

WHITEPAPER

LLMs für Executives

Wie große Sprachmodelle funktionieren – das Technik-Briefing für Geschäftsführung, Vorstand und Aufsichtsrat

Nicholas Krimmel

Juli 2026



Inhaltsverzeichnis

1	Executive Summary: Warum Sie dieses Briefing lesen sollten	2
1.1	Die Wissenslücke in den Gremien	2
1.2	Drei Gründe, warum LLM-Verständnis Chefsache ist	2
1.3	Was dieses Briefing leistet – und was nicht	2
2	Das Kernprinzip: eine Maschine, die das nächste Token schätzt	3
2.1	Was ein Sprachmodell tut – und was nicht	3
2.2	Tokens: die Betriebswährung der Sprachmodelle	4
2.3	Ziehen statt Nachschlagen: warum Antworten variieren	4
2.4	Warum „bessere Autovervollständigung“ die Sache unterschätzt	4
3	Der Blick ins Innere: Bedeutung als Geometrie, Attention als Mechanik	5
3.1	Embeddings: Bedeutung wird Geometrie	5
3.2	Self-Attention: der Mechanismus, der skaliert	5
3.3	Gewichte und Kontextfenster: das wichtigste Bild für die Führungsebene	6
4	Wie ein LLM entsteht: drei Phasen	7
4.1	Phase 1: Pretraining – Verdichtung von Billionen Tokens	7
4.2	Phase 2: Post-Training – aus dem Rohmodell wird ein Assistent	8
4.3	Phase 3: Der laufende Betrieb – Aktualität durch Anbindung	8
5	Warum die Modelle so schnell besser werden: drei Skalierungshebel	8
5.1	Hebel 1: Skalierung des Trainings – Fortschritt wurde planbar	9
5.2	Hebel 2: Qualität des Post-Trainings	9
5.3	Hebel 3: Rechenzeit pro Antwort – die Reasoning-Wende	9
5.4	Die ökonomische Folge: Preise fallen schneller als Vorlagen altern	10
6	Halluzinationen: warum sie entstehen – und warum das eine Prozessfrage ist	10
6.1	Erklärung 1: Die Anreiz-Ökonomie des Ratens	10
6.2	Erklärung 2: Die Fehlzündung im Schaltkreis	11
6.3	Die Konsequenz: Prozesse statt Produktversprechen	11
7	Die Fähigkeitslandkarte: was LLMs können – und was nicht	12
7.1	Die gezackte Grenze	12
7.2	Drei Studien, ein Muster: der Aufgabentyp entscheidet	12
7.3	Stark, schwach – und was das für Sie heißt	13
7.4	Der Faithfulness-Vorbehalt: Begründungen sind nicht der Rechenweg	13
8	Vom Modell zum Unternehmenswerkzeug: der Stack und die zehn Vorstandsfragen	13
8.1	Vier Schichten über dem Rohmodell	13
8.2	Die zehn Fragen für die nächste KI-Vorlage	14
9	Fazit und Glossar	15
9.1	Glossar: 14 Begriffe in Alltagssprache	17
	Literatur	19

1 Executive Summary: Warum Sie dieses Briefing lesen sollten

Sie müssen kein neuronales Netz trainieren können, um ein Unternehmen zu führen. Aber Sie entscheiden über KI-Budgets, geben KI-Vorlagen frei, verantworten KI-Risiken – und seit Kurzem knüpft der Gesetzgeber daran auch persönliche Pflichten. Dieses Briefing vermittelt das technische Grundverständnis großer Sprachmodelle (Large Language Models, LLMs), mit dem Sie in der nächsten Vorstands- oder Aufsichtsratssitzung die richtigen Fragen stellen – ohne Formeln, ohne Code, aber auch ohne Plattitüden.

1.1 Die Wissenslücke in den Gremien

Wie groß der Nachholbedarf ist, zeigen die Befragungen des Deloitte Global Boardroom Program – eine Beratererhebung, keine akademische Studie, aber die breiteste verfügbare Stichprobe zur KI-Reife von Gremien. In der ersten Welle (2024, 468 Board-Mitglieder und Führungskräfte aus 57 Ländern) stand KI bei 45 % der Boards noch gar nicht auf der Agenda, und 79 % der Befragten attestierten sich selbst begrenzte oder keine KI-Kenntnisse – nur 2 % stuften sich als sehr sachkundig ein [1]. Die zweite Welle (2025, 695 Befragte aus 56 Ländern) zeigt eine laufende Aufholjagd: noch 31 % ohne KI auf der Agenda, noch 66 % mit begrenzter oder keiner Kenntnis; zugleich denken 40 % der Gremien inzwischen über ihre eigene Zusammensetzung nach [2]. Die Richtung stimmt – die Wissenslücke bleibt.

1.2 Drei Gründe, warum LLM-Verständnis Chefsache ist

Erstens die Rechtspflicht. Artikel 4 der europäischen KI-Verordnung (Verordnung (EU) 2024/1689) gilt seit dem 2. Februar 2025 und verpflichtet Anbieter wie Betreiber von KI-Systemen, ein „ausreichendes Maß an KI-Kompetenz“ ihres Personals und der in ihrem Auftrag tätigen Personen sicherzustellen [3]. Die FAQ der Europäischen Kommission nennt die Leitungsebene nicht ausdrücklich [4]; nach Sinn und Zweck der Vorschrift schließt sie die Entscheidungsebene jedoch ein – wer über den KI-Einsatz entscheidet und ihn verantwortet, braucht das Verständnis dafür. Ab dem 2. August 2026 greifen zudem die nationalen Aufsichtsstrukturen.

Zweitens die Organhaftung. Mit dem am 6. Dezember 2025 in Kraft getretenen NIS-2-Umsetzungsgesetz haften Geschäftsleitungen betroffener Unternehmen persönlich dafür, die Risikomanagementmaßnahmen zu billigen und ihre Umsetzung zu überwachen; hinzu kommt eine ausdrückliche Schulungspflicht (§ 38 BSIG n. F.) [5]. KI-Systeme mit Zugriff auf Unternehmensdaten und -prozesse gehören zu genau dieser Risikolandschaft.

Drittens die Wirksamkeit. In der McKinsey-Befragung vom März 2025 (1 491 Unternehmen, Selbstauskunft) ist die aktive Aufsicht des CEO über die KI-Governance derjenige Einzelfaktor, der am stärksten mit einer messbaren Ergebniswirkung korreliert [6] – Korrelation, kein Kausalbeweis, aber ein deutlicher Fingerzeig.

1.3 Was dieses Briefing leistet – und was nicht

Die folgenden Kapitel bauen ein mentales Modell auf: von der Mechanik (Kapitel 2 bis 4) über Fortschritt, Fehler und Fähigkeitsgrenzen (Kapitel 5 bis 7) bis zum Unternehmenswerkzeug samt den zehn Vorstandsfragen (Kapitel 8). Jeder Fachbegriff wird beim ersten Auftreten erklärt; Kapitel 9 bündelt die wichtigsten im Glossar. Dieses Whitepaper schafft das technische Fundament, um über große Sprachmodelle auf Augenhöhe zu entscheiden. Was dieses Briefing nicht ist: Modellbau-Anleitung, Produktvergleich oder Rechtsberatung.

Die 5 wichtigsten Botschaften

- **Ein LLM ist keine Datenbank, sondern eine Wahrscheinlichkeitsmaschine:** Es berechnet Wortstück für Wortstück (Token für Token) die wahrscheinlichste Fortsetzung. Daraus folgen seine Stärken – und seine Fehlermuster.
- **Zwei Speicher, zwei Logiken:** Das „Weltwissen“ liegt eingefroren in den trainierten Gewichten; aktuell und firmenspezifisch wird ein LLM nur über das Kontextfenster. Deshalb sind Datenfundament und Retrieval strategische Themen, keine technischen Details.
- **Der Fortschritt hat drei planbare Hebel:** mehr Training, besseres Verhaltenstraining, mehr Rechenzeit pro Antwort. Fähigkeiten und Kosten verschieben sich im Quartalsrhythmus – einmalige Bewertungen veralten.
- **Halluzinationen sind eine Systemeigenschaft, keine Kinderkrankheit:** Die Antwort sind Prozesse – Quellenbindung, menschliche Freigabe, laufende Messung –, nicht das Warten auf das fehlerfreie Modell.
- **KI-Kompetenz ist Rechtspflicht und Führungsaufgabe:** Seit Februar 2025 europarechtlich gefordert, nach Sinn und Zweck bis in die Entscheidungsebene – und empirisch korreliert Führungsaufsicht am stärksten mit der Ergebniswirkung.

