

Übersicht KI-Anbieter & Modelle.

In diesem Paper geben wir einen kompakten Überblick über relevante KI-Anbieter und deren gängigste Modelle. Ziel ist es, die jeweiligen Stärken und typischen Einsatzbereiche verständlich einzuordnen und eine praxisnahe Orientierung für den sinnvollen Einsatz von KI-Modellen im Arbeitsalltag zu bieten.

Anbieter: Anthropic

Anthropic ist ein LLM-Anbieter mit Schwerpunkt auf strukturiertem Reasoning über lange Kontexte. Ihre Modelle eignen sich besonders für Aufgaben, bei denen Konsistenz, Argumentationslogik und die Verarbeitung sehr umfangreicher Texte im Vordergrund stehen.

Zentrale Einsatzgebiete

- **Dokumentenanalyse:** Verarbeitung und Analyse sehr großer Dokumente (z. B. Verträge, Compliance-Unterlagen, technische Dokumentationen mit mehreren hundert Seiten).
- **Coding & Debugging:** Analyse komplexer, dateiübergreifender Codebasen sowie Unterstützung bei Architektur- und Designentscheidungen.
- **Konzeptionelles Schreiben:** Erstellung längerer, argumentativ konsistenter Texte wie Fachartikel, Essays oder strategische Inhalte.

Grenzen

- **Zurückhaltendes Antwortverhalten:** Bei sensiblen oder unklaren Fragestellungen teilweise restriktiv.

- **Kostenstruktur & Nutzungslimits:** Überdurchschnittliche Tokenkosten bei gleichzeitig restriktiven Anfrage-Limits in der eigenen Anwendung.
- **Begrenzte Multimodalität:** Starker Fokus auf Text, eingeschränkte Fähigkeiten bei Bild-, Audio- oder Videoaufgaben.

Sprachmodelle von Anthropic

Der Fokus bei Claude liegt auf sauberem Reasoning, Programmierung und der Analyse sehr großer Dokumente. Neben Größenunterschieden bietet Anthropic explizite Reasoning-Varianten an.

Modell	Wofür allgemein bekannt	Konkrete typische Aufgaben im Arbeitsalltag
Claude Opus 4.5	Sehr hohe Reasoning- und Analysefähigkeit	• Tiefe Dokument- und Risikoanalysen • Komplexe Code-Reviews
Claude Opus 4.5 Reasoning	Maximale Denk- und Schlussfolgerungsfähigkeit	• Mehrstufige Problemlösungen • Strategische Bewertungen
Claude Sonnet 4.5	Gute Balance aus Leistung und Kosten	• Wissensdatenbanken • Präzise Unternehmensdokumente
Claude Sonnet 4.5 Reasoning	Fokus auf saubere Argumentationsketten	• Strukturierte Analysen • Entscheidungsunterstützung
Claude Haiku 4.5	Sehr schnell und effizient	• Echtzeit-Chatbots • Supportautomatisierung

Ältere Claude- Generationen	Solide, aber nicht mehr aktuell	• Routineaufgaben
--------------------------------	---------------------------------	-------------------

Für komplexe Analysen und anspruchsvolles Reasoning ist Claude Opus 4.5 die beste Wahl. Claude Sonnet 4.5 eignet sich gut für die tägliche Wissensarbeit, während Claude Haiku 4.5 für schnelle, operative Aufgaben ausgelegt ist. Ältere Claude-Modelle reichen für einfache Routineaufgaben aus, sind für neue Anwendungsfälle jedoch nur eingeschränkt empfehlenswert.

Anbieter: Perplexity

Perplexity AI ist der Anbieter einer KI-gestützte Rechercheplattform mit Fokus auf Echtzeit-Websuche und transparente Quellenangaben. Sie dient primär der faktenbasierten Informationsbeschaffung und weniger der kreativen Textproduktion.

Zentrale Einsatzgebiete

- **Web- und Marktrecherche:** Analyse aktueller Entwicklungen, Trends, Wettbewerber und Marktumfelder.
- **Quellenbasierte Arbeiten:** Bereitstellung von Zitaten, Referenzen und nachvollziehbaren Quellen für faktengetriebene Inhalte.
- **Vergleiche & Übersichten:** Strukturierte Gegenüberstellungen von Anbietern, Technologien oder Produkten.

