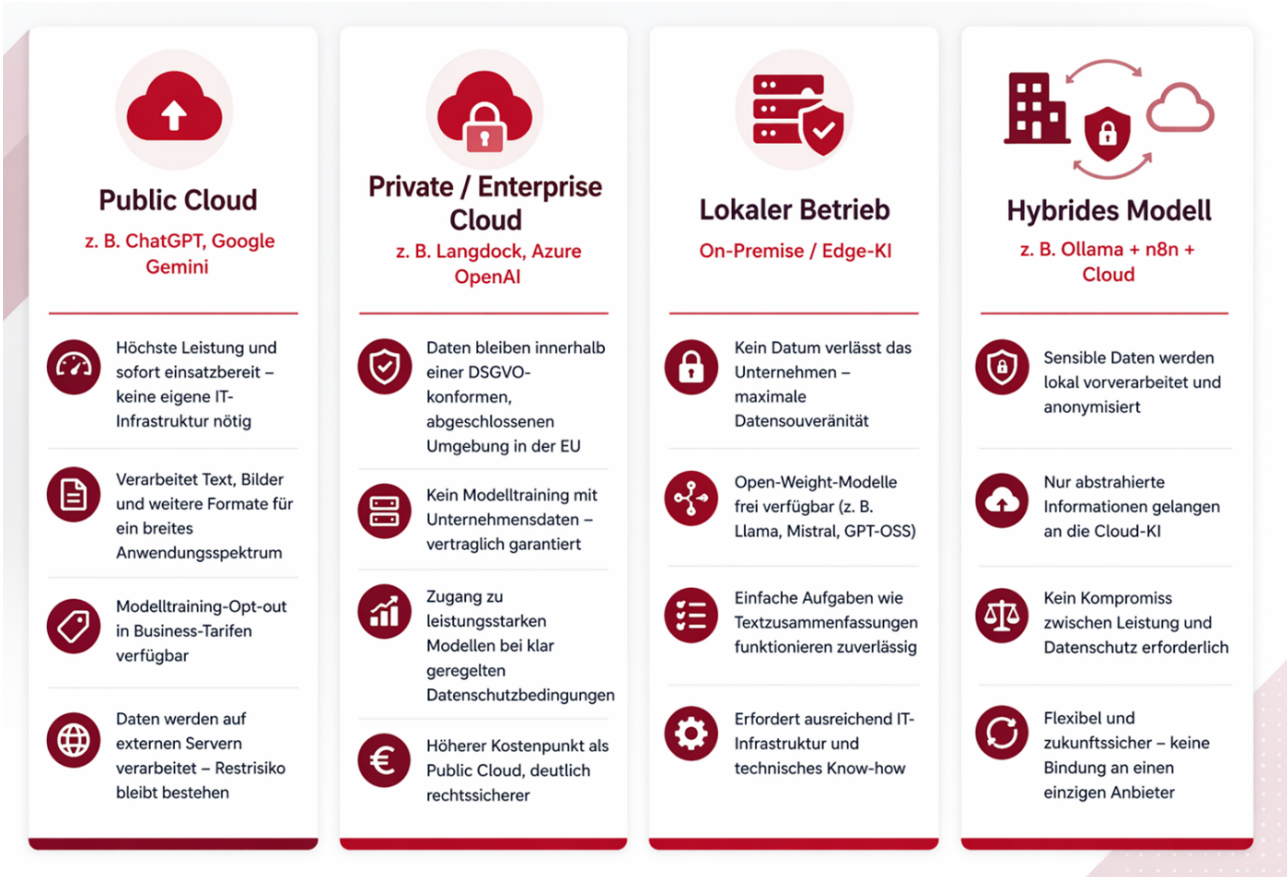

Abbildung 1: Möglichkeiten der LLM-Bereitstellung (erstellt mit KI)

