ChatGPT

Übersicht der aktuellen GPT-Modelle von OpenAI (Stand 2025)

Offizielle OpenAI-GPT-Modelle

OpenAI bietet derzeit mehrere **offizielle GPT-Modelle** an, die sich in Leistungsfähigkeit, Kontextgröße, Modalitäten (Eingabetypen/Ausgabetypen), sowie Preisstruktur und Verfügbarkeit unterscheiden. Im Wesentlichen sind dies die Modelle **GPT-3.5**, **GPT-4** (inklusive Varianten mit unterschiedlicher Kontextlänge), **GPT-4 Turbo** und **GPT-4o**. Im Folgenden werden diese Modelle vorgestellt und ihre technischen, preislichen und funktionalen Unterschiede erläutert:

- **GPT-3.5 (Turbo)** – Dieses Modell bildete die Grundlage für ChatGPT bei dessen Start Ende 2022. GPT-3.5 ist ein reines Textmodell mit ~175 Milliarden Parametern und wurde als verbesserte Version von GPT-3 entwickelt [1] [2]. Es kann nur Text verarbeiten und generieren (unimodal) und besitzt typischerweise einen **Kontextumfang von 4.096 Token**. Später führte OpenAI auch Varianten mit größerem Kontext ein (z. B. *gpt-3.5-turbo-16k* mit ~16.000 Token Kontext) [3]. Preislich ist GPT-3.5 sehr günstig – laut OpenAI etwa **0,002 USD pro 1.000 Token** (Prompt **und** Completion) [4]. In neueren Abrechnungsmodellen wird manchmal zwischen Eingabe- und Ausgabetoken unterschieden (z. B. ca. 0,0015 USD für Eingabe und 0,002 USD für Ausgabe pro 1.000 Token) [5]. Funktional bietet GPT-3.5 solide allgemeine Sprachfähigkeiten, hat aber gegenüber GPT-4 gewisse Einschränkungen in Hinsicht auf komplexes **Schlussfolgern**, **Kontextbehalten** und Genauigkeit. Beispielsweise wurde GPT-3.5 (175 Mrd. Parameter) nur auf Text trainiert und zeigt bei anspruchsvollen Aufgaben schwächere Leistungen als GPT-4 [2]. Es dient jedoch weiterhin als kostengünstige Option für viele Anwendungen und bildet das Rückgrat der kostenlosen ChatGPT-Version (mit Einschränkungen).

- **GPT-4 (Standard)** – GPT-4 ist OpenAIs **vierte Generation** von GPT und wurde im März 2023 vorgestellt. Es zeichnet sich durch **deutlich verbesserte Sprachverständnis- und Generierungsfähigkeiten** aus, insbesondere bei komplexen und nuancierten Eingaben [6] [7]. Technisch ist GPT-4 auch **multimodal**: Es kann neben Text **Bilder als Eingabe** verarbeiten und textliche Ausgaben dazu generieren (Bildverständnis) – eine Fähigkeit, die GPT-3.5 nicht hatte [8]. GPT-4 wurde auf einem wesentlich größeren Modell trainiert (geschätzt nahe **1 Billion Parametern** [2], OpenAI selbst nannte keine genaue Zahl) und erreicht damit **menschennähere Leistungen**. So berichtet OpenAI beispielsweise von etwa **40 % höherer Faktentreue** und deutlich verringertem Risiko für unerwünschte Inhalte im Vergleich zu GPT-3.5 [9]. GPT-4 bietet zudem eine **größere Kontextlänge**: Die reguläre Version erlaubt Eingaben von bis zu **8.000 Token**. Es existiert auch eine **GPT-4 (32k)**-Variante mit einem Kontextfenster von ~32.000 Token, was z. B. längere Dokumente in einer Anfrage ermöglicht. Preislich war GPT-4 zunächst sehr teuer: **ca. 0,03 USD pro 1.000 Eingabetoken und 0,06 USD pro 1.000 Ausgabetoken** für die 8k-Version (die 32k-Version etwa doppelt so teuer) [10]. Dieses hohe Preisniveau spiegelt die höhere Rechenlast und Fähigkeit von GPT-4 wider. GPT-4 war anfangs nur über eine Warteliste im API verfügbar und in ChatGPT nur für Plus-Abonnenten mit Limitierung (z. B. X Anfragen pro 3 Stunden). GPT-4 ist **reines Text-Ausgabemodell** – die Bild-Eingabe-Fähigkeit war zunächst nur in einer beschränkten Preview (etwa über die ChatGPT-Plugins/Vision) verfügbar. Audio-Ein- oder Ausgabe bot GPT-4 selbst nicht, nur via zusätzlicher Systeme (etwa Text-to-Speech bei ChatGPT).

