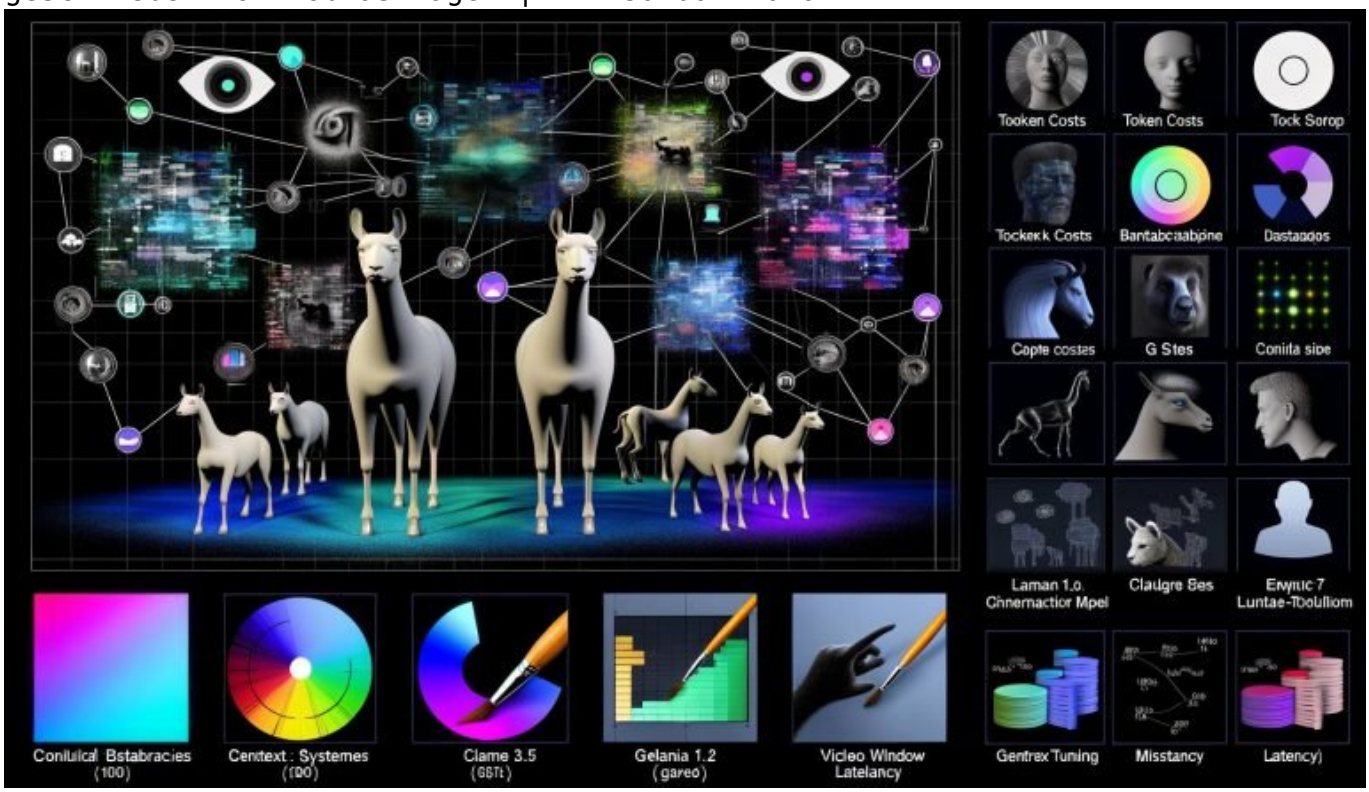
Welche AI gibt es: Marktüberblick für Online-Profis

Category: KI & Automatisierung
geschrieben von Tobias Hager | 12. Januar 2026



Welche AI gibt es 2025: Marktüberblick für Online-Profis

Du willst wissen, welche AI es gibt, was davon nur Buzzword-Bingo ist und welche Systeme dir heute real messbare Uplifts in SEO, Performance Marketing und Produktivität liefern? Willkommen bei 404: Wir sezieren den AI-Zoo, sortieren Modelle nach harten Kriterien wie Kontextfenster, Latenz, Tokenkosten und Datenschutz, und wir nennen Ross und Reiter ohne Werbesprech. Wenn du nach dieser Rundfahrt immer noch fragst “Welche AI gibt es?”, dann nur, um gezielt zu kaufen statt zu hoffen.

