Intelligenz Welt

Muse Spark Open

KI-Modellportrait · Quellenstand 2026-10-31

Meta

Meta (Anbieterlogo, kein modellspezifisches Logo) · Logo-Quelle / Rechtehinweis. Marken gehören ihren jeweiligen Inhabern. Redaktionelle Identifikation, keine Partnerschaft oder Empfehlung durch den Markeninhaber.

Die Open-Weight-Version von Metas Agentic- und Coding-Modell Muse Spark: 816 Milliarden Parameter gesamt, 42 Milliarden aktiv, 1-Million-Token-Kontext und multimodale Eingabe (Text, Bild, Video, Audio). Die Gewichte stehen unter Metas eigener Muse Model License mit Bedingungen; der Betrieb läuft self-hosted oder über Hosted-Inference-Provider, nicht über die Meta Model API.

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

Entwickler: Meta →

DAS MODELL EINFACH ERKLÄRT

Metas Open-Weight-Modell für agentisches Arbeiten und Coding, von der Workstation bis zum 8-GPU-Knoten

Muse Spark Open ist die offene Gewichtsversion von Metas Agentic- und Coding-Modell Muse Spark, veröffentlicht am 15. Oktober 2026. Das sparse Mixture-of-Experts-Modell hat 816 Milliarden Parameter, aktiviert davon 42 Milliarden pro Token, verarbeitet Text, Bild, Video und Audio und bietet ein Kontextfenster von einer Million Tokens. Es läuft nicht über die Meta Model API, sondern self-hosted oder über Hosted-Inference-Provider. Die eigene Muse Model License erlaubt weitgehende Nutzung inklusive Fine-Tuning, Destillation und Weitervertrieb, verlangt aber sichtbare Attribution und bei großen Model-as-a-Service-Angeboten ab 20 Mio. USD Jahresumsatz eine separate Vereinbarung mit Meta; sie ist keine OSI-gelistete Open-Source-Lizenz.

Offizielle Dokumentation: Muse Spark Open

Quellen geprüft: 31.10.2026. Herstellerangaben, kein eigener Modelltest.

Wofür eignet sich Muse Spark Open?

Praktisch relevant ist Muse Spark Open für Teams, die ein leistungsstarkes Agentic- und Coding-Modell unter eigener Kontrolle betreiben wollen: Prompts und Antworten bleiben in der eigenen Infrastruktur, und die Lizenz erlaubt Fine-Tuning und Destillation ausdrücklich. Für den Einstieg ohne eigene GPUs sind die Hosted-Provider der schnellere Weg; OpenRouter listet aktuell einheitlich 1,25/4,25 USD je Million Tokens über sechs Provider. Wer dauerhaft hohe Volumina fährt oder Datenhoheit braucht, rechnet auf Self-Hosting um: die NVFP4-Variante passt laut Meta-Doku auf einen 4-GPU-Blackwell-Knoten, für den Desktop bewirbt Meta zwei Mac Studios mit M5 Ultra über MLX. Für deutsche und europäische Nutzer relevant: die Lizenz enthält keine Regionsbeschränkung, Vertragspartner in der EU-Region ist Meta Platforms Ireland Limited; die Usage Policy gilt aber weltweit, und das Modell ist nicht für Minderjährige bestimmt.

Positionierung: Open Weight statt API-Modell

Muse Spark Open ist die offene Variante von Muse Spark und im Meta-Modellkatalog ausdrücklich nicht über die Meta Model API erreichbar. Meta positioniert es als das leistungsfähigste US-Open-Weight-Modell (Herstellerangabe) für agentische Workflows, komplexes Software Engineering und mehrstufiges Reasoning. Der Release richtet sich an zwei Gruppen: Betreiber mit eigener GPU-Infrastruktur, die volle Datenkontrolle wollen, und Entwickler, die über Hosted-Provider ohne Hardware-Aufwand starten. Anders als reine API-Modelle erlaubt die Gewichtsfreigabe Fine-Tuning über PyTorch TorchTitan und sogar Destillation, was die Lizenz ausdrücklich einschließt.

