M17b Vergleich von GPT-4o, GPT-4-o1 und GPT-4-o3

1 Technologische Merkmale

- **Modellarchitektur & Ansatz:**
  GPT-4o ist ein großer generativer Transformer (LLM), der als Haupt-ChatGPT-Modell dient. Es basiert auf der GPT-4-Architektur und wurde optimiert für hohe Allround-Intelligenz (Pricing | OpenAI). **GPT-4-o1 (OpenAI o1)** dagegen ist ein **„reflective" LLM**, das *Chain-of-Thought*-Techniken nutzt – es „denkt" zunächst intern in mehreren Schritten, bevor es antwortet (OpenAI o1 - Wikipedia). Dadurch kann es komplexe Probleme systematischer lösen (insbesondere in MINT-Bereichen) (OpenAI's STEM-Optimized 'Reasoning' Model, o1, Sees Daylight -- Pure AI). **GPT-4-o3 (OpenAI o3)** ist die nächste Generation dieser Reasoning-Modelle. Es baut auf o1 auf und zielt auf noch leistungsfähigeres logisches Denken ab, jedoch mit verschiedenen Größenvarianten (z. B. **o3-mini** als kleiner, schneller Ableger) (OpenAI o3-mini | OpenAI) (OpenAI o3-mini | OpenAI). Alle drei Modelle sind Transformer-basierte KI-Sprachmodelle; o1 und o3 sind jedoch besonders auf *tiefgehende Schlussfolgerungen* trainiert, während GPT-4o auf allgemeine Vielseitigkeit optimiert ist (OpenAI's STEM-Optimized 'Reasoning' Model, o1, Sees Daylight -- Pure AI) (Pricing | OpenAI).
- **Eingabe/Output-Fähigkeiten & Multimodalität:**
  GPT-4o ist **multimodal** und kann Text, Bilder und (über integrierte Tools) Audio verarbeiten. Es gilt als aktuell fortschrittlichstes multimodales Modell von OpenAI und besitzt **starke Visions-Fähigkeiten** (Azure OpenAI Service - Pricing | Microsoft Azure). (In ChatGPT Plus kann GPT-4o z. B. Bilder analysieren und per Sprachmodus Audio ausgeben.) **GPT-4-o1** beherrscht Text und – in der vollwertigen Version – ebenso Bildinputs (Visuelles Verständnis) (OpenAI o1 explained: Everything you need to know). Zum Start fehlte o1 jedoch noch die Audio-Unterstützung und Features wie Web-Browsing (OpenAI's STEM-Optimized 'Reasoning' Model, o1, Sees Daylight -- Pure AI) (OpenAI's STEM-Optimized 'Reasoning' Model, o1, Sees Daylight -- Pure AI). **GPT-4-o3** (insbesondere o3-mini) fokussiert primär auf Text; die **Visions-Funktion ist hier (noch) nicht aktiviert** (OpenAI o3-mini | OpenAI). Stattdessen liegt der Schwerpunkt auf strukturierten Textausgaben und Entwicklertools. O3-mini ist das erste kleine Reasoning-Modell mit **Function Calling** und strukturierten Ausgaben (OpenAI o3-mini | OpenAI). Insgesamt bietet GPT-4o die breiteste I/O-Palette (Text, Bild, teils Audio), während o1 und o3 sich (derzeit) auf Text und logische Ausgaben konzentrieren – Bilder werden bei Bedarf über o1 verarbeitet (OpenAI o3-mini | OpenAI).
- **Geschwindigkeit & Latenz:**
  GPT-4o wurde im Laufe von 2024 beschleunigt und reagiert etwa **doppelt so schnell wie frühere GPT-4-Versionen** (OpenAI o1 explained: Everything you need to know). Es

ist für Echtzeit-Dialoge in ChatGPT optimiert. **GPT-4-o1** benötigt durch das aufwändige Nachdenken etwas mehr Zeit pro Antwort (OpenAI o1 explained: Everything you need to know). Gerade bei **"o1-pro"** (der High-Compute-Variante) können Antworten länger dauern, da das Modell intensiver rechnet, um höchste Genauigkeit zu erzielen. **GPT-4-o3** adressiert die Latenz, insbesondere in der Mini-Variante: o3-mini ist **kleiner und schneller** als o1, bietet wählbare *Reasoning-Tiefen* (Low/Medium/High) je nach gewünschter Geschwindigkeit vs. Gründlichkeit (OpenAI o3-mini | OpenAI) (OpenAI o3-mini | OpenAI). In der Standardeinstellung (mittlere Denktiefe) liefert o3-mini ähnlich präzise Ergebnisse wie o1, aber **schneller** (OpenAI o3-mini | OpenAI). Für sehr komplexe Anfragen kann man o3 auch in einen High-Effort-Modus versetzen (auf Kosten höherer Latenz). Insgesamt: GPT-4o ist flott und interaktiv, o1 eher langsamer, und o3-mini kombiniert hohe Geschwindigkeit mit guter Denkleistung (o3-high wäre wiederum langsamer, aber besonders genau).

