

DeepSeek-V4.1-Flash

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.

Multimodales MoE-Modell mit 552 Mrd. Parametern, einer Million Token Kontext und stark komprimiertem KV-Cache.

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

Entwickler: [DeepSeek](#) →

Technische Einordnung

Neue Causal-Encoder-Decoder-Architektur: nur 8 Mrd. aktive Parameter beim Einlesen, 16 Mrd. beim Schreiben. Der globale KV-Cache liegt bei 890 Byte pro Token, etwa ein Viertel von V4-Flash. Verarbeitet Bilder und Text, gibt Text aus; die Reasoning-Stärke ist stufenlos von 1 bis 100 wählbar.

Zugang & Nutzung

DeepSeek-API (Modellname deepseek-flash, bedient von der Version 1017), Chat-App und Web; Gewichte auf Hugging Face unter MIT.

MIT (Gewichte und Karte auf Hugging Face).

Grenzen & richtig lesen

Alle Messwerte stammen vom Hersteller, mit maximaler Reasoning-Stufe und teils eigenem Harness. Die Basisversion und die API-Version 1017 sind nicht identisch mit der Erstveröffentlichung vom September.

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

10. September 2026; Update V4.1-Flash-1017 am 17. Oktober 2026.

Architektur

Causal Encoder-Decoder (CED), 40 Schichten; MoE mit 1 gemeinsamem und 384 gerouteten Experten, 6 aktiv pro Token.

Parameter

552 Mrd. gesamt; 8 Mrd. aktiv beim Einlesen (Prefill), 16 Mrd. beim Schreiben (Decode).

Kontextfenster

1 Mio. Token.

KV-Cache

890 Byte pro Token global, laut Hersteller etwa ein Viertel von V4-Flash und 1/437 von V1.

Modalität

Eingabe Bild und Text (eigener Vision-Encoder DeepSeek-ViT), Ausgabe Text.

Training

45 Bio. Token multimodaler Korpus; Kontext ab 34 Bio. Token auf 1 Mio. erweitert.

Reasoning

Stufenlos einstellbarer Reasoning-Aufwand (1 bis 100); Messwerte der Karte bei Stufe 100.

Herstellerwerte

Terminal-Bench 2.1: 90,6; CyberGym: 88,1; Codeforces: 3471; HLE mit Werkzeugen: 63,9 (Selbstauskunft).

API-Umstellung

Der API-Name deepseek-flash wird seit dem Update von V4.1-Flash-1017 bedient; die Vorgänger deepseek-v4-flash und -vision-exp gelten als abgelöst.

Angaben laut den unten verlinkten Primärquellen.

Quellen & Stand

1. [Hugging Face: DeepSeek-V4.1-Flash \(Modellkarte mit Architektur und Messwerten\)](#)

Geprüft: 2026-10-20

2. [DeepSeek: offizielle Website \(Hinweis zum Update V4.1-Flash-1017\)](#)

Geprüft: 2026-10-20

3. [DeepSeek API Docs: Modellnamen und Umstellung auf V4.1-Flash-1017](#)

Geprüft: 2026-10-20

4. [Hugging Face: deepseek-ai \(Modellliste mit Daten\)](#)

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-v4-1-flash>

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