Intelligenz Welt

Kimi K3.1 Pro

KI-Modellportrait · Quellenstand 2026-10-31



KIMI-Anbieterlogo von Moonshot AI (offizieller Branding-Guide); kein modellspezifisches Logo · Logo-Quelle / Rechtehinweis. Marken gehören ihren jeweiligen Inhabern. Redaktionelle Identifikation, keine Partnerschaft oder Empfehlung durch den Markeninhaber.

Kimi K3.1 Pro ist das erste Flaggschiff der K3.1-Serie von Moonshot AI: ein Open-Weight-Modell der 3-Billionen-Parameter-Klasse mit einem Kontextfenster von einer Million Tokens. Moonshot positioniert es für Coding-Agenten, 3D-Erstellung, Finanzrecherche und lange agentische Workflows; die API-Preise liegen unter denen des Vorgängers Kimi K3.

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

Entwickler: Moonshot AI →

DAS MODELL EINFACH ERKLÄRT

Kimi K3.1 Pro: Moonshots Open-Weight-Flaggschiff mit einer Million Tokens Kontext

Kimi K3.1 Pro ist das erste Modell der K3.1-Serie von Moonshot AI und steht seit dem 24.10.2026 über die Kimi API (OpenAI- und Anthropic-kompatibel) sowie in den Kimi-Produkten (Web, App, Work, Code) bereit. Es bietet ein Kontextfenster von 1.048.576 Tokens, immer aktives Reasoning mit steuerbarem Aufwand (low/high/max) und kostet in der API 2,00 USD je Million Eingabe- und 10,00 USD je Million Ausgabe-Tokens; eine Highspeed-Variante ist zwei- bis dreimal schneller und doppelt so teuer. Die vollständigen Gewichte sollen bis 04.11.2026 als Open Weights erscheinen.

Offizielle Dokumentation: Kimi K3.1 Pro

Quellen geprüft: 31.10.2026. Herstellerangaben, kein eigener Modelltest.

Wofür eignet sich Kimi K3.1 Pro?

Für die Praxis ist K3.1 Pro vor allem als Coding- und Agentenmodell mit sehr langem Kontext interessant: ganze Codebases, lange Dokumentenstapel und mehrstufige Workflows passen ohne Fragmentierung in ein Fenster. Moonshot selbst empfiehlt es für Programmier-Agenten wie Kimi Code, Codex und Claude Code. Die Standardvariante ist für Hintergrund- und Batchaufgaben die günstigere Wahl; wer interaktiv arbeitet (Live-Coding, Assistenten), greift zur

Highspeed-Variante. Preislich liegt das Standardmodell nach Herstellerangabe etwa ein Drittel unter Kimi K3, bei identischem Kontextumfang. Wichtig: Der Denkmodus lässt sich nicht abschalten; wer Latenz und Kosten drücken will, setzt reasoning_effort auf low statt auf kürzere Prompts. Prompt Caching (5-Minuten- oder 1-Stunden-TTL) senkt Kosten bei wiederholten langen Präfixen deutlich.

Positionierung und Architektur

Kimi K3.1 Pro ist das erste Modell der K3.1-Serie und nach Herstellerangabe das fähigste Open-Weight-Modell der 3-Billionen-Parameter-Klasse. Es entstand als nachtrainierte Version von Kimi K3 mit wesentlich mehr Post-Training-Rechenleistung; Moonshot beschreibt ein großskaliertes gemeinsames RL-Training über Hunderte Anwendungsdomänen, bei dem jede Stufe zusätzlicher Rechenleistung planbare Zuwächse liefern sollte. Das Ergebnis positioniert der Hersteller gegen die stärksten proprietären Modelle: Auf Terminal-Bench 4.0 erreicht K3.1 Pro 65,7 Prozent, knapp hinter Claude Opus 5.5 (66,4 Prozent) und vor GPT-6 Astra (58,2 Prozent). Alle Benchmark-Werte stammen aus Herstellerangaben und sind keine unabhängigen Tests.

