Intelligenz Welt

Kimi K3

KI-Modellportrait · Quellenstand 2026-10-31

KIMI

KIMI-Anbietericon aus dem offiziellen Brand-Guide, kein eigenes Modelllogo · Logo-Quelle / Rechtehinweis. Marken gehören ihren jeweiligen Inhabern. Redaktionelle Identifikation, keine Partnerschaft oder Empfehlung durch den Markeninhaber.

Kimi K3 ist ein offenes Multimodal-Modell von Moonshot AI mit 2,8 Billionen Parametern, einem Kontextfenster von einer Million Token und Fokus auf langlaufende Coding- und Wissensarbeitsaufgaben.

Quellen geprüft am 31. Oktober 2026 · Redaktion: Philipp Paulik

Entwickler: Moonshot AI →

DAS MODELL EINFACH ERKLÄRT

Offenes 2,8-Billionen-Modell mit Millionenkontext richtig einordnen.

Kimi K3 ist ein offenes Multimodal-Modell von Moonshot AI für langlaufendes Coding, Wissensarbeit und Reasoning. Es verarbeitet Text, Bilder und über die API auch Videos und erzeugt Text. Im API-Steckbrief stehen 1.048.576 Token Kontext; die maximale Ausgabe ist bis 1.048.576 Token einstellbar (Standard 131.072).

Offizielle Dokumentation: Kimi K3

Quellen geprüft: 31.10.2026. Herstellerangaben, kein eigener Modelltest.

Wofür eignet sich Kimi K3?

Kimi K3 richtet sich laut Moonshot AI an lange, agentische Aufgaben: Coding-Sessions über große Repositories, Recherche mit Werkzeugen, Dokument- und Bildverständnis. Als offenes Modell eignet es sich auch für Selbsthosting, wenn die nötige Hardware vorhanden ist. Ob es im eigenen Projekt passt, hängt von Ergebnisqualität, Antwortzeit, Kosten und dem verwendeten Agenten-Framework ab, nicht allein von der Parameterzahl. Wer das aktuell stärkste Kimi-Modell sucht, muss K3 mit dem neueren K3.1 Pro vergleichen.

Einordnung: Vorgänger-Flagship mit offenen Gewichten

Moonshot AI führt Kimi K3 inzwischen als previous-generation flagship: Seit dem 24.10.2026 ist Kimi K3.1 Pro das aktuelle Spitzenmodell. K3 bleibt über die Kimi API, Amazon Bedrock und die offenen Gewichte verfügbar. Wer K3 einsetzt, sollte die Migrationsschritte zu K3.1 Pro kennen und beide Modelle an den eigenen Aufgaben messen.

Offizielle Dokumentation: Kimi K3

Architektur: KDA, Attention Residuals und 896 Experten

Kimi Delta Attention ist laut Anbieter ein hybrider linearer Attention-Mechanismus; Attention Residuals sollen den Informationsfluss über lange Sequenzen und tiefe Modelle verbessern. Mit dem Stable-LatentMoE-Framework aktiviert das Modell 16 von 896 Experten. Moonshot beziffert die Skalierungseffizienz auf etwa das 2,5-Fache von Kimi K2. Das sind Herstellerangaben zur Architektur, keine unabhängige Prüfung.

Offizielle Dokumentation: Kimi K3

Denken ist immer an

Kimi K3 gibt reasoning_content zurück; ein Abschalten des Denkmodus ist nicht vorgesehen. Der Aufwand steuerbar über reasoning_effort mit den Stufen low, high und max (Standard max). Für Dialoge mit mehreren Runden und Tool-Aufrufen verlangt die API die vollständige Assistant-Nachricht zurück, inklusive reasoning_content und tool_calls. Wer nur content weiterreicht, verliert den Gedächtniszustand des Modells.

Offizielle Dokumentation: Kimi K3

Kontext, Cache und Kosten

Das Kontextfenster umfasst 1.048.576 Token, die Preislogik ist flach: keine Staffelung nach Kontextlänge. Kontext-Caching läuft automatisch mit einer TTL von fünf Minuten, optional einer Stunde; Cache-Schreibzugriffe werden separat nach TTL abgerechnet. Die Batch-API kostet laut Preisseite 60 % des Standardpreises. Moonshot nennt eine Cache-Trefferquote über 90 % in Coding-Workloads der eigenen API (Herstellerangabe).