- **Kontextlänge & Speicher:**
  Alle drei Modelle unterstützen sehr große Kontextfenster. **GPT-4o** kann Eingaben bis zu **128.000 Token** Kontext verarbeiten (Azure OpenAI Service - Pricing | Microsoft Azure) (deutlich mehr als ältere GPT-3.5-Modelle). **GPT-4-o1** wurde anfänglich mit ähnlichem Kontext (128K) eingeführt (OpenAI o1 explained: Everything you need to know), doch die **API-Variante von o1** bietet inzwischen sogar bis zu **200.000 Tokens Kontext** (Pricing | OpenAI). Das heißt, o1 kann extrem lange Dokumente/Chats im Prompt halten – nützlich für umfangreiche technische Analysen. Auch **GPT-4-o3** setzt diese Linie fort: o3-Modelle unterstützen laut OpenAI ebenfalls **Kontext bis 200K Token** (Pricing | OpenAI). Damit eignen sich o1/o3 hervorragend für Aufgaben, die sehr viel Input erfordern (z. B. lange Berichte, Codebasen). GPT-4o's 128K sind für die meisten Anwendungen schon groß, o1/o3 gehen im „Frontier"-Einsatz noch darüber hinaus.
- **Besondere Funktionen & Tools:**
  GPT-4o als ChatGPT-Modell hat breiten Feature-Support: es kann in ChatGPT z. B. auf Web-Browsing oder Code Interpreter (eingebetteter Python) zurückgreifen und unterstützt die API-Funktion *Function Calling*, um strukturierte Ergebnisse zu liefern. **GPT-4-o1** ist ebenfalls dafür ausgelegt, mit Tools zu arbeiten – OpenAI beschreibt o1 als **„Frontier"-Reasoning-Modell mit Tool-Unterstützung und strukturierten Outputs** (Pricing | OpenAI). Allerdings standen in der Preview-Phase einige GPT-4o-Features (z. B. Surfen im Web) in o1 nicht zur Verfügung (OpenAI o1 explained: Everything you need to know). **GPT-4-o3** schließt diese Lücke weiter: o3-mini unterstützt ab Start Function Calling, strukturierte Ausgaben und Streaming (OpenAI o3-mini | OpenAI). Außerdem wurde experimentell eine **Suchfunktion** integriert, sodass o3-mini bei Bedarf Web-Ergebnisse mit Zitaten liefern kann (OpenAI o3-mini | OpenAI). Unterm Strich: GPT-4o bietet das **vollständige ChatGPT-Erlebnis** mit allen Integrationen; o1/o3 sind spezialisierter auf reasoning-lastige Aufgaben, unterstützen aber zunehmend auch Entwickler-Features (API) und werden laufend erweitert.