Offizielle Dokumentation: Kimi K3.1 Pro

Coding, Agenten und 3D-Erstellung

Der Schwerpunkt liegt auf langen Coding- und Agentenaufgaben. Moonshot nennt Beispiele von Quickfixes in echten Codebases bis zu kompletten Projekt-Builds und berichtet von einem internen Test, in dem das Modell einen Lauf-Simulator mit 167 Missionen und 420 Reviews iterativ verbesserte, darunter das Erkennen und Korrigieren von 36 Fehlern in Muskel-Modellrechnungen. Im 3D-Bereich erreicht es auf BenchCAD (Rekonstruktion aus Mehrfachansichten per CAD-Code) 92,7 Prozent mit Werkzeugen und kann laut Hersteller Skripte für Blender und CAD-Programme direkt schreiben und ausführen. Auf kimi.ai und in der App ist mit 3D Make ein passendes Produkt-Feature live.

Offizielle Dokumentation: Kimi K3.1 Pro

Finanzrecherche und Faktentreue

Für Finanzworkflows meldet Moonshot einen Bestwert auf OfficeQA Pro v2 (64,4 Punkte), einem Benchmark für mehrstufige Analyse echter Finanzdokumente, sowie Zuwächse auf Frontier Finance und internen Finanz-Benchmarks. Parallel wurde die Faktentreue verbessert: In einem internen Stresstest sank die Claim-Level-Halluzinationsrate von 23,4 Prozent (K3) auf 8,3 Prozent; der Hersteller weist ausdrücklich darauf hin, dass diese Werte aus bewusst fehlerträchtig konstruierten Beispielen stammen und nicht den Alltagsfehlerraten entsprechen.

Offizielle Dokumentation: Kimi K3.1 Pro

Cyber-Kontrollen und Open-Weight-Prinzip

Moonshot beschreibt ungewöhnlich offen, wie die Cyber-Fähigkeit des Modells kontrolliert wurde: Cyberbezogene Daten wurden aus Mid-Training-Korpus und RL-Umgebungen identifiziert und entfernt. Auf dem internen xdev-bench-hard (Exploit-Entwicklung, pass@3) löst ein Checkpoint ohne diese Maßnahmen 90,5 Prozent der Aufgaben, das freigegebene K3.1 Pro 81,0 Prozent. Der Hersteller begründet die Open-Weight-Freigabe mit zwei Prinzipien: keine Gewichte, die die absolute Cyber-Fähigkeitsgrenze erweitern, und schrittweise Öffnung, damit mehr Teams defensive Code-Audits auf eigener Infrastruktur fahren können.

Offizielle Dokumentation: Kimi K3.1 Pro

Zugang, Preise und Highspeed-Variante

In der API wird K3.1 Pro nach einer Erstaufladung ab 1 US-Dollar freigeschaltet; Zero Data Retention ist verfügbar. Die Preise sind flach über das ganze Millionen-Token-Fenster: 2,00 USD je Million Eingabe-Tokens, 10,00 USD je Million Ausgabe-Tokens, 0,20 USD für Cache-Treffer; Cache-Schreiben kostet 2,00 USD (5-Minuten-TTL) beziehungsweise 4,00 USD (1-Stunden-TTL) je Million Tokens. Die Highspeed-Variante kimi-k3.1-pro-highspeed verdoppelt die regulären Raten (4,00/20,00 USD) bei zwei- bis dreifacher Geschwindigkeit, der Cache-Treffer-Preis bleibt gleich. In den Kimi-Produkten ist das Basismodell ab der kostenlosen Stufe eingeschränkt und ab Plus (15 USD/Monat bei Jahresabo) enthalten; Highspeed gibt es ab Pro (31 USD/Monat).

Offizielle Dokumentation: Kimi K3.1 Pro

Offene Gewichte und Lizenzlage

Kimi K3.1 Pro ist als Open-Weight-Modell angekündigt: Die vollständigen Gewichte sollen bis zum 04.11.2026 veröffentlicht werden. Zum Prüfzeitpunkt (31.10.2026) war das Hugging-Face-Repository moonshotai/Kimi-K3.1-Pro noch nicht öffentlich; die konkrete Gewichtelizenz ist damit noch nicht belegt und muss bei Release geprüft werden, bevor Aussagen zur Nutzung der Gewichte gemacht werden.