Offizielle Dokumentation: Muse Spark Open

Lizenz: großzügige Rechte, klare Auflagen

Die Muse Model License (Version Muse Spark Open, 15.10.2026) gewährt unentgeltlich nahezu alle Nutzungsrechte: verwenden, kopieren, modifizieren, fine-tunen, Derivate erstellen, distillieren, mergen, veröffentlichen, distribuieren, unterlizenzieren und verkaufen. Drei Auflagen sind praxisrelevant. Erstens muss jedes Produkt oder jeder Dienst, der das Modell oder Derivate nutzt, sichtbar 'Built with Muse Spark Open' anzeigen. Zweitens gilt eine Umsatzschwelle nur für Model-as-a-Service-Angebote: ab über 20 Mio. USD Bruttoumsatz (Lizenznehmer plus Affiliates, rollierend über 12 Monate) ist eine Benachrichtigung an Meta und eine separate schriftliche Vereinbarung nötig; interne Nutzung, Forschung und eigene Produkte ohne MaaS-Charakter sind ausgenommen. Drittens ist die Usage Policy in Exhibit A verbindlich, mit Verboten von Malware bis militärischer Nutzung und dem Hinweis, dass das Modell nicht für unter 18-Jährige bestimmt ist. Meta erhebt keinen Eigentumsanspruch an generierten Outputs. Eingestuft werden sollte das als Open-Weight-Freigabe mit Auflagen, nicht als OSI-gelistete Open-Source-Lizenz; Regionsausschlüsse enthält der Lizenztext nicht.

Offizielle Dokumentation: Muse Spark Open

Hardware und Betrieb

Zwei Checkpoints decken unterschiedliche Budgets ab. BF16 ist die Vollpräzisions-Referenz mit rund 1,65 TB Gewichten; Meta empfiehlt etwa 2 TB VRAM inklusive KV-Cache und Laufzeit-Overhead, als Beispielkonfiguration nennt die Doku acht B300-GPUs. NVFP4 speichert die meisten Gewichte in 4-Bit und komprimiert auf rund 537 GB; die Meta-Doku nennt einen einzelnen 4-GPU-NVIDIA-Blackwell-Knoten, die Model Card mindestens 700 GB VRAM. Ungewöhnlich für diese Modellklasse: Meta bewirbt den Desktop-Betrieb auf zwei Apple Mac Studios mit M5 Ultra über das MLX-Framework und beschreibt agentische Workflows komplett on-device, etwa iOS-Codeänderungen über Muse Code oder das Frame-by-Frame-Review von Videomaterial in DaVinci Resolve. Unterstützte Runtimes sind SGLang, vLLM, TensorRT-LLM, Hugging Face Transformers, llama.cpp, Unsloth und MLX, mit Optimierungen für AMD-, Apple-, Intel- und NVIDIA-Hardware.

Offizielle Dokumentation: Muse Spark Open

Benchmarks (Herstellerangaben)

Alle folgenden Werte stammen aus Metas Model Card beziehungsweise dem Evaluation-Methodology-Bericht und sind Herstellerangaben, keine unabhängigen Tests. Muse Spark Open erreicht 39 Punkte im Artificial Analysis Intelligence Index, 89,2 auf MCP-Atlas (Agentic Tool Use), 80,4 auf OSWorld-Verified (Computer Use), 72,8 auf Toolathlon-Verified, 76,4 auf Terminal-Bench 2.1, 60 auf DeepSWE 1.1 (mit Muse-Code-Harness), 89,0 auf GPQA Diamond und 61,0 auf Humanity's Last Exam mit Tools. Multimodal: 50,0 auf ZeroBench mit Tools, 68,3 auf IntPhys 2 und 61,5 auf Video-MME v2. Meta dokumentiert im Methodology-Bericht ungewöhnlich transparent, welche Werte aus eigenen Runs, offiziellen Leaderboards oder Selbstauskünften der Wettbewerber stammen, und weist selbst darauf hin, dass das Setup nicht auf Drittmodelle optimiert war. In mehreren Vergleichen liegen GLM 5.3 und Kimi K3 vorne; Meta führt diese Werte in derselben Tabelle.

Offizielle Dokumentation: Muse Spark Open

Sicherheit und Preparedness

Für den Open-Weight-Einsatz ohne vorgeschaltete System-Guardrails hat Muse Spark Open laut Meta ein zusätzliches Safety-Post-Training erhalten. Die Model Card meldet eine Angriffserfolgsrate von 0,5 auf dem internen StrongREJECT-v2-Benchmark (zum Vergleich: 62,9 bis 86,4 bei den gelisteten Offenmodellen, Herstellerangaben), eine Verweigerungsrate von 99,2 Prozent bei Jailbreaks auf dem internen Cyber-Benchmark sowie 98,1 Prozent Ehrlichkeit auf MASK und 0 Prozent Fehlverhalten auf dem Agentic-Misalignment-Benchmark. Geprüft wurde nach dem Meta Superintelligence Scaling Framework über Content Safety, Agentic Risk, Privacy und Preparedness (Chemie/Bio, Cyber, Loss of Control). Meta betont, dass die Bewertung nur den veröffentlichten Checkpoint unter den getesteten Bedingungen betrifft und nachgeschaltete Modifikationen die Sicherheit verschlechtern können, und empfiehlt eigene Guardrails samt Human-in-the-Loop für irreversible Aktionen.