2 Typische Einsatzszenarien

- **GPT-4o:** Dieses Modell wird für **Alltagsanwendungen** und generelle KI-Aufgaben eingesetzt. Typische Szenarien sind z. B. das **Verfassen von Texten** (Aufsätze, Blogbeiträge, kreative Geschichten), **Beantworten von Wissensfragen** und Zusammenfassungen, **Übersetzungen** oder auch einfach interaktive **Dialogassistenz**. Dank seiner breiten Wissensbasis und Spracheleganz eignet sich GPT-4o auch für **kreative Aufgaben** wie das Schreiben von Gedichten oder Marketing-Texten. In der Programmierung kann GPT-4o ebenfalls helfen (Code-Snippets generieren, Erklärungen liefern), wenngleich für sehr komplexe Programmieraufgaben die speziellen o-Modelle oft noch präziser sind. Kurz: GPT-4o ist das **Universalmodell für allgemeine Zwecke** – empfohlen für die meisten Routine-Texterstellungen und Dialoge (Pricing | OpenAI).
- **GPT-4-o1:** OpenAI o1 glänzt vor allem in **anspruchsvollen, mehrstufigen Aufgaben** und technisch-wissenschaftlichen Anwendungen (Pricing | OpenAI). Typische Einsatzszenarien sind z. B. **komplexe Mathematikprobleme**, Beweise oder logische Rätsel, bei denen schrittweises Argumentieren wichtig ist. Ebenso wird o1 für **Programmierung** genutzt, insbesondere zum **Debuggen oder Lösen schwieriger Coding-Challenges**, da es systematisch den Code durchdenken kann. Im Bereich **Wissenschaft** (Physik, Chemie, Biologie) kann o1 fachliche Fragen beantworten oder Probleme lösen – OpenAI berichtet, dass o1 auf PhD-Niveau in MINT-Fragen agieren kann (OpenAI's STEM-Optimized 'Reasoning' Model, o1, Sees Daylight -- Pure AI). Auch für Unternehmen, die **lange Analysen oder Berichte** durch das Modell prüfen lassen wollen, ist o1 geeignet (Stichwort: großer Kontext von 200k Tokens). Darüber hinaus bietet o1 erhöhte **Zuverlässigkeit in sensiblen Bereichen**: z. B. für medizinische oder rechtliche Auskünfte, wo man Wert auf korrekte Schritte legt, kann o1 vorteilhaft sein. Insgesamt kommt GPT-4-o1 immer dann zum Einsatz, wenn **gründliches logisches Denken** gefragt ist und einfache KI-Antworten an ihre Grenzen stoßen.
- **GPT-4-o3:** Da GPT-4-o3 die Weiterentwicklung von o1 ist, deckt es ähnliche Szenarien ab, jedoch mit Verbesserungen in **Geschwindigkeit und Zugänglichkeit**. **OpenAI o3-mini** (die kompakte Variante) wird empfohlen für **technische Domänen** wie **Coding, mathematische Aufgaben und logische Problemlösung** – überall dort, wo Präzision und zugleich schnelle Antworten benötigt werden (OpenAI o3-mini | OpenAI) (OpenAI o3-mini | OpenAI). Durch die effizientere Architektur eignet sich o3-mini sogar für **Echtzeit-Dialoge in technischen Support- oder Lern-Assistenten**, da es weniger Latenz hat, aber dennoch verlässliche Zwischenschritte liefert. Große Unternehmen oder Forschungsprojekte könnten ein vollausgebautes o3-Modell (mit hoher Reasoning-Tiefe) für **hochkomplexe Analysen** einsetzen – etwa KI-gestützte Forschung, die mehrere Hypothesen durchdenken soll. Ein weiterer Anwendungsfall ist der **API-Einsatz bei spezialisierten Apps**: Entwickler könnten o3-mini nehmen, um z. B. **Code-Autokomplettierungen oder mathematische Tutor-Systeme** zu bauen, da es eine gute Balance aus Leistung und Kosten hat. Zudem testet OpenAI mit o3 die Integration von **aktueller Websuche**, was es attraktiv für **Faktenfragen mit aktuellen Bezügen** macht (OpenAI o3-mini | OpenAI). Zusammengefasst dient GPT-4-o3 heute vor allem als **schneller Problemlöser** für MINT-Aufgaben und wird perspektivisch überall dort Verwendung finden, wo man die Stärken von o1 mit mehr Effizienz kombinieren möchte.

3 Vor- und Nachteile in den Einsatzszenarien

GPT-4o (Allround-ChatGPT): *Vorteile:* Sehr **breites Allgemeinwissen** und vielseitige Sprachfähigkeiten (gut für offene Dialoge, kreatives Schreiben etc.). Dank Multimodalität kann es **Bilder verstehen** und vielfältige Aufgaben (Übersetzen, Zusammenfassen, Texte in Ton umwandeln) in einem Tool erledigen (Azure OpenAI Service - Pricing | Microsoft Azure). Zudem ist es inzwischen recht **schnell** und in den meisten Anwendungen ausreichend präzise. Nicht zuletzt ist GPT-4o **weit verbreitet und leicht zugänglich** (ChatGPT-Plus, demnächst auch als Standard in vielen Apps) (OpenAI o1 explained: Everything you need to know). *Nachteile:* Bei **komplexer Problemlösung** (z. B. anspruchsvolle Mathematik oder mehrstufiges logisches Schließen) erreicht GPT-4o nicht die Tiefe der spezialisierten Modelle – es kann zu Fehlern oder Halluzinationen kommen, wo o1/o3 systematischer vorgehen würden (OpenAI's STEM-Optimized 'Reasoning' Model, o1, Sees Daylight -- Pure AI). In sicherheitskritischen Fällen ist GPT-4o zwar gut, aber **weniger robust gegen Jailbreaks** als o1 (OpenAI's STEM-Optimized 'Reasoning' Model, o1, Sees Daylight -- Pure AI). Außerdem hat GPT-4o gewisse Einschränkungen: Das Modell hält sich strikt an erlernte Sicherheitsvorgaben, was teils zu Ablehnungen führen kann, und es fehlen ihm die neuesten Spezialoptimierungen für z. B. detailliertes Planen. Auch kostet es im API-Einsatz mehr als kleinere Modelle. Zusammengefasst ist GPT-4o **überall „gut bis sehr gut", aber nicht in jeder Nische das Optimum**.

