

KUNDEN-CHEATSHEET · KEIN TECHNIK-HANDBUCH

LLMs für Entscheider

Was Sprachmodelle sind, wie sie geprägt werden,
und wo ihre blinden Flecken liegen.

12

Kapitel

15–20

Min. Lesedauer

4

Anbieter im Vergleich

8

Punkte Checkliste

Sie müssen kein Modell trainieren können, um Entscheidungen über den Einsatz von KI-Sprachmodellen zu treffen. Aber Sie sollten wissen, wo diese Systeme strukturell schwach sind – nicht weil sie schlecht gebaut wären, sondern weil bestimmte Schwächen aus der Funktionsweise selbst folgen und sich nicht wegtrainieren lassen.

KAPITEL 1

Was ein LLM eigentlich tut

Ein Sprachmodell sagt keine „Wahrheit“ voraus, sondern das statistisch wahrscheinlichste nächste Wort, gegeben alles, was vorher stand. Es wurde mit Billionen von Wörtern aus Texten, Büchern, Webseiten und Code trainiert und hat daraus Muster gelernt – Zusammenhänge zwischen Begriffen, Argumentationsstrukturen, Stilformen.

Wichtig für Entscheider: Es gibt keine eingebaute Instanz, die „weiß“, was wahr ist. Es gibt nur eine sehr gute Mustererkennung, die in den meisten Alltagsfällen mit der Wahrheit übereinstimmt – aber eben nicht immer, und ohne dass das Modell selbst den Unterschied zuverlässig markiert.

▲ Kernproblem

Ein LLM kann falsche Aussagen mit derselben sprachlichen Sicherheit formulieren wie richtige. Sprachliche Souveränität ist kein Indikator für inhaltliche Richtigkeit.

KAPITEL 2

Wie die Prägung entsteht — vier Phasen

01 Pretraining (Rohmaterial)

Das Modell liest riesige Mengen Text aus dem Internet, Büchern und Fachliteratur. Es lernt Sprachmuster, nicht kuratiertes Wissen. Was im Internet häufig vorkommt, prägt stärker – unabhängig davon, ob es richtig, repräsentativ oder seriös ist.

02 Das Ungleichgewicht in den Rohdaten

Die Textmenge zu einem Thema spiegelt nicht automatisch dessen wissenschaftliche oder gesellschaftliche Bedeutung wider. Kontroverse Themen erzeugen überproportional viel Text, weil Streit mehr Schreibenanlass bietet als Konsens.

03 Fine-Tuning / RLHF (Nachjustierung)

Nach dem Pretraining bewerten Menschen Modellantworten nach Kriterien wie Genauigkeit, Nützlichkeit, Sicherheit. Hier fließen explizite Wertentscheidungen der Entwicklerfirma ein – nicht neutrale, sondern bewusst gesetzte Kriterien.

04 System-Prompts und Verhaltensrichtlinien

Zusätzlich bekommt das Modell Verhaltensregeln mit: Wann soll es einen wissenschaftlichen Konsens klar benennen, wann mehrere Positionen nebeneinanderstellen? Diese Regeln werden von Menschen bei der Entwicklerfirma festgelegt.

KAPITEL 3

Die zentralen Schwächen

3.1 Halluzination

Modelle erfinden Fakten, Quellen, Zitate oder Zahlen, wenn sie eine Lücke im Gelernten mit einer plausibel klingenden Antwort füllen. Das geschieht nicht als „Fehler“, den das Modell bemerkt, sondern als normale Funktionsweise – es gibt keinen internen Wahrheitscheck.

- **Besonders riskant bei:** Zahlen, Rechtsquellen, Zitaten, aktuellen Ereignissen, Nischenthemen.
- **Gegenmittel:** Prüfpflicht bei allem, was in eine Entscheidung, Publikation oder einen Vertrag einfließt.

3.2 Fehlende Introspektion

Ein Modell kann nicht zuverlässig sagen, warum es etwas geschrieben hat oder welche Quellen es besonders geprägt haben. Erklärungen zu seinem eigenen Verhalten sind selbst wieder generierter Text – plausibel klingend, aber nicht notwendig zutreffend.

3.3 Verzerrung durch Trainingsdaten (Bias)

Weil das Internet nicht neutral ist, sind es die Trainingsdaten auch nicht:

- **Sprachliche Dominanz:** Englischsprachige, US-/UK-geprägte Inhalte sind überrepräsentiert.
- **Zeitliche Schieflage:** Ältere, digitalisierte Texte sind übervertreten gegenüber aktuellen Wissensbeständen.
- **Soziale Schieflage:** Wer im Internet viel schreibt, prägt die Trainingsdaten überproportional.
- **Stereotype:** Historisch gewachsene Vorurteile können reproduziert werden, wenn nicht gezielt gegengesteuert wird.

3.4 Künstliche Ausgewogenheit ("False Balance")

Modelle sind häufig darauf trainiert, bei kontroversen Themen mehrere Perspektiven darzustellen. Bei Themen mit klarem Sachstand kann „Ausgewogenheit“ den Konsens künstlich relativieren.