Offizielle Dokumentation: Muse Spark Open

Grenzen und offene Punkte

Das maximale Output-Limit ist offiziell nicht belegt und wird hier bewusst nicht spekuliert. Nicht alle der über 100 Trainingssprachen sind evaluiert; Meta warnt vor Qualitätseinbußen außerhalb des stark unterstützten Sprachkerns. Bei sehr kurzen oder niedrig aufgelösten Videos (wenige Sekunden, rund 100 px) meldet das Modell mitunter, es könne das Video nicht sehen; Meta empfiehlt mindestens 448 px auf der kurzen Seite. Quantisierte Inferenz kann in Randfällen von der Vollpräzision abweichen. Benchmark-Werte sind Herstellerangaben mit dokumentierter, aber eben herstellereigener Methodik. Und wer das Modell in Produkten einsetzt, muss die Usage Policy durchsetzen, die Verantwortung für sichere Deployments liegt ausdrücklich beim Lizenznehmer.

Offizielle Dokumentation: Muse Spark Open

Ein- und Ausgaben im Überblick.

Eingabe

Self-Hosting ohne Tokenpreis; Hosted-Listenpreis 1,25 USD je 1M Input-Tokens über OpenRouter-Provider (Stand 31.10.2026), Cache-Read 0,15 USD

Ausgabe

Hosted-Listenpreis 4,25 USD je 1M Output-Tokens über OpenRouter-Provider (Stand 31.10.2026)

Kontext / Grenzen

1.048.576 Tokens (1M) bei Inferenz

Maximale Ausgabe

Nicht offiziell belegt

Was kostet Muse Spark Open über die API?

Muse Spark Open ist nicht über die Meta Model API buchbar, es gibt also keinen Meta-Tokenpreis. Self-Hosting verursacht keine Tokenkosten, dafür Hardwarekosten (NVFP4: 4-GPU-Blackwell-Knoten laut Meta-Doku; BF16: rund 2 TB VRAM). Hosted-Listenpreise über OpenRouter-Provider lagen am 31.10.2026 einheitlich bei 1,25 USD Input, 4,25

USD Output und 0,15 USD Cache-Read je 1M Tokens; einzelne Provider können abweichen, tatsächlich gezahlte Preise liegen laut OpenRouter durch Caching und Rabatte oft darunter.

Vor dem Einsatz prüfen

- ['Von Muse Spark (Meta Model API)', 'Die API-Versionen muse-spark-1.1 bis 1.3 und Muse Spark Open teilen Familie und Kontextgröße, aber nicht das Bereitstellungsmodell: Es gibt keinen API-Endpoint für die Open-Version. Ein Wechsel bedeutet Umstieg auf Self-Hosting oder einen Hosted-Provider, eigene Verantwortung für Guardrails und Prompt-Disziplin bei den empfohlenen Inferenzparametern (temperature 1.0, top_p 1.0, reasoning_effort high/xhigh für komplexe Aufgaben).']
- ['Von anderen Open-Weight-Modellen', 'Beim Wechsel etwa von GLM-, Kimi- oder Qwen-Offenmodellen sind drei Punkte zu prüfen: die Muse Model License mit ihrer Attribution- und MaaS-Klausel ersetzt die bisherige Lizenz (Derivate erben die Auflagen), das Serving-Setup profitiert von den mitgelieferten Helion-Kernels in SGLang, und der Reasoning-Aufwand wird über reasoning_effort statt anbieterspezifischer Denk-Schalter gesteuert.']

Häufige Fragen zu Muse Spark Open

Ist Muse Spark Open Open Source?

Nein im OSI-Sinn. Meta veröffentlicht die Gewichte unter der eigenen Muse Model License, die Nutzung, Modifikation, Fine-Tuning, Destillation und Weitervertrieb erlaubt, aber Auflagen enthält: sichtbare Attribution 'Built with Muse Spark Open', eine Umsatzschwelle mit Separatvertrag für große Model-as-a-Service-Angebote und eine verbindliche Usage Policy. Die Lizenz ist nicht bei der OSI gelistet; korrekt ist die Bezeichnung Open Weight.