GPT-4-o1 (Reasoning-Modell): *Vorteile:* Herausragend in **Leistungsfähigkeit auf schwierigen Aufgaben** – o1 erreicht z. B. ~86 % in Mathematik-Olympiade-Benchmarks (vs. 13 % bei GPT-4o) (OpenAI o1 explained: Everything you need to know). Es kann **komplizierte Fragen schrittweise und korrekt lösen**, was in Bereichen wie Programmierung (Algorithmus-Challenges) oder Wissenschaft enorme Mehrwerte bringt (OpenAI's STEM-Optimized 'Reasoning' Model, o1, Sees Daylight -- Pure AI). Auch im Bereich **Safety/Alignment** hat o1 Vorteile: Es wurde darauf trainiert, die OpenAI-Regeln aktiv in seine Gedankenschritte einzubeziehen, was zu höherer **Robustheit gegen Fehlverhalten** führt (OpenAI's STEM-Optimized 'Reasoning' Model, o1, Sees Daylight -- Pure AI). Tests zeigen, dass o1 riskante Anfragen viel konsequenter ablehnt bzw. sicher beantwortet (z. B. in einem Sicherheits-Benchmark erzielte o1 eine 0,92 Nicht-Unsicherheits-Rate vs. 0,713 bei GPT-4o) (OpenAI o1 explained: Everything you need to know) (OpenAI o1 explained: Everything you need to know). Darüber hinaus kann o1 mit seinem großen Kontext **sehr lange Inhalte** verarbeiten – ideal, um z. B. technische Dokumentationen oder ganze Bücher zu analysieren. *Nachteile:* o1 ist **langsamer und rechenintensiver** – im Chat merkt man die Verzögerung durch das „Nachdenken" (OpenAI o1 explained: Everything you need to know), was in Echtzeit-Anwendungen hinderlich sein kann. Es ist zudem **stark auf STEM-Themen optimiert**, sodass es bei alltäglichen Smalltalk-Themen oder kreativ-literarischen Aufgaben manchmal steifer oder weniger ausdrucksstark wirkt (das kleinere o1-mini hatte z. B. Schwächen bei breit gefächertem Weltwissen) (OpenAI's STEM-Optimized 'Reasoning' Model, o1, Sees Daylight -- Pure AI). Ferner ist die **Verfügbarkeit** eingeschränkt: o1 war lange nur für Plus/Pro-Nutzer oder ausgewählte API-Kunden

zugänglich (OpenAI o1 explained: Everything you need to know). Auch die Kosten sind hoch – insbesondere über die API ist o1 deutlich teurer als GPT-4o (OpenAI o1 explained: Everything you need to know). Schließlich muss man beachten, dass o1 in frühen Versionen nicht alle Features (wie Surfen) bot, wobei neuere Iterationen diese Lücken teils schließen. Fazit: GPT-4-o1 ist **unschlagbar bei kniffligen Aufgaben** und in Sachen Verlässlichkeit/Sicherheit top, erkauft dies aber mit Geschwindigkeitseinbußen, höherem Preis und etwas geringerer Allgemeinflexibilität.