- **GPT-4 Turbo** – Unter diesem Namen führte OpenAI im **November 2023** (auf der DevDay-Konferenz) eine optimierte Version von GPT-4 ein [11]. GPT-4 Turbo stellt gewissermaßen **„GPT-4 Version 1.1“** dar – es ist *vergleichbar intelligent wie GPT-4*, aber **günstiger** und **mit erweiterten Fähigkeiten**. Konkret erhielt GPT-4 Turbo eine massiv vergrößerte Kontextlänge von **bis zu 128.000 Token** (128K) [12] – das entspricht über 300 Seiten Text, die in eine einzelne Anfrage passen [13]. Zudem hatte GPT-4 Turbo zum Start einen **Aktualitätsvorsprung**: Sein Wissensstand reichte bis April 2023 oder sogar darüber hinaus (später wurde der Cutoff auf **Ende 2023** aktualisiert) [14], während das ursprüngliche GPT-4-Modell in ChatGPT noch mit einem Wissensstand von Sept 2021 operierte. **Preislich** war GPT-4 Turbo deutlich attraktiver: etwa **1 Cent pro 1.000 Eingabetoken und 3 Cent pro 1.000 Ausgabetoken**, also rund **3× günstiger (Input)** bzw. **2× günstiger (Output)** als GPT-4 [15]. OpenAI beschreibt GPT-4 Turbo als *„günstigere, bessere Version von GPT-4“* [16]. Es beherrscht weiterhin Text und Bild-Eingaben (OpenAI bot auch *GPT-4 Turbo with Vision* an) und war dank Optimierungen schneller in der Antwort. In ChatGPT Plus wurde GPT-4 Turbo Ende 2023 schrittweise als Backend für die GPT-4-Option ausgerollt, ohne die UI-Kennung zu ändern – d. h. viele Nutzer bemerkten das Upgrade nur an geänderten Antworten oder Kenntnissen (z. B. aktualisierter Knowledge Cutoff) [14] [17]. GPT-4 Turbo eignet sich gut für Aufgaben wie Code-Generierung und „normale“ Textaufgaben mit hohem Kontextbedarf [18], während das originale GPT-4 in bestimmten Benchmarks und komplexen Reasoning-Aufgaben minimal besser sein konnte. Insgesamt stellte GPT-4 Turbo jedoch einen großen praktischen Fortschritt in Nutzbarkeit dar.

- **GPT-4o** – *GPT-4o* (das **„o“ steht für „omni“**) ist das **aktuelle Flaggschiff-Modell** von OpenAI und wurde im **Mai 2024** angekündigt [19]. GPT-4o geht noch einen Schritt weiter als GPT-4 Turbo und ist ein wahrhaft **multimodales** Modell: Es kann **Text, Bilder, Audio und sogar Video** als Eingabe verarbeiten und **Text, Bild** und **Audio** als Ausgabe generieren [20]. Es vereint also Sprachverständnis, Bildbeschreibung/Analyse und Sprachsynthese in einem System, was wesentlich natürlichere Mensch-Computer-Interaktionen erlaubt [20]. Beispielsweise kann GPT-4o Sprachbefehle fast in Echtzeit verstehen (ca. 232–320 ms Latenz, ähnlich menschlicher Reaktionszeit) [21], Bilder interpretieren und kommentieren, oder Audio ausgeben (z. B. gesprochene Antworten). Trotz dieser breiteren Fähigkeiten erreicht GPT-4o auf rein textuellen Aufgaben **mindestens die Leistung von GPT-4 Turbo** (insbesondere in Englisch und im Codieren), zeigt **Verbesserungen bei nicht-englischen Sprachen** sowie **deutlich bessere Fähigkeiten in visuellem und audio-Verständnis** [22] [23]. Beeindruckenderweise gelang dies OpenAI mit **höherer Effizienz**: GPT-4o ist spürbar schneller als GPT-4 und **etwa 50 % günstiger** in der API-Nutzung als sein direkter Vorgänger [24]. (OpenAI gibt an, dass GPT-4o nur halb so viel kostet wie GPT-4 Turbo für API-User [24]. Tatsächlich liegen die Tokenpreise von GPT-4o etwa bei **0,005 USD pro 1.000 Eingabetoken** und **0,015 USD pro 1.000 Ausgabetoken**, gegenüber 0,01 $/0,03 $ bei GPT-4 Turbo [25].) GPT-4o besitzt ebenfalls einen großen Kontext (vermutlich vergleichbar 128K) und wurde hinsichtlich **Sprache, Code, visuellem Verständnis und Echtzeitfähigkeit** zum neuen Standard erklärt. OpenAI stellte GPT-4o der Allgemeinheit zur Verfügung – **ChatGPT nutzt nun GPT-4o** sogar für kostenlose Nutzer (mit strikten Limits), während Plus/Enterprise-Kunden höhere Limits und zusätzliche Features erhalten [26]. GPT-4o markiert somit einen weiteren großen Schritt, indem es *fast alle Modalitäten* in einem Modell vereint und dabei **Geschwindigkeit und Kosten optimiert**.

Nachfolgend eine **Vergleichstabelle** der wichtigsten Kennzahlen der offiziellen GPT-Modelle:

Modell	Kontextgröße	Modalitäten	API-Preis (Input/Output)	Besonderheiten / Nutzung
GPT-3.5 Turbo	4k (optional 16k) Tokens	Nur Text (Eingabe & Ausgabe)	~$0.002 / 1k Token (gesamt) (ca. 0,0015$ In / 0,002$ Out) [5]	Standardmodell für ChatGPT (Free), sehr günstig und schnell, Knowledge-Cutoff i.d.R. 2021.
GPT-4	8k Tokens (32k Variante)	Text; Bild-Eingabe, Text-Ausgabe (Vision)	$0.03 / $0.06 pro 1k Token (8k-Kontext) [25] (32k: $0.06 / $0.12) [27]	Deutlich bessere Qualität, komplexes Reasoning, Bildverständnis. Nur für Plus/Entwicklung (2023) verfügbar.
GPT-4 Turbo	bis 128k Tokens [12]	Text; Bild-Eingabe (Vision)	$0.01 / $0.03 pro 1k Token [15]	Verbesserte GPT-4-Version (Nov 2023): 3×/2× günstiger, aktuelleres Wissen, in ChatGPT Plus als GPT-4 implementiert.
GPT-4o	sehr groß (≈128k Tokens)	Text, Bild, Audio (Ein- **und** Ausgabe)	~$0.005 / $0.015 pro 1k Token [24] [25]	„Omni"-Modell (Mai 2024): Multimodal (inkl. Sprache), 50 % günstiger als Turbo, schneller; neuer Standard in ChatGPT.

