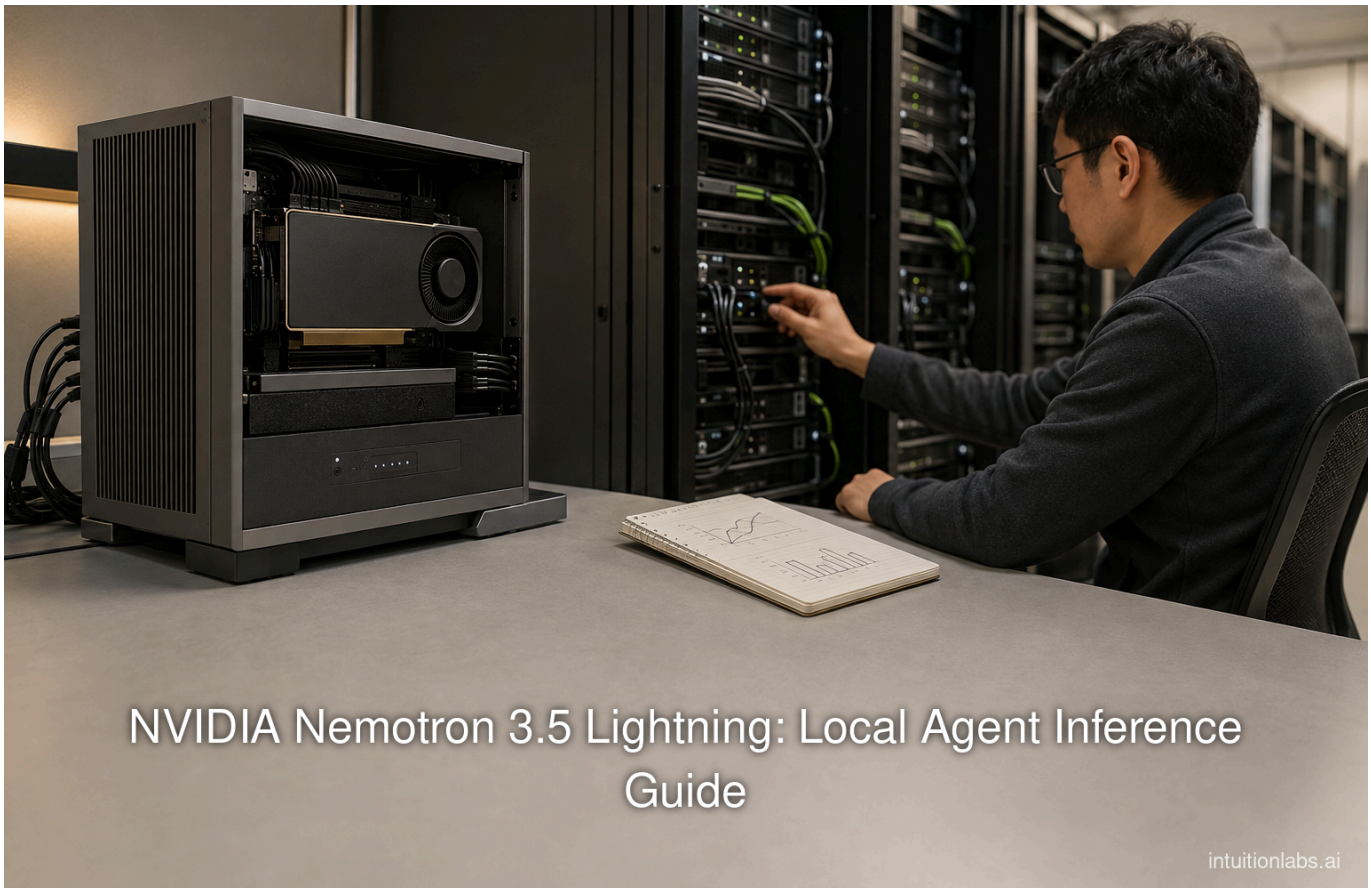
# NVIDIA Nemotron 3.5 Lightning: Local Agent Inference Guide

IntuitionLabs.ai · Adrien Laurent · 2026-09-05 · nemotron 3.5 lightning · nvidia nemotron · local agent inference · on-device ai agents · dgx spark · edge ai inference · nvfp4 quantization · ai inference hardware · nemo switchyard




## Executive Summary

NVIDIA released **Nemotron 3.5 Lightning** on **August 11, 2026**, an open weight, **30 billion parameter** Mixture-of-Experts model that activates only **3 billion parameters** per token and is purpose built to run agent tool-calling and validation steps locally rather than in the cloud (see Product and Platform Architecture below). The model ships in a full precision **BF16** checkpoint (about **66GB**) and a quantized **NVFP4** checkpoint (about **22GB**), the latter supported on a single **DGX Spark (GB10)** or a single **H100** GPU, and more broadly across NVIDIA's Blackwell, Hopper, and Ampere architectures (see Table 1 below). At the consumer level, the 22GB NVFP4 checkpoint is smaller than a single **GeForce RTX 5090's 32GB of GDDR7 memory**, but usable context and runtime memory depend on the serving configuration (nvidia.com), and a DGX Spark desktop unit sized for this workload retails for **$4,699** as of the most recent tech-press pricing observed (notebookcheck.net).