- Kompletter Marktüberblick: Welche AI gibt es, welche Klassen existieren

und wofür taugen sie wirklich
- Foundation Models, Multimodalität und Voice/Video-Stacks technisch erklärt – ohne Agenturzaubertricks
- SEO- und Content-Workflows mit AI, die skalieren, statt Duplicate-Content-Schrott zu produzieren
- Architektur-Entscheidungen: RAG vs. Fine-Tuning, Vektor-Datenbanken, Agenten und Guardrails
- Deployment-Optionen von API-SaaS bis Self-Hosted unter DSGVO, inklusive Datenresidenz und DLP
- Kennzahlen, die zählen: Kontextlänge, Zeit-zu-erstem-Token, Durchsatz, Kosten pro Aufgabe
- Schritt-für-Schritt-Blueprints, um “Welche AI gibt es” in “Welche AI setze ich morgen ein” zu übersetzen
- Realistische Grenzen, Risiken und wie du sie mit Guardrails, Evals und Observability abfederst

Die Frage “Welche AI gibt es” klingt naiv, ist aber die einzig sinnvolle, wenn du Budget, Risiko und Ergebnis im Griff behalten willst. “Welche AI gibt es” ist nicht nur ein Katalogthema, sondern eine Architekturfrage, die entscheidet, ob deine Kampagnen schneller lernen als die deiner Konkurrenz. “Welche AI gibt es” zwingt dich, Modelle, Modalitäten und Anbieter anhand von harten Parametern zu bewerten, statt dich auf bunte Demos zu verlassen. “Welche AI gibt es” bedeutet auch: Welche API-Verträge, welche SLAs und welche Datenverarbeitungsbedingungen bekommst du im Ernstfall durch die Rechtsabteilung. “Welche AI gibt es” ist somit die richtige Einstiegsfrage, aber die falsche Endstation, wenn du dich nicht mit Skalierung, Betrieb und Sicherheit beschäftigst. Und genau da setzen wir an, ohne Rücksicht auf Marketing-Mythen oder Tool-FOMO.

Bevor wir in Kategorien, Metriken und Stacks eintauchen, sortieren wir den Markt entlang der Praxis: Text-LLMs, multimodale Modelle, Vision- und Speech-Stacks, Generative Media für Bild und Video, Agent-Infrastruktur und Analytics-Layer. Dazu gehören die üblichen Verdächtigen wie GPT-4o, Claude 3.5, Gemini 1.5, Llama 3.1, Mistral Large, aber auch Nischenkönige wie Stable Diffusion XL, Flux, Runway Gen-3, Pika, ElevenLabs, Whisper und Deepgram. Wir klären, welche AI in welchem Use Case die Nase vorn hat, wie du Latenzbudgets planst, welche Kostenfallen dir die Marge auffressen und wie du mit RAG deinem Modell echtes Firmenwissen einflüsterst. Wir trennen sauber zwischen Spielzeug, Showcase und produktionsreifer Lösung. Und wir zeigen, wie du aus “Welche AI gibt es” eine belastbare Roadmap machst, die deine KPIs nicht verschönert, sondern verbessert.

Welche AI gibt es: Kategorien, Modelle und Marketing-Use-

Cases

Beginnen wir mit dem groben Raster, das dir im Tagesgeschäft Zeit spart und Fehlkäufe verhindert. Erstens: Large Language Models, kurz LLMs, für Textverständnis, Textgenerierung, Code und Planung sind die Arbeitspferde deines Stacks. Zweitens: Multimodale Modelle, die Text, Bild, Audio und Video in einer Architektur verknüpfen, lösen komplexere Aufgaben wie visuelle Analysen, Voice-Chat und kreative Komposition. Drittens: Vision-Modelle für OCR, Objekt- und Szenenerkennung, die strukturierte Daten aus chaotischem Bildmaterial extrahieren und deine Automationen füttern. Viertens: Speech-Modelle für ASR und TTS, also automatische Spracherkennung und Text-to-Speech, die Callcenter, Voicebots und Video-Vertonung skalieren. Fünftens: Generative Media für Bild- und Videoproduktion, die kreative Assets durch Prompting statt Briefing erzeugen und Varianten in Minuten statt Tagen liefern. Dieses Raster beantwortet “Welche AI gibt es” nicht nur taxonomisch, sondern funktionsbasiert, und das ist die Perspektive, die in Budgetgesprächen zählt.

