MODELLE

DeepSeek V4 Flash 0731 - Daten, Benchmarks, Preise

DeepSeeks Effizienzmodell im Stand vom 31. Juli 2026: MIT-Lizenz, Preise am unteren Marktrand und ein unabhängig gemessener Sprung gegenüber dem Vorgänger.

Stand: 04.08.2026 · 3 Min. Lesezeit

DeepSeek hat sein Effizienzmodell V4 Flash zum 31. Juli 2026 in einem aktualisierten Stand veröffentlicht, und eine unabhängige Messung sieht diesen Stand deutlich vor seinem Vorgänger. Hinter der Reihe steht das 2023 in Hangzhou gegründete Unternehmen DeepSeek. Bemerkenswert ist die Kombination aus frei verfügbaren Gewichten unter MIT-Lizenz und Listenpreisen am unteren Rand des Marktes. Was der Datumsstand tatsächlich ändert, zeigt sich zum Stichtag allerdings weniger in den technischen Eckdaten als in einer einzelnen Messreihe.

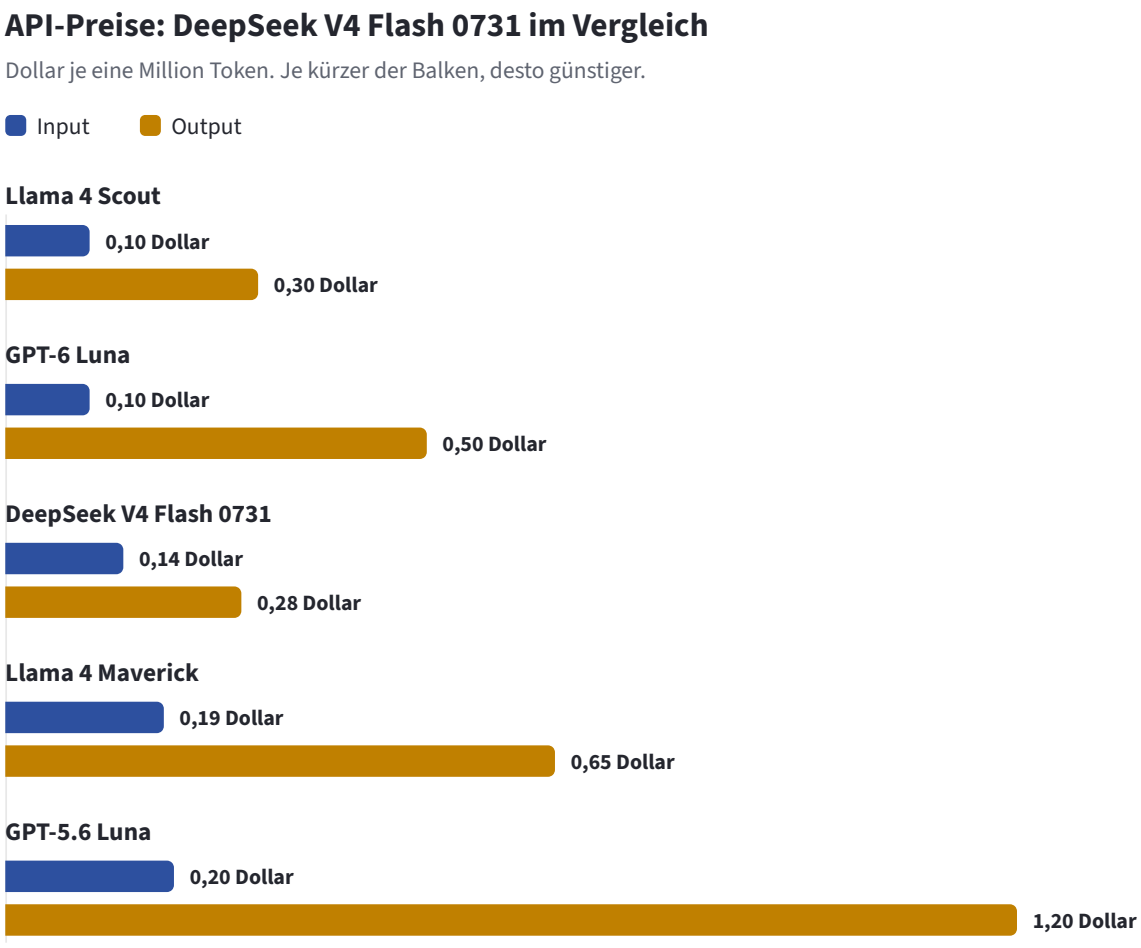
DeepSeek V4 Flash 0731 - Überblick und technische Daten

