Intelligenz Welt

# Nemotron 3 Ultra 550B-A55B

KI-Modellportrait · Quellenstand 2026-10-31

NVIDIA-Anbieterlogo, kein modellspezifisches Logo für Nemotron 3 Ultra · Logo-Quelle / Rechtehinweis. 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 →

DAS MODELL EINFACH ERKLÄRT

## NVIDIAs größtes offenes Modell ist für Agenten gebaut, nicht für den heimischen PC.

Nemotron 3 Ultra 550B-A55B ist ein offenes Sprachmodell von NVIDIA mit 550 Mrd. Parametern, von denen pro Schritt 55 Mrd. aktiv sind. Es verarbeitet Text und erzeugt Text, das Kontextfenster reicht laut Modellkarte bis zu 1 Mio. Token. Die Gewichte stehen unter der OpenMDW-1.1-Lizenz frei zum Download, der Betrieb erfordert allerdings mindestens vier bis acht Rechenzentrums-GPUs.

Offizielle Dokumentation: Nemotron 3 Ultra 550B-A55B

Quellen geprüft: 31.10.2026. Herstellerangaben, kein eigener Modelltest.

## Wofür eignet sich Nemotron 3 Ultra 550B-A55B?

Das Modell richtet sich an Teams, die lang laufende KI-Agenten, aufwendige Recherche oder Analysen über sehr lange Dokumente betreiben und dafür eigene Infrastruktur oder einen Hosting-Partner nutzen. Für Einzelpersonen und kleine Projekte ist es wegen der Hardwareanforderungen praktisch nur über gehostete Dienste erreichbar. Wer nur eine Chat- oder Textlösung sucht, fährt mit kleineren Modellen meist günstiger und einfacher.

### Offene Gewichte, aber Rechenzentrum vorausgesetzt

NVIDIA veröffentlicht die Gewichte frei, nennt aber als Minimum 8x B200 für BF16 beziehungsweise 4x B200 für die NVFP4-Variante. Das ist Rechenzentrums-Hardware; auf einem Arbeitsplatzrechner läuft das Modell nicht. Alternativen sind gehostete Angebote von NVIDIA-Partnern oder die Test-API auf build.nvidia.com.

Offizielle Dokumentation: Nemotron 3 Ultra 550B-A55B

### Reasoning lässt sich steuern

Über das Chat-Template schaltet enable_thinking die Reasoning-Spur ein oder aus, medium_effort senkt den Denkaufwand, und ein reasoning_budget begrenzt die Länge der Denkspur. Für Kosten und Antwortzeit ist das die wichtigste Stellschraube; die Standardeinstellung erzeugt zuerst eine Denkspur und dann die Antwort.

Offizielle Dokumentation: Nemotron 3 Ultra 550B-A55B

### Ein Release, vier Checkpoints

Neben dem nachtrainierten BF16-Modell gibt es eine NVFP4-quantisierte Variante für höheren Durchsatz, das reine Basismodell und ein GenRM-Modell, das NVIDIA beim RLHF-Training einsetzt. Die Varianten haben eigene Modellkarten und Hardwareanforderungen; Werte sind nicht zwischen ihnen übertragbar.

Offizielle Dokumentation: Nemotron 3 Ultra 550B-A55B

### Lizenz: OpenMDW-1.1

Die Lizenz der Linux Foundation erlaubt Nutzung, Veränderung und Weitergabe ohne Feldbeschränkung und stellt erzeugte Ausgaben ausdrücklich frei. Einzige Bedingungen: Lizenztext und Herkunftshinweise bleiben bei Weitergabe erhalten, und wer wegen Patent- oder Urheberrechten klagt, verliert die eigenen Rechte. Die OSI-Prüfung der Lizenzversion 1.1 wurde im August 2026 eingereicht; ein Prüfergebnis ist daraus nicht abzuleiten.

Offizielle Dokumentation: Nemotron 3 Ultra 550B-A55B

### Hersteller-Benchmarks richtig lesen

Alle veröffentlichten Werte stammen aus NVIDIAs eigener Auswertung mit dem Nemo Evaluator SDK. NVIDIA meldet unter anderem SWE-Bench Verified 70,7, RULER bei 1 Mio. Kontext 94,7 und MMLU-Pro 86,8 für den BF16-Checkpoint. Die Vergleichsmodelle und Einstellungen wählt der Hersteller selbst; die Werte sind Anhaltspunkte, keine unabhängige Messung.

