

Kimi K3.1 Pro

KI-Modellportrait · Quellenstand 2026-10-31



KIMI-Anbieterlogo von Moonshot AI (offizieller Branding-Guide); kein modellspezifisches Logo · [Logo-Quelle / Rechteinweis](#). 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](#) →

Technische Einordnung

Kimi K3.1 Pro ist laut Moonshot AI eine nachtrainierte Version von Kimi K3 mit deutlich mehr Post-Training-Rechenleistung; das RL-Training lief gemeinsam über Hunderte Anwendungsdomänen. Das Modell denkt immer mit: Der Denkmodus lässt sich nicht abschalten, aber über das API-Feld `reasoning_effort` in den Stufen low, high und max (Standard max) steuern. Es verarbeitet Text, Bilder und Videos als Eingabe und gibt Text aus; die maximale Ausgabelänge lässt sich per `max_completion_tokens` auf bis zu 1.048.576 Tokens setzen (Standard 131.072). Neben der Standardvariante gibt es `kimi-k3.1-pro-highspeed` mit nach Herstellerangabe zwei- bis dreifacher Ausgabegeschwindigkeit zum doppelten Tokenpreis. Von den Hersteller-Benchmarks hebt Moonshot Terminal-Bench 4.0 (65,7 Prozent, knapp hinter Claude Opus 5.5 mit 66,4 Prozent), BenchCAD für 3D-Rekonstruktion (92,7 Prozent mit Werkzeugen) und OfficeQA Pro v2 für Finanzdokumentenanalyse (64,4 Punkte, Bestwert) hervor. Für die Open-Weight-Freigabe hat Moonshot nach eigenen Angaben cyberbezogene Trainingsdaten entfernt, um die offensive Cyber-Fähigkeit bewusst unter dem zu halten, was dieselbe Trainingspipeline sonst liefern würde. Die vollständigen Modellgewichte sollen bis zum 4. November 2026 veröffentlicht werden; zum Prüfzeitpunkt (31.10.2026) war das Hugging-Face-Repository noch nicht öffentlich und die Gewichtelizenz damit noch nicht belegt.

Zugang & Nutzung

API über die Kimi API Plattform (api.moonshot.ai, OpenAI- und Anthropic-kompatible Endpunkte); als Flaggschiff wird es nach einer erfolgreichen Erstaufladung ab 1 US-Dollar freigeschaltet, das kumulative Aufladungsvolumen bestimmt Kontingente. Zero Data Retention ist für API-Kunden verfügbar. Zusätzlich in den Kimi-Produkten nutzbar: kimi.ai und Kimi-App, Kimi Work (Desktop ab Version 3.1.0) und Kimi Code; Mitgliedschaften von Free (eingeschränkt) bis Ultra, Highspeed ab Pro. Kimi-Enterprise-Admins müssen K3.1 Pro für ihre Organisation aktivieren (bei Start standardmäßig aus).

Open-Weight-Modell; die Gewichte werden laut Ankündigung bis 04.11.2026 veröffentlicht. Die konkrete Gewichtelizenz war zum Prüfzeitpunkt 31.10.2026 noch nicht veröffentlicht (Hugging-Face-Repository moonshotai/Kimi-K3.1-Pro noch nicht öffentlich) und muss bei Integration nachgeprüft werden. API- und Produktnutzung richten sich nach den Kimi-Bedingungen.

Grenzen & richtig lesen

Denkmodus immer aktiv (reasoning_effort low/high/max, Standard max). Feste Sampling-Parameter: temperature=1.0, top_p=0.95, n=1, presence_penalty=0, frequency_penalty=0. Bildeingabe unterstützt keine öffentlichen Bild-URLs (nur Base64 oder Datei-Upload via ms://). Web-Search-Tool wird laut Doku überarbeitet und ist derzeit nicht für Produktivworkflows empfohlen. Flaggschiff-Modelle erfordern eine Erstaufladung (mindestens 1 US-Dollar); Rate Limits hängen vom Kontingent-Tier ab.

Benchmark-Belege

Für dieses Portrait ist noch kein ausreichend zugeordnetes Ergebnis erfasst. Fehlende Werte sind keine Null. Wir übernehmen keine Werte aus ähnlich benannten Modell- oder Benchmark-Versionen.

Steckbrief

Anbieter

Moonshot AI (Kimi)

Veröffentlicht

24.10.2026

Parameterklasse

3-Billionen-Klasse (Herstellerangabe), nachtrainiert aus Kimi K3

Kontextfenster

1.048.576 Tokens (1 Mio.), flache Preise ohne Kontext-Staffelung

Maximale Ausgabe

bis 1.048.576 Tokens, Standard 131.072 (max_completion_tokens)

Modalitäten

Eingabe Text, Bild, Video; Ausgabe Text

Reasoning

Immer aktiv; reasoning_effort low/high/max, Standard max

API-Kennung

kimi-k3.1-pro; Highspeed-Variante kimi-k3.1-pro-highspeed

API-Preis Standard

2,00 USD je 1 Mio. Eingabe-Tokens, 10,00 USD je 1 Mio. Ausgabe-Tokens, Cache-Treffer 0,20 USD, Cache-Schreiben 2,00 USD (5 Min.) bzw. 4,00 USD (1 Std.)

API-Preis Highspeed

4,00 USD Eingabe, 20,00 USD Ausgabe, Cache-Treffer 0,20 USD je 1 Mio. Tokens (jeweils USD, vor Steuern)

Terminal-Bench 4.0

65,7 % (Herstellerangabe; Claude Opus 5.5: 66,4 %, GPT-6 Astra: 58,2 %)

BenchCAD (3D, mit Werkzeugen)

92,7 % (Herstellerangabe)

OfficeQA Pro v2 (Finanzen)

64,4 Punkte, Hersteller spricht von Bestwert (Herstellerangabe)

Faktenreue

Claim-Level-Halluzinationsrate interner Stresstest von 23,4 % (K3) auf 8,3 % gesenkt (Herstellerangabe, Hallu Bench Hard mit Werkzeugen)

Gewichte

Open-Weight angekündigt, Veröffentlichung bis 04.11.2026; Lizenz zum Prüfzeitpunkt noch offen

Angaben laut den unten verlinkten Primärquellen.

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: 智能前沿，更进一步 \(offizielle Release-Meldung, 24.10.2026\)](#)
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.

KI Fachblatt · Redaktion: Philipp Paulik · Quellenstand: 2026-10-31

Vollständiges Webportrait: <https://kifachblatt.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](#)