

Nemotron 3.5 Super VL NVFP4

KI-Modellportrait · Quellenstand 2026-10-19

N3

Monogramm: kein eindeutig zugeordnetes Modelllogo in der geprüften Auswahl.

Ein Modell für Text, Bilder und Video mit quantisierten Gewichten.

Quellen geprüft am 19. Oktober 2026 · Redaktion: Philipp Paulik

Entwickler: [NVIDIA](#) →

Technische Einordnung

Die Modellkarte beschreibt Mamba-2-, MoE- und Attention-Schichten plus C-RADIO-Vision-Encoder. Dieses Portrait bezieht sich auf die NVFP4-Variante, nicht pauschal auf BF16 oder die gesamte Nemotron-Familie.

Zugang & Nutzung

NVIDIA-gehostete API und verlinkte Modellgewichte.

OpenMDW-1.1 für das Modell; der API-Testdienst hat eigene NVIDIA-Trialbedingungen.

Grenzen & richtig lesen

Benchmark-Werte sind Herstellerangaben. Präzision, Agentensystem und Testaufbau verändern die Ergebnisse; es wurden hier keine eigenen Tests durchgeführt.

Benchmark-Belege

Herstellerbericht, keine eigene Messung. Die Zahlen sind keine direkt vergleichbare Modellrangliste. Messdatum und vollständige Testkonfiguration sind in den referenzierten Tabellenzeilen nicht ausgewiesen.

TEST / MODELL	WERT	KONTEXT
GPQA Diamond	89,90 % laut Herstellertabelle	NVFP4; Prompt, Agent, Budget und Wiederholungen in dieser Tabellenzeile nicht vollständig spezifiziert. Nicht für direkte Modellranglisten geeignet. Quelle: NVIDIA · Abruf 19.10.2026
SWE-bench Verified Verified	71,07 % laut Herstellertabelle	NVFP4; Prompt, Agent, Budget und Wiederholungen in dieser Tabellenzeile nicht vollständig spezifiziert. Nicht für direkte Modellranglisten geeignet. Quelle: NVIDIA · Abruf 19.10.2026
Terminal-Bench 2.1 2.1	60,80 % laut Herstellertabelle	NVFP4; Prompt, Agent, Budget und Wiederholungen in dieser Tabellenzeile nicht vollständig spezifiziert. Nicht für direkte Modellranglisten geeignet. Quelle: NVIDIA · Abruf 19.10.2026

Steckbrief

Version

NVIDIA-Nemotron-3.5-Super-VL-120B-A12B-NVFP4

Veröffentlicht

15. Oktober 2026

Parameter gesamt / aktiv

120 Mrd. / 12 Mrd.

Kontextfenster

Bis 1 Mio. Token

Eingabe → Ausgabe

Text, Bild, Video → Text

Gewichtspräzision

NVFP4, gemischte Präzision

Angaben laut den unten verlinkten Primärquellen.

Quellen & Stand

1. [NVIDIA: Nemotron 3.5 Super VL Modellkarte](#)

Geprüft: 2026-10-19

Redaktionelle Zusammenfassung mit KI-Unterstützung, verantwortet von Philipp Paulik. Kein eigener Leistungstest. Änderbare Anbieterangaben gelten für den angegebenen Quellenstand.

Entwickler-Kontext des Anbieters

Ein öffentliches Entwicklerprojekt des Anbieters: [NVIDIA-NeMo/Speech](#). Das ist kein Nachweis, dass dieses Repository die Gewichte oder den Quellcode des hier porträtierten Modells enthält.

18604 GitHub-Stars · 3651 Forks · Repository-Lizenz laut API: Apache-2.0.

Snapshot: 2026-10-19T13:28:52+00:00 · letzter Push: 2026-10-19T12:28:18Z. Beachtung und Repository-Aktivität, keine Modellqualität oder Unternehmensgröße. [API-Quelle](#). Lokal gespeichert; kein Live-Abruf beim Lesen. SDK-/Code-Lizenzen gelten nicht automatisch für Modelle oder APIs.