GPT-4-o3 (Fortgeschrittenes Reasoning-Modell): *Vorteile:* o3 kombiniert viele Stärken von o1 mit Verbesserungen. Die **Genauigkeit bei technischen Aufgaben** ist sehr hoch – o3-mini erreicht mit mittlerem Aufwand bereits die Leistung von o1 in Mathe, Coding und Wissenschaft (OpenAI o3-mini | OpenAI). Gleichzeitig bietet es **geringere Latenz** und höhere Effizienz, besonders in der Mini-Variante, was es alltagstauglicher macht (z. B. für interaktive Beratungsbots in Technikfragen). Entwickler profitieren von den von Anfang an integrierten Features (Function Calling, strukturierte Ausgabe), was o3 **leicht in Anwendungen integrierbar** macht (OpenAI o3-mini | OpenAI). Ein großer Vorteil ist auch die **Skalierbarkeit nach Bedarf**: Durch die wählbaren Reasoning-Modi kann man je nach Szenario Geschwindigkeit oder Gründlichkeit priorisieren (OpenAI o3-mini | OpenAI). Außerdem democratisiert OpenAI o3-mini etwas die Nutzung – erstmals bekam sogar die kostenlose ChatGPT-Version Zugang zu einem Reasoning-Modell (OpenAI o3-mini | OpenAI). *Nachteile:* Noch ist GPT-4-o3 **relativ neu** – die vollausgereifte große o3-Version (ohne „mini“) ist Stand Anfang 2025 ggf. noch nicht allgemein verfügbar, d.h. man arbeitet vorwiegend mit o3-mini als Vorgeschmack. Dieses kleinere Modell ist zwar flink, hat aber (wie o1-mini) einen **engeren Wissensfokus** – für breit angelegte Konversations-KI mit emotionaler Intelligenz oder Kreativität ist weiterhin GPT-4o besser geeignet. Aktuell **unterstützt o3 keine Bild-Eingaben** (OpenAI o3-mini | OpenAI), was einen Nachteil gegenüber GPT-4o und o1 darstellt, falls multimodale Fähigkeiten gefragt sind. In puncto **Kosten** liegt o3-mini zwar deutlich unter GPT-4o, allerdings erreicht es (insbesondere im High-Effort-Modus) noch nicht die absolute Spitzenleistung von einem hypothetischen großen o3-Modell – sprich, wer maximale Qualität will, müsste auf zukünftige o3-Varianten warten oder hohen Rechenaufwand investieren (im Extremfall wurden für „o3 High“ interne Kosten von bis zu $20k pro komplexer Aufgabe berichtet (- … | Dan Leteky)). Zusammengefasst bietet GPT-4-o3 derzeit **exzellente Leistung bei besserer Effizienz**, ist aber in der vollständigen Ausbaustufe noch im Kommen und hat in Randbereichen (Multimodalität, Kreativität) gewisse Limitierungen.

4 Besonderheiten im Vergleich

- **Leistungsfähigkeit & Genauigkeit:** In logisch-mathematischen Aufgaben liegen GPT-4-o1 und o3 klar vor GPT-4o. O1 hat gezeigt, dass es **menschliches Expertenniveau** in vielen STEM-Benchmarks erreicht (z. B. Top-500 bei der USAMO-Qualifikation und übertrifft PhD-Level in Wissenschaftsfragen) (OpenAI's STEM-Optimized 'Reasoning' Model, o1, Sees Daylight -- Pure AI). GPT-4o ist zwar stark, aber eher generalistisch – in einem direkten Vergleich schnitt o1 bei einem schwierigen Mathematiktest z. B.

sechsmal besser ab (OpenAI o1 explained: Everything you need to know). GPT-4-o3 baut die Führung weiter aus: interne Tests (ARC-AGI Evaluations) zeigten, dass ein **o3-High** Modell bis zu 88 % der Aufgaben löst, während o1 (High) nur ~32 % schaffte (- … | Dan Leteky). Wichtig ist jedoch der Kontext: Bei **kreativen oder offenen Dialogen** kann GPT-4o oft geschmeidiger und breiter antworten, da o1/o3 stark formalisiert denken. Hier behält GPT-4o die Nase vorn in Natürlichkeit und Vielfalt der Antworten. Kurz: Für *Fachprobleme* sind o1/o3 überlegen, für *Generelles* ist GPT-4o meist ausreichend bzw. passender.