2 Das Kernprinzip: eine Maschine, die das nächste Token schätzt

2.1 Was ein Sprachmodell tut – und was nicht

Der Kern lässt sich in einem Satz sagen: Ein großes Sprachmodell erhält den bisherigen Text und berechnet daraus für jedes mögliche nächste **Token** – ein Wortstück, dazu gleich mehr – eine Wahrscheinlichkeit. Dann wird ein Token gemäß dieser Verteilung ausgewählt, an den Text angehängt, und die Berechnung beginnt von vorn. Aus dieser Schleife entsteht jede Chat-Antwort und jeder generierte Vertragsentwurf: Token für Token.

Das Modell schlägt dabei nichts in einer Datenbank nach; es gibt keinen Ort, an dem „die richtige Antwort“ gespeichert wäre – nur eine gigantische, trainierte Rechenvorschrift, die aus dem bisherigen Text die Verteilung für die Fortsetzung erzeugt. Wenn im Folgenden trotzdem Worte wie „wissen“ oder „verstehen“ fallen, stehen sie bewusst in Anführungszeichen: Gemeint ist immer statistisch verdichtetes Verhalten, kein Bewusstsein.

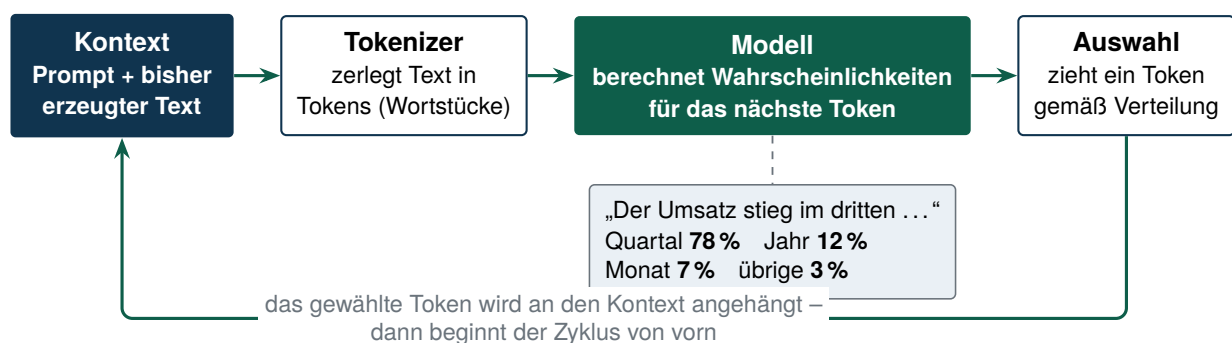


Abbildung 1 Die Erzeugungsschleife eines Sprachmodells (schematisch; die Beispiel-Wahrscheinlichkeiten sind illustrativ). Jede noch so lange Antwort entsteht Token für Token aus dieser Schleife.

2.2 Tokens: die Betriebswährung der Sprachmodelle

Sprachmodelle verarbeiten Text nicht als Buchstaben und nicht als ganze Wörter, sondern als **Tokens**: Zeichenketten-Bausteine, die häufige Wörter am Stück und seltene Wörter in Teilstücken abbilden. Der heute übliche Ansatz dafür, die Byte-Pair-Kodierung (BPE), stammt aus der maschinellen Übersetzung: Er baut das Vokabular aus den häufigsten Zeichenfolgen der Trainingsdaten zusammen [7]. Als Faustregel nennt OpenAI für englischen Text rund 4 Zeichen bzw. etwa 0,75 Wörter pro Token [8]; Anthropic gibt für seine Modelle etwa 3,5 englische Zeichen pro Token an [9]. Beide Regeln gelten mit Vorbehalt: Sie sind sprach- und tokenizerabhängig – deutsche Texte benötigen wegen Komposita und Flexionen spürbar mehr Tokens pro Wort, und jede neue Tokenizer-Generation verschiebt die Umrechnung. An Tokens hängen drei praktische Konsequenzen:

■ Drei Token-Konsequenzen für die Praxis

- **Abrechnung:** API-Preise werden pro Token berechnet (üblich: je Million Tokens, Ein- und Ausgabe getrennt). Gleicher Inhalt kostet auf Deutsch mehr Tokens als auf Englisch – US-Kostenschätzungen sind nicht eins zu eins übertragbar.
- **Kapazitätsgrenzen:** Das Kontextfenster (Kapitel 3) wird in Tokens gemessen. Ob ein Vertragskonvolut „hineinpasst“, ist eine Token-, keine Seitenfrage.
- **Blinde Flecken:** Das Modell „sieht“ Tokens, keine Buchstaben. Scheinbar triviale Aufgaben wie Buchstabenzahlen können deshalb fehlschlagen, während anspruchsvolle Textarbeit gelingt.

2.3 Ziehen statt Nachschlagen: warum Antworten variieren

Das Modell liefert eine Wahrscheinlichkeitsverteilung – welche Fortsetzung tatsächlich gewählt wird, entscheidet ein Zufallszug (**Sampling**). Ein Regler namens **Temperatur** steuert, wie stark dabei vom Wahrscheinlichsten abgewichen wird: nahe null wird die Ausgabe gleichförmiger und vorhersagbarer, höher wird sie variantenreicher. Dass dieselbe Frage zweimal gestellt zwei verschieden formulierte Antworten liefert, ist also kein Defekt, sondern eingebauter Mechanismus – und bei kreativen Aufgaben (Varianten, Entwürfe, Ideen) ausdrücklich erwünscht.

Governance-relevant wird das bei Prüfung und Abnahme: Identische Wiederholbarkeit ist nicht der Normalfall eines Sprachmodells – selbst bei Temperatur null verbleiben technische Quellen kleiner Abweichungen. Wer beim doppelten Testlauf zwei unterschiedlich formulierte, aber inhaltlich korrekte Antworten erhält, sieht kein „instabiles System“, sondern den vorgesehenen Mechanismus. Abnahmetests müssen deshalb gegen inhaltliche Kriterien prüfen, nicht auf Wortgleichheit – und produktive Systeme brauchen Protokollierung von Ein- und Ausgaben.

2.4 Warum „bessere Autovervollständigung“ die Sache unterschätzt

Die gängige Analogie – „das ist doch nur eine große Autovervollständigung“ – beschreibt die Mechanik korrekt und verfehlt trotzdem den Punkt. Denn wer das nächste Token über beliebige Texte hinweg *gut* vorhersagen will, muss enorme Regelmäßigkeiten von Sprache und Welt verdichten: Grammatik und Stilebenen, Fachterminologie, typische Vertragslogik, Ursache-Wirkungs-Muster. Genau deshalb ist die simple Mechanik so mächtig.

Den Beleg lieferte 2020 das Modell GPT-3 mit seinen 175 Milliarden Parametern – bis heute die einzige offiziell publizierte Größenangabe eines Spitzenmodells: Allein durch Skalierung

entstand die Fähigkeit, neue Aufgaben aus wenigen Beispielen im Prompt zu übernehmen, ohne dass ein einziges Gewicht verändert wurde – sogenanntes In-Context-Learning [10]. Diese Eigenschaft macht Sprachmodelle im Unternehmen so vielseitig – und erklärt, warum gutes Anleiten (Prompting) einen messbaren Unterschied macht.

3

Der Blick ins Innere: Bedeutung als Geometrie, Attention als Mechanik

Zwischen „Text kommt rein“ und „Wahrscheinlichkeiten kommen raus“ liegt die eigentliche Maschinerie. Zwei Konzepte genügen, um sie auf Führungsebene zu verstehen: Embeddings und Attention.



Abbildung 2 Die Verarbeitungsstrecke im Inneren (vereinfacht): Vom Text über Tokens und Embeddings durch viele Attention-Schichten zur Wahrscheinlichkeitsverteilung.

3.1 Embeddings: Bedeutung wird Geometrie

Am Eingang des Modells wird jedes Token in einen **Embedding-Vektor** übersetzt – eine lange Zahlenliste, die man sich als Punkt in einem Raum mit sehr vielen Dimensionen vorstellen kann. Die entscheidende Eigenschaft dieses Raums: Nachbarschaft bedeutet Bedeutungsverwandtschaft. „Rechnung“, „Faktura“ und „Invoice“ liegen beieinander; „Gänseblümchen“ liegt weit entfernt.

Dass dies keine schiefe Metapher ist, sondern messbare Realität, zeigte Anthropic im Mai 2024: Mit einem Analyseverfahren (Dictionary Learning) identifizierte das Unternehmen in seinem Produktionsmodell Claude 3 Sonnet Millionen interpretierbarer Merkmale („Features“) – Aktivierungsmuster, die für Konzepte wie die Golden Gate Bridge, Sicherheitslücken in Code oder Verzerrungen stehen; ähnliche Konzepte liegen geometrisch benachbart [11]. Der Kausaltest machte das Ergebnis berühmt: Verstärkten die Forscher das Golden-Gate-Feature künstlich, baute das Modell die Brücke in nahezu jede Antwort ein – der Eingriff in die Geometrie ändert das Verhalten kausal. Zur Einordnung: Das ist Forschung des Anbieters selbst, nach eigener Darstellung der erste detaillierte Blick in ein modernes Produktionsmodell; Methode und Interventionsbelege sind aber offen publiziert.

