

Nemotron 3 Ultra 550B-A55B

KI-Modellportrait · Quellenstand 2026-10-31



NVIDIA-Anbieterlogo, kein modellspezifisches Logo für Nemotron 3 Ultra · [Logo-Quelle / Rechteinweis](#). Marken gehören ihren jeweiligen Inhabern. Redaktionelle Identifikation, keine Partnerschaft oder Empfehlung durch den Markeninhaber.

NVIDIAs größtes offenes Sprachmodell der Nemotron-3-Familie: 550 Mrd. Parameter, 55 Mrd. aktiv, gebaut für lang laufende KI-Agenten und komplexes Reasoning.

Quellen geprüft am 31. Oktober 2026 · Redaktion: Philipp Paulik

Entwickler: [NVIDIA](#) →

Technische Einordnung

Nemotron 3 Ultra ist laut NVIDIA das letzte und leistungsstärkste Modell der Nemotron-3-Familie. Die Architektur ist ein hybrider Latent-Mixture-of-Experts-Aufbau aus Mamba-2-, MoE- und Attention-Schichten mit Multi-Token-Prediction; vortrainiert wurde mit einem NVFP4-Rezept. NVIDIA veröffentlicht vier getrennte Checkpoints: das nachtrainierte BF16-Modell, eine NVFP4-quantisierte Variante, das Basismodell und ein GenRM-Modell für RLHF. Dieses Portrait behandelt die nachtrainierten Varianten BF16 und NVFP4; Werte des Basismodells oder von GenRM werden nicht übernommen.

Zugang & Nutzung

Offene Modellgewichte über Hugging Face (BF16, NVFP4, Base und GenRM); laut NVIDIA-Pressemitteilung außerdem über ModelScope, OpenRouter und als NIM-Microservice. Eine gehostete Test-API gibt es auf build.nvidia.com unter den NVIDIA API Trial Terms of Service. Selbstbetrieb über vLLM, SGLang oder TensorRT-LLM; laut Modellkarte mindestens 8x B200 (BF16) beziehungsweise 4x B200 (NVFP4).

OpenMDW License Agreement 1.1 (Steward: Linux Foundation). Die Lizenz erlaubt Nutzung, Änderung und Weitergabe ohne Einsatzbeschränkung, auch kommerziell; bei Weitergabe sind Lizenztext und Herkunftshinweise zu behalten. Erzeugte Ausgaben unterliegen keinen Lizenzpflichten. Eine Klausel beendet die Rechte bei aktiven Patent- oder Urheberrechtsklagen gegen die Modellmaterialien. Die Linux Foundation hat OpenMDW-1.1 im August 2026 der OSI zur Prüfung vorgelegt; eine OSI-Anerkennung ist daraus nicht abzuleiten. Der gehostete Testdienst auf build.nvidia.com hat eigene Nutzungsbedingungen.

Grenzen & richtig lesen

Alle Benchmark- und Durchsatzwerte sind Herstellerangaben aus NVIDIAs eigener Auswertung (Nemo Evaluator SDK), keine unabhängigen Messungen. Der Betrieb erfordert Rechenzentrums-Hardware, keine Consumer-GPUs. Die 1-Mio.-Token-Konfiguration muss am Server ausdrücklich freigeschaltet werden; die dokumentierten Beispiele nutzen standardmäßig 262.144 Token. TensorRT-LLM unterstützt das Modell laut Modellkarte derzeit nur auf Blackwell-Hardware. NVIDIA weist zudem auf unausgewogene Demografie-Anteile in Trainingsdaten hin und empfiehlt eigene Bias-Prüfungen. Ein offizielles Ausgabelimit und ein offizieller Tokenpreis werden von NVIDIA nicht genannt.

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

Version

NVIDIA-Nemotron-3-Ultra-550B-A55B; Checkpoints BF16, NVFP4, Base-BF16 und GenRM getrennt veröffentlicht

Veröffentlicht

4. Juni 2026 auf Hugging Face; angekündigt am 1. Juni 2026

Parameter gesamt / aktiv

550 Mrd. gesamt, 55 Mrd. aktiv (MoE)

Architektur

LatentMoE: Hybrid aus Mamba-2-, MoE- und Attention-Schichten mit Multi-Token-Prediction; NVFP4-Vortrainingsrezept

Kontextfenster

Bis 1.000.000 Token; Serverbeispiele starten mit 262.144, 1 Mio. erfordert ausdrückliche Freischaltung

Eingabe -> Ausgabe

Text -> Text; Reasoning-Spur optional ein- und ausschaltbar

Sprachen

Englisch, Französisch, Spanisch, Italienisch, Deutsch, Japanisch, Hindi, Koreanisch, Brasilianisches Portugiesisch, Chinesisch

Vortraining und Datenstände

Ca. 20 Bio. Token Vortraining; Pre-Training-Daten bis September 2025, Post-Training-Daten bis Mai 2026

Mindesthardware (Herstellerangabe)

BF16: 8x GB200/B200/GB300/B300, 16x H100 oder 8x H200. NVFP4: 4x GB200/B200/GB300/B300 oder 8x H100

Lizenz

OpenMDW-1.1 (Linux Foundation), kommerzielle und nicht-kommerzielle Nutzung laut Modellkarte zulässig

Hersteller-Durchsatzangabe

Bis zu 5,9-facher Inferenzdurchsatz gegenüber GLM-5.1-754B-A40B, 4,8-fach gegenüber Kimi-K2.6-1T-A32B und 1,6-fach gegenüber Qwen-3.5-397B-17B (8k Eingabe / 64k Ausgabe, Herstellerangabe)

Angaben laut den unten verlinkten Primärquellen.

Quellen & Stand

1. [NVIDIA: NVIDIA-Nemotron-3-Ultra-550B-A55B-BF16 Modellkarte \(Hugging Face\)](#)
Geprüft: 2026-10-31
2. [NVIDIA: NVIDIA-Nemotron-3-Ultra-550B-A55B-NVFP4 Modellkarte \(Hugging Face\)](#)
Geprüft: 2026-10-31
3. [NVIDIA Research: NVIDIA Nemotron 3 Ultra \(veröffentlicht 04.06.2026\)](#)
Geprüft: 2026-10-31
4. [NVIDIA: nemotron-3-ultra-550b-a55b Modellkarte auf build.nvidia.com \(NIM\)](#)
Geprüft: 2026-10-31
5. [NVIDIA Newsroom: Enterprise Software Leaders Build AI Agents With NVIDIA \(01.06.2026\)](#)
Geprüft: 2026-10-31
6. [OpenMDW License Agreement, Version 1.1 \(Lizenztext, Linux Foundation\)](#)
Geprüft: 2026-10-31
7. [OSI License Review: Einreichung von OpenMDW-1.1 durch die Linux Foundation \(August 2026\)](#)
Geprüft: 2026-10-31

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.

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

Vollständiges Webportrait: <https://kifachblatt.de/ki-modelle/nemotron-3-ultra-550b-a55b>

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