Wirtschaftliche Kriterien: Kostenstrukturen im Blick

Die Einführung von KI muss ökonomisch nachhaltig sein. Dabei lohnt es sich, nicht nur den monatlichen Abo-Preis zu betrachten, sondern die Gesamtkosten über die gesamte Nutzungsdauer – Fachleute sprechen hier von der sogenannten „Total Cost of Ownership“ (TCO):

- **Variable vs. Fixe Kosten:** Viele Anbieter bieten monatliche Pauschalpreise pro Nutzer an, das schafft Planungssicherheit. Alternativ gibt es nutzungsbasierte Modelle, bei denen nur das abgerechnet wird, was tatsächlich verarbeitet wird. Das ist für gelegentliche Nutzung günstiger, erfordert jedoch ein konsequentes Monitoring: Werden große Datenmengen abgefragt, können die Kosten schnell und unerwartet steigen.

- **Infrastruktur-Investitionen:** Wer KI auf eigenen Servern betreibt, spart Lizenzkosten, trägt dafür aber einmalige Investitionskosten für Hardware sowie laufende Ausgaben für Strom, Kühlung und IT-Personal. KMU sollten vorab berechnen, ab welcher Nutzungsintensität sich ein eigener Server gegenüber Cloud-Gebühren amortisiert.

- **Vermeidung des Vendor Lock-in:** Wer seine Prozesse stark auf die Sonderfunktionen eines einzelnen Anbieters ausrichtet, verliert bei Preiserhöhungen den Handlungsspielraum. Modelle, die über standardisierte Schnittstellen ansprechbar sind, ermöglichen einen flexiblen Wechsel, etwa von einer US-Cloud zu einem lokalen Modell, falls sich rechtliche oder wirtschaftliche Rahmenbedingungen ändern.

- **Effizienz durch Model Routing:** Nicht jede Aufgabe erfordert das leistungsstärkste Modell. Einfache Routineaufgaben wie E-Mail-Zusammenfassungen lassen sich kostengünstig an kleinere, ressourcenschonende Modelle delegieren. Nur für komplexe Analysen wird auf leistungsstarke und entsprechend teurere Modelle zurückgegriffen. Diese Strategie kann die KI-Kosten eines Unternehmens erheblich reduzieren.

Funktionale & Technische Profile: Das richtige Werkzeug für den Job

Die Auswahl des geeigneten LLM stellt für viele KMU eine der zentralen praktischen Fragen dar und sie lässt sich nicht pauschal beantworten. Jedes Modell verfügt über ein spezifisches funktionales Profil, das sich aus Architektur, Trainingsschwerpunkt und Integration ergibt.

Die nachfolgende Übersicht ordnet die aktuell relevantesten öffentlich verfügbaren Modelle nach ihrem primären Anwendungsbereich ein. Sie dient als erste Orientierungshilfe und ersetzt keine unternehmensindividuelle Evaluierung, bildet jedoch eine fundierte Grundlage für eine informierte Entscheidung (siehe Tabelle 1).

Kein Tool ist universell überlegen, vielmehr hat sich die Spezialisierung weiter ausdifferenziert.

- ChatGPT bleibt der flexibelste Allrounder mit starkem Fokus auf Workflows und Integration.

- Claude hat sich deutlich in Richtung High-End-Analyse, Schreiben und Programmierung entwickelt und liefert hier oft die qualitativ besten Ergebnisse.

- Gemini ist längst mehr als ein Google-Tool und überzeugt insbesondere bei datengetriebenen Aufgaben und komplexem Reasoning.

- Perplexity ist die erste Wahl für Recherche mit nachvollziehbaren Quellen und entwickelt sich zunehmend zu einer Multi-Model-Plattform.

- Grok punktet weiterhin bei Echtzeitdaten und Trends, erweitert seinen Nutzen jedoch durch sehr große Kontextfenster und die Fähigkeit, umfangreiche Datenmengen zu verarbeiten.

- Mistral AI ergänzt dieses Feld als leistungsfähige und kosteneffiziente Alternative mit starkem API-Fokus. Besonders hervorzuheben ist die Möglichkeit der Datenverarbeitung innerhalb europäischer Infrastrukturen.