Wenn du “Welche AI gibt es” aus Marketing-Sicht fragst, brauchst du konkrete Use-Cases statt Buzzwords. Für SEO sprechen wir über Entitäten-Extraktion, Snippet-Optimierung, interne Verlinkung, Schema-Markup und Logfile-Analysen mit LLM-Unterstützung. Für Content reden wir über Outline-Engines, Brand-Kompass via System-Prompts, Programmatic SEO mit RAG und Redaktions-Workflows mit automatischer Fact-Checking-Schleife. Für Performance Marketing zählen Anzeigentexte, Keyword-Clustering, Landingpage-Varianten, Bid-Management-Hinweise aus Attribution-Daten und kreative Tests in Social Ads. Für CRM und E-Mail sind Segmentierung, Copy-Personalisierung, Betreffzeilen-Optimierung und Response-Klassifikation starke Hebel. Diese Aufgaben sind nicht sexy, aber sie zahlen auf KPIs ein, und genau daran misst du den Sinn von “Welche AI gibt es” im Alltag.

Der Markt ist groß, aber nicht unüberschaubar, wenn du Messgrößen definierst. Achte bei LLMs auf Kontextfenster in Tokens, auf Zeit-zu-erstem-Token, auf Durchsatzlimits pro Minute und auf Kosten pro 1.000 Tokens, getrennt nach Prompt und Output. Prüfe bei Bild- und Video-Generatoren Renderzeit, Auflösungen, Stiltreue, Seed-Reproduzierbarkeit und Lizenzbedingungen, besonders für kommerzielle Nutzung. Beurteile Voice-Stacks nach WER (Word Error Rate), Latenz in Millisekunden, Streaming-Fähigkeit und Stimmklonen-Qualität inklusive rechtlicher Absicherung. Schaue bei jedem Anbieter auf Datenverarbeitung: Training auf deinen Inputs ja oder nein, Datenresidenz EU oder global, Modellisolierung und Audit-Logs. So wird aus “Welche AI gibt es” ein Nachfragebogen, der deinen Einkauf schützt und deine Roadmap liefert.

Foundation Models und

Multimodalität: LLMs, Vision und Speech im technischen Vergleich

Bei LLMs dominieren 2025 etwa eine Handvoll Klassen: proprietäre Spitzenreiter wie GPT-4o und Claude 3.5 für komplexe Reasoning-Aufgaben, Enterprise-Varianten wie Azure OpenAI mit Netzwerkisolation und Audit, sowie offene Modelle wie Llama 3.1 und Mistral Large für Self-Hosting und Kostenkontrolle. Wichtige Metriken sind Kontextlänge, typischerweise 128k bis 1M Tokens, die Fehleranfälligkeit bei sehr langen Kontexten (“lost in the middle”) und die Fähigkeit zu Tool-Use via Function Calling. Du willst deterministische Ergebnisse? Dann nutze “temperature” niedrig, verwende Systemprompts als Policy-Layer und setze Evals, um Regressionen nach Modellupdates zu erkennen. Für Code- und Data-Tasks sind spezialisierte Codex-Nachfolger oder Code-Llama-Ableger oft stabiler in Struktur und Syntax. Für skaliertes Schreiben ist der Trade-off klar: Mid-Tier-Modelle sind billiger und schnell genug, wenn du sie mit RAG, Vorlagen und Validierungen absicherst.