Anbieter	DeepSeek
Familie	DeepSeek V4 Flash
Erschienen	31. Juli 2026
Lizenz	MIT
Weights	open weights, Selbstbetrieb möglich
Nutzbar über	DeepSeek, OpenRouter
Preis ab	0,14 USD je Mio. Input-Token
Architektur	Mixture of Experts
Parameter gesamt	284 Milliarden
Parameter aktiv	13 Milliarden
Kontextfenster	1.048.576 (rund 1,05 Mio.) Token
Maximaler Output	384.000 Token

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

Input	Text
Output	Text
Wissensstand	April 2026
Prompt Caching	Ja Die Preisliste unterscheidet zwischen Eingabe-Token mit und ohne Cache-Treffer.
Kennung	deepseek-v4-flash-0731

Preise



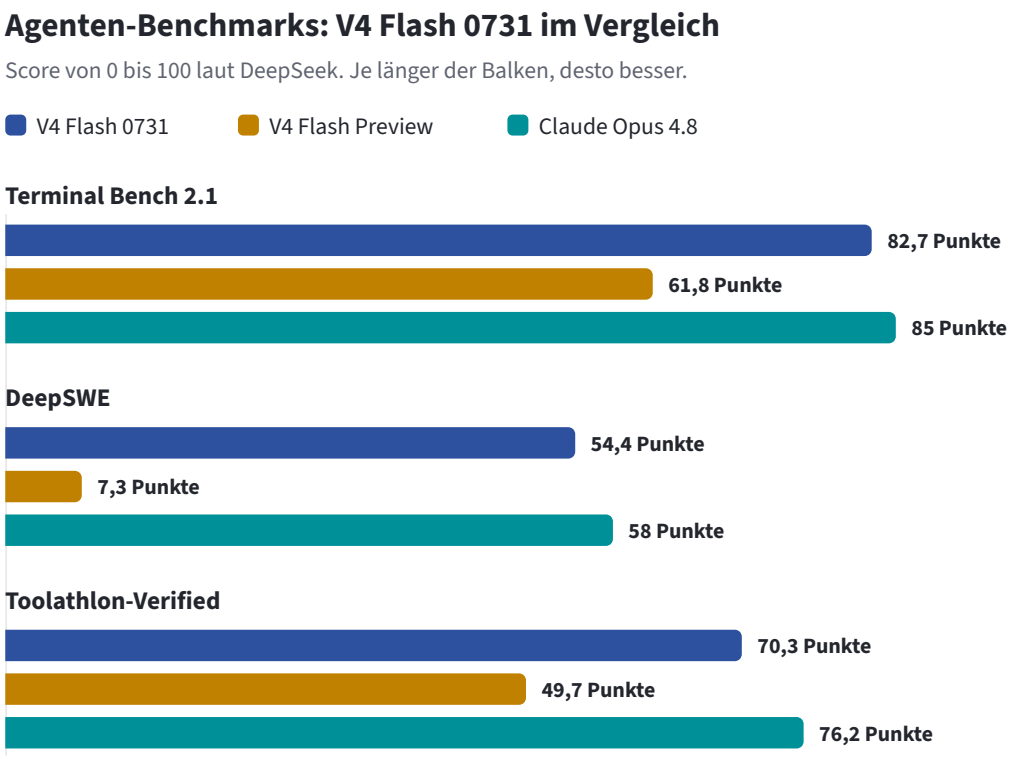
Quelle: Anbieter-Preislisten, Stand: 7. Oktober 2026 · Grafik: AI NEWS pro

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

Stufe	Input	Output	Cache (Lesen)	Stand
standard (DeepSeek)	0,14	0,28	-	4. August 2026

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

Benchmarks



Quelle: DeepSeek (Herstellerangabe), Stand: 31. Juli 2026 · Grafik: AI NEWS pro

Suite	Wert	Gemessen von	Bedingungen	Stand
GDPval-AA v2 (Fassung v2)	1.559	unabhängig (Artificial Analysis)	Elo-Wert, nicht Prozent	31. Juli 2026

Stand der Daten: 4. August 2026

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

DeepSeek V4 Flash 0731

Die vier Ziffern im Namen folgen einem Datumsschema. 0731 steht für den 31. Juli 2026, den Tag der Veröffentlichung - das Modell ist also keine neue Hauptversion, sondern der aktualisierte Stand der bestehenden Reihe V4 Flash. Deren bisherige Fassung führt das Portal seither als Vorgänger.

> **Häufiger Irrtum**
>
> Die Ziffernfolge 0731 ist ein Datumskürzel für den 31. Juli 2026 - keine Angabe zur Modellgröße. Auch die Kurzform "731" ist keine offizielle Bezeichnung.

Technisch ist der Stand 0731 ein Mixture-of-Experts-Modell. Je Rechenschritt ist nur ein kleiner Teil der gespeicherten Gewichte beteiligt, gut vier Prozent des Gesamtbestands. Genau dieser Zuschnitt macht die Flash-Reihe zu dem, was ihr Name verspricht - ein Modell, das auf Durchsatz und niedrige Betriebskosten hin gebaut ist.