Grenzen

- **Begrenzte Kreativleistung:** Weniger geeignet für kreatives Schreiben oder umfangreiche Textproduktionen.
- **Kontext- und Token-Limits:** Einschränkungen bei der Verarbeitung sehr langer Dokumente.
- **Abhängigkeit von Online-Quellen:** Qualität der Ergebnisse ist stark von der Verfügbarkeit und Güte öffentlicher Inhalte abhängig.

Modi von Perplexity AI

Perplexity ist kein eigenes KI-Grundmodell, sondern lediglich eine KI-gestützte Such- und Analyseplattform. Der Kern von Perplexity ist die Kombination aus Large Language Models und Echtzeit-Recherche, ergänzt um ein internes Retrieval- und Ranking-System. Ziel ist es, aktuelle, belegbare Antworten mit Quellen bereitzustellen.

Im Unterschied zu klassischen Chatbots steht bei Perplexity Recherche, Verifikation und Aktualität im Vordergrund, nicht primär kreative Textgenerierung.

Modell	Wofür allgemein bekannt	Konkrete typische Aufgaben im Arbeitsalltag
Perplexity (Standard)	KI-gestützte Suche mit automatischen Quellenangaben	• Schnelle Faktenabfragen • Markt- und Wettbewerbsrecherchen • Überblick über neue Themen oder Trends
Perplexity Pro	Zugriff auf leistungsstärkere Modelle und längere Kontexte	• Tiefgehende Fach- und Branchenrecherchen • Analyse umfangreicher Themenfelder

		• Regelmäßige Wissensarbeit mit hohem Aktualitätsbedarf
--	--	---

Perplexity nutzt im Hintergrund verschiedene führende Modelle (u. a. von OpenAI, Anthropic oder Google) und entscheidet je nach Anfrage, welches Modell und welche Quellen geeignet sind. Die eigentliche Mehrleistung entsteht durch die Verknüpfung von LLM-Antworten mit Live-Quellen, nicht durch ein einzelnes Modell. Perplexity eignet sich daher besonders für aktuelle, faktenbasierte Wissensarbeit, bei der Quellenangaben wichtig sind.

Anbieter: Google

Google ist der Anbieter des multimodalen KI-Modells Gemini, sie legen den Schwerpunkt auf die Verarbeitung von Texten, Bildern und Videos sowie auf die Integration des Google-Ökosystems.

Zentrale Einsatzgebiete

- **Bild- und Videoanalyse:** Generierung, Bearbeitung und Auswertung visueller Inhalte.
- **Video-Zusammenfassungen:** Transkription und inhaltliche Analyse längerer Videoformate.
- **Multimodale Workflows:** Verarbeitung großer Informationsmengen in Kombination aus Text, Bild und Video.
- **Google-Integration:** Nutzung direkt in Google Docs, Gmail und Drive zur Unterstützung bestehender Arbeitsprozesse.

Grenzen

- **Halluzinationen bei Wissenslücken:** Plausible, aber nicht immer korrekte Antworten bei unklaren Fragestellungen.
- **Begrenzte fachliche Tiefe:** In spezialisierten Themenfeldern weniger präzise als stark auf Textanalyse fokussierte Modelle.
- **Kontextgröße ≠ Genauigkeit:** Große Kontextfenster garantieren keine konsistente Analysequalität.

Sprachmodelle von Google

Gemini ist sehr beliebt bei Multimedialen Aufgaben und ist im Vergleich zu den Modellen der anderen Anbieter sehr effizient und kostengünstig.

Modell	Wofür allgemein bekannt	Konkrete typische Aufgaben im Arbeitsalltag
Gemini 3 Pro	Aktuelles Spitzenmodell für Analyse und Multimodalität	• Analyse komplexer Texte, Bilder und Diagramme • Recherche anspruchsvoller Themen
Gemini 3 Flash	Sehr schneller Allrounder	• Berichte • Meeting-Zusammenfassungen
Gemini 2.5 Pro	Sehr stark in Multimedia und Analyse	• Einfache Bild- und Diagrammanalyse
Gemini 2.5 Flash	Schnell & kosteneffizient	• Sehr kurze Antwortzeiten bei geringeren Kosten • Geeignet für einfache Text-, Bild- und Analyseaufgaben mit hohem Durchsatz

Für anspruchsvolle Analyse- und Rechercheaufgaben eignet sich Gemini Pro am besten. Für schnelle Alltagsaufgaben und Dokumentenarbeit ist Gemini Flash die

passende Wahl. Ältere oder kleinere Gemini-Varianten reichen für einfache Routineaufgaben aus, spielen bei komplexen Anforderungen jedoch eine untergeordnete Rolle.