3.2 Self-Attention: der Mechanismus, der skaliert

Das Herzstück der heutigen Architektur – des **Transformers** – ist die **Self-Attention**: Beim Verarbeiten jedes Tokens berechnet das Modell, wie stark jedes andere Token im Kontext für diesen Schritt gewichtet werden soll. Ein Beispiel aus dem Geschäftsalltag: „Der Vertrag, den die Tochtergesellschaft nach monatelanger Verhandlung im März unterzeichnete, er läuft im Dezember aus.“ Damit das Modell beim Wort „er“ richtig weiterrechnet, muss es den Bezug zum weit entfernten „Vertrag“ herstellen – nicht zu „Tochtergesellschaft“ oder „März“. Genau solche Gewichtungen lernt die Attention, über beliebige Distanzen im Text.

Vorgestellt wurde der Transformer 2017 im Google-Paper „Attention Is All You Need“: eine Architektur, die allein auf Attention setzt und auf die zuvor üblichen rekurrenten Netze vollständig verzichtet [12]. Der zweite – ökonomisch entscheidende – Vorteil: Weil die Tokens nicht mehr nacheinander durchgeschleust werden müssen, lässt sich die Verarbeitung massiv parallelisieren – ideal für Grafikprozessoren (GPUs) und der Grund, warum sich diese Architektur auf immer größere Daten- und Rechenbudgets skalieren ließ. Das Paper zählt über 180 000 Zitationen (Semantic Scholar, Stand Juli 2026) [12] und gehört laut einer Nature-Analyse von 2025 zu den zehn meistzitierten wissenschaftlichen Arbeiten des 21. Jahrhunderts [13].

Im Inneren eines LLM sind Dutzende solcher Attention-Schichten hintereinandergeschaltet; jede verfeinert die Zwischendarstellung weiter. Die **Parameter** – die im Training justierten Zahlen, auch **Gewichte** genannt – gehen bei Spitzenmodellen in die Milliarden. Wie viele es heute genau sind, ist öffentlich nicht bekannt: Außer der GPT-3-Angabe von 175 Milliarden [10] haben die Anbieter für GPT-4/GPT-5, Claude oder Gemini keine offiziellen Parameterzahlen publiziert; kursierende Zahlen sind Spekulation oder Leaks. Für Ihre Entscheidungen ist das verschmerzbar: Relevant sind gemessene Leistung und Kosten im eigenen Anwendungsfall, nicht die Rohgröße (Kapitel 4).

3.3 Gewichte und Kontextfenster: das wichtigste Bild für die Führungsebene

Wenn Sie nur ein Bild aus diesem Kapitel mitnehmen, dann dieses: Ein LLM hat zwei grundverschiedene „Speicher“.

- **Die Gewichte** sind das eingefrorene Ergebnis des Trainings: Milliarden Zahlen, in denen die Regelmäßigkeiten der Trainingsdaten verdichtet sind. Sie ändern sich im laufenden Betrieb nicht; ihr Wissensstand endet am Trainingsstichtag (**Wissens-Cutoff**).
- **Das Kontextfenster** ist das Arbeitsgedächtnis der laufenden Anfrage: die Frage, der Gesprächsverlauf, mitgelieferte Dokumente und Daten. Es ist flüchtig – nach der Sitzung ist es leer – und der einzige Weg, auf dem Aktualität und Firmenwissen in die Antwort gelangen.

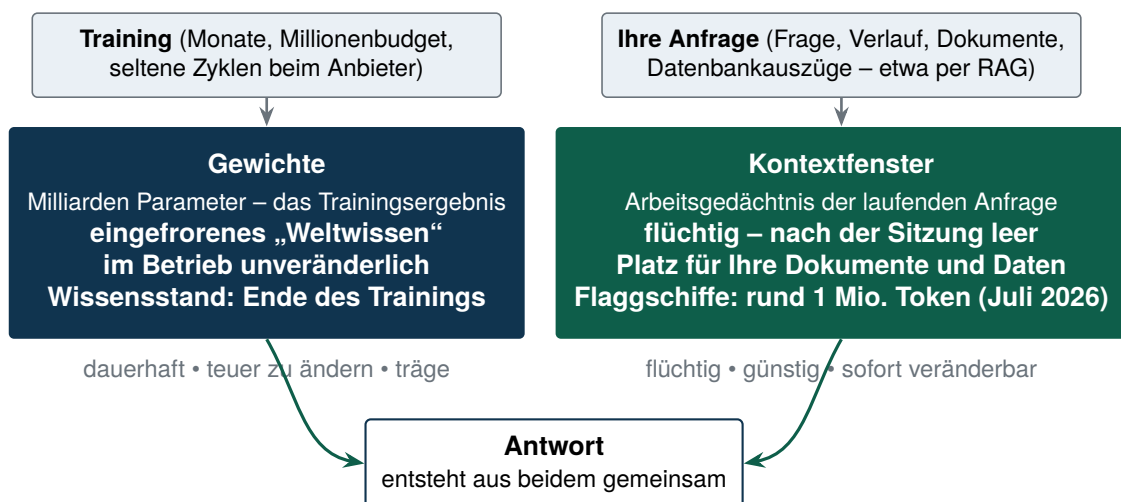


Abbildung 3 Zwei Speicher, zwei Logiken: eingefrorene Gewichte und flüchtiges Kontextfenster (Kontextgrößen: Stand Juli 2026 [15]).

Die Kontextfenster sind in kurzer Zeit dramatisch gewachsen: Stand Juli 2026 fassen die Flaggschiff-Modelle aller drei großen Anbieter – OpenAI, Anthropic und Google – jeweils rund 1 Million Token; kleinere, günstigere Modelle teils 200 000 bis 400 000 [15]. Nach der englischen

Faustregel entspricht das mehreren Hunderttausend Wörtern [8] – ganze Aktenordner in einer einzigen Anfrage; auf Deutsch entsprechend weniger.

Ein großes Fenster ist allerdings Transportkapazität, kein Verständnisversprechen. Liu et al. zeigten 2024 in einer viel beachteten Studie, dass Modelle Informationen am Anfang und am Ende eines langen Kontexts deutlich zuverlässiger nutzen als in der Mitte – die Leistung fällt dort messbar ab („Lost in the Middle“) [14]. Neuere Modelle mildern den Effekt, strukturell verschwunden ist er nicht – sorgfältige Auswahl und Platzierung der Inhalte (Retrieval) bleibt also auch im Zeitalter der Millionen-Token-Fenster wertvoll.

„Gewichte sind das, was Sie beim Anbieter einkaufen. Der Kontext ist das, was Ihr Unternehmen daraus macht. Fast jede strategische KI-Frage – Daten, Datenschutz, Souveränität, Kosten – ist am Ende eine Frage über diese zwei Speicher.“

— Nicholas Krimmel

4

Wie ein LLM entsteht: drei Phasen

Vom leeren Netz zum brauchbaren Assistenten führen zwei Trainingsphasen – und eine dritte Phase, die im Betrieb für Aktualität sorgt.

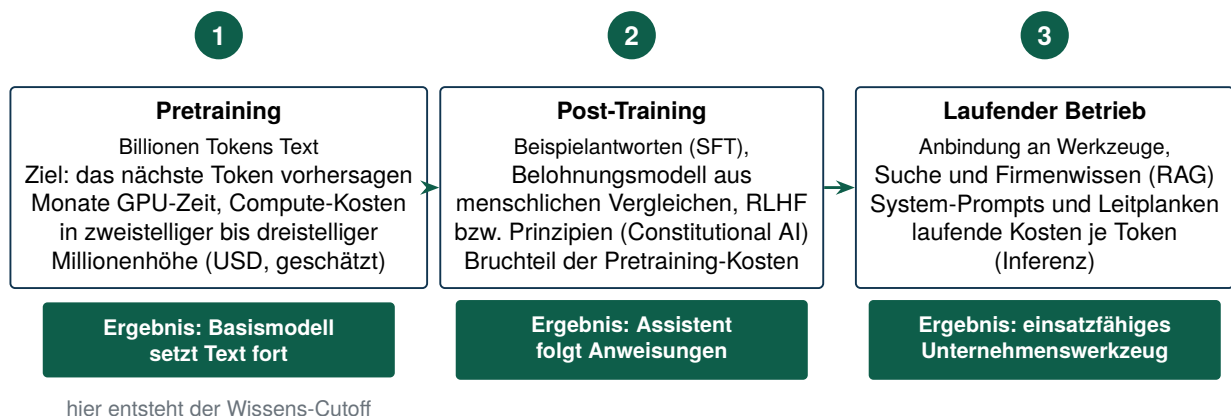


Abbildung 4 Drei Phasen vom Rohmodell zum Werkzeug. Kostenanker für Phase 1: geschätzte 78 Mio. USD Trainings-Compute für GPT-4, 191 Mio. USD für Gemini Ultra [16].