- **Verfügbarkeit & Zugang:** GPT-4o ist das Standardmodell hinter ChatGPT (für Plus-Nutzer seit 2023, und via API breit zugänglich), daher **am verfügbarsten**. Viele Dienste (z. B. Bing Chat) nutzen GPT-4-Varianten, was GPT-4o allgegenwärtig macht. GPT-4-o1 wurde anfänglich nur im **Preview** an zahlende Nutzer (ChatGPT Plus/Team) ausgerollt (OpenAI o1 - Wikipedia) und im Dezember 2024 vollständig veröffentlicht (OpenAI o1 - Wikipedia) (OpenAI o1 - Wikipedia) – seitdem ist es für ChatGPT Pro-Abonnenten (mit o1-pro) und teils Plus-Nutzer verfügbar. Über die API ist o1 als kostenintensives Modell verfügbar, jedoch eher für Enterprise-Tier-Kunden oder auf Anfrage (OpenAI positioniert es als *Frontier Model*). GPT-4-o3 befindet sich noch in Einführung: Die **kleine Version o3-mini** ist seit Januar 2025 für ChatGPT Plus/Pro und sogar eingeschränkt für Free-User nutzbar (OpenAI o3-mini | OpenAI). Größere o3-Modelle sind perspektivisch zu erwarten, aber aktuell vor allem für ausgewählte Entwickler (API-Tiers 3–5) freigeschaltet (OpenAI o3-mini | OpenAI). Insgesamt gilt: GPT-4o **breit verfügbar**, o1 **limitiert (Pro/User mit Berechtigung)**, o3 **im Übergang** (Mini-Version offen für viele, Full-Version noch begrenzt).
- **Kosten (API-Nutzung):** Die unterschiedlichen Ausrichtungen spiegeln sich deutlich in den API-Preisen wider. GPT-4o ist deutlich teurer als gängige GPT-3.5-Modelle, aber immer noch günstiger als das spezialisierte o1. Als Richtwert kostet GPT-4o etwa **$10 pro 1 Mio. Ausgabe-Tokens** (Pricing | OpenAI). OpenAI o1 liegt etwa bei **$60/Mio Tokens (Output)** (Pricing | OpenAI) – also etwa das 6-fache von GPT-4o, was seine Premium-Nische unterstreicht (OpenAI o1 explained: Everything you need to know). OpenAI o3-mini hingegen ist sehr kostengünstig: ca. **$4.40/Mio Tokens** (Pricing | OpenAI), also weniger als die Hälfte von GPT-4o, was es attraktiv für häufige Aufrufe macht. Diese Preise bedeuten: Für einfache Anwendungen ist GPT-4o in puncto Preis-Leistung gut, während o1 wirklich nur dort eingesetzt wird, wo die Mehrleistung die Mehrkosten rechtfertigt. O3-mini bietet einen **neuen Kompromiss** – viel Reasoning-Power für wenig Geld – und dürfte daher in 2025 an Bedeutung gewinnen. Unternehmen müssen also zwischen **Kosten und erforderlicher Genauigkeit** abwägen: Warum $60 zahlen, wenn $10 genügen – oder $10 zahlen, wenn auch $4 reichen? (- … | Dan Leteky).
- **Zusammenfassung der Unterschiede:** GPT-4o, GPT-4-o1 und GPT-4-o3 sind **komplementäre Varianten** des ChatGPT-Modellspektrums. GPT-4o ist der *Allrounder* – multimodal, allgemeinwissend, schnell und relativ günstig – ideal für die meisten Chats und Texte. GPT-4-o1 ist der *Spezialist* – langsam aber gründlich, teuer aber dafür in der Lage, Aufgaben auf expertenhaftem Niveau zu lösen (vor allem in Mathematik, Coding,

Naturwissenschaft) (OpenAI's STEM-Optimized 'Reasoning' Model, o1, Sees Daylight -- Pure AI). GPT-4-o3 schließlich versucht, das Beste aus beiden Welten zu vereinen: *leistungsfähig wie o1*, aber *effizienter und zugänglicher*, besonders in der Mini-Version (OpenAI o3-mini | OpenAI). Im direkten Vergleich untereinander zeigt sich, dass keine Variante „perfekt“ für alles ist – es kommt auf den Anwendungsfall an. Für einen kreativen Dialog oder breite Wissensfragen nimmt man GPT-4o; für einen kniffligen Programmierwettbewerb oder wissenschaftlichen Problemlöser greift man zu o1/o3. OpenAI’s Strategie 2024/25 spiegelt dies wider: **verschiedene Modelle für verschiedene Bedürfnisse** bereitzustellen (Pricing | OpenAI) – von schnellem, preiswertem Denken (o3-mini) bis zu maximaler Denktiefe (o1/o3-high) – wobei jedes Modell klar definierte Stärken und Schwächen im Vergleich zu den anderen aufweist.

Quellen:
Offizielle OpenAI-Blogposts und Dokumentationen von 2024/2025 sowie Berichte zum o1- und o3-Release

- Learning to reason with LLMs | OpenAI
- (OpenAI's STEM-Optimized 'Reasoning' Model, o1, Sees Daylight -- Pure AI)
- (OpenAI o1 explained: Everything you need to know)
- (OpenAI o3-mini | OpenAI)
- (Pricing | OpenAI)
- siehe referenzierte Zitate).