TOOL	AM BESTEN GEEIGNET FÜR	TYPISCHE ANWENDUNGSFÄLLE	STÄRKEN	HINWEIS
CHATGPT	Allrounder, kreative Inhalte, Programmierung, Workflows	Schreiben, Programmieren, Automatisierungen, Datenanalyse, multimodale Aufgaben	Sehr vielseitig; starke Multimodalität; Custom GPTs; Workflow-Automatisierung; gute Tool-Integration	Bleibt der flexibelste Allrounder mit starkem Ökosystem
GROK	Echtzeitdaten, große Kontexte, Trends	Social Media Analys, Trendbeobachtung, große Textmengen analysieren	Sehr großer Kontext (bis ~1M+ Tokens); Echtzeitdatenzugriff; schnelle Reaktionen	Stärker bei Datenvolumen als bei Struktur/Qualität
GEMINI	Datenanalyse, Google-Integration, komplexe Reasoning-Aufgaben	Arbeiten in Docs/Sheets; datengetriebene Analysen; Unternehmensworkflows	Sehr große Kontextfenster (bis ~1M+ Tokens); starke Integration; gute Performance bei Benchmarks	Besonders stark bei datengetriebenen Aufgaben und Teams
CLAUDE	Tiefenanalyse, Schreiben, Programmierung	Verträge, Codebasen, wissenschaftliche Texte, komplexe Inhalte	Sehr große Kontextfenster (bis ~1M Tokens); sehr natürlicher Schreibstil; starke Coding-Performance; Agenten-Fähigkeiten	Führend bei Textqualität und komplexen Analysen
PERPLEXITY	Recherche, Faktenprüfung, Multi-Model-Zugriff	Webrecherche, Vergleich von Quellen, schnelle Zusammenfassungen	Quellenbasiert; Zugriff auf mehrere Modelle; aktuelle Webdaten	Beste Wahl für nachvollziehbare Recherche
MISTRAL	Kosteneffiziente API-Nutzung, europäische Cloud-Alternative	Automatisierungen, RAG, Integration in Anwendungen	Sehr gute Preis-Leistung; schnelle Modelle; EU-Anbieter; API-first Ansatz; flexibel kombinierbar	Gute Alternative zu US-Anbietern, besonders für kosten- & daten-sensible Anwendungen

↑ **Tabelle 1:** Sprachmodelle im Vergleich

Fazit

Die Auswahl eines Large Language Models ist für KMU keine rein technische Entscheidung, sie ist eine strategische. Rechtliche Compliance, Datenschutz, wirtschaftliche Nachhaltigkeit und funktionale Eignung müssen gleichzeitig betrachtet und gegeneinander abgewogen werden.

Es gibt kein universell richtiges Modell. Was zählt, ist die Passung zum eigenen Geschäftsmodell, zur Sensibilität der verarbeiteten Daten und zur vorhandenen IT-Infrastruktur. Wer heute mit einem einfachen Cloud-Tool beginnt, sollte morgen in der Lage sein, auf ein datenschutzkonformeres oder kostengünstigeres Modell zu wechseln, ohne seine gesamten Prozesse neu aufzubauen. Flexibilität und Anbieterunabhängigkeit sind daher keine Komfortmerkmale, sondern unternehmerische Notwendigkeiten.

Der Markt entwickelt sich rasant. Modelle werden leistungsfähiger, europäische Alternativen reifen und hybride Architekturen machen es zunehmend möglich, höchste Rechenleistung mit vollständiger Datensouveränität zu verbinden. KMU, die jetzt eine durchdachte Grundlage schaffen - rechtlich abgesichert, wirtschaftlich kalkuliert und technisch flexibel - positionieren sich nicht nur für den aktuellen Stand der Technik, sondern für die nächsten Jahre.

Gleichzeitig gilt: Wer auf den perfekten Moment wartet, wartet zu lang. Der Einstieg muss nicht vollständig sein, er muss stattfinden. Ein erstes Experiment mit einem öffentlich verfügbaren Tool, eine konkrete Aufgabe, die damit erprobt wird, schafft mehr Erkenntnisgewinn als jede theoretische Analyse. Praktische Erfahrung ist in einem so dynamischen Feld der schnellste Weg zur richtigen Entscheidung.