Gesamtparameter	Aktive Parameter
Alle gespeicherten Gewichte eines Modells. Sie bestimmen, wie viel Speicher der Betrieb belegt.	Der Teil, der je Rechenschritt tatsächlich mitarbeitet. Er bestimmt die Rechenlast und damit die Kosten je Anfrage.

Für den Betrieb heißt das aber auch, dass der schmale Aktivanteil nur die Rechenlast senkt, nicht den Platzbedarf. Wer das Modell selbst betreibt, hält immer den vollen Parameterbestand vor.

Eingabe wie Ausgabe sind auf Text beschränkt. Bilder, Audio oder Dokumente nimmt das Modell nicht direkt an - das unterscheidet es von vielen aktuellen Modellen. Die Gewichte hat DeepSeek unter der MIT-Lizenz freigegeben, und das ist mehr als das übliche Open Weights-Versprechen. MIT ist eine von der Open Source Initiative anerkannte Lizenz ohne kommerzielle Schranken und ohne Sonderklauseln für große Nutzer. Das Modell lässt sich damit auch im eigenen Haus betreiben.

Einordnung

Vermessen hat den neuen Stand das Analysehaus Artificial Analysis, und zwar unabhängig vom Anbieter, in der Suite GDPval-AA v2. Die Suite stellt agentische Aufgaben aus realer Wissensarbeit und bewertet die Ergebnisse auf einer Elo-Skala, wie man sie vom Schach kennt, statt in Prozent. Der Abstand zum vorherigen Stand der Reihe beträgt 370 Elo-Punkte. Dieser Sprung, nicht der absolute Wert, ist der Kern der Messung - auch die Messenden selbst haben ihn als den eigentlichen Befund herausgestellt.

Zwei Einschränkungen gehören daneben. Ein Spitzenwert ist das nicht, teurere Modelle liegen in derselben Messreihe weiterhin darüber. Und es bleibt eine einzelne Suite - weitere unabhängig geprüfte Werte liegen für diesen Stand zum Stichtag nicht vor.

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

Der zweite Pfeiler der Reihe ist der Preis. Im Modellbestand dieses Portals hat V4 Flash 0731 zum Stichtag den niedrigsten belegten Listenpreis. Selbst die Sparstufe von OpenAIs aktueller GPT-Familie kostet bei der Eingabe knapp das Anderthalbfache, bei der Ausgabe mehr als das Vierfache. Der ausgewiesene Eingabepreis gilt für Token ohne Cache-Treffer. Dass es einen Cache gibt, belegt die Preisliste damit selbst, bezifferte Cache-Preise liegen zum Stichtag aber nicht vor.

Zusammen mit dem aktuellen Kontextfenster in der Millionen-Token-Klasse ergibt sich daraus ein klares Einsatzprofil. Das Modell taugt für Aufgaben in großer Stückzahl, bei denen der Stückpreis über die Machbarkeit entscheidet - etwa Zusammenfassen oder Auswerten über große Textbestände hinweg. Die MIT-Lizenz öffnet zusätzlich den Weg zum Selbstbetrieb. Wo dagegen Bildverarbeitung gefragt ist oder ein breit belegtes Leistungsbild über viele Aufgabenfelder, gibt die Datenlage zum Stichtag schlicht keine Antwort.

Verwandte Modelle

Der direkte Vorgänger ist das bisherige DeepSeek V4 Flash. In der genannten Messreihe lag es deutlich unter dem neuen Stand und hat als aktuelle Fassung der Reihe damit ausgedient. Einen Nachfolger gibt es nicht - der Stand 0731 ist zum Stichtag die aktuelle Fassung.

Preislich am nächsten steht ihm GPT-5.6 Luna, die Sparstufe der aktuellen OpenAI-Familie - geschlossen gehalten, dafür mit Bildeingabe. Kimi K3 von Moonshot AI verfolgt einen ähnlichen Offenheits-Ansatz aus China, allerdings unter einer eigenen Anbieterlizenz mit Umsatzschranke statt MIT. Qwen 3.8 Max von Alibaba Cloud ist multimodal und deutlich teurer, die angekündigte Freigabe seiner Gewichte stand zum Stichtag noch aus.

Quellen

1. Artificial Analysis: DeepSeek V4 Flash 0731 im Intelligence Index · https://artificialanalysis.ai/articles/deepseek-v4-flash-0731-scores-50-on-the-artificial-analysis-intelligence-index-10-points-above-previous-deepseek-v4-flash
2. OpenRouter: DeepSeek V4 Flash 0731 - Preise und Benchmarks · https://openrouter.ai/deepseek/deepseek-v4-flash-0731
3. DeepSeek: Agenten-Benchmarks: V4 Flash 0731 im Vergleich · Stand 31.07.2026 · https://huggingface.co/deepseek-ai/DeepSeek-V4-Flash-0731

Themen: Modelle · DeepSeek · Open Source

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