4.1 Phase 1: Pretraining – Verdichtung von Billionen Tokens

Im **Pretraining** lernt das Netz eine einzige Aufgabe: das jeweils nächste Token in riesigen Textmengen vorherzusagen – Webtexte, Bücher, Code, zunehmend auch lizenzierte und synthetische Daten. Niemand etikettiert dabei richtige Antworten; der Text selbst ist das Lehrmaterial (selbstüberwachtes Lernen). Das Ergebnis ist ein **Basismodell**: ein System, das Texte druckreif fortsetzen kann, aber noch kein Assistent ist – es beantwortet Fragen auch mal mit einer Gegenfrage-Liste, weil so etwas im Internet eben häufig auf Fragen folgt.

Diese Phase ist der Kapitalblock der Branche: Der Stanford AI Index beziffert die reinen Rechenkosten auf geschätzte 78 Mio. USD für das Training von GPT-4 und geschätzte 191 Mio. USD für Gemini Ultra – Compute-Schätzungen nach der Methodik von Epoch AI, ohne Personal, Fehlversuche und Vorlauftforschung [16]. Und der Einsatz wächst weiter: Der Trainingsaufwand

der jeweils größten Sprachmodelle steigt seit 2020 um rund den Faktor 5 pro Jahr (Epoch AI, Datenstand Februar 2026) [17].

Für Ihre Praxis am wichtigsten ist eine Nebenwirkung: Weil das Training zu einem Stichtag endet, hat jedes Modell einen **Wissens-Cutoff**. Alles danach – Ihr letzter Quartalsbericht, die neue Preisliste – existiert für die Gewichte nicht; Aktualität muss in Phase 3 über den Kontext nachgeliefert werden.

4.2 Phase 2: Post-Training – aus dem Rohmodell wird ein Assistent

Dass ChatGPT und seine Nachfolger sich wie hilfsbereite Assistenten verhalten, ist das Ergebnis des **Post-Trainings**. Die von OpenAI 2022 publizierte InstructGPT-Pipeline hat drei Stufen [18]: Zuerst schreiben Menschen vorbildliche Antworten, mit denen das Modell nachtrainiert wird (**Supervised Fine-Tuning**, SFT). Dann bewerten Menschen Paare von Modellantworten („welche ist besser?“); daraus lernt ein **Belohnungsmodell**, menschliche Präferenzen vorherzusagen. Schließlich wird das Sprachmodell per Verstärkungslernen darauf optimiert, diese Belohnung zu maximieren (**RLHF**, Reinforcement Learning from Human Feedback).

Der für die Führungsebene entscheidende Befund steht in der InstructGPT-Arbeit selbst: Die Antworten eines Modells mit nur 1,3 Milliarden Parametern wurden von menschlichen Bewertern denen des über 100-mal größeren GPT-3 (175 Milliarden) vorgezogen [18]. **Verhalten schlägt Rohgröße.**

Anthropic entwickelte mit **Constitutional AI** eine prinzipienbasierte Variante: Statt allein auf menschliche Einzelurteile zu setzen, kritisiert und überarbeitet das Modell eigene Antworten anhand eines explizit niedergelegten Prinzipienkatalogs (einer „Verfassung“); das Feedback für das Verstärkungslernen erzeugt dabei eine KI (RLAIF) [19]. Für Sie relevant: Verhaltens- und Wertefragen sind dokumentierte Trainingsentscheidungen – man kann Anbieter im Auswahlprozess danach fragen.

■ Die Governance-Lektion aus Phase 2

Modellqualität ist kein Naturprodukt der Größe, sondern zu großen Teilen das Ergebnis von Post-Training-Entscheidungen des Anbieters. Konsequenz für Ihre Beschaffung: Bewerten Sie Modelle nicht nach Größenversprechen oder Benchmark-Marketing, sondern durch Tests im eigenen Anwendungsfall mit eigenen Dokumenten und Qualitätskriterien (Evaluationen).

4.3 Phase 3: Der laufende Betrieb – Aktualität durch Anbindung

Die dritte Phase ist streng genommen kein Training mehr: Im Betrieb wird das eingefrorene Modell mit der Gegenwart verbunden – über System-Prompts, Websuche, Retrieval auf geprüfte Firmenquellen (RAG) und Werkzeuge, mit denen das System Aufgaben ausführt (Agenten; Kapitel 8). Ein aktuelles, firmenkundiges LLM-System ist immer ein *Verbund* aus eingefrorenem Modell und lebendiger Anbindung.

5 Warum die Modelle so schnell besser werden: drei Skalierungshebel

Ist der rasante Fortschritt ein Strohfeuer – oder planbar? Er speist sich aus drei unterscheidbaren Hebeln, die kombinierbar sind und bislang alle weiter wirken.

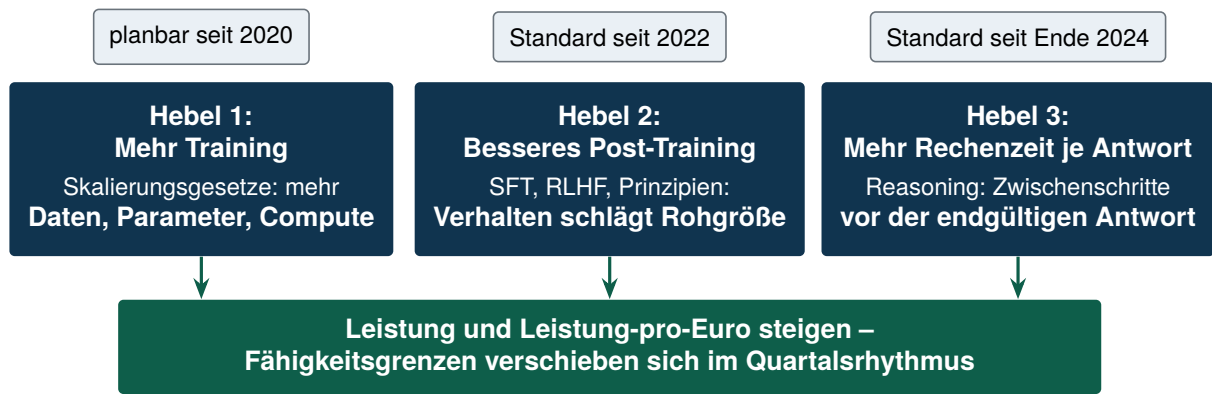


Abbildung 5 Drei Skalierungshebel wirken kombiniert: Training, Post-Training und Rechenzeit zur Antwortzeit („Test-Time Compute“).

5.1 Hebel 1: Skalierung des Trainings – Fortschritt wurde planbar

2020 zeigte ein OpenAI-Team empirisch, dass der Vorhersagefehler von Sprachmodellen mit wachsender Modellgröße, Datenmenge und Rechenleistung einem glatten Potenzgesetz folgt – über viele Größenordnungen hinweg, ohne Sättigungsknick im untersuchten Bereich [20]. Das klingt technisch, war aber ökonomisch ein Wendepunkt: **Fortschritt wurde planbar**. Wer ein Vielfaches an Rechenleistung investiert, kann die Verbesserung grob vorhersagen – eine Investitionslogik, wie sie kaum eine andere Forschungsrichtung je geboten hat, und die Grundlage der Milliardenwetten der Anbieter.

2022 präzisierte DeepMind die Rezeptur: Modellgröße und Datenmenge gehören im Gleichgewicht skaliert – wer die Parameterzahl verdoppelt, sollte auch die Trainingsdaten verdoppeln. Das Modell Chinchilla (70 Milliarden Parameter, trainiert auf 1,4 Billionen Tokens, also rund 20 Tokens je Parameter) übertraf ein viermal größeres Schwestermodell [21]. Die „20-Tokens-je-Parameter“-Regel ist dabei eine historische Heuristik, kein Gesetz: Heute trainieren Anbieter kleinere Modelle bewusst weit über diesen Punkt hinaus („Übertraining“), weil im Massenbetrieb die Inferenzkosten dominieren – ein kleineres, länger trainiertes Modell ist pro Anfrage billiger.

5.2 Hebel 2: Qualität des Post-Trainings

Der zweite Hebel ist das in Kapitel 4 beschriebene Post-Training: Dieselbe Rohintelligenz wird durch bessere Beispieldaten, Belohnungsmodelle und Prinzipien nutzbarer, sicherer und präziser im Anweisungsbefolgen [18, 19] – zu einem Bruchteil der Kosten von Hebel 1. Er erklärt, warum sich Modelle derselben Generation im Alltag spürbar unterscheiden.