3.5 Regulatorische und unternehmenspolitische Filter

Jeder Anbieter unterliegt rechtlichen, kulturellen und geschäftlichen Zwängen, die sich im Modellverhalten niederschlagen – von impliziter Vorsicht bis zu expliziter Themenvermeidung. Das ist keine Ausnahme, sondern die Regel.

3.6 Nicht-Aktualität

Ein Modell kennt nur, was bis zu einem bestimmten Stichtag in seine Trainingsdaten eingeflossen ist. Ohne aktive Internetsuche sind Informationen zu aktuellen Ereignissen unzuverlässig oder veraltet.

3.7 Selbstbewusstes Auftreten trotz Unsicherheit

Modelle sind sprachlich darauf trainiert, flüssig und kohärent zu klingen – auch wenn die zugrunde liegende Faktenlage dünn ist. Diese rhetorische Sicherheit ist keine verlässliche Kalibrierung der tatsächlichen Gewissheit.

KAPITEL 4

Anbietervergleich

Jedes Modell trägt Spuren seiner Entstehung. Die folgende Übersicht ordnet grobe, öffentlich beobachtbare Tendenzen ein – keine exakte Messung, da Innenansichten fehlen.

Anbieter	Strukturelle Prägung	Beobachtbare Tendenz
Anthropic (Claude)	Gegründet mit explizitem „AI-Safety“-Fokus	Vorsichtig, differenzierend, teils überabgesichert; Tendenz zu Multi-Perspektiven-Antworten auch dort, wo eine klare Antwort möglich wäre
OpenAI (GPT)	Breite kommerzielle Nutzerbasis, Consumer- bis Enterprise-Kontext	Optimiert auf breite Akzeptanz; historisch Tendenz zu übermäßiger Bestätigung gegenüber Nutzeraussagen („Sykophanz“)
Meta (Llama)	Offene Modellgewichte, Weiterverwendung durch Dritte	Weniger einheitlich abgesichertes Grundverhalten, da Feinschliff oft bei nachgelagerten Anwendern liegt
DeepSeek u. a.	Regulatorisches Umfeld mit expliziten Inhaltsvorgaben	Nachweisbare Vermeidung bestimmter politischer Themen (z. B. Tiananmen 1989, Taiwan-Status); bei technischen Themen oft kaum Unterschied zu westlichen Modellen

Wichtige Einordnung

Es gibt keinen neutralen Referenzpunkt, von dem aus sich „Bias-Freiheit“ messen ließe. Jeder Anbieter hat eine Schiefelage – nur die Art und die Sichtbarkeit unterscheiden sich. Bei westlichen Anbietern ist die Schiefelage meist diffus und schwer belegbar. Bei regulatorisch stark kontrollierten Anbietern ist sie oft konzentrierter und leichter nachweisbar.

Praktische Konsequenzen

- **Faktenprüfung institutionalisieren:** LLM-Output wie den Entwurf eines nicht immer zuverlässigen Praktikanten behandeln – brauchbar als Ausgangspunkt, nicht als Endprodukt.
- **Modellwahl bewusst treffen:** Je nach Anwendungsfall unterschiedliche Modelle testen; bei sensiblen Anwendungen die Herkunft des Anbieters mitdenken.
- **Keine Illusion von Objektivität:** „Das Modell sagt..." ist keine neutrale Autorität, sondern das Ergebnis vieler menschlicher Entscheidungen.
- **Aktualität prüfen:** Bei zeitkritischen Themen aktiv nachfragen, ob das Modell Internetzugriff hat.
- **Vier-Augen-Prinzip bei Grenzfällen:** Bei ethisch, rechtlich oder politisch heiklen Fragen LLM-Antworten als einen Input unter mehreren behandeln.
- **Kritisch bei zu glatten Antworten sein:** Wirkt eine Antwort ungewöhnlich entschieden oder ausgewogen, lohnt der Blick, ob das der Sachlage oder dem Trainingsprogramm entspricht.

▲ Merksatz

Ein Sprachmodell ist ein sehr kompetenter, sehr selbstsicher klingender Assistent ohne eigenes Urteilsvermögen über Wahrheit – geprägt von den Daten, mit denen es lernte, und den Menschen, die entschieden haben, wie es sich verhalten soll. Nachfragen und Gegenprüfen sind nicht optional, sondern fester Bestandteil der Nutzung.

KAPITEL 6

Praxisbeispiele — Risiko nach Komplexität

Nicht jede Anwendung ist gleich riskant. Faustregel: Risiko steigt mit (a) der Tragweite der Entscheidung, (b) der Schwierigkeit, einen Fehler zu erkennen, und (c) der Sensibilität der eingegebenen Daten.

