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.

GPT-4.1-Serie vs. GPT-5-Serie

Modell	Kerneigenschaft (in einem Satz)	Konkretes Alltags-Beispiel (einfach verständlich)
GPT-4.1	Starker Allrounder für Coding, lange Kontexte (bis zu 1M Tokens), gute Instruktions-Folge	„Suche in einem ganzen Buch nach allen Erwähnungen eines Fachbegriffs und fasse Kapitel 3 zusammen“ — macht das zuverlässig.
GPT-4.1-mini	Kompakter, schneller, günstigere API-Option für viele Aufgaben	„Automatisch E-Mails klassifizieren und Schlagwörter vorschlagen“ — schnell + günstig.
GPT-4.1-nano	Sehr klein und extrem schnell; ideal für einfache Klassifikation/Autovervollständigung	„Kurzantworten oder Kategorie-Tags für Support-Tickets“ — sehr günstig pro Anfrage.
GPT-5	Größere „Denktiefe“ und Router/Thinking-Modi; besser bei komplexen Analysen, Multimodalität	„Mehrstufiges Recherche-Projekt: Web-Abfragen + Code zum Auswerten von Daten + Ergebnisbericht“ — kann Tools/Plugins nutzen und länger denken.
GPT-5-mini	Mini-Variante von GPT-5 — schneller, für viele Nutzer-Anfragen, behält wichtige Verbesserungen	„Schnelle interne Zusammenfassungen großer Dokumente für Team-Chats“.
GPT-5-nano	Nano-Variante für sehr hohes Volumen / einfache Tasks; günstigste GPT-5-Option	„Automatische Stichwort-Extraktion in hohen Volumina (Logs, Chats)“.

- Quellen/Belegbeispiele: Offizielle OpenAI-Ankündigung zu GPT-4.1 und deren Benchmarks;

Offizielle GPT-5-Seite mit Beispielen zu Thinking/Router.

o-Serie: o3 und o4-mini — praxisorientierte Unterschiede

Modell	Typischer Einsatz (Warum man es wählt)	Konkretes Beispiel
o3	Höchstleistung beim logischen, multimodalen Problemlösen; starkes Reasoning, gut für Forschung & Programmier-Debugging	„Analysiere ein technisches Paper, teste Hypothesen mit Python-Schnipseln und generiere reproduzierbare Plots“ — gut für Forschungsteams.
o4-mini	Optimierte, kosteneffiziente Reasoning-Variante; sehr schnell und gut bei Mathe/Coding mit Toolzugriff	„Löse viele Mathe-Aufgaben (AIME-ähnlich) unter Verwendung eines eingebetteten Python-Interpreters“ — sehr gutes Preis/Leistungsergebnis.

- Quellen (Beispiele in den Tabellen stammen von den offiziellen Produkt-Ankündigungen):
  - OpenAI – Introducing GPT-4.1 (Details zu 4.1 / mini / nano und Benchmarks).
  - OpenAI – Introducing GPT-5 (Beispiele: Thinking, Router, Einsatzszenarien).
  - OpenAI – Introducing o3 and o4-mini (Reasoning, Tool-Use, AIME / Benchmarks).
- Praxis-Unterschied kurz: Wenn du tiefes, mehrstufiges Nachdenken und flexible Tool-Ketten brauchst (z. B. Web-Suche, Datei-Auswertung, Python), wähle o3; wenn Sie hohe Nachfrage, viele Anfragen und trotzdem starkes Reasoning wollen, ist o4-mini oft besser (kosteneffizient). Beispiele und Messungen (z. B. AIME / Coding Benchmarks) stammen aus der OpenAI-Ankündigung zu o3/o4-mini.