On NVIDIA's own benchmarks, the model scores **81.94 (BF16) / 81.62 (NVFP4)** on MMLU Pro and **75.44 / 75.57** on GPQA Diamond (see Table 2 below), and NVIDIA's launch blog claims **86% accuracy while completing agent tasks 30% faster than Qwen3.6 35B** on its PinchBench agentic benchmark (see Benchmark Performance below). Independent verification tells a more mixed story: the OpenRouter hosting marketplace measured GPQA Diamond accuracy of **63.0% to 68.8%** across third-party providers, seven to twelve points below NVIDIA's self-reported figure (openrouter.ai), and independent testing by Thoughtworks found the model's native speculative decoding delivered **1.46 to 1.96 times the throughput** of unaccelerated decoding, cutting self-hosted cost per million output tokens from **$0.477 to $0.250** (thoughtworks.com).

The model's release is paired with **NeMo Switchyard**, an open source routing library that NVIDIA says can cut agent task cost to nearly one-third of using a frontier cloud model alone by sending only the hardest steps to larger models (blogs.nvidia.com), a design grounded in NVIDIA's own 2025 research finding that most agent subtasks are narrow and repetitive enough for small models to handle reliably (arxiv.org). This local deployment model sits within a broader market shift: **Fortune Business Insights** projects edge inference will account for **70.76% of the global AI inference market in 2026**, itself valued at **$117.80 billion** (fortunebusinessinsights.com), while **Grand View Research** projects the edge AI market overall will grow from **$30.0 billion in 2026 to $118.7 billion by 2033** (grandviewresearch.com).

Nemotron 3.5 Lightning is a distinct product category from data-center accelerators offered by **Cerebras**, **SambaNova**, and **Groq**, which target large-scale training and high-concurrency inference rather than single-GPU local deployment (cerebras.ai). Organizations evaluating it for production agentic workloads should treat NVIDIA's benchmark table as an upper bound, validate independently on their own task distribution, and note that no independently published head-to-head comparison against a frontier cloud model on an agentic benchmark yet exists as of September 2026.

## Introduction and Background

On **August 11, 2026**, NVIDIA published **Nemotron 3.5 Lightning**, an open weight language model built specifically for the "execution layer" of long running AI agents rather than for open ended chat (developer.nvidia.com) (developer.nvidia.com). The model is a **30 billion parameter** Mixture-of-Experts (MoE) design that activates only **3 billion parameters** per token, released alongside a companion routing library called **NeMo Switchyard** that decides, per request, whether a small local model or a larger cloud model should handle a given agent step (developer.nvidia.com) (blogs.nvidia.com).

This report addresses the query "nvidia nemotron 3.5 lightning local agent inference" by answering three practical questions for engineering teams evaluating on-device or on-premises agent deployments as of September 2026: what the model actually is architecturally, what hardware is required to run it locally, and where the published evidence (vendor benchmarks and independent measurements) supports NVIDIA's efficiency claims versus where it does not. IntuitionLabs, a life-sciences and AI consultancy that positions its work around governed, specialist AI implementation for pharmaceutical and biotechnology organizations (intuitionlabs.ai), has previously covered wafer-scale and dataflow inference chips from Cerebras, SambaNova, and Groq in a comparative report (intuitionlabs.ai); this report does not restate that comparison but extends it to a newer question: whether a small, locally hosted model, rather than a specialized accelerator, is the more relevant unit of comparison for agentic workloads in 2026.

The premise behind Nemotron 3.5 Lightning is that most steps inside a long running AI agent (tool calls, format validation, retrieval formatting, subagent delegation) are narrow, repetitive, and do not require a frontier-scale model, a thesis NVIDIA's own research organization argued in a 2025 position paper on small language models (SLMs) for agentic systems (arxiv.org). Nemotron 3.5 Lightning is the productized answer to that thesis: a model small enough to run on a single workstation-class GPU, fast enough to serve interactive agent loops, and paired with a routing mechanism that escalates to larger models only when needed. The sections below examine the architecture, the hardware matrix, the benchmark record (distinguishing NVIDIA's own numbers from independently measured ones), and the broader local inference hardware landscape the model competes and cooperates with.