● Geringes Risiko

Breiter Einsatz sinnvoll

- E-Mails, Memos, Formulierungshilfen, Übersetzungen
- Brainstorming, Gliederungen, Strukturierung
- Code-Unterstützung mit Developer-Review
- Zusammenfassen eigener, geprüfter Dokumente

Ein Mensch liest und verantwortet das Ergebnis, bevor es wirkt.

● Mittleres Risiko

Prüfpflicht einführen

- Recherche zu Fakten, Studien, Rechtslage
- Entwürfe von Vertragsklauseln, Stellenanzeigen
- Datenauswertung zur Vorbereitung
- Chatbots mit engem, vordefiniertem Themenbereich

Fehler sind nicht auf den ersten Blick erkennbar. Vier-Augen-Prinzip verpflichtend.

● Hohes Risiko

Nur mit strengen Leitplanken

- Automatisierte Personalentscheidungen ohne menschliche Prüfung
- Medizinische, rechtliche, steuerliche Auskünfte ungefiltert
- Personaldaten in öffentliche KI-Tools ohne AVV
- Vollautomatische Veröffentlichung ohne Prüfung

KI liefert Vorschläge – Menschen bleiben verantwortlich, auch haftungsrechtlich.

KAPITEL 7

Kosten — grobe Einordnung

☁ Cloud-Abo

Ca. 20–30 € / Nutzer / Monat.
Kein technischer Aufwand, sofort startklar. Daten verlassen das Haus.

⚡ API-Nutzung

Abrechnung nach Tokens.
Sinnvoll für eigene Workflows.
Kostenrisiko bei unkontrolliertem Volumen.

🖥 Self-Hosting

Für Datenschutz / Offline-Betrieb.
Leistungsfähige GPUs nötig. Kein Klick-Setup – ein Projekt.

Faustregel: Self-Hosting lohnt sich bei großer Nutzungsmenge oder wenn Daten das Haus aus rechtlichen Gründen nicht verlassen dürfen. Für die meisten KMU ist ein Cloud-Anbieter mit belastbarem AVV der pragmatischere Weg.

KAPITEL 8 & 9

Datenschutz & Regulierung

▲ Grundregel

Alles, was Sie in ein KI-System eingeben, betrachten Sie so, als würden Sie es einem externen Dienstleister übergeben – weil das technisch und rechtlich meist genau das ist.

- **Personenbezogene Daten (DSGVO):** Ohne Auftragsverarbeitungsvertrag (AVV) dürfen in der Regel keine personenbezogenen Daten an den KI-Anbieter gehen.
- **Business-Tarife:** Bieten meist AVV sowie Zusagen, dass Eingaben nicht zum Training genutzt werden – aktiv prüfen, nicht annehmen.
- **Serverstandort:** Verarbeitung außerhalb EU/EWR kann zusätzliche Anforderungen auslösen.
- **Art. 22 DSGVO:** Vollautomatisierte Entscheidungen über Personen sind eingeschränkt – menschliche Letztentscheidung bei Personalprozessen in der Regel Pflicht.
- **EU AI Act:** Stuft Anwendungen nach Risiko ein; Personalauswahl und Leistungsbeurteilung gelten als Hochrisiko mit zusätzlichen Dokumentationspflichten.
- **BetrVG § 87:** Mitbestimmung des Betriebsrats bei KI zur Leistungs-/Verhaltenskontrolle – frühzeitig einbinden.

▲ Kein Ersatz für Rechtsberatung

Bei konkreten Anwendungsfällen – besonders Personalprozessen, Kundendaten oder öffentlicher Kommunikation – vor Einführung Datenschutz- und Fachrecht prüfen.

Checkliste vor Einführung eines KI-Tools

- ☐ Welche Daten werden eingegeben – dürfen diese laut DSGVO an diesen Anbieter gehen?
- ☐ Gibt es einen AVV? Wo werden Daten verarbeitet und gespeichert?
- ☐ Werden Eingaben zum Training genutzt – lässt sich das abschalten?
- ☐ Wie schwerwiegend ist ein Fehler im schlimmsten Fall – wer prüft, bevor es wirkt?
- ☐ Ist eine menschliche Letztentscheidung vorgesehen, wo Personen betroffen sind?
- ☐ Ist der Betriebsrat einzubinden?
- ☐ Ist die Anwendung nach EU AI Act Hochrisiko?
- ☐ Cloud, API oder Self-Hosting – passt die Wahl zu Datenschutz, Nutzungsmenge und Fachpersonal?

Haben Sie Fragen zur KI-Einführung in Ihrem Unternehmen? Wir beraten Sie
herstellerneutral – von der Datenschutzprüfung bis zur technischen Umsetzung.

www.sys-cologne.de
0221 / 165399300

Dieses Dokument wurde mit KI-Unterstützung (Claude, Anthropic) erarbeitet und redaktionell zusammengestellt – passend zum Thema, das es behandelt. Die rechtlichen Ausführungen sind eine allgemeine Einordnung und ersetzen keine Rechtsberatung im Einzelfall. · Sys-Cologne IT-Services e.K. · © 2026