MODELLE

GPT-5.6 Luna - technische Daten, Benchmarks und Preise

GPT-5.6 Luna im Datenblatt: die unterste Stufe der GPT-5.6-Familie mit technischen Daten, Preisen und Einordnung - am 30. Juli 2026 um 80 Prozent verbilligt.

Stand: 04.08.2026 · 3 Min. Lesezeit

GPT-5.6 Luna ist die günstigste der drei Stufen von OpenAIs GPT-5.6-Familie, vorgesehen für Aufgaben in großer Stückzahl: Klassifikation, Absichtserkennung, Moderation und einfache Zusammenfassungen. Bemerkenswert ist weniger diese Rolle als der Preis. Am 30. Juli 2026 hat OpenAI die Sätze für Eingabe wie Ausgabe um 80 Prozent gesenkt, und das Ausmaß ist so ungewöhnlich, dass man die Herstellerdokumentation zweimal liest - dort aber steht es genau so. Die dokumentierte Ausstattung bleibt dabei dieselbe wie bei den deutlich teureren Geschwistern Terra und Sol - die gemessene Leistung nicht: Vor allem beim Erinnern über lange Kontexte fällt Luna weit hinter die beiden zurück.

GPT-5.6 Luna - Überblick und technische Daten

Anbieter	OpenAI
Familie	GPT-5.6
Erschienen	nicht belegt Die Herstellerdokumentation nennt kein Erscheinungsdatum.
Lizenz	Proprietär, ausschließlich über Schnittstellen nutzbar.
Weights	closed weights, kein Selbstbetrieb
Nutzbar über	OpenAI, OpenRouter
Preis ab	0,20 USD je Mio. Input-Token
Architektur	nicht veröffentlicht Vom Anbieter nicht veröffentlicht.
Parameter gesamt	nicht veröffentlicht Vom Anbieter nicht veröffentlicht.

AI NEWS pro - KI-News ohne Hype! · © 2026 SOFTCLICK Labs · Stand des Beitrags: 04.08.2026

Kontextfenster	1.050.000 (rund 1,05 Mio.) Token Das Kontextfenster von 1.050.000 Token umfasst Ein- und Ausgabe. Für die Eingabe allein nennt der Anbieter eine Obergrenze von 922.000 Token. Anfragen über 272.000 Eingabe-Token kosten den doppelten Eingabe- und den anderthalbfachen Ausgabepreis.
Maximale Eingabe	922.000 Token
Maximaler Output	128.000 Token
Input	Text, Bild
Output	Text
Wissensstand	16. Februar 2026
Reasoning	Wahlweise Aufwandsstufen von none bis max, voreingestellt ist medium. Mit der Stufe none arbeitet das Modell ohne Reasoning.
Tool Calls	Ja
Structured Output	Ja
Prompt Caching	Ja
Kennung	gpt-5.6-luna

Einordnung in die Familie

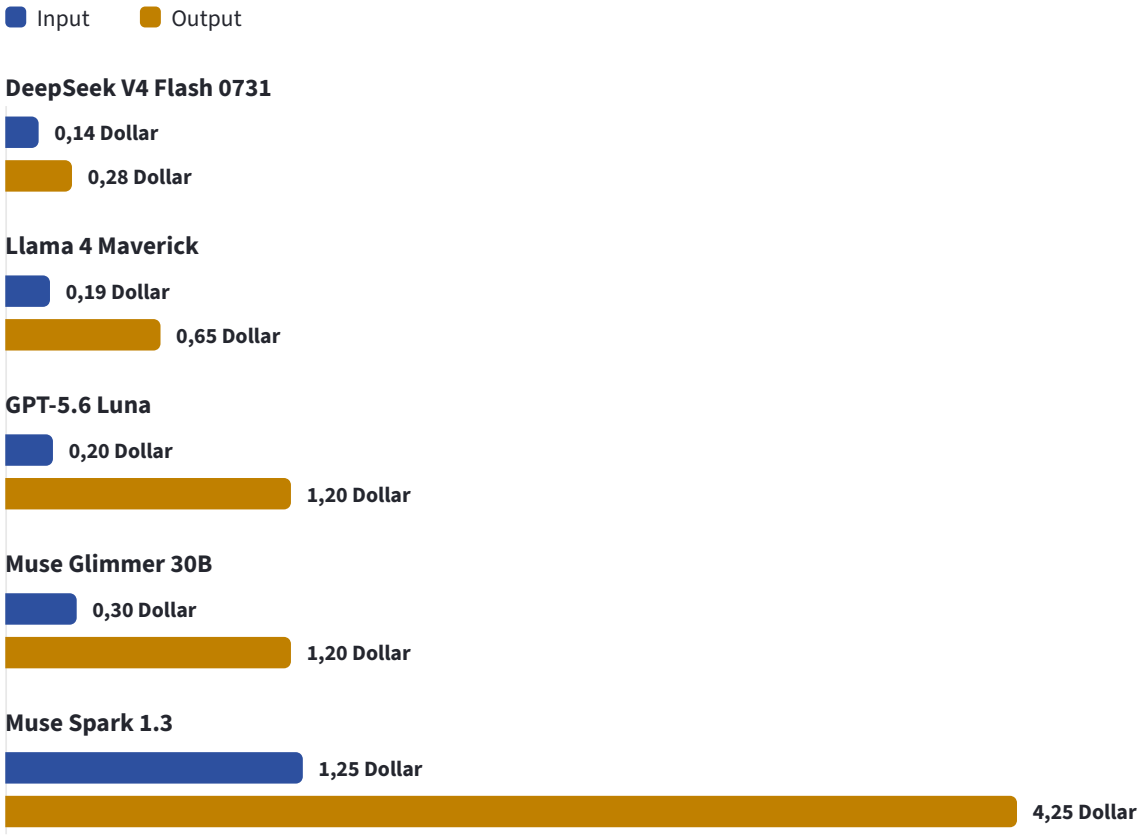
Günstigste der drei Stufen, auf Durchsatz statt Tiefe ausgelegt. Vom Anbieter für Klassifikation, Absichtserkennung, Moderation und einfache Zusammenfassungen positioniert. Technisch identisch mit Sol und Terra - bei einem Zwanzigstel des Eingabepreises von Sol.