5.3 Hebel 3: Rechenzeit pro Antwort – die Reasoning-Wende

Der dritte Hebel ist der jüngste. Im September 2024 stellte OpenAI mit o1 eine Modellreihe vor, die per Verstärkungslernen gelernt hat, vor der endgültigen Antwort eine lange interne Kette von Zwischenschritten zu erzeugen – und deren Leistung sowohl mit mehr Training als auch mit mehr Rechenzeit *pro Antwort* steigt [22]. Solche **Reasoning-Modelle** „denken“ nicht im menschlichen Sinn: Sie erzeugen zusätzliche Zwischentokens, eine Art internen Schmierzettel, der die Trefferquote bei komplexen Aufgaben nachweislich erhöht. Qualität wird damit zur Stellgröße pro Anfrage: mehr Rechenzeit für die schwierige Rechtsfrage, weniger für die Routinemail.

Im Januar 2025 folgte der Paukenschlag: Das chinesische Labor DeepSeek zeigte mit R1, dass solches Reasoning-Verhalten allein durch Verstärkungslernen entstehen kann – ohne von Menschen

vorgeschriebene Beispiel-Denkketten – und legte Methode samt Modellgewichten offen. Die Arbeit durchlief als erstes großes LLM eine unabhängige Fachbegutachtung und erschien im September 2025 als Titelgeschichte in *Nature* [23]. Die Börse reagierte historisch: Am 27. Januar 2025 verlor Nvidia 17 % bzw. 589 Milliarden USD Marktkapitalisierung an einem einzigen Handelstag – der größte Tagesverlust der Börsengeschichte [24] –, weil Anleger schlagartig neu bewerteten, wie viel Kapital Spitzenleistung erfordert.

Stand 2026 ist Reasoning kein Sondermodus mehr, sondern Standard: GPT-5 führte im August 2025 die GPT- und die o-Serie mit einem automatischen Verteiler (Router) zusammen, Claude arbeitet mit durchgängig aktivem adaptivem Denkmodus, und die Gemini-3-Generation ist durchgängig als Thinking-Modell ausgelegt [15].

Sprünge oder Messartefakte? Die Emergenz-Debatte

2022 dokumentierte ein Google-Team „emergente Fähigkeiten“: Leistungen, die in kleineren Modellen fehlen und ab einer gewissen Größe scheinbar sprunghaft auftreten [25]. 2023 hielt eine Stanford-Arbeit dagegen – ausgezeichnet als NeurIPS Outstanding Paper: Viele dieser Sprünge sind Artefakte von Alles-oder-nichts-Metriken; misst man stetig, wachsen die Fähigkeiten glatt [26]. Die Management-Synthese: Die zugrunde liegende Kompetenz wächst stetig mit der Skalierung – aber *wann* eine Fähigkeit die Schwelle zur Praxistauglichkeit überschreitet, bleibt schwer prognostizierbar. Konsequenz: Bewerten Sie Use Cases nicht einmalig, sondern quartalsweise neu.

5.4 Die ökonomische Folge: Preise fallen schneller als Vorlagen altern

Die drei Hebel wirken doppelt: Die Spitze wird besser, und einmal erreichte Qualität wird rapide billiger. Die Inferenzkosten für ein konstantes Leistungsniveau (GPT-3.5-Niveau) fielen von November 2022 bis Oktober 2024 um mehr als das 280-Fache [27]. Ein Business Case, der vor achtzehn Monaten an den Kosten scheiterte, kann heute tragfähig sein – rechnen Sie abgelehnte Vorhaben periodisch neu.

6 Halluzinationen: warum sie entstehen – und warum das eine Prozessfrage ist

Als **Halluzination** bezeichnet man flüssig formulierte, plausibel klingende, aber sachlich falsche oder erfundene Ausgaben – das erfundene Urteil samt Aktenzeichen, die nie erschienene Studie. Der Begriff ist etabliert, aber unglücklich: Das Modell erlebt nichts und täuscht auch niemanden absichtlich; es erzeugt die wahrscheinlichste Fortsetzung – auch dort, wo ehrliches Passen die bessere Antwort wäre. Die Forschung unterscheidet mehrere Spielarten, von der Abweichung von der Quelle bis zur frei erfundenen Tatsache [28]. Für die Führungsebene zählt die Ursachenfrage; zwei Erklärungsebenen ergänzen sich.

6.1 Erklärung 1: Die Anreiz-Ökonomie des Ratens

Eine Forschungsarbeit von OpenAI aus dem Jahr 2025 – als Anbieterforschung bislang nicht unabhängig begutachtet – formalisiert eine ökonomische Erklärung: Die üblichen Trainings- und Bewertungsverfahren **belohnen Raten und bestrafen ehrliches Eingestehen von Unsicherheit** [29]. In den meisten Benchmarks bringt „Das weiß ich nicht“ null Punkte, ein selbstbewusster Tipp dagegen manchmal einen Treffer. Modelle werden so systematisch zu „guten Prüflingen“ erzogen,

die im Zweifel tippen statt passen. Der Reformvorschlag der Autoren: Benchmarks so umbauen, dass kalibrierte Unsicherheit belohnt wird.

6.2 Erklärung 2: Die Fehlzündung im Schaltkreis

Die zweite Ebene liefert Anthropic's Interpretability-Programm, das im März 2025 mit sogenannten Attributionsgraphen eine Art Schaltplan-Rekonstruktion für sein Produktionsmodell vorlegte – ebenfalls Anbieterforschung, aber mit kausalen Interventionsexperimenten belegt [30]. Der Befund: **Verweigern ist der Default**. Im Modell ist ein Schaltkreis standardmäßig aktiv, der auf „das kann ich nicht beantworten“ hinausläuft. Erkennt das Modell eine bekannte Entität – etwa den Basketballspieler Michael Jordan –, unterdrückt ein „bekannte Entität“-Signal diese Bremse, und die Antwort fließt. Fragt man nach dem frei erfundenen „Michael Batkin“, bleibt die Bremse aktiv: Das Modell lehnt ab. Aktivierten die Forscher die „bekannt“-Merkmale künstlich, behauptete das Modell prompt, Batkin sei Schachspieler – die Halluzination ließ sich experimentell ein- und ausschalten.

Die natürliche **Fehlzündung** folgt demselben Muster: Ein Name ist aus den Trainingsdaten vertraut, belastbare Fakten dazu fehlen – das „bekannt“-Signal löst die Bremse trotzdem, und das Modell füllt die Lücke mit Plausiblen. Deshalb treffen Halluzinationen so oft die Ränder des Wissens: die wenig dokumentierte Person, das Nischenurteil, die interne Kennzahl.

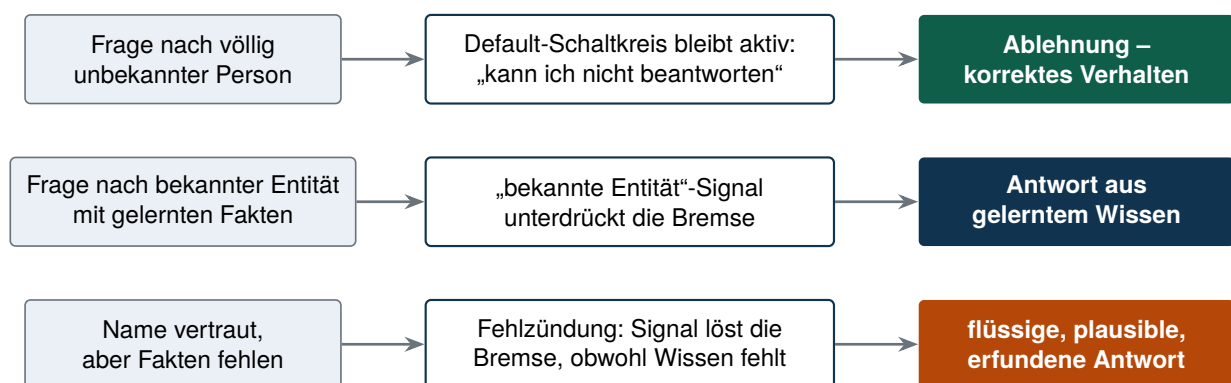


Abbildung 6 Halluzination als Fehlzündung: Verweigern ist der Default; ein „bekannte Entität“-Signal schaltet ihn ab – im Fehlerfall auch dann, wenn die Fakten fehlen (nach Anthropic-Interpretability-Befunden [30]).