Anbieter: OpenAI

OpenAI ist Anbieter des vielseitigen Allround-Modells ChatGPT für Kommunikation, Kreativität und schnelles Prototyping. Es deckt einen breiten Anwendungsbereich ab und eignet sich besonders für den täglichen Einsatz.

Zentrale Einsatzgebiete

- **Alltagskommunikation:** E-Mails, Social-Media-Posts, Textentwürfe und Zusammenfassungen.
- **Kreativität & Ideenfindung:** Brainstorming, Kampagnen- und Konzeptentwicklung.
- **Prototyping:** Schnelle Erstellung von Code-Snippets, kleinen Tools oder MVPs in Kombination mit Text und Bildern.
- **Transkription & Sprachverarbeitung:** Sehr zuverlässige Umwandlung von Audio in Text, auch bei längeren Aufnahmen, Fachsprache oder mehreren Sprechern.

Grenzen

- **Faktenzuverlässigkeit:** Anfällig für Halluzinationen bei spezifischen Daten, Statistiken oder Rechercheaufgaben.

- **Logische Konsistenz:** Bei komplexen mehrstufigen Aufgaben können Widersprüche auftreten.
- **Modellabhängige Qualität:** Leistungsfähigkeit variiert je nach eingesetzter Modellvariante und Denkmodus.

Sprachmodelle von OpenAI

Die Modelle unterscheiden sich sowohl nach Leistungsstufe (Full / mini / nano) als auch nach Denkmodus (Fast / Standard / Thinking).

Modell	Wofür allgemein bekannt	Konkrete typische Aufgaben im Arbeitsalltag
GPT-5.2	Sehr hohe Logik- und Analyseleistung	• Markt- und Wettbewerbsanalysen • Strategiepapiere und Whitepapers • Ausarbeitung komplexer Konzepte
GPT-5.2 Thinking	Explizites, mehrstufiges Reasoning für komplexe Fragestellungen	• Strategische Planung • Entscheidungs- und Risikoanalysen
GPT-5	Leistungsstarkes Allround-Modell der 5.x-Reihe	• Hochwertige Text- und Analysearbeit
GPT-5 Thinking	Reasoning-Variante für strukturierte Problemlösungen	• Mehrstufige Analysen • Logische Herleitungen
GPT-5 mini	Gute Balance aus Qualität, Geschwindigkeit und Kosten	• Berichte und Blogartikel • Zusammenfassungen • Feedback-Analyse

GPT-4o	Multimodal (Text, Bild, Audio), stabil und breit eingesetzt	• Bild- und Screenshot-Analyse • Transkription und Zusammenfassung • Support-Workflows
GPT-4.1	Bewährtes Vorgängermodell für Text und Analyse	• Längere Texte • Strukturierte Recherche
GPT-3.5 (Legacy)	Einfaches, günstiges Basismodell	• Chatbots • Kurze Texte

Für höchste Analyseleistung und komplexe Fragestellungen eignen sich GPT-5.2 und GPT-5.2 Thinking. Für die tägliche Wissensarbeit sind GPT-5 oder GPT-5 mini meist ausreichend, da sie Qualität, Geschwindigkeit und Kosten gut ausbalancieren. Multimodale Aufgaben wie die Kombination von Text, Bild oder Audio lassen sich am besten mit GPT-4o umsetzen. Wenn Kosten und Geschwindigkeit im Vordergrund stehen, sind mini- oder nano-Varianten sinnvoll. Für einfache Routineaufgaben reichen in der Regel ältere Generationen wie GPT-4 oder GPT-3.5 aus.

Anbieter: Mistral AI

Mistral AI ist ein europäischer Anbieter offener, leistungsfähiger Sprachmodelle. Die Modelle können lokal oder in eigener Cloud-Umgebung betrieben und an spezifische Anforderungen angepasst werden. Sie eignen sich besonders für Unternehmen oder Entwickler:innen, die Kontrolle über Daten und Modifikationen benötigen.

Zentrale Einsatzgebiete

- **Eigene KI-Integration & Anpassung:** Modelle können in eigene Systeme integriert und für spezifische Anwendungsfälle feinjustiert werden.
- **Multilinguale Anwendungen:** Unterstützung zahlreicher natürlicher Sprachen und Programmiersprachen für Übersetzungen, Dokumentation oder Textanalyse.
- **Prototyping & Entwicklungsunterstützung:** Einsatz für Code-Snippets, interne Tools, Chatbots oder kundenspezifische Assistenzsysteme.

