MODELLE

Kimi K3 - technische Daten, Benchmarks und Preise

Moonshot AI hat mit Kimi K3 ein Mixture-of-Experts-Modell im Billionen-Parameter-Maßstab vorgelegt, dessen Gewichte zum Herunterladen bereitstehen. Das Datenblatt zu Architektur, Lizenz und Preisen - von Kimi Delta Attention bis zur kommerziellen Schranke der Kimi K3 License.

Stand: 04.08.2026 · 4 Min. Lesezeit

Der Pekinger Anbieter Moonshot AI hat Mitte Juli 2026 mit Kimi K3 die nächste Generation seiner Kimi-Reihe vorgestellt - ein Mixture-of-Experts-Modell im Billionen-Parameter-Maßstab, das Text und Bilder verarbeitet. Elf Tage nach dem Produktstart hat der Anbieter auch die Gewichte zum Herunterladen freigegeben. Für ein KI-Publikum liegt die eigentliche Nachricht in der Architektur, denn die neu eingeführte Kimi Delta Attention senkt den Speicherbedarf langer Kontexte und zielt damit genau auf das Fenster im Millionen-Token-Bereich. Beim Preis hat sich Moonshot allerdings deutlich vom Vorgänger entfernt - K3 kostet rund das Fünffache von K2.6.

Kimi K3 - Überblick und technische Daten

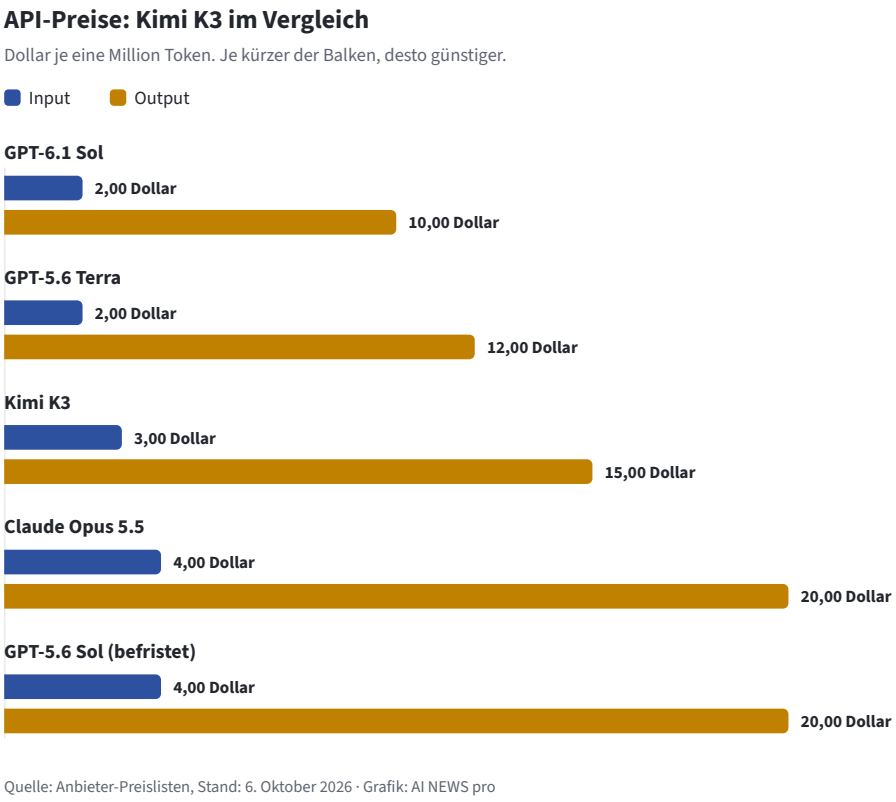
Anbieter	Moonshot AI
Familie	Kimi K
Erschienen	16. Juli 2026 Start am 16. Juli 2026 über Kimi Code und die Kimi-App. Die Gewichte folgten am 27. Juli 2026.
Lizenz	Kimi K3 License
Weights	open weights seit 27. Juli 2026, Selbstbetrieb möglich
Nutzbar über	Moonshot AI, OpenRouter
Preis ab	3,00 USD je Mio. Input-Token
Architektur	Mixture of Experts
Parameter gesamt	2,8 Billionen
Experten	896 (davon 16 aktiv)

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

Kontextfenster	1.000.000 (rund 1 Mio.) Token
Maximaler Output	131.072 Token Kein vom Anbieter festgeschriebener Höchstwert. 131.072 ist die Voreinstellung der Schnittstelle und damit die praktisch wirksame Grenze, nicht dasselbe wie eine dokumentierte Obergrenze.
Input	Text, Bild Bildverarbeitung ist nativ, nicht nachgerüstet.
Output	Text
Kennung	kimi-k3

- Kimi Delta Attention (KDA) im Verbund mit MLA senkt den Speicherbedarf des KV-Caches gegenüber voller Attention deutlich.
- Attention Residuals (AttnRes) als zweite Architekturneuerung gegenüber K2.