Vorsicht vor der Schlagzeile „Halluzinationen gelöst“

Beide Erklärungslinien zeigen Wege zur *Reduktion*: ehrlichere Bewertungsmaßstäbe, gezielte Eingriffe in erkannte Schaltkreise, bessere Kalibrierung. Keine zeigt eine Abschaltung. Solange ein System Wahrscheinlichkeitsverteilungen fortschreibt, bleibt die überzeugend formulierte Falschaussage möglich. Vorlagen, die ein „halluzinationsfreies“ System versprechen, verdienen Ihre kritischsten Nachfragen.

6.3 Die Konsequenz: Prozesse statt Produktversprechen

Wenn Halluzination eine Systemeigenschaft ist, lautet die richtige Führungsreaktion nicht „warten, bis die Technik fehlerfrei ist“, sondern: Prozesse bauen, die mit einer bekannten Fehlerart professionell umgehen – so, wie das Rechnungswesen mit Erfassungsfehlern umgeht.

- **Quellenbindung:** Antworten auf geprüfte Firmenquellen stützen (RAG) und Belegstellen mitliefern lassen – das verlagert die Aufgabe vom „Erinnern“ zum Zusammenfassen vorliegender Dokumente.
- **Menschliche Freigabe an Wirkschwellen:** Überall dort, wo Ausgaben Außenwirkung oder Rechtsfolgen entfalten, prüft ein Mensch – Human-in-the-Loop als Architekturprinzip, nicht als Appell.
- **Messen statt hoffen:** Halluzinations- und Fehlerraten je Anwendungsfall regelmäßig messen (Evaluationen), Grenzwerte definieren, Abweichungen eskalieren.
- **Aufgabenwahl:** Anwendungsfälle bevorzugen, in denen Fehler billig und sichtbar sind (Entwürfe, Recherche mit Beleg) – nicht teuer und unsichtbar (ungeprüfte Auskünfte).

■ Die richtige Vorstandsfrage

Nicht: „Halluziniert das System?“ (Die ehrliche Antwort lautet immer: gelegentlich.) Sondern: „Welche Fehlerrate messen wir in diesem Anwendungsfall, welche Rate ist akzeptabel, und welcher Prozessschritt fängt den Rest ab?“

7

Die Fähigkeitslandkarte: was LLMs können – und was nicht

7.1 Die gezackte Grenze

Die vielleicht wichtigste empirische Studie zur Arbeitswirkung stammt von Dell’Acqua und Kollegen: 758 Beraterinnen und Berater der Boston Consulting Group bearbeiteten realistische Aufgaben, randomisiert mit und ohne KI-Unterstützung. Innerhalb der Fähigkeitsgrenze der Technologie erledigten die KI-Nutzer im Schnitt 12,2 % mehr Aufgaben, waren 25,1 % schneller und lieferten über 40 % höhere Qualität. Bei einer Aufgabe *außerhalb* dieser Grenze – konstruiert mit einer verdeckten Falle – lag die KI-Gruppe dagegen 19 Prozentpunkte *unter* der Vergleichsgruppe [31]. Die Autoren prägten dafür das Bild der **Jagged Frontier**, der gezackten Grenze: Die Fähigkeitsgrenze verläuft nicht entlang gefühlter Schwierigkeit; zwei ähnlich wirkende Aufgaben können diesseits und jenseits liegen. Wer der Maschine blind vertraut, wird ausgerechnet dort schlechter, wo sie schwach ist – und merkt es nicht.

7.2 Drei Studien, ein Muster: der Aufgabentyp entscheidet

Derselbe Dreiklang zieht sich durch die belastbarsten Feldbefunde:

- **+55,8 %:** In einem kontrollierten Experiment lösten Entwickler eine klar umrissene Programmieraufgabe mit GitHub Copilot deutlich schneller als die Kontrollgruppe [32] – der Idealfall der Technologie.
- **+15 %:** Im realen Kundensupport eines Fortune-500-Konzerns steigerte ein Assistent die Produktivität im Schnitt um 15 %, mit den größten Gewinnen bei Berufseinsteigern [33] – solide, aber kein Vervielfacher.
- **–19 %:** Sehr erfahrene Open-Source-Entwickler, die in ihren eigenen, über Jahre gewachsenen Codebasen arbeiteten, waren mit KI-Werkzeugen 19 % *langsamer* – und glaubten dennoch, schneller gewesen zu sein (METR 2025). Ein Nachtrag der Autoren von 2026 relativiert die Zahl deutlich [34].

Die Spannweite von plus 55,8 % bis minus 19 % ist keine Widersprüchlichkeit der Forschung, sondern die Kernbotschaft: **Der Aufgabenzuschnitt entscheidet über Gewinn oder Verlust**, und die Selbstwahrnehmung der Nutzer ist dabei kein verlässlicher Kompass. Die Zuordnung von Aufgaben zu Systemen ist damit eine Managemententscheidung.

7.3 Stark, schwach – und was das für Sie heißt

Stark bei	Schwach bei	Was das für Sie heißt
Sprache: Entwerfen, Zusammenfassen, Umformulieren, Übersetzen	Faktenauskunft ohne mitgelieferte Quelle (Halluzinationsrisiko)	Erstfassungen automatisieren; Fakten nur mit Quellenbindung (RAG) und definierter Prüfung
Struktur: Extrahieren, Klassifizieren, Mustererkennung in Text und Daten	präziser Arithmetik und strenger Logik ohne Werkzeuganbindung	Rechnen, Abgleichen und Prüfen an Werkzeuge und Datenbanken delegieren, nicht an das Sprachmodell selbst
Breite: Querschnittswissen, Varianten- und Ideenbildung, Perspektivwechsel	Aktualität: Wissensstand endet am Trainings-Cutoff	Aktuelles und Firmenspezifisches gehört in den Kontext (Retrieval), nicht in die Erwartung an das Modell
Tempo und Skala: Sekunden je Entwurf, parallelisierbar, rund um die Uhr	Determinismus: gleiche Frage, variierende Antworten	Für audit-relevante Prozesse: Protokollierung, inhaltliche Prüfkriterien, Stichproben und klare Freigaben

Tabelle 1 Die Fähigkeitslandkarte in Kurzform – als Ausgangspunkt für den Aufgabenzuschnitt im eigenen Haus.

7.4 Der Faithfulness-Vorbehalt: Begründungen sind nicht der Rechenweg

Ein letzter, für Kontrollgremien wichtiger Punkt: Wenn ein Modell seine Antwort begründet oder ein Reasoning-Modell Zwischenschritte zeigt, liegt es nahe, das als Einblick in den tatsächlichen Entscheidungsweg zu lesen. Die Interpretability-Forschung mahnt zur Vorsicht: Anthropic dokumentiert Fälle, in denen die ausgegebene Begründung nachweislich nicht dem intern rekonstruierten Rechenweg entspricht – das Modell konstruiert eine plausible Erklärung, mitunter rückwärts vom nahegelegten Ergebnis her [30]. Begründungstexte sind also Plausibilisierung, kein Audit-Beweis; Verlässlichkeit entsteht durch externe Verifikation – Quellenabgleich, Tests, Stichproben, Vier-Augen-Prinzip.

8

Vom Modell zum Unternehmenswerkzeug: der Stack und die zehn Vorstandsfragen

8.1 Vier Schichten über dem Rohmodell

Zwischen dem Modell eines Anbieters und einem produktiven Unternehmenssystem liegen mehrere Schichten – und fast der gesamte Gestaltungsspielraum Ihrer Organisation:

- 1. System-Prompt:** eine ständige, für Nutzer unsichtbare Rahmenanweisung, die Rolle, Tonalität, Zuständigkeit und Tabuzonen festlegt – die billigste und schnellste Steuerungsebene.
- 2. RAG (Retrieval-Augmented Generation):** Vor der Antwort werden passende Passagen aus geprüften Firmenquellen gesucht und in das Kontextfenster gelegt; das Modell fasst zusammen, statt sich zu „erinnern“ – Aktualität, Quellenbindung und Zugriffskontrolle entstehen hier.

3. **Werkzeuge und Agenten:** Das System darf definierte Aktionen ausführen – suchen, rechnen, Tickets anlegen, Entwürfe ablegen – in Schleifen und mit Leitplanken. Hier entsteht der größte Nutzen und das größte Missbrauchspotenzial.
4. **Fine-Tuning:** Nachtraining der Gewichte mit eigenen Beispielen, um Stil oder Spezialverhalten dauerhaft zu verankern – wirkungsvoll, aber wartungsintensiv und seltener nötig, als Anbieter suggerieren.