Checkliste: LLM-Auswahl für KMU

Rechtliche Compliance

- ☑ Wird das Modell in der EU betrieben und liegt ein Auftragsverarbeitungsvertrag vor?
- ☑ Ist das Modelltraining mit eigenen Daten vertraglich ausgeschlossen?
- ☑ Fällt der Anwendungsfall unter eine Hochrisiko-Kategorie des EU AI Acts?
- ☑ Ist der Anbieter DPF-zertifiziert und ist das Restrisiko akzeptabel?

Datensicherheit & Bereitstellungsmodell

- ☑ Wie sensibel sind die Daten, die verarbeitet werden?
- ☑ Reicht eine Public Cloud, oder ist Private Cloud bzw. lokaler Betrieb erforderlich?
- ☑ Ist eine hybride Architektur sinnvoll, um Leistung und Datenschutz zu verbinden?
- ☑ Gibt es klare Regelungen zu Datenspeicherung und Aufbewahrungsdauer?

Wirtschaftlichkeit

- ☑ Passt das Kostenmodell zur Nutzungsintensität (Pauschalabo oder nutzungsbasiert)?
- ☑ Sind alle Folgekosten berücksichtigt - Hardware, Strom, IT-Personal?
- ☑ Besteht ein Vendor-Lock-in-Risiko durch proprietäre Funktionen?
- ☑ Können einfache Aufgaben an kostengünstigere Modelle delegiert werden?

Funktionale Eignung

- ☑ Ist das Modell für den konkreten Anwendungsfall nachweislich geeignet?
- ☑ Werden multimodale Fähigkeiten oder große Kontextfenster benötigt?
- ☑ Kann das Modell über APIs mit bestehender Software kommunizieren?

Zukunftssicherheit

- ☑ Ist ein Anbieterwechsel ohne großen Aufwand möglich?
- ☑ Unterstützt das Modell standardisierte Schnittstellen?
- ☑ Ist die Lösung skalierbar, wenn die Nutzung im Unternehmen wächst?

Verfasst von

ALEXANDER RAUSCHER ist wissenschaftlicher Mitarbeiter in der Abteilung Digitalisierung in der Produktion am Fraunhofer Institut für Werkzeugmaschinen und Umformtechnik in Chemnitz. Im Mittelstand-Digital Zentrum Chemnitz fokussiert er Themen rund um generative Künstliche Intelligenz, darunter Tools und Use Cases für den Einsatz im Unternehmen.
alexander.rauscher@digitalzentrum-chemnitz.de

Weitere Informationen

Das Mittelstand-Digital Zentrum Chemnitz gehört zu Mittelstand-Digital. Mit dem Mittelstand-Digital Netzwerk unterstützt das Bundesministerium für Wirtschaft und Energie die Digitalisierung in kleinen und mittleren Unternehmen und dem Handwerk.

WAS IST MITTELSTAND-DIGITAL?

Das Mittelstand-Digital Netzwerk bietet mit den *Mittelstand-Digital Zentren* und der Initiative *IT-Sicherheit in der Wirtschaft* umfassende Unterstützung bei der Digitalisierung mit dem Schwerpunkt Künstliche Intelligenz. Kleine und mittlere Unternehmen profitieren von konkreten Praxisbeispielen und passgenauen, anbieterneutralen Angeboten zur Qualifikation und IT-Sicherheit. Das Bundesministerium für Wirtschaft und Energie ermöglicht die kostenfreie Nutzung der Angebote von Mittelstand-Digital. Weitere Informationen finden Sie unter www.mittelstand-digital.de.

Mittelstand-Digital
Zentrum
Chemnitz

Gefördert durch:

Bundesministerium
für Wirtschaft
und Energie

Mittelstand-
Digital

aufgrund eines Beschlusses
des Deutschen Bundestages