

DeepSeek-V4-Pro

KI-Modellportrait · Quellenstand 2026-10-20



DeepSeek-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.

Flaggschiff der V4-Serie: 1,6 Bio. Parameter (49 Mrd. aktiv), eine Million Token Kontext, Max-Modus für schwere Reasoning-Aufgaben.

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

Entwickler: [DeepSeek](#) →

Technische Einordnung

Hybrid-Attention aus Compressed Sparse Attention (CSA) und Heavily Compressed Attention (HCA): Bei 1 Mio. Token Kontext braucht das Modell laut Hersteller nur 27 Prozent der Rechenoperationen und 10 Prozent des KV-Cache von V3.2. Trainiert mit dem Muon-Optimizer auf über 32 Bio. Token; Spezialexperten werden per On-Policy-Destillation zu einem Modell konsolidiert.

Zugang & Nutzung

DeepSeek-API (Modellname deepseek-v4-pro) sowie Gewichte auf Hugging Face (Pro und Pro-Base, FP4/FP8 gemischt).

MIT (Gewichte und Karte auf Hugging Face).

Grenzen & richtig lesen

Herstellerangaben; der beschriebene Vorsprung beruht auf firmeninternen Messungen. Der 0813-Checkpoint (13. August 2026) ist ein Zwischenstand, kein neues Modell.

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

22. Juni 2026 (V4-Serie); Checkpoint V4-Pro-0813 am 13. August 2026.

Parameter

1,6 Bio. gesamt, 49 Mrd. aktiv pro Token.

Kontextfenster

1 Mio. Token.

Effizienz

27 Prozent der Rechenoperationen und 10 Prozent des KV-Cache von V3.2 bei 1 Mio. Token (Herstellerangabe).

Training

Über 32 Bio. Token Vortraining; zweistufige Nachtrainierung mit domänenspezifischen Experten und On-Policy-Destillation.

Modi

V4-Pro-Max als maximale Reasoning-Stufe; V4-Flash-Max erreicht laut Hersteller ähnliche Reasoning-Werte bei höherem Denkbudget.

Herstellerwerte

In der Vergleichstabelle der V4.1-Karte: HLE 42,7, Terminal-Bench 2.1: 87,9, CyberGym: 83,3, Codeforces: 3348 (Selbstauskunft).

Angaben laut den unten verlinkten Primärquellen.

Quellen & Stand

1. [Hugging Face: DeepSeek-V4-Pro \(Modellkarte der V4-Serie\)](#)
Geprüft: 2026-10-20
2. [Hugging Face: deepseek-ai \(Modellliste mit Daten\)](#)
Geprüft: 2026-10-20
3. [Hugging Face: DeepSeek-V4.1-Flash \(Modellkarte mit Architektur und Messwerten\)](#)
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/deepseek-v4-pro>

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