Multimodale Modelle wie GPT-4o, Gemini 1.5 und Claude 3.5 Sonnet interpretieren Screenshots, PDFs, Tabellen und Diagramme, generieren Bildbeschreibungen und beantworten Fragen zu UI-Flows oder Heatmaps. Achte auf Bild-Token-Pricing, denn Vision-Kontext kann die Rechnung explodieren lassen, wenn du unkomprimierte Seiten-Scans in Serie verarbeitest. Praktisch sind integrierte Tools wie OCR, Tabellen-Parser und Chart-Reasoning, die vorher separate Pipelines brauchten. Für Marketing relevant: Analyse von Kreativvarianten, visuelle Konsistenz für Brand-Guidelines und automatisierte QA für Landingpages, inklusive Erkennung von kaputten Hero-Breakpoints in Responsive-Layouts. Multimodalität bringt dir Geschwindigkeit und Konsistenz, wenn du Prompts als Templates mit Variablen führst und Bild-Assets in Versionen dokumentierst.

Speech-Stacks teilen sich in ASR, TTS und Voice-Cloning. Whisper, Deepgram und AssemblyAI liefern Streaming-ASR mit niedriger Latenz, Domain-Anpassung via Vokabular und PII-Redaktion im Flug. Für TTS sind ElevenLabs und Play.ht stark in Natürlichkeit, Mehrsprachigkeit und SSML-Kontrolle, also Stimmlage, Pausen, Betonungen und phonemische Feinheiten. Wer rechtssicher arbeiten will, braucht Einwilligungen für Stimmklone, klare Nutzungsrechte und interne Richtlinien, die Missbrauch verhindern. Für Live-Erlebnisse zählt Latenz unter 300ms, sonst wirkt das Gespräch wie schlechter Funk. Kombiniert mit LLM-Reasoning entstehen Voice-Agents, die nicht nur vorlesen, sondern Prozesse auslösen, Daten nachschlagen und im CRM dokumentieren. Damit wird AI nicht nur Assistent, sondern Schnittstelle zum Stack.

Welche AI gibt es für SEO und Content: Tools, Workflows und harte Grenzen

Für SEO gibt es grob drei AI-Schichten: Research, Produktion und Optimierung. Research umfasst Entitäten-Mining, SERP-Analysen, Topic-Cluster-Planung und Intent-Klassifikation mit LLMs, die semantische Nähe besser abbilden als klassische Keyword-Tools. Produktion heißt nicht "Artikel raushauen", sondern strukturierte Briefings, Outline-Generierung, Quellensammlung, Drafts mit Zitatankern und automatische Prüfung von Claims per RAG gegen deine Quellen. Optimierung bedeutet Snippet-A/B-Tests, interne Linkvorschläge, Schema.org-Markups für Artikel, FAQ und Produktvarianten, sowie Logfile-Analysen, die Crawlerpfade und Statuscodes in Klartext überführen. Tools wie Surfer, Frase, MarketMuse und Clearscope liefern On-Page-Signale, aber die Magie entsteht durch deine Pipeline, nicht durch die Tool-UI. Die harte Grenze: Ohne belastbare Quellen wird dein Output schnell dünn, generisch und riskant.

Im Content-Ops-Alltag zählt Wiederholbarkeit, nicht Genie. Nutze Systemprompts als Brand-Style-Guide mit Tonalität, VO-Beispielen, verbotenen Formulierungen und rechtlichen Dos & Don'ts. Kapsle Wissen via RAG mit einer Vektor-Datenbank wie pgvector, Milvus oder Pinecone und kontrolliere Retrieval mit Hybrid-Search, also Dense- und BM25-Signalen, um Fact-Drift zu minimieren. Setze Evals wie Ragas oder DeepEval ein, um Halluzinationsrate, Zitationsquote und stilistische Konsistenz zu messen. Führe Review-Gates ein, in denen Claims mit Quellen verknüpft sein müssen, bevor Inhalte live gehen. Miss Erfolg an organischem Wachstum, SERP-Stabilität und Zeit-zu-Publikation, nicht an Output-Menge. Skalierung ohne Qualität ist nur Lärm, den Google schneller erkennt, als dir lieb ist.