Offizielle Dokumentation: Kimi K3.1 Pro

Ein- und Ausgaben im Überblick.

Eingabe

Text, Bild, Video

Ausgabe

Text

Kontext / Grenzen

1.048.576 Tokens

Maximale Ausgabe

bis 1.048.576 Tokens (Standard 131.072)

Was kostet Kimi K3.1 Pro über die API?

Preisstand 31.10.2026 (Hersteller-Preisseite, geprüft am selben Tag): Standard 2,00 USD Input / 10,00 USD Output je 1 Mio. Tokens, Cache-Treffer 0,20 USD, Cache-Write 2,00 USD (5 min) bzw. 4,00 USD (1 h). Highspeed 4,00/20,00 USD. Alle Preise USD vor Steuern; Mitgliedschaften und API werden getrennt abgerechnet.

Vor dem Einsatz prüfen

- Von Kimi K3: API-Aufruf bleibt gleich, nur das model-Feld auf kimi-k3.1-pro setzen; Standardpreise liegen laut Hersteller etwa ein Drittel unter K3 (K3: 3,00/15,00 USD je 1 Mio. Tokens). Denkmodus ist bei K3.1 Pro immer aktiv, reasoning_effort steuert den Aufwand.
- Feste Sampling-Parameter beachten: temperature, top_p, n und die Penalty-Parameter sind fix und sollten aus bestehenden Requests entfernt werden.

Häufige Fragen zu Kimi K3.1 Pro

Was kostet Kimi K3.1 Pro in der API?

2,00 USD je Million Eingabe-Tokens und 10,00 USD je Million Ausgabe-Tokens, Cache-Treffer 0,20 USD je Million. Die Highspeed-Variante kostet 4,00 beziehungsweise 20,00 USD. Alle Angaben in USD vor Steuern, Stand 31.10.2026.

Wie groß ist das Kontextfenster?

1.048.576 Tokens (eine Million), ohne Preisstaffelung nach Kontextlänge. Die maximale Ausgabe ist per max_completion_tokens auf bis zu 1.048.576 Tokens einstellbar, Standard sind 131.072.

Kann man den Denkmodus abschalten?

Nein. K3.1 Pro denkt immer; über reasoning_effort (low, high, max; Standard max) lässt sich der Reasoning-Aufwand und damit Latenz und Tokenverbrauch steuern.

Was ist der Unterschied zu Kimi K3.1 Pro Highspeed?

Gleiche Fähigkeiten bei zwei- bis dreifacher Ausgabegeschwindigkeit, aber doppelte reguläre Tokenpreise. Für interaktive Arbeit wie Live-Coding gedacht; die Standardvariante ist günstiger.

Ist Kimi K3.1 Pro Open Source?

Es ist als Open-Weight-Modell angekündigt: Die Gewichte sollen bis 04.11.2026 veröffentlicht werden. Die konkrete Lizenz war zum Prüfzeitpunkt noch nicht veröffentlicht und muss bei Release geprüft werden.

Welche Eingaben unterstützt das Modell?

Text, Bilder und Videos. Bilder werden als Base64 oder Datei-Upload übergeben, öffentliche Bild-URLs werden nicht unterstützt; Ausgabe ist Text.

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 K3.1 Pro Tech Blog: Advancing Open Intelligence (Moonshot AI)
   Geprüft: 2026-10-31
2. Kimi K3.1 Pro Pricing: Membership Plans and API Costs (Moonshot AI)
   Geprüft: 2026-10-31
3. Kimi K3.1 Pro - Kimi API Platform (Quickstart, Limits, Parameter)
   Geprüft: 2026-10-31
4. Kimi API Quickstart / Modellliste (platform.moonshot.ai)
   Geprüft: 2026-10-31
5. Kimi K3.1 Pro: □□□□□□□□□ (offizielle Release-Meldung, 24.10.2026)
   Geprüft: 2026-10-31
6. Moonshot AI - offizielle Website
   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-k3-1-pro/
Statischer Export, kein Live-Datenblatt. Preise, Zugriff und Anbieterangaben können sich ändern. Quellen und ihre jeweiligen Abrufstände stehen oben. Impressum · Transparenz