AI NEWS pro - KI-News ohne Hype! · © 2026 SOFTCLICK Labs · Stand des Beitrags: 04.08.2026

Preise

API-Preise: GPT-5.6 Luna im Vergleich

Dollar je eine Million Token. Je kürzer der Balken, desto günstiger.



Quelle: Anbieter-Preislisten, Stand: 7. Oktober 2026 · Grafik: AI NEWS pro

Stufe	Input	Output	Cache (Lesen)	Stand
standard (OpenAI)	0,20	1,20	0,02	4. August 2026

Beträge in US-Dollar je einer Million Token, nach Anbieterangabe.

standard: Am 30. Juli 2026 um 80 Prozent gesenkt, von zuvor 1,00 Eingabe und 6,00 Ausgabe. Die Senkung betraf beide Seiten.

AI NEWS pro - KI-News ohne Hype! · © 2026 SOFTCLICK Labs · Stand des Beitrags: 04.08.2026

Benchmarks

Terminal-Bench 2.1: GPT-5.6 im Vergleich

Anteil gelöster Aufgaben in Prozent. Je länger der Balken, desto besser.

Quelle: OpenAI (Herstellerangabe), Stand: 9. Juli 2026 · Grafik: AI NEWS pro

Stand der Daten: 6. Oktober 2026

GPT-5.6 Luna

Die GPT-5.6-Familie ist in drei Stufen gestaffelt, und Luna bildet das untere Ende. Sol soll nach der Zuordnung des Anbieters komplexe agentische und wissenschaftliche Aufgaben übernehmen, Terra den Programmier- und Agentenalltag. Luna bleibt das Massengeschäft - auf Durchsatz ausgelegt statt auf Tiefe.

Gestaffelt wird dabei über die Leistung, nicht über die Ausstattung. Das aktuelle Kontextfenster, die Ausgabegrenze, die Bildeingabe, der Wissensstand und der Funktionsumfang von Reasoning über Tool Calls bis zu Structured Output - all das ist bei den drei Stufen gleich dokumentiert. Wer die günstigste wählt, verzichtet auf dem Papier auf nichts - was er in den Ergebnissen hergibt, zeigen erst die Messwerte.

OpenAI führt die drei Stufen als eigenständige Leistungsklassen, die sich unabhängig voneinander weiterentwickeln können. Was genau dahintersteckt, bleibt allerdings unveröffentlicht - weder Architektur noch Parameterzahl sind dokumentiert. Ein Erscheinungsdatum nennt die Herstellerdokumentation ebenfalls nicht; aus der Startberichterstattung ist die allgemeine Verfügbarkeit der Familie seit dem 9. Juli 2026 belegt.

Einordnung

Interessant ist Luna überall dort, wo die Rechnung über die Stückzahl läuft. Moderation, Klassifikation und Absichtserkennung sind Daueraufgaben, die millionenfach anfallen und pro Anfrage fast nichts kosten dürfen. Auf sie zielt der Preisabstand, denn Eingaben kosten bei Luna ein Zehntel des Terra-Satzes und ein

AI NEWS pro - KI-News ohne Hype! · © 2026 SOFTCLICK Labs · Stand des Beitrags: 04.08.2026

Zwanzigstel dessen, was Sol verlangt.