Offizielle Dokumentation: Kimi K3

Multimodal: Text, Bild und Video hinein, Text heraus

Über die Kimi API nimmt das Modell Text, Bilder (URL oder Base64) und Videos (Datei-Upload) entgegen; die Ausgabe ist Text. Die Hugging-Face-Modellkarte weist als Modalität Text und Bild aus. Auf Amazon Bedrock ist Video als Eingabe nicht ausgewiesen. Audioeingabe, Sprachdialog und Bild- oder Videoerzeugung sind nicht Teil dieses Steckbriefs.

Offizielle Dokumentation: Kimi K3

Lizenz und Selbsthosting

Die Kimi K3 License erlaubt Nutzung, Veränderung und Vertrieb inklusive kommerzieller Nutzung. Zwei Schwellen gelten: Model-as-a-Service-Anbieter ab 20 Mio. USD Jahresumsatz brauchen eine separate Vereinbarung; sehr große Produkte (über 100 Mio. monatlich aktive Nutzer oder über 20 Mio. USD Monatsumsatz) müssen Kimi K3 prominent ausweisen. Interne Nutzung und Nutzung über Moonshot-Produkte oder zertifizierte Partner sind davon ausgenommen. Für das Selbsthosting empfiehlt der Anbieter vLLM, SGLang oder TokenSpeed und Supernode-Konfigurationen mit mindestens 64 Beschleunigern.

Offizielle Dokumentation: Kimi K3

Benchmarks als Herstellerangaben lesen

Im eigenen Vergleich nennt Moonshot unter anderem GPQA Diamond 93,5, Terminal-Bench 2.1 88,3 und SWE-Marathon 42,0 für K3 mit Aufwand max. Gleichzeitig schreibt der Anbieter im Tech-Blog, die Gesamtleistung liege hinter den stärksten proprietären Modellen Claude Fable 5 und GPT-5.6 Sol, vor den übrigen getesteten Modellen. Harness, Aufwand und Stichprobendaten unterscheiden sich je Benchmark; die Werte ersetzen keinen Test an den eigenen Aufgaben.

Offizielle Dokumentation: Kimi K3

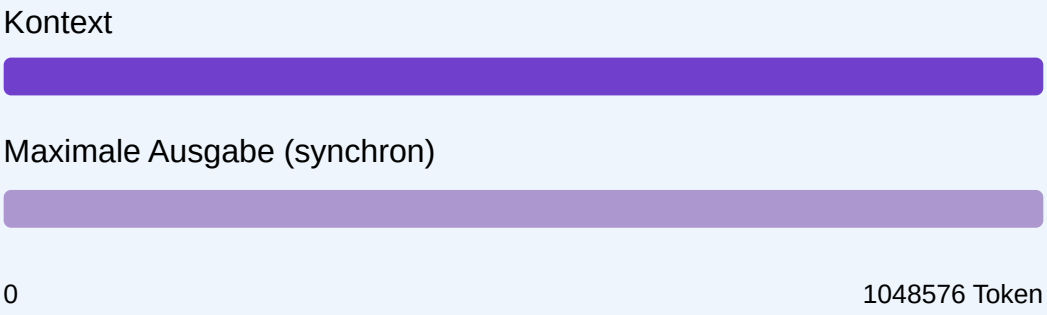
Kontextfenster ist nicht Ausgabelimit.

1.048.576

Token Kontextfenster

1.048.576

maximale Ausgabe-Token, synchron



Zwei unterschiedliche Grenzen auf gemeinsamer Nullskala, nicht addieren. Offizielle Dokumentation: Kimi K3

Was kostet Kimi K3 über die API?

Standard-Eingabe

3 USD

je 1 Mio. Texttoken · Basisrate

Standard-Ausgabe

15 USD

je 1 Mio. Texttoken · Basisrate

Flache Pay-as-you-go-Preise in USD je 1 Mio. Token: Eingabe (Cache-Miss) 3,00; Eingabe (Cache-Treffer) 0,30; Ausgabe 15,00. Keine Staffelung nach Kontextlänge. Cache-Schreibzugriffe werden nach TTL (5 min / 1 h) separat

abgerechnet; die konkreten TTL-Beträge stehen in einer clientseitig gerenderten Tabelle auf der Preisseite und konnten beim Abruf nicht belegt werden. Batch-API: 60 % des Standardpreises. Preise ohne Steuern. Freischaltung nach erster Aufladung (mindestens 1 USD); die kumulierte Aufladung bestimmt Tier und Ratenlimits (Concurrency, RPM, TPM, TPD). Keine Abopreise. Stand: 31.10.2026.