Quellenhinweis: Die obigen Preise und Angaben basieren auf OpenAI-Dokumentation und Ankündigungen (z. B. DevDay-Blogpost [15], GPT-4o-Systemkarte [24]) sowie bestätigten Community-Beiträgen. Sie geben den Stand von 2024/2025 wieder und können sich ändern. GPT-4 und Nachfolger werden kontinuierlich weiterentwickelt – so mentionierte OpenAI auch Zwischenversionen wie *GPT-4 1106*, *GPT-4 0613* etc., die neue Funktionen (z. B. **Function Calling**, **steuere Output-Formate**) brachten [28].

Inoffizielle und interne Modellvarianten

Neben den offiziellen, dokumentierten Modellen existieren bei OpenAI auch **interne oder nicht-öffentlich dokumentierte Modellvarianten**. Diese werden teils intern zu Testzwecken genutzt, teils in speziellen Produkten (wie ChatGPT Enterprise) eingebunden, ohne dass sie als separate Modelle öffentlich benennbar wären. Hier ein Überblick über solche Varianten:

- **Versionierte Snapshots und Test-Modelle:** OpenAI versieht seine Modelle im API mit Datumsstempeln, wenn neue Versionen ausgerollt werden. Beispielsweise gab es *gpt-3.5-turbo-0301* und *-0613* (März bzw. Juni 2023 Versionen mit neuen Features) sowie *gpt-4-0314* und *-0613* etc. Diese Snapshots sind intern separate Modellcheckpoints. Auch *GPT-4 Turbo* wurde initial als Preview (*gpt-4-1106-preview*) bereitgestellt [29]. Solche Varianten sind offiziell, aber oft nur in Entwicklerkreisen bekannt. Ähnlich gab es experimentelle Ableger wie **gpt-3.5-turbo-instruct**, die speziell für Anweisungen/Instruktionen optimiert waren (knappere, relevantere

Antworten) [30] . Diese war eine kurz verfügbare Test-Variante von GPT-3.5, um Nutzerfeedback für „Anweisungs-Stil“ zu sammeln.

- **„o“-Modellreihe (OpenAI o-series):** Im Zuge von ChatGPT Enterprise/Pro führte OpenAI intern sogenannte **o-Modelle** ein (z. B. *OpenAI o1*, *o3*, *o4-mini* etc.). Dies sind spezielle **Reasoning-Modelle**, die für **schwer lösbare Probleme** und mehrstufiges Denken entwickelt wurden [31] . Sie kommen teilweise im Hintergrund zum Einsatz, wenn ChatGPT bestimmte Tools nutzt oder komplexe Teilaufgaben löst. Beispielsweise erwähnt OpenAI, dass *o3* und *o4-mini* in Domänen wie Forschung, Strategie, Coding, Mathematik helfen können, während *o1* sogar einen einsehbaren Gedankengang (eine Art transparentes Chain-of-Thought) anzeigen kann [32] . Diese Modelle sind **nicht über die öffentliche API direkt verfügbar** und auch in der UI nicht als eigene Auswahl gelistet. Vielmehr werden sie im Rahmen von Team- und Enterprise-Abos von ChatGPT unter der Haube eingesetzt, wenn der Nutzer bestimmte „erweiterte Reasoning“-Modi aktiviert. Die ChatGPT **Team- und Enterprise-Pläne** beschreiben z. B. Zugriff auf *OpenAI o3-mini, o3-mini-high, o1 pro mode* etc. [33] [34] – was darauf hindeutet, dass es unterschiedliche **Größen** dieser o-Modelle gibt (mini, high etc. = kleineres oder größeres Modell mit unterschiedlicher Leistungsfähigkeit). Diese *o*-Modelle sind weitgehend undokumentiert für Entwickler, deuten aber an, dass OpenAI neben den großen generellen Modellen parallel spezialisierte **Intermediate-Modelle** für Problemlösen und Tool-Integration pflegt.

- **GPT-4.1, GPT-4.5 und weitere**: In einigen OpenAI-Dokumentationen und Partnerdiensten tauchen Hinweise auf **GPT-4.1** und sogar *GPT-4.5* auf [35] . So listet etwa Azure OpenAI schon ein *GPT-4.1* Modell [36] . Dies lässt vermuten, dass OpenAI intern eine Art **„GPT-4 Nachfolgegeneration“** (4.1 könnte man als GPT-4 Turbo oder leicht weiterentwickelte GPT-4-Version verstehen) sowie eine mögliche **Zwischengeneration GPT-4.5** testet. Offiziell veröffentlicht war Anfang 2025 noch nichts unter dem Namen GPT-4.5; der Eintrag auf OpenAIs Webseite unter „Latest Advancements“ [35] könnte auf Forschungsarbeit hindeuten. Es ist bekannt, dass OpenAI kontinuierlich sein Hauptmodell verbessert – GPT-4o selbst könnte man als GPT-4.2 interpretieren. Die genaue Nomenklatur ist aber intern. Auch Begriffe wie *GPT-4.1 mini/nano* deuten an, dass OpenAI **verschiedene Modellgrößen** aus derselben Familie testet, um z. B. sehr schnelle, kleinere Modelle („nano“) für einfachere Aufgaben anzubieten [37] . Solche Varianten wurden allerdings (Stand jetzt) nicht als separate Produkte angekündigt – sie fließen entweder in neue Releases (Turbo, 4o) ein oder bleiben speziellen Anwendungsfällen vorbehalten.