Offizielle Dokumentation: Nemotron 3 Ultra 550B-A55B

## Kontext und Ausgabe.

**1.000.000**

Token Kontextfenster

## Was kostet Nemotron 3 Ultra 550B-A55B über die API?

NVIDIA nennt keinen offiziellen Preis pro Token. Die Gewichte sind kostenlos unter OpenMDW-1.1; Kosten entstehen über eigene GPU-Infrastruktur (Hersteller-Minimum: 4x B200 für NVFP4) oder über Hosting- und Cloud-Partner mit eigenen Preisen. Die Test-API auf build.nvidia.com läuft unter den NVIDIA API Trial Terms of Service. Tokenpreise von Drittanbietern wie OpenRouter sind keine NVIDIA-Angaben und werden hier nicht übernommen.

## Vor dem Einsatz prüfen

- Checkpoint genau festhalten: BF16, NVFP4, Base und GenRM sind getrennte Artefakte mit eigenen Karten und Hardwareanforderungen.
- Reasoning-Modus explizit konfigurieren (enable_thinking, medium_effort, reasoning_budget) und Antwortzeit sowie Kosten mit eigenen Aufgaben messen.
- Kontextkonfiguration prüfen: Standard 262.144 Token; 1.048.576 Token nur mit ausdrücklicher Server-Freischaltung und passendem Speicher.
- Serving-Stack vorab festlegen: vLLM und SGLang laut Karte getestet; TensorRT-LLM derzeit nur auf Blackwell-Hardware.

## Häufige Fragen zu Nemotron 3 Ultra 550B-A55B

### Was ist Nemotron 3 Ultra 550B-A55B?

Das größte offene Sprachmodell von NVIDIAs Nemotron-3-Familie: 550 Mrd. Parameter gesamt, 55 Mrd. aktiv, veröffentlicht am 4. Juni 2026. Es ist für lang laufende KI-Agenten, Werkzeugnutzung und lange Dokumente gedacht.

### Ist Nemotron 3 Ultra wirklich offen?

NVIDIA veröffentlicht Gewichte, Trainingsdaten und Rezepte unter OpenMDW-1.1. Die Lizenz erlaubt kommerzielle Nutzung und Weitergabe, verlangt bei Weitergabe aber Lizenztext und Herkunftshinweise. Der Begriff Open Source trifft auf die Lizenz zu; die formale OSI-Anerkennung der Version 1.1 war zum Prüfstand noch als eingereicht dokumentiert.

### Kann ich Nemotron 3 Ultra auf eigener Hardware betreiben?

Nur mit Rechenzentrums-Hardware: laut Modellkarte mindestens 8x B200 (BF16) oder 4x B200 (NVFP4), alternativ 16x H100 oder 8x H200 für BF16. Consumer-GPUs reichen nicht. Ohne eigene Infrastruktur bleiben Hosting-Partner oder die Test-API auf build.nvidia.com.

### Wie groß ist das Kontextfenster von Nemotron 3 Ultra?

Bis zu 1.000.000 Token laut Modellkarte. Die dokumentierten Server-Beispiele starten mit 262.144 Token; die volle Länge erfordert eine ausdrückliche Konfiguration am Server. Ein offizielles Limit für Ausgabe-Token nennt NVIDIA nicht.

### Was kostet Nemotron 3 Ultra?

Die Gewichte sind kostenlos. NVIDIA veröffentlicht keinen Tokenpreis; wer das Modell nutzt, zahlt entweder eigene GPU-Infrastruktur oder die Preise eines Hosting-Anbieters. Die NVIDIA-Test-API unterliegt eigenen Testbedingungen.

Alle Antworten beziehen sich auf die verlinkte Herstellerdokumentation und den angegebenen Stand. Kein unabhängiger Benchmark, keine Garantie für Verfügbarkeit oder Ergebnisqualität.

## 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.


Intelligenz Welt · Redaktion: Philipp Paulik · Quellenstand: 2026-10-31
Vollständiges Webportrait: https://intelligenz-welt.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