Preise



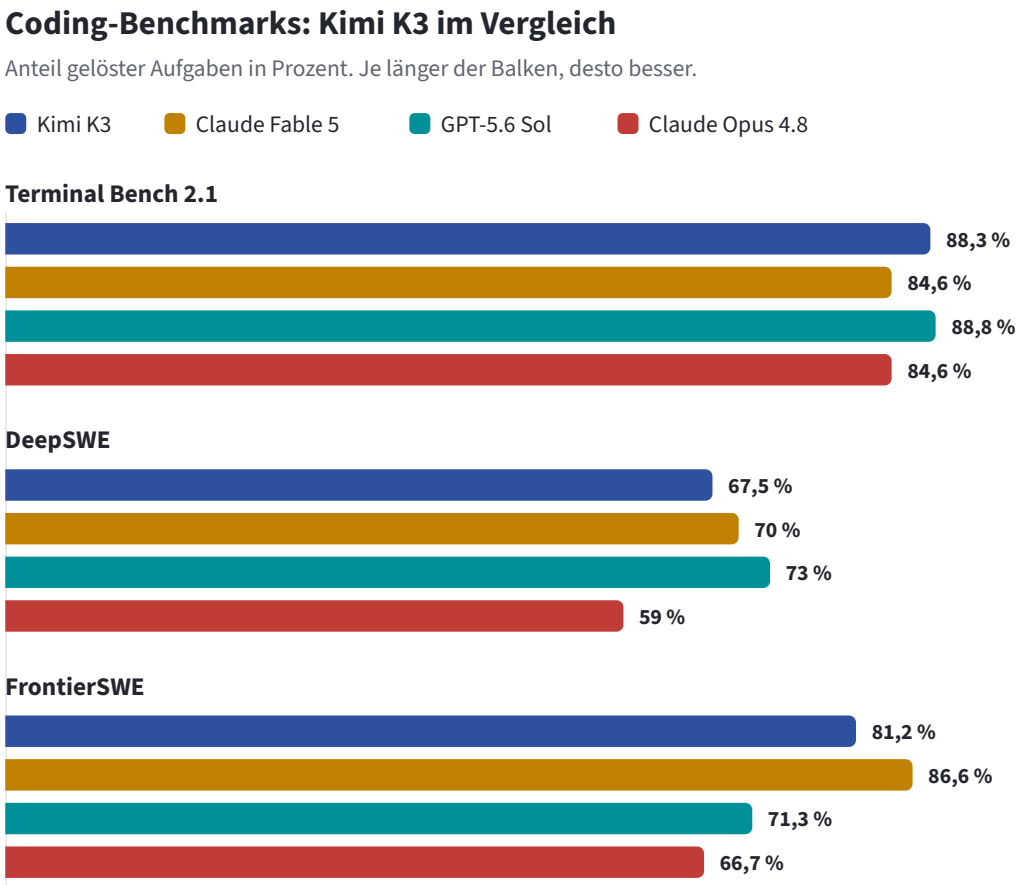
AI NEWS pro - KI-News ohne Hype! · © 2026 SOFTCLICK Labs · Stand des Beitrags: 04.08.2026

Stufe	Input	Output	Cache (Lesen)	Stand
standard (Moonshot AI)	3,00	15,00	-	4. August 2026

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

standard: Rund das Fünffache des Vorgängers K2.6.

Benchmarks



Quelle: Moonshot AI (Herstellerangabe), Stand: 16. Juli 2026 · Grafik: AI NEWS pro

Suite	Wert	Gemessen von	Bedingungen	Stand
GPQA Diamond (Fassung diamond)	93,5 %	Anbieter (Moonshot AI)	Bedingungen nicht angegeben	4. August 2026

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

Stand der Daten: 4. August 2026

Kimi K3

Moonshot AI ist 2023 in Peking gegründet worden und führt seine Sprachmodelle unter dem Namen Kimi. K3 ist der Nachfolger von Kimi K2 und setzt sich von ihm durch zwei Architekturneuerungen ab. Die erste ist Kimi Delta Attention, kurz KDA, die im Verbund mit MLA den Speicherbedarf des KV-Caches gegenüber voller Attention deutlich senkt. Die zweite trägt den Namen Attention Residuals, kurz AttnRes.

Als Mixture-of-Experts-Modell folgt K3 einem Prinzip, das die schiere Größe relativiert. Das Netz ist in viele Experten unterteilt, von denen bei jeder Anfrage nur ein kleiner Teil mitrechnet - die Gesamtzahl der Parameter beschreibt also die Kapazität des Modells, nicht die Last eines einzelnen Durchlaufs. Wie viele Parameter dabei tatsächlich aktiv sind, ist bislang nicht belegt. Bekannt ist nur, wie viele der Experten jeweils anspringen, und daraus allein lässt sich die Zahl nicht ableiten.

Auf der Eingabeseite verarbeitet K3 neben Text auch Bilder, und zwar nativ statt über ein nachgerüstetes Zusatzmodul. Die Ausgabe bleibt Text. Gestartet ist das Modell am 16. Juli 2026 über Kimi Code und die Kimi-App, die Gewichte hat Moonshot am 27. Juli auf Hugging Face nachgereicht - Produktstart und offene Verfügbarkeit sind hier also zwei verschiedene Daten.