- **Spezielle API-Modelle (Codex, Embeddings, Moderation):** Obwohl nicht direkt GPT-4, sei der Vollständigkeit halber erwähnt, dass OpenAI separate Modelle für bestimmte Aufgaben bereitstellt. *Codex* (z. B. code-davinci-002) war ein auf Code-Generierung spezialisiertes Modell, basierend auf GPT-3, das zeitweise als eigene API angeboten wurde. Seit GPT-4 hat OpenAI jedoch Code-Fähigkeiten ins Hauptmodell integriert (GPT-4 und 4o sind sehr gut im Coden), sodass Codex als separates Modell eingestellt wurde. Ebenso gibt es embedding-Modelle (z. B. *text-embedding-ada-002*), die Vektorembeddings erzeugen – sie gehören zur GPT-Familie in weitem Sinne, werden aber meist nicht als „GPT-x“ bezeichnet. Für Moderation stellt OpenAI eigene classifier-Modelle bereit. Diese **Sondermodelle** sind hier nur am Rande erwähnt, da der Fokus auf den *ChatGPT*- bzw. generativen Sprachmodellen liegt.

Zusammenfassend verfügt OpenAI intern über ein **Ökosystem von Modellvarianten**, die über das hinausgehen, was der normale API-Nutzer sieht. Viele dieser Varianten (verschiedene *Snapshots*, *o*-Reihe, *mini*-Modelle) sind **nicht öffentlich dokumentiert**, tauchen aber in Developer-Foren oder der OpenAI-Dokumentation indirekt auf. Beispielsweise wunderte sich ein Nutzer, welches Modell wohl sein *Custom GPT* nutzt – die Vermutung im Forum war, dass es auf jeden Fall ein *großes GPT-4o-Modell* ist,

möglicherweise eine spezielle Version optimiert für diese Funktion [38]. OpenAI kommuniziert solche internen Modellwechsel selten proaktiv, um den Nutzern ein nahtloses Erlebnis zu bieten, ohne sie mit technischen Details zu überfrachten.

Unklare Modellangaben in ChatGPT-UI und Custom GPTs

Viele Nutzer stellen fest, dass in der **ChatGPT-Benutzeroberfläche** (sowohl beim Standard-Chat als auch bei *Custom GPTs*) **nicht klar ersichtlich** ist, welches Modell gerade im Einsatz ist. Beispielsweise kann ein ChatGPT Plus-Nutzer zwar zwischen „GPT-3.5“ und „GPT-4“ wählen – aber ob hinter „GPT-4“ gerade das originale GPT-4-Modell, GPT-4 Turbo oder sogar schon GPT-4o steckt, wird **nicht explizit angezeigt**. Ebenso gibt es in der neuen *Custom GPT*-Funktion keine offensichtliche Modellanzeige. Die Gründe dafür liegen in OpenAIs Produktdesign und dem dynamischen Modell-Upgrade-Prozess:

- **Modell-Upgrades ohne UI-Änderung:** OpenAI handhabt neue Modellversionen oft als *stilles Upgrade* für bestehende Optionen. Als GPT-4 Turbo erschien, wurde es für ChatGPT-Plus Nutzer einfach anstelle des bisherigen GPT-4 verwendet, ohne dass die Schaltfläche umbenannt wurde [14] [17]. Ähnlich wurde vermutlich GPT-4o im Hintergrund ausgerollt. OpenAI bewirbt diese Verbesserungen zwar in Blogposts, belässt aber die UI-Bezeichnung „GPT-4“ für die Auswahl, um Verwirrung zu vermeiden. Aus Sicht der Produktstrategie ist *GPT-4* eine Kategorie (High-End-Modell), und was genau dahinter steckt, kann sich ändern, ohne dass der Nutzer bei jedem Update neu wählen muss. Das führt allerdings dazu, dass **die UI keine transparente Modellkennung bietet**.

- **Custom GPTs nutzen immer das beste verfügbare Modell:** Laut OpenAI-Dokumentation werden Custom GPTs (benutzerdefinierte ChatGPT-Bots, die man über ChatGPT Plus/Team erstellen kann) **derzeit alle mit GPT-4o betrieben** [39]. Das heißt, ein Custom GPT greift im Hintergrund auf das neueste große Modell zu. Für den Nutzer sieht es aber nicht anders aus – es gibt kein Dropdown „Wähle Modell“ in der GPT-Erstellung. Teilweise stand in frühen Versionen an einigen Stellen „Powered by GPT-3.5“ als Platzhalter, was zu Verwirrung führte [40], aber faktisch wurde dennoch GPT-4 genutzt. Die fehlende Anzeige kann verwirrend sein, insbesondere wenn ein Custom GPT eventuell etwas andere Antworten liefert als der normale GPT-4o-Chat – manche Nutzer bemerkten z. B. einen einfacheren Sprachstil bei Custom GPTs und fragten, ob vielleicht ein kleineres Modell genutzt wird [41]. OpenAI hat hierzu (Stand Anfang 2025) noch keine klare Kennzeichnung implementiert.

