Modellbeschreibung

Wann welches Modell?

- **GPT-5:** Höchste Qualität/“Denktiefe” – komplexe Aufgaben, Tools, Agenten, Enterprise.
- **5-mini:** Sehr schnelle “Allrounder”; fast so gut wie “groß”, günstiger.
- **5-nano:** Ultraschnell/günstig für einfache, hohe Volumen-Aufgaben.
- **GPT-4.1:** Sehr langer Kontext (1 Mio Token), ”4er-Flaggschiff”; gut für Code, große Dateien.
- **4.1-mini:** Starke Preis/Leistungs-Wahl, Vision + Text, viel Volumen.
- **4.1-nano:** Minimalste Kosten/Latenz, simple Aufgaben, Klassifikation.
- **o4-mini:** Vision, Audio, Text — Alltags-Produktionsmodell mit schnellen Antworten.
- **o3:** Solide, super breiter Einsatz – kleine Apps, klassische Chat-Bots.

Legende (Skala 1–5)

- □ = Modell-Fähigkeit / Intelligenz. 1 = Smalltalk & einfache Umformulierung; 3 = solide Analyse, Code-Bugfixes; 5 = sehr komplexe Aufgaben (Forschung, Olympiade-Mathe).
- ⚡ = gefühlte Geschwindigkeit (Time-to-First-Token, Durchsatz). Mehr = schneller.
- □ = Reasoning-Stärke (mehrstufiges Denken/Planen). Mehr = „denkt länger“ und robuster.
- □ = Tool-/Multimodal-Fähigkeiten (Web, Code, Dateien, Bilder, Interpreter). Mehr = breiteres Tooling und stabilere Nutzung.
- □ = Kontextfenster (Tokens). Wir zeigen sowohl Icons (relativ) als auch die typische Obergrenze in Zahlen.
- □ = Kosten (relativ). Mehr = teurer.
- □ = Latenz (relativ). Mehr = langsamer/höhere Verzögerung.
- □ = Sicherheit / Halluzinations-Reduktion. Mehr = sicherer/stabiler.

Beispiel zur Einordnung:

- 1 Icon ≈ Einstiegsniveau / sehr günstig / sehr schnell (aber begrenzter Umfang)
- 3 Icons ≈ Allrounder
- 5 Icons ≈ Spitzenklasse / teuer / kann lange „denken“

Skalenvergleich (1–5, mehr Icons = stärker/teurer/langsamer je nach Metrik)

Modell	**Fähigkeit**	**Speed**	**Reasoning**	**Tools**	**Kontext**	**Kosten**	**Latenz**	**Sicherheit**
GPT-5	⭐⭐⭐⭐⭐	⚡⚡	⭐⭐⭐⭐⭐	⭐⭐⭐⭐⭐	⭐⭐⭐⭐ (≈400k)	⭐⭐⭐⭐	⭐⭐⭐	⭐⭐⭐⭐⭐
GPT-5 mini	⭐⭐⭐⭐	⚡⚡	⭐⭐⭐⭐	⭐⭐⭐⭐⭐	⭐⭐⭐⭐	⭐⭐	⭐⭐	⭐⭐⭐⭐
GPT-5 nano	⭐⭐⭐	⚡⚡⚡⚡	⭐⭐⭐	⭐⭐⭐⭐	⭐⭐⭐⭐	⭐	⭐	⭐⭐⭐
GPT-4.1	⭐⭐⭐⭐	⚡⚡⚡	⭐⭐⭐⭐	⭐⭐⭐⭐⭐	⭐⭐⭐⭐⭐ (≈1M)	⭐⭐⭐	⭐⭐⭐	⭐⭐⭐⭐
GPT-4.1 mini	⭐⭐⭐	⚡⚡⚡⚡	⭐⭐⭐	⭐⭐⭐⭐⭐	⭐⭐⭐⭐⭐ (≈1M)	⭐⭐	⭐⭐	⭐⭐⭐⭐
GPT-4.1 nano	⭐⭐⭐	⚡⚡⚡⚡⚡	⭐⭐	⭐⭐⭐⭐	⭐⭐⭐⭐⭐ (≈1M)	⭐	⭐	⭐⭐⭐
OpenAI o3	⭐⭐⭐⭐	⚡⚡	⭐⭐⭐⭐⭐	⭐⭐⭐⭐⭐	⭐⭐⭐ (≈128k)	⭐⭐⭐⭐⭐	⭐⭐⭐⭐⭐	⭐⭐
OpenAI o4-mini	⭐⭐⭐⭐⭐	⚡⚡⚡⚡	⭐⭐⭐⭐	⭐⭐⭐⭐⭐	⭐⭐⭐ (≈128k)	⭐⭐⭐	⭐⭐	⭐⭐⭐