Dass auch die günstigste Stufe Bilder entgegennimmt, passt zu diesem Zuschnitt. Wer Inhalte prüft oder sortiert, bekommt selten nur Text vorgelegt - Strecken, die beides bewerten müssen, kommen so mit einer einzigen Schnittstelle aus.

Diesen Abstand geschaffen hat erst die Preisrunde vom 30. Juli. OpenAI hat die unterste Stufe um 80 Prozent verbilligt, die mittlere um 20 Prozent, die oberste gar nicht - vor der Senkung lag Terra beim Zweieinhalbfachen des Luna-Preises, seither liegt es beim Zehnfachen. Wer bisher zwischen den beiden unteren Stufen schwankte, rechnet seit Ende Juli unter anderen Vorzeichen.

In die Durchsatzrechnung gehört auch das Prompt Caching. Bereits verarbeitete Eingaben kosten beim Wiederverwenden ein Zehntel des regulären Eingabesatzes - für Daueraufgaben, die jede Anfrage mit denselben Anweisungen beginnen, ist das der Hebel zwischen Listenpreis und tatsächlicher Rechnung.

Auch das Kontextfenster spielt in diese Rechnung hinein. Es fasst in der aktuellen Größe umfangreiche Dokumente am Stück, und weil es in allen drei Stufen gleich groß ist, zwingen lange Eingaben allein niemanden in eine teurere Stufe. Verlassen sollte man sich darauf allerdings nicht: Beim Erinnerungsvermögen über lange Kontexte erreicht Luna in den Messwerten 41,3 Prozent, während Terra (89,6) und Sol (91,5) nah beieinanderliegen. Wer lange Dokumente nicht nur einlesen, sondern präzise auswerten will, stößt hier an Lunas deutlichste Grenze.

Die weiteren Grenzen zieht der Anbieter mit seiner eigenen Staffelung: Für komplexere Arbeit verweist die Familie nach oben zu Terra und Sol, auch wenn Luna Reasoning als Funktion mitbringt.

> **Gleiches Datenblatt heißt nicht gleiche Leistung**
>
> Die Messwerte trennen die Stufen deutlich: Bei schwerer Mathematik (FrontierMath) erreicht Luna 58,5 gegenüber 68,3 bei Terra und 83,0 bei Sol, beim Erinnern über lange Kontexte bricht der Wert auf 41,3 Prozent ein. Für Massenaufgaben nach Lunas Zuschnitt fällt das kaum ins Gewicht - für alles darüber verweist die Staffelung zu Recht nach oben.

Zu haben ist Luna wie die ganze Familie nur über Schnittstellen. Offene Gewichte gibt es nicht, ein Betrieb auf eigener Hardware scheidet aus. Neben OpenAIs eigener Schnittstelle führt auch OpenRouter das Modell.

Verwandte Modelle

Die Arbeitsteilung mit den Geschwistern ist oben schon angeklungen. GPT-5.6 Sol steht am oberen Ende und soll nach der Zuordnung des Anbieters die schweren Fälle übernehmen - ihr Eingabepreis liegt beim Zwanzigfachen von Luna, und die Preisrunde vom 30. Juli hat sie als einzige Stufe zunächst unverändert gelassen. GPT-5.6 Terra sitzt in der Mitte und ist für alltägliche Programmier-, Reasoning- und Agentenaufgaben gedacht, zum zehnfachen Luna-Satz, Ende Juli um 20 Prozent verbilligt.

AI NEWS pro - KI-News ohne Hype! · © 2026 SOFTCLICK Labs · Stand des Beitrags: 04.08.2026

Jenseits des OpenAI-Katalogs bedienen andere Anbieter dieselbe Preisregion. DeepSeek V4 Flash etwa listet Sätze, die nochmals unter denen von Luna liegen. Ein Qualitätsvergleich zwischen den beiden ist allerdings nicht belastbar, denn die vorliegenden Messwerte stammen aus unterschiedlichen Messaufbauten und lassen sich nicht direkt gegeneinanderstellen.

Quellen

1. OpenAI: GPT-5.6 Luna - Modellseite (Herstellerdokumentation) · https://developers.openai.com/api/docs/models/gpt-5.6-luna
2. OpenAI: Advancing the price-performance frontier with GPT-5.6 · https://openai.com/index/advancing-the-price-performance-frontier-with-gpt-5-6/
3. OpenAI: Preisangabe GPT-5.6 Luna (standard) · Stand 06.10.2026 · https://developers.openai.com/api/docs/changelog
4. OpenAI: Terminal-Bench 2.1: GPT-5.6 im Vergleich · Stand 09.07.2026 · https://openai.com/index/gpt-5-6/

Themen: Modelle · OpenAI

AI NEWS pro - KI-News ohne Hype! · © 2026 SOFTCLICK Labs · Stand des Beitrags: 04.08.2026