Einordnung

Der greifbarste Nutzen von K3 liegt im Selbstbetrieb. Wer Modelle im eigenen Haus laufen lässt - wegen Vorgaben zur Datenhaltung oder zur Kostenkontrolle -, bekommt ein Modell dieser Größenordnung zum Herunterladen. Die Lizenz verdient dabei einen genauen Blick. Moonshot nennt sie Kimi K3 License, und sie ist eine eigene Anbieterlizenz, die von der Open Source Initiative nicht anerkannt ist - Hugging Face führt sie schlicht als "other". Inhaltlich ähnelt sie der MIT-Lizenz, trägt aber eine kommerzielle Schranke: Wer ein Model-as-a-Service-Geschäft betreibt und mit verbundenen Unternehmen in zwölf zusammenhängenden Monaten mehr als 20 Millionen US-Dollar umsetzt, braucht eine gesonderte Vereinbarung mit Moonshot. Für interne Nutzung und kleinere Anbieter ändert das nichts, für Inferenz-Dienstleister ab einer gewissen Größe ist es der entscheidende Absatz des Lizenztexts.

Anzuerkennen ist, dass Moonshot selbst durchgehend von Open Weights spricht und den Begriff Open Source vermeidet. Der Vorgänger K2 lief noch unter einer als "modified MIT" bezeichneten Lizenz - diese Bezeichnung hat der Anbieter bei K3 aufgegeben.

Open Weights	Open Source
Die Gewichte stehen zum Herunterladen bereit, das Modell lässt sich selbst betreiben. Was darüber hinaus erlaubt ist, bestimmt die Lizenz des Anbieters.	Eine von der Open Source Initiative anerkannte Lizenz sichert Nutzung, Änderung und Weitergabe für jeden Zweck zu - auch kommerziell, ohne Umsatzschranke.

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

Beim Preis hat Moonshot die Richtung gewechselt. Der API-Zugang kostet rund das Fünffache des Vorgängers K2.6, und eine Steigerung in dieser Höhe ist mehr als eine Anpassung - sie sortiert das Modell in eine andere Preisklasse ein. Wer den Vorgänger vor allem wegen seines Preises gewählt hat, rechnet bei K3 neu.

Das Kontextfenster im Millionen-Token-Bereich hängt direkt mit der Architektur zusammen. Lange Kontexte kosten vor allem Speicher, weil der KV-Cache mit der Kontextlänge wächst, und genau an diesem Engpass setzt Kimi Delta Attention an. Offen bleibt allerdings, wie viel eines solchen Fensters in der Praxis trägt. Das beantworten erst unabhängige Messungen, und die gab es für K3 bis Anfang August 2026 nicht.

Damit ist auch die Benchmark-Lage beschrieben, denn belastbare Vergleichswerte sind rar. Moonshot hat zum Start einen Wert für die naturwissenschaftliche Prüfungsreihe GPQA genannt, allerdings ohne die Messbedingungen offenzulegen - und ohne Klärung, ob sich die Zahl auf die bereinigte Diamond-Teilmenge oder auf die Vollfassung der Prüfung bezieht. Ein Anbieterwert ohne Bedingungen und ohne eindeutige Zuordnung lässt sich mit nichts vergleichen. Ähnliches gilt für die Effizienz: Solange die Zahl der je Durchlauf aktiven Parameter nicht belegt ist, bleibt offen, wie sich die Rechenlast eines K3-Aufrufs zu anderen Mixture-of-Experts-Modellen verhält.

Verwandte Modelle

Innerhalb der Reihe ist Kimi K2 der Vorgänger, von dem sich K3 durch die beiden Architekturneuerungen und die geänderte Lizenzbezeichnung absetzt. Dessen Stufe K2.6 ist der Preisvergleichspunkt - ihr API-Zugang liegt Anfang August 2026 bei rund einem Fünftel des K3-Preises.

Unter dem K3-Namen führt Moonshot zwei weitere Ausführungen. K3 Max ist für Dialog und Agenten gedacht, K3 Swarm Max für Parallelverarbeitung in großem Umfang.

Außerhalb des eigenen Hauses ist DeepSeek V4 Flash 0731 der nächste Verwandte - ebenfalls ein Mixture-of-Experts-Modell aus China mit offenen Gewichten und einem Kontextfenster derselben Größenordnung. Qwen 3.8 Max von Alibaba setzt gleichfalls auf Mixture of Experts und ein ähnlich großes Fenster, hält die Gewichte aber unter Verschluss - genau darin liegt der Unterschied zu K3.

Quellen

1. MarkTechPost: Moonshot AI Releases Kimi K3 · https://www.marktechpost.com/2026/07/16/moonshot-ai-releases-kimi-k3-a-2-8-trillion-parameter-open-moe-model-with-kimi-delta-attention-and-1m-context/
2. OpenRouter: Kimi K3 - Preise und Benchmarks · https://openrouter.ai/moonshotai/kimi-k3

Themen: Modelle · Moonshot AI · Open Weights

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