- **Keine Versionsinfo für Endnutzer:** Anders als bei API-Entwicklern, die exakt angeben, welches Modell sie verwenden, hat die ChatGPT-UI **kein Feld „Modellversion“**. Früher gab es nur die grobe Auswahl (GPT-3.5 vs GPT-4). Mittlerweile könnten theoretisch *o*-Modelle oder ähnliches beteiligt sein (z. B. ChatGPT Enterprise kann im Hintergrund je nach Aufgabe zwischen GPT-4o und o-Modellen wechseln). Um die UI nicht zu überfrachten, blendet OpenAI diese technische Info aus. Das Ergebnis ist aus Nutzersicht **Intransparenz**: Man muss den Aussagen des Modells oder indirekten Hinweisen entnehmen, was läuft.

- **Beispiel GPT-4 vs GPT-4 Turbo:** Mitte/Ende 2023 merkten Nutzer an zwei Indizien, dass ihr ChatGPT-„GPT-4“ wohl auf GPT-4 Turbo umgestellt wurde: erstens antwortete es auf Fragen nach dem Wissens-Cutoff plötzlich mit **April 2024 oder Dez 2023** statt 2021 [14], und zweitens nahm die Reaktionsgeschwindigkeit leicht zu. Offiziell wurde das Modell aber weiterhin nur als „GPT-4“ betitelt. Ein Reddit-Beitrag fasste zusammen: *Einige sehen April 2024 als Cutoff, andere Dez 2023 – beides ist das neue GPT-4 Turbo; wer April 2023 sieht, hat noch das alte Modell* [14]. Diese versteckte Umstellung ohne UI-Indikator ist genau, was viele als **untransparent** empfinden.

Kurz gesagt: Bei ChatGPT (vor allem Plus/Enterprise) **weiß man nicht immer auf Anhieb**, welches Modell gerade antwortet. OpenAI wechselt die Modelle serverseitig aus, um Verbesserungen bereitzustellen oder Kosten/Leistung zu optimieren, lässt aber die Frontend-Anzeige unverändert. Auch *Custom GPTs* zeigen nirgends „Model XYZ“, sie laufen einfach auf dem aktuell vorgesehenen Backend (derzeit GPT-4o). Das kann für Power-User frustrig sein, die gern wüssten, ob sie nun z. B. von den GPT-4o-Updates profitieren. Bisher muss man dafür auf indirekte Hinweise zurückgreifen oder selbst nachforschen (siehe nächster Abschnitt).

Modellidentifikation: Wie findet man heraus, welches Modell läuft?

Obwohl die Oberfläche es nicht anzeigt, gibt es ein paar **Möglichkeiten für Nutzer, das zugrundeliegende Modell zu identifizieren** – zumindest näherungsweise:

- **Das Modell selbst fragen:** Man kann versuchen, den Chatbot direkt zu fragen, z. B. *„Which model are you?“* oder *„What is your knowledge cutoff and training data?“*. Oft sind Modelle angewiesen, diese Frage ausweichend zu beantworten (etwa „Ich bin ein KI-Sprachmodell von OpenAI…“). Allerdings enthüllte sich GPT-4 Turbo in ChatGPT wie erwähnt über die Knowledge-Cutoff-Angabe (weil GPT-4 Turbo eine aktualisierte Wissensbasis hatte). Manchmal gibt das Modell auch freiwillig seinen Namen an. Ein Community-Tipp: *„Man kann immer seinen GPT direkt fragen, und es wird vielleicht mit seinem Modellnamen antworten.“* [42]. Dies ist jedoch nicht garantiert verlässlich – insbesondere kennen Modelle ihren internen Codenamen nicht unbedingt, sondern nur das, was in ihren Daten darüber stand. GPT-4 Turbo wusste z. B. nicht von sich aus, dass es „Turbo“ heißt, verriet aber durch das Wissensdatum indirekt seine Natur. GPT-4o könnte auf Nachfrage hin erwähnen, dass es multimodal ist. Solche Fragen können Hinweise liefern, ersetzen aber keine eindeutige Auskunft.

- **API-Antwort/Dev-Tools überprüfen:** Die zuverlässigste Methode ist technisch: **Netzwerkaufrufe** der ChatGPT-Webapp analysieren. Wenn man z. B. im Browser die Entwicklertools öffnet und eine Anfrage in ChatGPT schickt, kann man den laufenden **XHR-Request** abfangen. Im Streaming-Endpoint findet sich in der JSON-Antwort ein Feld namens `model_slug`, welches den Modellnamen enthält [43]. Ein Entwickler hat eine Schritt-für-Schritt-Anleitung veröffentlicht, wie man per Chrome Dev-Tools und JSON-Formatter den `model_slug` eines Custom GPTs ausliest [44] [45]. Dort zeigte sich dann z. B. eindeutig, ob `gpt-4o` oder `gpt-4-turbo` im Einsatz ist. Dies erfordert allerdings etwas Technikaffinität. Ein ähnlicher Trick besteht darin, eine Konversation per Share-Link oder API mitzuschneiden. Für Average-User ist das nicht sofort zugänglich, aber **es funktioniert zuverlässig**, da der Backend-Call das tatsächliche Modell enthält.