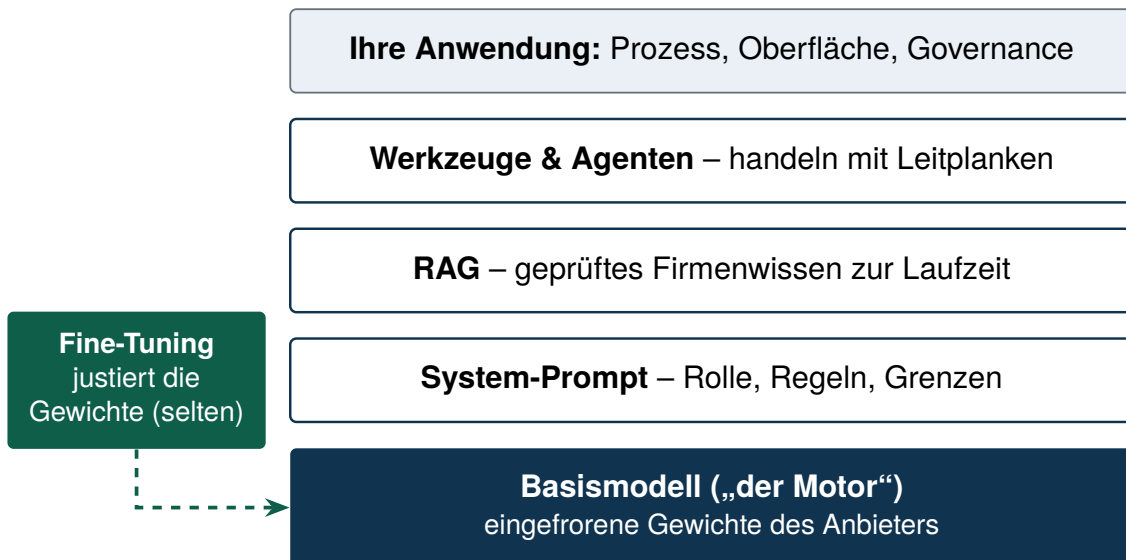


Abbildung 7 Der Anwendungs-Stack: Das Modell ist der Motor – Wert, Sicherheit und Compliance entstehen in den Schichten darüber, die Ihre Organisation verantwortet.

Der Kernsatz dieses Kapitels: **Das Modell ist der Motor – das Auto baut Ihre Organisation.** Zwei Unternehmen mit identischem Modell erzielen völlig unterschiedliche Ergebnisse, weil Daten, Prozesse, Leitplanken und Kompetenz den Unterschied machen. Wie aus Piloten ein skalierender Betrieb wird, ist die nächste, betriebliche Frage.

8.2 Die zehn Fragen für die nächste KI-Vorlage

Damit ist das mentale Modell einsatzbereit: Die folgenden zehn Fragen prüfen jede KI-Vorlage auf die Punkte, an denen Projekte tatsächlich scheitern; jede ist einem Themenfeld zugeordnet.

So setzen Sie die Fragen ein

Nehmen Sie die zehn Fragen der folgenden Tabelle in die Standardgliederung für KI-Entscheidungsvorlagen auf. Vorlagen, die Frage 1 (Nutzen) und Frage 10 (menschliche Kontrolle) nicht beantworten können, sind nicht entscheidungsreif.

Nr.	Frage an die Vorlage	Themenfeld
1	Welches Geschäftsproblem lösen wir – und woran messen wir den Erfolg in Euro, nicht in Demos?	Business Case
2	Auf welcher Datenbasis arbeitet das System, und wer verantwortet deren Qualität und Pflege?	Datenfundament
3	In welche Risikoklasse fällt der Anwendungsfall nach der KI-Verordnung, und welche Pflichten folgen daraus?	EU AI Act
4	Welche personenbezogenen Daten fließen in Prompts, Kontext und Protokolle – und auf welcher Rechtsgrundlage?	Datenschutz
5	Was passiert bei gezieltem Missbrauch – von Prompt Injection bis Datenabfluss – und wie begrenzen wir den Schaden?	Sicherheit
6	Wer kann das System kompetent bedienen und beaufsichtigen, und wie bauen wir diese Kompetenz nachweisbar auf?	AI Literacy
7	Machen oder kaufen wir – und wie kommen wir wieder heraus (Anbieterwechsel, Datenportabilität, Exit)?	Werkzeugwahl
8	Wie kommt der Pilot in den Regelbetrieb: Wer betreibt, überwacht und misst das System dauerhaft – mit welchen Kennzahlen?	Skalierung
9	Welches Team setzt um – und ist es klein und fokussiert genug, um in Wochen zu liefern statt in Jahren?	Microteaming
10	An welchen Stellen prüft ein Mensch, bevor eine Ausgabe Außenwirkung entfaltet – und wie ist das technisch erzwungen?	KI-Agenten

Tabelle 2 Zehn Vorstandsfragen an jede KI-Vorlage – mit dem jeweils zugehörigen Themenfeld.

9

Fazit und Glossar

Das mentale Modell ist beisammen – genug, um jede KI-Vorlage souverän zu hinterfragen; genau das ist die Aufgabe der Führungsebene.

LLMs in 7 Sätzen für die nächste Vorstandssitzung

1. Ein großes Sprachmodell erzeugt Text, indem es Token für Token die wahrscheinlichste Fortsetzung berechnet und daraus zieht – es ist eine Wahrscheinlichkeitsmaschine, keine Datenbank.
2. Sein „Weltwissen“ liegt eingefroren in den trainierten Gewichten mit festem Wissensstichtag; alles Aktuelle und Firmenspezifische kommt nur über das Kontextfenster hinein.
3. Es entsteht in Phasen: Vortraining macht es sprachmächtig, Verhaltenstraining mit menschlichem Feedback macht es zum Assistenten – Verhalten schlägt rohe Größe.
4. Besser wird es über drei planbare Hebel – mehr Training, besseres Verhaltenstraining, mehr Rechenzeit pro Antwort –, weshalb Fähigkeits- und Kostenannahmen quartalsweise zu erneuern sind.
5. Halluzinationen entstehen, weil das Training flüssiges Raten belohnt und die „Weiß-ich-nicht“-Bremse fehlzünden kann; als Systemeigenschaft werden sie durch Prozesse beherrscht, nicht durch Warten auf das fehlerfreie Modell.
6. Die Fähigkeitsgrenze verläuft gezackt: Passende Aufgaben bringen messbar zweistellige Produktivitätsgewinne, falsche machen Teams langsamer und Fehler plausibler – der Aufgabenzuschnitt ist die eigentliche Managemententscheidung.
7. Das Modell ist der Motor, nicht das Auto: Wert, Sicherheit und Compliance entstehen in den Schichten darüber – und die verantwortet Ihre Organisation.

9.1 Glossar: 14 Begriffe in Alltagssprache

Begriff	Erklärung
Token	Kleinste Verarbeitungseinheit: ein Wortstück von wenigen Zeichen. Grundlage für Abrechnung und Kapazitätsgrenzen.
Parameter / Gewichte	Die Milliarden Zahlen, in denen das Trainingsergebnis gespeichert ist – das „Langzeitgedächtnis“, nur durch Training veränderbar.
Kontextfenster	Arbeitsgedächtnis einer Anfrage: alles, was das Modell gerade „vor Augen“ hat. Flüchtig und in Tokens bemessen.
Embedding	Zahlenvektor eines Tokens; ähnliche Bedeutung liegt benachbart.
Attention / Transformer	Der Mechanismus, mit dem jedes Token alle anderen gewichtet – und die darauf aufgebaute, parallelisierbare Architektur.
Pretraining	Erste Trainingsphase: Vorhersage des nächsten Tokens auf riesigen Textmengen. Erzeugt das Basismodell – und den Wissens-Cutoff.
Post-Training / RLHF	Zweite Phase: Beispielantworten und menschliches Feedback formen aus dem Basismodell einen Assistenten, der Anweisungen folgt.
Halluzination	Flüssig formulierte, aber sachlich falsche oder erfundene Ausgabe – systembedingte Nebenwirkung der Wahrscheinlichkeitsmechanik.
RAG	Retrieval-Augmented Generation: geprüfte Quellpassagen werden vor der Antwort in den Kontext gelegt – für Aktualität und Belegbarkeit.
Fine-Tuning	Nachtraining der Gewichte mit eigenen Beispielen; verankert Stil oder Spezialverhalten dauerhaft. Wirksam, aber wartungsintensiv.
Agent	LLM-System, das in Schleifen Werkzeuge aufruft, um mehrschrittige Aufgaben zu erledigen – nur mit Leitplanken und Protokollen zu betreiben.
Inferenz	Der laufende Betrieb: Jede Antwort kostet Rechenleistung (je Token abgerechnet).
Reasoning-Modell	Modellmodus, der vor der Antwort zusätzliche Zwischenschritte erzeugt – stärker bei komplexen Aufgaben, teurer je Antwort.
Prompt	Die gesamte Eingabe: Nutzerfrage, Verlauf und System-Prompt.

„Sie müssen kein Modell bauen können, um es zu führen. Aber Sie müssen wissen, dass Sie eine Wahrscheinlichkeitsmaschine führen – denn dann fragen Sie nach Daten statt Demos, Prozessen statt Versprechen, Messung statt Magie.“

— Nicholas Krimmel

Kontakt & Umsetzung

Möchten Sie die Inhalte dieses Whitepapers in Ihrem Unternehmen umsetzen?

