

Kimi K3

KI-Modellportrait · Quellenstand 2026-10-31



KIMI-Anbietericon aus dem offiziellen Brand-Guide, kein eigenes Modelllogo · [Logo-Quelle / Rechteinweis](#). 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](#) →

Technische Einordnung

Das Modell nutzt Kimi Delta Attention (KDA) und Attention Residuals sowie eine Mixture-of-Experts-Architektur mit 16 von 896 aktiven Experten. Die vollständigen Gewichte sind unter der hauseigenen Kimi K3 License veröffentlicht. Die Einordnung als Vorgänger-Flagship stammt vom Anbieter: Seit Kimi K3.1 Pro (Oktober 2026) ist K3 die vorherige Generation, bleibt aber im API-Katalog und bei Amazon Bedrock verfügbar.

Zugang & Nutzung

Kimi API (Modell-ID kimi-k3; OpenAI- und Anthropic-kompatible Endpunkte), Amazon Bedrock (seit 18.09.2026) sowie selbst gehostet über die offenen Gewichte (Hugging Face/GitHub; vLLM, SGLang, TokenSpeed). Der API-Zugang wird nach erster Aufladung (mindestens 1 USD) freigeschaltet; die kumulierte Aufladung bestimmt Tier und Ratenlimits.

Gewichte und Code stehen unter der proprietären Kimi K3 License: Nutzung, Kopie, Modifikation, Verbreitung und kommerzielle Nutzung sind erlaubt. Zwei Schwellenregeln: Wer Model-as-a-Service betreibt und mit Affiliates über 20 Mio. USD Umsatz in zwölf Monaten erzielt, braucht eine separate Vereinbarung mit Moonshot AI. Produkte mit mehr als 100 Mio. monatlich aktiven Nutzern oder mehr als 20 Mio. USD Monatsumsatz müssen "Kimi K3" prominent in der Oberfläche ausweisen.

Grenzen & richtig lesen

Kontextfenster (1.048.576 Token) und maximale Ausgabe (max_completion_tokens bis 1.048.576, Standard 131.072) sind getrennte Grenzen. Denken ist immer aktiv; reasoning_effort kennt low, high und max (Standard max). Bei mehrteiligen Dialogen muss die vollständige Assistant-Nachricht inklusive reasoning_content und tool_calls zurückgereicht werden. Videoeingabe läuft über den Datei-Upload der Kimi API; auf Amazon Bedrock ist Video als Eingabe nicht ausgewiesen. Alle Benchmark-Werte sind Herstellerangaben, keine eigenen Tests.

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

API-ID

kimi-k3

Eingabe → Ausgabe

Text, Bild, Video (API-Upload) → Text

Kontextfenster

1.048.576 Token

Maximale Ausgabe

bis 1.048.576 Token (Standard 131.072)

Architektur

MoE, 2,8 Bio. Parameter gesamt, 104 Mrd. aktiviert, 16 von 896 Experten

Veröffentlichung

16.07.2026 angekündigt, Gewichte seit 27.07.2026 offen

API-Preis

Eingabe 3 USD, Cache-Treffer 0,30 USD, Ausgabe 15 USD je Mio. Token

Gewichtslizenz

Kimi K3 License (eigene Open-Weight-Lizenz mit Schwellenregeln)

Quantisierung

MXFP4-Gewichte, MXFP8-Aktivierungen (quantization-aware trainiert)

Angaben laut den unten verlinkten Primärquellen.

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.

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

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