Pragmatischer Workflow, der wirklich trägt, sieht so aus:

- Briefing: Ziel, Persona, SERP-Ziele, Entitäten und No-Go-Claims definieren.
- Recherche: Top-Quellen scrapen, archivieren, mit Embeddings indexieren und bewertbar machen.
- Outline: Struktur mit H-Tags, Fragen, Visual-Slots und internen Linkzielen erzeugen.
- Draft: LLM mit RAG füttern, Zitate als Anker ausgeben, Fakten automatisch verifizieren.
- Optimierung: Snippets, Schema, interne Links und Medienvarianten generieren und prüfen.
- Review: Menschliche QA, Plagiatscheck, Compliance-Check, dann Publishing mit Versionierung.

Wenn du "Welche AI gibt es" für Content fragst, ist die bessere Anschlussfrage: "Welche Pipeline brauche ich, damit es in 6 Monaten nicht implodiert." Nur so wird aus Copy-KI ein Wettbewerbsvorteil statt einer

Duplicate-Content-Fabrik.

Architekturentscheidungen: RAG, Fine-Tuning, Agenten und Vektor-Datenbanken

Viele Teams springen sofort zu Fine-Tuning, obwohl Retrieval-Augmented Generation, kurz RAG, in 80 Prozent der Fälle die robustere Wahl ist. RAG trennt Modellwissen von Unternehmenswissen, indem es relevante Dokumente zur Laufzeit einblendet, statt das Modell dauerhaft umzuschreiben. Das sorgt für Aktualität, Auditierbarkeit und schnellere Rollbacks, wenn sich Fakten ändern. Fine-Tuning lohnt sich bei Stil-Adaption in großem Stil, bei strukturierten Aufgaben wie Extraktion oder bei Domänenjargon, den das Basismodell systematisch verfehlt. Eine saubere Architektur erlaubt beides: RAG als Default, Feintuning als Turbo für wiederholbare Aufgaben mit klaren Evaluationskriterien.

Die Vektor-Datenbank ist keine Deko, sondern kritischer Pfad. Wähle Embedding-Modelle, die zu deinen Inhalten passen, etwa text-embedding-3-large, E5 oder bge-m3, und sorge für Chunking, das semantisch sinnvoll schneidet statt nur nach Zeichenlänge. Nutze Re-Ranking mit Cross-Encodern, um Präzision zu erhöhen, und implementiere Citations, damit jede Antwort mit Quellen verknüpft ist. Cache Antworten, aber invalidiere aggressiv, wenn Quellen aktualisiert werden. Logge jeden Retrieval-Schritt, denn ohne Observability tappst du im Dunkeln, wenn Antworten driften. Und beseitige PII im Index mit DLP-Filtern, sonst diskutierst du mit der Rechtsabteilung länger als dir lieb ist.

Agenten sind der Versuch, Planung, Tool-Use und Ausführung zu orchestrieren, ohne jedes Mal Prompt-Matryoshkas zu bauen. Frameworks wie OpenAI Assistants, Anthropic Workbench, LangChain, LlamaIndex, AutoGen oder CrewAI ermöglichen Tool-Aufrufe, Dateihandling, Gedächtnis und Multi-Agent-Kollaboration. Der Haken ist Komplexität, die du mit Guardrails und Evals abfedern musst, sonst bauen dir Agenten fröhlich Endlosschleifen. Ein tragfähiger Startpunkt ist ein Single-Agent mit Funktionsaufrufen für Suche, Retrieval, Schreiben, Validierung und Publishing. Erst wenn das stabil ist, lohnt sich Multi-Agent-Koordination mit Rollen wie Researcher, Writer, Editor und QA. So kommst du von “Welche AI gibt es” zu “Welche Orchestrierung kann ich zuverlässig betreiben”.

RAG-Blueprint zum Nachbauen:

- Quellen erfassen, normalisieren, in Chunks schneiden und mit Embeddings indexieren.
- Hybrid-Search kombinieren, Re-Ranking auf Top-K anwenden, Zitationspflicht erzwingen.
- Prompt-Vorlagen mit klaren Output-Schemas nutzen, z. B. JSON oder Markdown-Konventionen.