Nicholas Krimmel

Web: www.nicholaskrimmel.com

Nicholas Krimmel

„Ich helfe Unternehmen, durch kleine, KI-gestützte Teams große Ergebnisse zu erzielen.“

Rechtliche Hinweise: Dieses Whitepaper wurde mit größter Sorgfalt erstellt (Stand: Juli 2026). Es dient ausschließlich der allgemeinen Information und stellt keine Rechts-, Steuer- oder Anlageberatung dar. Alle genannten Studien und Quellen sind im Literaturverzeichnis ausgewiesen; für die Inhalte externer Quellen wird keine Gewähr übernommen. Genannte Marken- und Produktnamen sind Eigentum der jeweiligen Inhaber.

© 2026 Nicholas Krimmel. Alle Rechte vorbehalten. Die Weitergabe dieses Whitepapers in unveränderter Form ist ausdrücklich erwünscht; auszugsweise Vervielfältigung nur mit Quellenangabe.

Literatur

- [1] Deloitte Global Boardroom Program (2024). *AI slow to gain prominence in the boardroom*. Befragung von 468 Board-Mitgliedern und Führungskräften in 57 Ländern. <https://www.prnewswire.com/news-releases/deloitte-ai-slow-to-gain-prominence-in-the-boardroom-302269876.html>
- [2] Deloitte Global Boardroom Program (2025). *Progress on AI in the boardroom – but room to accelerate*. Zweite Befragungswelle, 695 Befragte in 56 Ländern. <https://www.deloitte.com/global/en/issues/trust/progress-on-ai-in-the-boardroom-but-room-to-accelerate.html>
- [3] Verordnung (EU) 2024/1689 des Europäischen Parlaments und des Rates (KI-Verordnung), Artikel 4: KI-Kompetenz. Anwendbar seit dem 2. Februar 2025.
- [4] Europäische Kommission. *AI literacy – Questions & Answers*. FAQ zur KI-Kompetenzpflicht nach Art. 4 KI-VO. <https://digital-strategy.ec.europa.eu/en/faqs/ai-literacy-questions-answers> (abgerufen im Juli 2026).
- [5] Bundesamt für Sicherheit in der Informationstechnik (2025). *NIS-2-Umsetzungsgesetz in Kraft*. Pressemitteilung, Dezember 2025. https://www.bsi.bund.de/DE/Service-Navi/Presse/Pressemitteilungen/Presse2025/251205_NIS-2-Umsetzungsgesetz_in_Kraft.html
- [6] McKinsey & Company (2025). *The State of AI: How organizations are rewiring to capture value*. Globale Befragung, n = 1 491 (Selbstauskunft), März 2025. <https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-how-organizations-are-rewiring-to-capture-value>
- [7] Sennrich, R., Haddow, B. & Birch, A. (2016). *Neural Machine Translation of Rare Words with Subword Units*. ACL 2016, S. 1715–1725. DOI: 10.18653/v1/P16-1162.
- [8] OpenAI. *What are tokens and how to count them?* OpenAI Help Center. <https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them> (abgerufen im Juli 2026).
- [9] Anthropic. *Glossary*. Claude-Entwicklerdokumentation. <https://platform.claude.com/docs/en/docs/about-claude/glossary> (abgerufen im Juli 2026).
- [10] Brown, T. B. et al. (2020). *Language Models are Few-Shot Learners*. NeurIPS 2020. arXiv:2005.14165.
- [11] Anthropic (2024). *Mapping the Mind of a Large Language Model*. 21.5.2024. <https://www.anthropic.com/research/mapping-mind-language-model>; Fachbeitrag: Templeton, A. et al. (2024). *Scaling Monosemanticity*. Transformer Circuits. <https://transformer-circuits.pub/2024/scaling-monosemanticity/>
- [12] Vaswani, A. et al. (2017). *Attention Is All You Need*. NeurIPS 2017. arXiv:1706.03762.
- [13] Pearson, H. et al. (2025). *The most-cited papers of the twenty-first century*. Nature, 640, 588–592 (Analyse, April 2025).
- [14] Liu, N. F. et al. (2024). *Lost in the Middle: How Language Models Use Long Contexts*. Transactions of the ACL, 12, 157–173. DOI: 10.1162/tacl_a_00638.
- [15] OpenAI, Anthropic & Google (2026). Offizielle Modelldokumentationen (Kontextfenster und Modell-Modi, Stand Juli 2026): <https://platform.openai.com/docs/models>; <https://platform.claude.com/docs/en/docs/about-claude/models/overview>; <https://ai.google.dev/gemini-api/docs/long-context>
- [16] Stanford HAI (2024). *The AI Index Report 2024*. Trainingskosten als Compute-Schätzungen nach Epoch-Methodik (Cottier et al., arXiv:2405.21015). <https://hai.stanford.edu/news/ai-index-state-ai-13-charts>

- [17] Epoch AI (2026). *Machine Learning Trends*. Datenstand: Februar 2026. <https://epoch.ai/trends>
- [18] Ouyang, L. et al. (2022). *Training language models to follow instructions with human feedback*. NeurIPS 2022. arXiv:2203.02155.
- [19] Bai, Y. et al. (2022). *Constitutional AI: Harmlessness from AI Feedback*. Anthropic. arXiv:2212.08073 (Preprint).
- [20] Kaplan, J. et al. (2020). *Scaling Laws for Neural Language Models*. arXiv:2001.08361 (Preprint; Konsens-Referenz).
- [21] Hoffmann, J. et al. (2022). *Training Compute-Optimal Large Language Models*. NeurIPS 2022. arXiv:2203.15556 („Chinchilla“).
- [22] OpenAI (2024). *Learning to reason with LLMs*. 12.9.2024. <https://openai.com/index/learning-to-reason-with-llms/>
- [23] DeepSeek-AI; Guo, D. et al. (2025). *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*. Nature, 645, 633–638 (17.9.2025). DOI: 10.1038/s41586-025-09422-z (Preprint: arXiv:2501.12948).
- [24] CNBC (2025). *Nvidia sheds almost \$600 billion in market cap, biggest drop ever*. 27.1.2025. <https://www.cnbc.com/2025/01/27/nvidia-sheds-almost-600-billion-in-market-cap-biggest-drop-ever.html>
- [25] Wei, J. et al. (2022). *Emergent Abilities of Large Language Models*. Transactions on Machine Learning Research. arXiv:2206.07682.
- [26] Schaeffer, R., Miranda, B. & Koyejo, S. (2023). *Are Emergent Abilities of Large Language Models a Mirage?* NeurIPS 2023, Outstanding Main Track Paper. arXiv:2304.15004.
- [27] Stanford HAI (2025). *The AI Index Report 2025*. <https://hai.stanford.edu/ai-index/2025-ai-index-report>
- [28] Ji, Z. et al. (2023). *Survey of Hallucination in Natural Language Generation*. ACM Computing Surveys, 55(12), Art. 248. DOI: 10.1145/3571730.
- [29] Kalai, A. T., Nachum, O., Vempala, S. S. & Zhang, E. (2025). *Why Language Models Hallucinate*. OpenAI-Forschungsarbeit (nicht peer-reviewed). arXiv:2509.04664.
- [30] Lindsey, J. et al. (2025). *On the Biology of a Large Language Model*. Anthropic, Transformer Circuits, 27.3.2025. <https://transformer-circuits.pub/2025/attribution-graphs/biology.html>
- [31] Dell’Acqua, F. et al. (2023). *Navigating the Jagged Technological Frontier*. Harvard Business School Working Paper 24-013; erschienen in Organization Science (2025). DOI: 10.1287/orsc.2025.21838.
- [32] Peng, S., Kalliamvakou, E., Cihon, P. & Demirel, M. (2023). *The Impact of AI on Developer Productivity: Evidence from GitHub Copilot*. arXiv:2302.06590.
- [33] Brynjolfsson, E., Li, D. & Raymond, L. (2025). *Generative AI at Work*. The Quarterly Journal of Economics, 140(2), 889–942. DOI: 10.1093/qje/qjae044.
- [34] Becker, J., Rush, N., Barnes, E. & Rein, D. (2025). *Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity*. METR. arXiv:2507.09089. <https://metr.org/blog/2025-07-10-early-2025-ai-experienced-os-dev-study/> Nachtrag der Autoren, 24.02.2026: geändertes Design, Selektionseffekte in der Aufgabenauswahl eingeräumt; neue Schätzung –18 % (KI –38 % bis +9 %), von den Autoren selbst als „only very weak evidence“ und als Untergrenze eingestuft. <https://metr.org/blog/2026-02-24-uplift-update/> (abgerufen 27.07.2026).