Kann ich Muse Spark Open kommerziell nutzen?

Ja, in weitem Umfang. Die Lizenz erlaubt kommerzielle Nutzung inklusive Verkauf und Unterlizenzierung. Erst wenn Sie das Modell selbst als Model as a Service anbieten und mit Affiliates über 20 Mio. USD Bruttoumsatz in zwölf Monaten liegen, brauchen Sie eine separate schriftliche Vereinbarung mit Meta. Interne Nutzung, Forschung und eigene Produkte ohne MaaS-Charakter sind von dieser Schwelle ausgenommen. Die Usage Policy mit Verbotsliste gilt immer.

Gibt es eine API von Meta für Muse Spark Open?

Nein. Laut Meta-Modellkatalog läuft Muse Spark Open nicht über die Meta Model API. Sie hosten selbst (SGLang, vLLM, TensorRT-LLM, llama.cpp, MLX und weitere) oder nutzen Hosted-Inference-Provider wie Baseten, Fireworks AI, Groq, Modal, Nebius, OpenRouter oder Together AI.

Was kostet Muse Spark Open?

Beim Self-Hosting nur die Hardware: Meta nennt einen 4-GPU-Blackwell-Knoten für NVFP4 beziehungsweise rund 2 TB VRAM für BF16. Über OpenRouter-Provider lag der Listenpreis am 31.10.2026 bei 1,25 USD pro Million Input-Tokens und 4,25 USD pro Million Output-Tokens (Cache-Read 0,15 USD). Es gibt keinen offiziellen Meta-Preis, weil Meta das Modell nicht selbst hostet.

Kann Muse Spark Open auf einem Mac laufen?

Laut Meta ja: das Modell wurde für zwei Apple Mac Studios mit M5 Ultra über das MLX-Framework optimiert (je bis zu 512 GB Unified Memory, zusammen etwa 1 TB). Meta beschreibt agentische Workflows komplett on-device, etwa Codeänderungen an einer iOS-App über Muse Code. Das ist eine Herstellerangabe, kein unabhängiger Test; ein einzelner Mac Studio reicht nach Metas Darstellung wegen des Speicherbedarfs nicht.

Wie groß ist das Kontextfenster?

1.048.576 Tokens (eine Million) bei der Inferenz, laut Meta-Doku und Model Card. Ein offizielles Maximum für die Ausgabelänge ist nicht belegt.

Darf ich Muse Spark Open distillieren oder fine-tunen?

Ja. Die Muse Model License schließt Modifikation, Fine-Tuning, abgeleitete Werke, Destillation und Merge ausdrücklich in die Rechtegewährung ein. Für Derivate gelten dieselben Auflagen, insbesondere die Attribution 'Built with Muse Spark Open', und Meta weist darauf hin, dass Modifikationen die getestete Sicherheitslage verändern können.

Gibt es regionale Einschränkungen für die Nutzung?

Im Lizenztext selbst finden sich keine Regionsausschlüsse; Vertragspartner ist Meta Platforms Ireland Limited für die EU-Region beziehungsweise Meta Platforms, Inc. außerhalb. Die verbindliche Usage Policy enthält jedoch Verbotskategorien mit Bezug zu US-Regulierungen (etwa ITAR), und die Einhaltung anwendbaren Rechts einschließlich Handelsrecht liegt beim Nutzer.

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

Quellen & Stand

1. Introducing Muse Spark Open, The Most Capable US Open-Weight Model to Date (Meta AI Research)
   Geprüft: 2026-10-31
2. meta-models/Muse-Spark-Open-816B-A42B - Model Card inkl. Muse Model License (Hugging Face)
   Geprüft: 2026-10-31
3. Muse Model License, Version Muse Spark Open (LICENSE.md, Hugging Face)
   Geprüft: 2026-10-31
4. Muse Spark Open - Meta Model API Dokumentation
   Geprüft: 2026-10-31
5. Models - Meta Model API Dokumentation (Modellkatalog)
   Geprüft: 2026-10-31
6. Muse Spark Open Evaluation Methodology (Meta AI Research)
   Geprüft: 2026-10-31
7. Meta: Muse Spark Open - API Pricing & Providers (OpenRouter)
   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/muse-spark-open/
Statischer Export, kein Live-Datenblatt. Preise, Zugriff und Anbieterangaben können sich ändern. Quellen und ihre jeweiligen Abrufstände stehen oben. Impressum · Transparenz