- Halluzinations-Evals laufen lassen, Antworten mit Score unter Schwelle verwerfen oder neu schreiben.
- Observability mit Langfuse, Humanloop oder PromptLayer einbauen, inklusive Kosten- und Latenz-Logs.

Deployment, Datenschutz und Kosten: DSGVO-konforme AI in der Praxis

Die schönste Demo ist wertlos, wenn du sie rechtlich nicht live bekommst. DSGVO bedeutet Datenminimierung, Zweckbindung, Rechtsgrundlage und Betroffenenrechte, die du auch mit AI erfüllen musst. Lies die Datenverarbeitungsbedingungen deiner Anbieter: Trainieren sie auf deinen Inputs, wie lange speichern sie Logs, in welcher Region liegen die Daten und gibt es Modell-Isolation. Enterprise-Optionen wie Azure OpenAI, Google Vertex AI oder AWS Bedrock bieten VNET-Isolation, Private Endpoints und EU-Residency, kosten aber mehr und brauchen Setup-Disziplin. Self-Hosting mit offenen Modellen auf NVIDIA-GPUs, etwa A100 oder H100, oder auf Managed-GPU-Angeboten ist machbar, erfordert jedoch MLOps-Kompetenz und Security-Hygiene. Am Ende zählt: Kannst du einer Audit-Frage in 48 Stunden sauber antworten, oder nicht.

Kosten steuerst du über Architektur, nicht über Rabatte. Verwende Mid-Tier-Modelle für Massenaufgaben und hebe Premium-Modelle für High-Stakes-Outputs auf, die tatsächlich Marge bewegen. Komprimiere Kontexte aggressiv, nutze Summaries und tool-gestützte Extraktion statt ganzer PDF-Dumps. Messe Zeit-zu-erstem-Token und Ende-zu-Ende-Latenz, nicht nur Modell-Millisekunden, denn Nutzer springen bei trägen Interfaces ab. Caching, Batching und Streaming sind Pflicht, wenn du UX und Kosten parallel optimieren willst. Und ja, Bild- und Video-Generierung frisst Budget schneller als jede Copy-Pipeline, also plane Renderzyklen und Qualitätsstufen wie ein Producer, nicht wie ein Prompt-Poet.

Compliance-Checkliste für den Go-Live:

- Datenflüsse dokumentieren, Kategorien und Speicherorte benennen, Aufbewahrungsfristen definieren.
- DPIA durchführen, Risiken wie Datenleck, Modell-Missbrauch und Bias bewerten und mitigieren.
- Provider-DPA prüfen, Trainings-Opt-out aktivieren, EU-Residency und Logging-Kontrolle sicherstellen.
- Guardrails einbauen: PII-Redaktion, Toxicity-Filter, Prompt-Injection-Schutz wie Rebuff oder NeMo Guardrails.
- Incident-Prozess festlegen, Rollback-Strategien testen und Verantwortlichkeiten klären.

Fazit: Vom Fragezeichen zur produktionsreifen AI-Roadmap

Die Frage “Welche AI gibt es” ist legitim, aber sie ist nur der Startschuss für Entscheidungen, die Technik, Recht und Geschäft zusammendenken. Wer Modelle nach Demos kauft, zahlt später mit Latenz, Fehlerraten und Compliance-Kopfschmerzen. Wer Kategorien versteht, Metriken vergleicht und Architekturen sauber aufsetzt, baut sich eine Pipeline, die zuverlässig liefert und skalierbar bleibt. Der Markt ist laut, aber die Signale sind klar: RAG vor Feintuning, Observability vor Vanity-Metriken, Mid-Tier-Modelle vor All-In auf Premium, und Guardrails als Standard, nicht als Zusatz.

Wenn du heute mit einem strukturierten Raster in Auswahl, Architektur und Betrieb gehst, übersetzt du “Welche AI gibt es” in “Welche AI performt in meinem Stack, unter meinen Regeln, zu meinen Kosten.” Genau das trennt modische Experimente von Wettbewerbsvorteilen. Der Rest ist nur Promo. Und Promo rankt selten langfristig.