Preisstand 31.10.2026. Offizielle Dokumentation: Kimi K3

Vor dem Einsatz prüfen

- Modell-ID kimi-k3 und den verwendeten Endpunkt (Kimi API oder Bedrock) festhalten; K3.1 Pro als Nachfolger an denselben Aufgaben vermessen.
- reasoning_effort explizit setzen und bei Multiturn-Dialogen die vollständige Assistant-Nachricht inklusive reasoning_content zurückreichen.
- Cache-TTL, Cache-Schreibkosten und Batch-Rabatt getrennt im Budget erfassen; bei Selbsthosting die Schwellen der Kimi K3 License prüfen.

Häufige Fragen zu Kimi K3

Was ist Kimi K3?

Ein offenes Multimodal-Modell von Moonshot AI mit 2,8 Billionen Parametern, 1-Million-Token-Kontext und Fokus auf langlaufende Coding- und Wissensarbeitsaufgaben. Die API-Modell-ID lautet kimi-k3; Kimi bezeichnet zugleich die Anwendung und die Modellfamilie.

Ist Kimi K3 das aktuelle Kimi-Flagship?

Nein. Moonshot AI bezeichnet K3 als previous-generation flagship. Das aktuelle Spitzenmodell ist Kimi K3.1 Pro (24.10.2026). K3 bleibt über API, Bedrock und offene Gewichte verfügbar.

Wie groß sind Kontext und Ausgabe von Kimi K3?

Das Kontextfenster umfasst 1.048.576 Token. Die maximale Ausgabe (max_completion_tokens) ist bis 1.048.576 Token einstellbar und steht standardmäßig auf 131.072. Kontext und Ausgabe sind getrennte Grenzen; Reasoning-Token zählen mit.

Was kostet Kimi K3 über die API?

Laut Anbieter 3,00 USD je Million Eingabe-Token (0,30 USD bei Cache-Treffern) und 15,00 USD je Million Ausgabe-Token, flach ohne Kontextlängen-Staffelung. Cache-Schreibzugriffe und Batch-Verarbeitung (60 % des Standardpreises) verändern die Rechnung. Stand: 31.10.2026.

Darf Kimi K3 kommerziell selbst gehostet werden?

Ja, unter der Kimi K3 License. Ab bestimmten Größenordnungen gelten Zusatzpflichten: Model-as-a-Service ab 20 Mio. USD Jahresumsatz erfordert eine separate Vereinbarung; Produkte über 100 Mio. monatlich aktive Nutzer oder 20 Mio. USD Monatsumsatz müssen Kimi K3 prominent ausweisen. Interne Nutzung ist ausgenommen.

Alle Antworten beziehen sich auf die verlinkte Herstellerdokumentation und den angegebenen Stand. Kein unabhängiger Benchmark, keine Garantie für Verfügbarkeit oder Ergebnisqualität.

Quellen & Stand

1. Kimi API Platform: Kimi K3 (Quickstart)
   Geprüft: 2026-10-31
2. Kimi K3 Tech Blog: Open Frontier Intelligence
   Geprüft: 2026-10-31
3. MoonshotAI/Kimi-K3: README und Modellzusammenfassung
   Geprüft: 2026-10-31
4. moonshotai/Kimi-K3: Hugging-Face-Modellkarte
   Geprüft: 2026-10-31
5. Kimi K3 License (Volltext)
   Geprüft: 2026-10-31
6. Kimi API Platform: Model Inference Pricing
   Geprüft: 2026-10-31
7. Amazon Bedrock: Kimi K3 Model Card
   Geprüft: 2026-10-31

Redaktionelle Zusammenfassung mit KI-Unterstützung, verantwortet von Philipp Paulik. Kein eigener Leistungstest. Änderbare Anbieterangaben gelten für den angegebenen Quellenstand.

Intelligenz Welt · Redaktion: Philipp Paulik · Quellenstand: 2026-10-31
Vollständiges Webportrait: https://intelligenz-welt.de/ki-modelle/kimi-k3/
Statischer Export, kein Live-Datenblatt. Preise, Zugriff und Anbieterangaben können sich ändern. Quellen und ihre jeweiligen Abrufstände stehen oben. Impressum · Transparenz