- **Interface-Details/indirekte Hinweise:** Wie zuvor erwähnt, können **Knowledge Cutoff** oder offensichtliche neue Fähigkeiten auf ein Modell hindeuten. Wenn z. B. plötzlich Bild-Uploads möglich sind und vom Bot verstanden werden (was im Oktober 2023 in ChatGPT Plus eingeführt wurde), war klar, dass nun GPT-4 mit Vision bzw. GPT-4 Turbo am Werk war, da GPT-3.5 keine Bilder versteht. Ebenso deutet das Auftreten von Audio-Funktionen (Spracherkennung, Sprachausgabe in ChatGPT) auf GPT-4o hin, weil erst dieses Modell Audio wirklich integriert hat. OpenAI hat in mobilen Apps teils kleine Hinweise ergänzt – Berichten zufolge sollte in der GPT-Beschreibung auf dem Handy evtl. das Modell genannt sein, jedoch war das nicht konsistent verfügbar [46]. Bis eine offizielle Modellanzeige kommt (OpenAI testet so etwas offenbar), bleibt Nutzern nur, anhand solcher Merkmale zu folgern.

- **Über die API selbst bestimmen:** Wenn man statt der ChatGPT-UI die **OpenAI-API** nutzt, hat man natürlich volle Kontrolle über die Modellwahl. In einem API-Call gibt man explizit `model: gpt-3.5-turbo` oder `gpt-4` etc. an, und die API-Antwort enthält ebenfalls den Modellnamen, mit dem die Completion erzeugt wurde. Für Developer ist das der sicherste Weg – hier *weiß* man genau, welches Modell läuft, da man es selbst gewählt hat. Allerdings sind nicht alle Modelle jederzeit frei verfügbar (z. B. GPT-4o im API erfordert ggf. ein Upgrade oder bestimmte Berechtigungen zu Beginn). Nichtsdestotrotz: Wer programmatisch auf die OpenAI-API zugreift, kann auf diese Weise verifizieren, welches Modell genutzt wurde. Für ChatGPT-Webnutzer bleibt, wie beschrieben, praktisch nur die Entwickler-Console oder die Selbst-Auskunft des Bots.

Zusammengefasst: **Zuverlässig erfährt man das laufende Modell nur durch technische Tricks oder via API.** OpenAI hat diese Info absichtlich abstrahiert. Dennoch teilen Community-Mitglieder ihre Workarounds, z. B. per Browser-Devtools den `model_slug` auszulesen [45]. Dieses Feld offenbarte in Tests, dass *alle Custom GPTs aktuell auf GPT-4o laufen* (teils mit spezifischen Snapshot-Bezeichnungen wie `gpt-4o-2024-08-06` etc.), und dass keine 3.5er Modelle mehr dafür verwendet werden [38] [47]. Mit solchen Mitteln kann man also Gewissheit erlangen.

Unterschiede bei Web-UI, API und Custom GPTs hinsichtlich Modellwahl

Je nachdem, **wie man OpenAIs GPT-Modelle nutzt** – über die ChatGPT-Weboberfläche, über die API, oder als Custom GPT – gibt es Unterschiede, welche Modelle zur Auswahl stehen und wie die Modellnutzung gesteuert wird:

- **ChatGPT Web-App (UI):** In der Standard-Webanwendung von ChatGPT haben Benutzer nur begrenzte Auswahlmöglichkeiten. *Kostenlose Nutzer* konnten lange ausschließlich GPT-3.5 verwenden. Mit der Einführung von GPT-4o hat OpenAI jedoch begonnen, auch freien Nutzern eingeschränkten Zugriff auf GPT-4o zu geben (da GPT-4o günstiger ist) [26]. Allerdings gelten strenge Limits (z. B. langsamere Generierung oder Kapazitätsgrenzen), um die Kosten zu kontrollieren. *Plus-Abonnenten* können in der UI zwischen „GPT-3.5“ und „GPT-4“ umschalten. Hinter „GPT-4“ verbirgt sich das jeweils beste verfügbare High-End-Modell (aktuell GPT-4o). Plus-Nutzer haben höhere Nutzungslimits als Free-Nutzer und Zugriff auf zusätzliche Features (z. B. Plugins/Tools, Bilduploads, Sprachmodus). *Enterprise-Nutzer* erhalten nochmals erweiterte Kontingente – die Enterprise-Planbeschreibung spricht von **„höheren Nachrichtengrenzen für GPT-4o und erweitertem Kontextfenster“** [48]. Dies bedeutet, dass Enterprise-Kunden z. B. längere Unterhaltungen/Dateien mit GPT-4o führen können, vermutlich also die vollen 128k Tokens Kontext nutzen dürfen, während Plus-Nutzer evtl. auf 8k oder 32k Kontext in der UI beschränkt sind. Außerdem können Enterprise-Kunden zwischen mehreren „Reasoning-Modi“ wechseln, die intern unterschiedliche Modelle (o1, o3, o4-mini etc.) nutzen [34]. Normaluser sehen diese Optionen gar nicht. Zusammengefasst: In der Web-App wählt man als Plus-User grob das Modellniveau (3.5 vs 4), als Enterprise-User evtl. verschiedene Lösungsstrategien, aber man hat **keine Feinauswahl** etwa „bitte GPT-4 Turbo statt GPT-4o“. Die Zuordnung passiert automatisch gemäß Nutzer-Tier.