Warum so bewertet? (Kurzprofile + konkrete, leicht verständliche Beispiele)

- GPT-5 (Flaggschiff, 400k Kontext, volle Tools): Für „Agenten“-Aufgaben mit langen Tool-Ketten. Beispiel: „Baue mir ein kleines Frontend, lies meine Firmenrichtlinie (200 Seiten) und verdrahte beides zu einem Daten-Eingabe-Flow“—GPT-5 bewirbt bessere Tool-Ketten („long chains of tool calls“) und führt neben Vision auch Web/File-Suche; Preise: $1.25/$10 pro 1M Token; Kontext: 400k.
- GPT-5-mini / GPT-5-nano (gleiche API-Features, günstiger/schneller): Für „viel Volumen, ordentliche Qualität“. Beispiel: „Tausende Support-Chats schnell vorsortieren, bei kniffligen Fällen Tools nutzen“. Preise und Kontext (400k) laut.
- GPT-4.1 (1M-Kontext, stark bei Code/Instruktionen): Für „lange Dokumente + präzise Format-Befehle“. Beispiel: „Fasse 3 komplette Forschungsberichte zusammen (insg. ~600k Tokens) und generiere ein sauberes XML nach Schema“.
- GPT-4.1-mini (schneller/billiger; 1M-Kontext): Für „Alltags-Automationen“ wie „Markdown-Berichte generieren, Code-Diffs nach Schema erzeugen“. Mini reduziert Latenz und Kosten vs 4o deutlich.
- GPT-4.1-nano (sehr schnell/sehr günstig; 1M-Kontext): Für „Autovervollständigen/klassifizieren“. Beispiel: „Auto-Tagging von E-Mails in Near-Realtime“. Offiziell: schnellstes/ günstigstes 4.1-Modell; MMLU 80.1% etc.
- o3 (Reasoning-Spezialist; „denkt länger“, volle Tools in ChatGPT): Für „mehrstufige, nicht offensichtliche Antworten“. Beispiel: „Webrecherche + Python-Forecast + Diagramm + Begründung in einem Rutsch“. OpenAI beschreibt o3 als stärkstes Reasoning-Modell mit weniger schweren Fehlern als o1; agentische Tool-Nutzung.
- o4-mini (kleiner, schneller Reasoner; 128k Kontext): Für „viele technisch-knackige Fragen, wenig Latenz“. Beispiel: „AIME-ähnliche Mathe, kurz rechnen (ggf. mit Python), Antwort darlegen“. Offiziell: bestes kleines Reasoning-Modell; in ChatGPT/ API mit Tool-Zugriff; 128k Kontext (wie o3).

Quelle:

- https://openai.com/gpt-5
- https://openai.com/index/gpt-4-1/
- https://openai.com/index/introducing-o3-and-o4-mini/
- https://openai.com/api/pricing/?Tag=Heat%25252520Map&utm_source=chatgpt.com|Pricing-Seite
- https://help.openai.com/en/articles/10491870-o4-mini-in-chatgpt-faq?utm_source=chatgpt.com

Praktische Mini-Beispiele (wann welches Modell)

- „Bitte fasse 300 Seiten Verträge mit Zitaten zusammen und extrahiere Kennzahlen.“ → GPT-4.1 / 4.1 mini (wegen 1M-Kontext). GPT 4.1
- „Schwere Mathe-/Programm-Knobelei mit Web-/Code-Toolkette (Plot bauen, Ergebnis prüfen).“ → o3; budgetschonender: o4-mini. o3 & o4-mini Einführung, o4-mini help
- „Produktforschung + Schreiben + Grafiken in einem End-to-End-Agenten.“ → GPT-5. GPT 5 für Entwickler
- „Schnelle UI-Antworten, Autovervollständigen, massenhaft Klassifikation.“ → 4.1 nano oder 5 nano (sehr günstig/schnell). GPT 4.1
- „Tool-reiche Assistenten (Browsing, Dateien, Vision) mit gutem Preis-Leistungs-Verhältnis.“ → 4.1 mini, 5 mini oder o4-mini je nach Reasoning-Tiefe und Budget. GPT 4.1, o4-mini help