Grenzen

- **Technischer Aufwand:** Open-Weight-Modelle erfordern Erfahrung bei Deployment, Infrastruktur und Anpassung.
- **Infrastrukturbedarf:** Große Modelle benötigen leistungsfähige GPUs; kleiner Teams kann dies limitieren.
- **Keine Echtzeit-Aktualität:** Modelle haben Trainings-Cutoffs und verfügen standardmäßig über keinen Webzugriff, was aktuelle Informationen einschränkt.

Sprachmodelle von Mistral AI

Mistral ist ein Open-Source-LLM aus Paris, welcher unterschiedlichste Modelle anbietet. Es gibt sehr kostenbewusste Varianten und auch Spezialmodelle für Code etc.

Modell	Wofür allgemein bekannt	Konkrete typische Aufgaben im Arbeitsalltag
Mistral Large	Hohe Analyse- & Logikleistung innerhalb der Mistral-Familie	- komplexe Projektanalysen - Text-zu-Struktur-Extraktion - Schreiben von Konzepten
Mistral Medium	Gute Balance, vielseitig, kosteneffizient	- Prozessdokumentation - Erstellung von Richtlinien - Standardberichte
Mistral Small	Sehr günstig, gut für einfache Texte	- E-Mail-Antwortvorschläge - einfache Erklärtexpte - interne Notizen
Mistral Nemo	Open-Source-orientiert, für On-Premise-Anwendungen geeignet	- interne Wissenssysteme - sichere DSGVO-Workflows - Dokument-Zusammenfassungen
Ministral 8B	Kompaktes Modell mit guter Leistung	- Datenextraktion - kurze Zusammenfassungen - regelbasierte Texttransformation
Pixtral Large	Modernes Multimodal-Modell (Open-Source-nah)	- Bildanalyse - Visualisierungen beschreiben - KI-gestützte Qualitätskontrolle
Pixtral	Vielseitiger Allrounder für Text + Bild	- Auswertung von Screenshots - Erstellen kurzer Bildbeschreibungen
Codestral	Spezialisiert auf Programmierung	- Code-Generierung - automatisches Schreiben von Testfällen - Erkennung von Bugs

Devstral Small	Kleine Variante für Entwickler-Operativarbeit	• Fehlersuche • Code-Erklärungen • Erstellen kleiner Skripte
-----------------------	---	--

Mistral-Modelle sind z. B. dann interessant, wenn Kosten niedrig gehalten werden sollen, Open-Source wichtig ist oder Spezialfälle (wie Code) im Vordergrund stehen.

Kurzfazit

Auf Basis praktischer Tests und realer Anwendungsszenarien hat sich bei uns eine klare Arbeitsteilung zwischen verschiedenen KI-Tools herauskristallisiert. Diese Aufteilung versteht sich als Erfahrungswert und kann je nach Anforderungen und Workflows variieren.

- **ChatGPT** nutzen wir für Strategieentwicklung, Kurzanalysen und schnelle Alltagsaufgaben. Es dient als zentrales Werkzeug für den Großteil der täglichen Arbeit.
- **Claude** setzen wir für längere, durchdachte Texte, umfangreiche Dokumentanalysen sowie Code-Reviews, Bugfixing und Codeerstellung ein.
- **Gemini** kommt bei multimodalen Aufgaben zum Einsatz, insbesondere bei Bildgenerierung sowie Bild- und Videoanalysen.
- **Perplexity** nutzen wir für fundierte Recherchen und Quellanalysen, wenn belastbare und aktuelle Informationen benötigt werden.
- **Mistral** setzen wir gezielt für OCR-Parsing und die Verarbeitung schwer lesbarer oder nicht maschinenlesbarer Dokumente ein.

Diese Arbeitsteilung hat sich in der täglichen Anwendung als effizient erwiesen und trägt zu klaren Zuständigkeiten zwischen den Tools bei. Gleichzeitig zeigen die bisherigen Erfahrungen, dass sich die spannendsten Anwendungsfälle gerade erst herausbilden. Die optimale Kombination kann je nach Kontext unterschiedlich aussehen und wird sich mit der weiteren Entwicklung der Modelle kontinuierlich weiterentwickeln.