- **OpenAI API:** Über die API hat man die **freieste Modellwahl** (eingeschränkt nur durch Zugangsrechte). Entwickler können genau das Modell angeben, das sie möchten – sei es *gpt-3.5-turbo*, *gpt-3.5-turbo-16k*, *gpt-4*, *gpt-4-32k*, *gpt-4-turbo*, *gpt-4o* usw., sofern ihr Account Berechtigung für das jeweilige Modell hat. Hier kann man also bewusst z. B. aus Kosten- oder Performancegründen GPT-3.5 einsetzen, auch wenn GPT-4o verfügbar wäre. Die **Preise** unterscheiden sich entsprechend (API-Nutzung wird ja tokenbasiert berechnet, wie oben

dargestellt). Wichtig: Die API erfordert, dass man ein Modell auswählt – es gibt **keine Automatik**, die einem je nach Anfrage ein anderes Modell zuteilt. Allerdings stellt OpenAI neue Modelle manchmal unter neuen Namen bereit; Entwickler müssen dann ggf. umstellen (z. B. von `gpt-4` auf `gpt-4-turbo` als dieser erschien, um die Vorteile zu nutzen – oder sie bleiben beim alten, wenn gewünscht). Hier hat man also volle Transparenz und Kontrolle, aber auch die Verantwortung, das richtige Modell zu wählen. Unterschiede in Knowledge Cutoff oder Fähigkeiten muss der Entwickler selbst berücksichtigen. Zusammengefasst kann die API **feinkörniger** genutzt werden: ein Chatbot-Betreiber könnte entscheiden, den Großteil einfacher Anfragen mit GPT-3.5 abzudecken (Kosten sparen) und nur bei schwierigen Fällen per Programmierung auf GPT-4o umzuschalten. In der Web-App ist solches Routing vom Nutzer kaum beeinflussbar.

- **Custom GPTs (ChatGPT GPT-Builders):** Custom GPTs sind ein Feature innerhalb der ChatGPT-Oberfläche (verfügbar ab Plus/Team-Plan), mit dem man eigene Chatbot-Personas/Knowledge Bases erstellen kann. Bezüglich **Modellwahl** hat der Benutzer dort **keine direkte Auswahlmöglichkeit** – alle Custom GPTs laufen standardmäßig auf dem leistungsfähigsten Modell, das OpenAI bereitstellt. Wie oben erwähnt, ist das derzeit GPT-4o. In Zukunft könnte OpenAI hier Optionen hinzufügen (die Community fragte schon nach der Möglichkeit, auch kleinere Modelle für Custom GPTs zu wählen [49]). Doch Stand jetzt gilt: **Custom GPT = GPT-4o** im Hintergrund. Einzige Ausnahme: Wenn ein Nutzer keinen zahlenden Account hat, kann er gar keine eigenen GPTs erstellen; in der kostenlosen Version gibt es dieses Feature nicht. Das heißt, alle Custom GPT-Instanzen liegen ohnehin bei zahlenden (Plus/Team) Kunden und nutzen damit das GPT-4-Level. Entsprechend gibt es auch **keine Kostenunterschiede** zwischen verschiedenen Custom GPTs – OpenAI rechnet die Nutzung vermutlich wie normale GPT-4o-Nutzung ab (derzeit sind Custom GPTs in der Beta-Phase ohne separate Tokenabrechnung, aber das könnte sich ändern, wenn sie in API-Nutzung übergehen). Funktional könnte OpenAI intern jedoch entscheiden, Custom GPTs ein etwas anderes Modell zuzuweisen, falls das sinnvoll ist (z. B. ein GPT-4o, das speziell auf Wissensabruf optimiert ist). Es gab Spekulationen, ob Custom GPTs eventuell *GPT-4o-mini* verwenden, weil die Antworten manchmal simpler wirkten [41]. Offizielle Doku stellte aber klar: *"Creating a GPT is currently powered by GPT-4o."* [39]. Unterschiede bestehen allenfalls darin, dass der **Systemprompt** bei Custom GPTs anders gelagert ist (je nach vom Ersteller definiertem Persona/Wissen) – das kann die Ausdrucksweise beeinflussen, nicht unbedingt das Modell selbst.

- **Unterschiede in Knowledge Cutoff und Tools:** Noch ein Punkt: Die Web-App-Modelle (ChatGPT) erhalten gelegentlich **Updates oder Limits**, die in der API nicht gelten. Zum Beispiel wurde im Frühjahr 2023 der Browsing-Zugang für GPT-4 zeitweise deaktiviert („GPT-4 kann vorerst nicht browsen"), was Web-User betraf. In der API gab es diese live-Browsing-Funktion zu dem Zeitpunkt gar nicht, dafür aber andere Workarounds. Mit ChatGPT Enterprise kamen **erweiterte Tools** (wie Dateiupload, Websuche, Canvas) exklusiv in der UI hinzu, die aber nur mit bestimmten Modellen funktionieren (siehe o-Serie Tabelle im Help Center: *OpenAI o1 pro kann z. B. kein Web oder File, GPT-4o kann alles* [50] [51]). Diese Unterschiede in der **Modellnutzung je nach Plattform** bedeuten: Der Funktionsumfang, den ein Modell entfalten kann, hängt auch vom Interface ab. Über die API hat man die rohe Modellfähigkeiten, aber muss eigene Tools anflanschen; in ChatGPT UI sind einige Tools integriert, aber wiederum an bestimmte Modellvarianten gebunden. Ein Enterprise-Nutzer nutzt evtl. andere Modelldialekte für Toolgebrauch, ohne es zu merken, als ein API-User, der denselben Task programmiert.

