

MiniMax M3.1 Flash

KI-Modellportrait · Quellenstand 2026-10-20



MINIMAX

MiniMax-Logo (Unternehmenslogo; ein separates Modelllogo liegt nicht vor) · [Logo-Quelle / Rechteinweis](#). Marken gehören ihren jeweiligen Inhabern. Redaktionelle Identifikation, keine Partnerschaft oder Empfehlung durch den Markeninhaber.

Schnelle, kostenoptimierte Variante der M3-Linie für Agenten, Programmierung und Computer Use.

Quellen geprüft am 20. Oktober 2026 · Redaktion: Philipp Paulik

Entwickler: [MiniMax](#) →

Technische Einordnung

MiniMax meldet 73,7 Prozent auf DeepSWE v1.1, 61,7 auf OSWorld V2 und 82 Prozent auf CyberGym, dazu Spitzen über 200 Ausgabetokens pro Sekunde. Alle Werte sind Herstellerangaben ohne unabhängige Prüfung. Preis laut Hersteller: 0,15 USD pro Million Eingabetokens, 0,60 pro Million Ausgabetokens, 0,015 pro Million gecachter Eingabetokens; Reasoning-Tokens zählen als Ausgabe.

Zugang & Nutzung

MiniMax API, M Plan und MiniMax Code; außerdem über Vercel AI Gateway und AnyRouter erreichbar.

Keine Gewichtslicenz: Die Gewichte sind angekündigt, aber zum Abrufzeitpunkt (20.10.2026) nicht veröffentlicht; die Hugging-Face-Karte antwortete mit HTTP 401.

Grenzen & richtig lesen

Benchmarkwerte, Geschwindigkeit und die Angabe „70 Prozent weniger Kosten“ stammen vom Hersteller. Die Parameterzahl von 428 Mrd. stammt aus einer Zweitquelle, nicht aus einer öffentlichen Modellkarte.

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

Veröffentlicht

19. Oktober 2026 (API, M Plan, MiniMax Code)

Kontextfenster

1 Mio. Token

Eingaben

Text, Bild, Video; Reasoning immer aktiv, fünf Stufen

Parameter

428 Mrd. gesamt laut Entwicklerhandbuch der Vorschau (Zweitquelle)

Preis

0,15 / 0,60 / 0,015 USD pro 1 Mio. Tokens (Eingabe / Ausgabe / Cache)

Gewichte

Angekündigt, noch nicht veröffentlicht (Stand 20.10.2026)

Space Bunny

Frühe Fassung lief anonym auf OpenRouter; laut MiniMax über 200 Bio. Tokens verarbeitet.

Herstellerwerte

DeepSWE v1.1: 73,7 %, OSWorld V2: 61,7, CyberGym: 82 % (Selbstauskunft)

Angaben laut den unten verlinkten Primärquellen.

Quellen & Stand

1. [MiniMax M3.1 Flash: Redefining Agent Performance and Cost](#)

Geprüft: 2026-10-20

2. [MiniMax launches M3.1 Flash model, says it cuts AI agent costs by 70% \(SiliconANGLE\)](#)

Geprüft: 2026-10-20

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: 2026-10-20

Vollständiges Webportrait: <https://kifachblatt.de/ki-modelle/minimax-m3-1-flash>

Statischer Export, kein Live-Datenblatt. Preise, Zugriff und Anbieterangaben können sich ändern. Quellen und ihre jeweiligen Abrufstände stehen oben. [Impressum](#) · [Transparenz](#)