Fazit: Web-App, API und Custom GPTs bieten unterschiedliche Grade von Kontrolle vs. Komfort. **Die Web-App** abstrahiert die Modellwahl stark – der Nutzer wählt höchstens die „Klasse" (Standard vs. Advanced), und OpenAI kümmert sich um das optimale Modell dahinter. **Die API** gibt volle Wahlfreiheit,

verlangt aber vom Nutzer/Entwickler das Wissen um Modelle und ihre Unterschiede (und ein separates API-Budget, da ChatGPT-Abos und API-Kosten getrennt sind). **Custom GPTs** wiederum nehmen dem Nutzer die Modellentscheidung komplett ab und setzen immer das leistungsfähigste Modell ein, damit sich der Ersteller auf Inhalte konzentrieren kann. Für Organisationskunden gibt es interne Varianten (o-Series), die in der UI wählbar sind, aber ebenfalls nicht die klassischen Modellnamen tragen.

Bei allen Varianten gilt: OpenAI versucht, **offizielle Dokumentation** (wie API-Dokumente, Blogposts und Supportartikel) bereitzustellen, um über neue Modelle und Änderungen zu informieren. Dennoch bleibt es für den Endanwender manchmal eine Herausforderung nachzuvollziehen, *„Welches GPT arbeitet hier gerade für mich?“*. Mit den obigen Informationen und Tipps kann man die aktuellen Modelle und ihre Verwendung besser einordnen und – wenn nötig – gezielt herausfinden [43], welches Modell im Hintergrund aktiv ist.

Quellen: Offizielle OpenAI-Dokumentationen und Ankündigungen (OpenAI Blog, Help Center, Pricing) [20] [12] sowie Erkenntnisse der Community (Foren, Reddit) [14] [38]. Diese liefern die Basis der obigen Übersicht und werden empfohlen, um bei zukünftigen Änderungen auf dem Laufenden zu bleiben.

---

[1] [2] [3] [6] [7] [8] [9] [30] GPT-3.5 vs. GPT-4: Biggest differences to consider | TechTarget
https://www.techtarget.com/searchenterpriseai/tip/GPT-35-vs-GPT-4-Biggest-differences-to-consider

[4] GPT4 and GPT-3.5-turb API cost comparison and understanding
https://community.openai.com/t/gpt4-and-gpt-3-5-turb-api-cost-comparison-and-understanding/106192

[5] Pricing for GPT-3.5 Turbo on Azure is not updated. - Learn Microsoft
https://learn.microsoft.com/en-us/answers/questions/1327201/pricing-for-gpt-3-5-turbo-on-azure-is-not-updated

[10] [25] [27] ChatGPT 3.5 vs ChatGPT 4 - Key Differences to Consider
https://www.kommunicate.io/blog/chatgpt-4-vs-chatgpt-3-5-key-differences/

[11] [12] [13] [15] [28] [29] New models and developer products announced at DevDay | OpenAI
https://openai.com/index/new-models-and-developer-products-announced-at-devday/

[14] [17] The new GPT-4 Turbo is now available to paid ChatGPT users. : r/OpenAI
https://www.reddit.com/r/OpenAI/comments/1c1v4v2/the_new_gpt4_turbo_is_now_available_to_paid/

[16] GPT-4 Turbo in the OpenAI API | OpenAI Help Center
https://help.openai.com/en/articles/8555510-gpt-4-turbo-in-the-openai-api

[18] Gpt-4 vs gpt-4-turbo-preview - OpenAI Developer Community
https://community.openai.com/t/gpt-4-vs-gpt-4-turbo-preview/693031

[19] [20] [21] [22] [23] [24] Hello GPT-4o | OpenAI
https://openai.com/index/hello-gpt-4o/

[26] GPT-4o - Wikipedia
https://en.wikipedia.org/wiki/GPT-4o

[31] [32] [50] [51] Using OpenAI o-series models and GPT-4o models on ChatGPT | OpenAI Help Center
https://help.openai.com/en/articles/9824965-using-openai-o-series-models-and-gpt-4o-models-on-chatgpt

[33] [34] [48] ChatGPT Pricing | OpenAI
https://openai.com/chatgpt/pricing/

[35] [37] Pricing | OpenAI
https://openai.com/api/pricing/

36 Azure OpenAI Service - Pricing
https://azure.microsoft.com/en-us/pricing/details/cognitive-services/openai-service/

38 41 47 Does anyone know what model the custom GPTs are using? : r/OpenAI
https://www.reddit.com/r/OpenAI/comments/1es5uxw/does_anyone_know_what_model_the_custom_gpts_are/

39 Which open-ai model is used with custom GPTs
https://community.openai.com/t/which-open-ai-model-is-used-with-custom-gpts/1112864

40 What models are my custom GPTs using? - GPT builders
https://community.openai.com/t/what-models-are-my-custom-gpts-using/776983

42 43 44 45 46 How to Identify the OpenAI Model Used in Your Custom GPTs | by Dimitri Pletschette | Medium
https://dimitripletschette.medium.com/how-to-identify-the-openai-model-used-on-your-custom-gpts-c39a091444b8

49 Model use for custom GPTs - OpenAI Developer Community
https://community.openai.com/t/model-use-for-custom-gpts/1139861