## Product and Platform Architecture

Nemotron 3.5 Lightning uses what NVIDIA calls a hybrid **Latent Mixture-of-Experts (LatentMoE)** architecture, interleaving Mamba-2 state space layers and MoE layers with a smaller number of standard attention layers (web.archive.org). According to the model's training recipe published on GitHub, the network has **52 layers** with a hidden size of **2,688**, routing tokens across **128 routed experts (top-6 selected)** plus one always-active shared expert (github.com). The model natively supports a context window of **262,144 tokens (256K)**, which NVIDIA's documentation states can be extended further, with a formally listed maximum of **1 million tokens** on the NIM API reference page (web.archive.org) (docs.api.nvidia.com). It is explicitly a **text-only** model: NVIDIA's own documentation states it is "not a multimodal image, audio, or video model," intended for text reasoning, coding, tool calling, and agentic workflows (web.archive.org).

NVIDIA ships the model in two precision formats. The reference **BF16** checkpoint occupies approximately **66GB**; a quantized **NVFP4** checkpoint compresses this to roughly **22GB**, which NVIDIA states delivers "up to 4x output speed" relative to the full precision weights (github.com). NIM deployment documentation puts the effective in-memory model cache footprint at roughly **19GB for the NVFP4 profile up to about 63GB for the BF16 profile**, depending on the serving configuration (web.archive.org). The model is released under the **OpenMDW License Agreement, version 1.1**, an open weights license (huggingface.co).

For latency-sensitive agent loops, the model includes a native **Multi-Token Prediction (MTP)** speculative decoding head that accelerates local inference without a separate draft model. NVIDIA states the model can be served through the standard local inference tooling ecosystem, including **LM Studio, llama.cpp, Ollama, and Unsloth**, rather than only through NVIDIA's own NIM microservices (developer.nvidia.com).

## Local Deployment: Hardware Requirements and Options

The central claim behind "local agent inference" is that Nemotron 3.5 Lightning fits on hardware a single engineer or a single on-premises rack can own outright, rather than requiring multi-GPU cloud clusters. NVIDIA's own model card states the NVFP4 checkpoint supports **single-GPU deployment on either one DGX Spark (GB10) unit or one H100** (huggingface.co), with broader hardware compatibility spanning NVIDIA's **Blackwell** generation (DGX Spark/GB10, GB200, GeForce RTX 5090), **Hopper** (H100, H200), and **Ampere** via W4A16 quantization (huggingface.co). The heavier BF16 reference checkpoint requires more headroom: NVIDIA's NGC catalog listing specifies **one H100 80GB (or one A100 80GB)**, or alternatively **one RTX 5090 via a llama.cpp GGUF Q4_K_M quantization** for consumer hardware (catalog.ngc.nvidia.com). NVIDIA's own deployment guidance also notes a

practical constraint on the GB10-based DGX Spark: because the BF16 profile shares a **unified memory pool with the host system**, operators are advised to reduce GPU memory utilization settings to leave headroom and avoid out-of-memory failures (web.archive.org).

At the system level, the **DGX Spark** desktop unit that NVIDIA positions as a reference platform for this model is built around the GB10 Grace Blackwell Superchip with **128GB of coherent unified LPDDR5x memory** and up to **1 petaFLOP of FP4 compute**, and NVIDIA markets it explicitly as "designed to build and run autonomous agents" locally (nvidia.com) (nvidia.com). Independent tech press coverage puts DGX Spark's retail price at **$4,699**, up from an original **$3,999** at launch, as of that report's observation date (notebookcheck.net). On the consumer side, the **GeForce RTX 5090** carries **32GB of GDDR7 memory**, and available KV-cache headroom is configuration- and context-dependent room for a KV cache, and NVIDIA rates the card at **3,352 AI TOPS** using its fifth-generation Tensor Cores (nvidia.com) (nvidia.com). For genuinely embedded or robotics-adjacent edge deployment, NVIDIA's **Jetson Thor** module offers **128GB of LPDDR5X memory** and up to **2,070 FP4 TFLOPS (sparse)** inside a much smaller power envelope than a desktop GPU (nvidia.com) (nvidia.com).

*Table 1* below summarizes the deployment matrix NVIDIA publishes for the two released checkpoints.

| Checkpoint | Disk size | Minimum single-GPU option | Supported architectures | Notes |
|---|---|---|---|---|
| **NVFP4** (quantized) | ~22GB, cache footprint ~19GB (see above) | 1x DGX Spark (GB10) or 1x H100 (see above) | Blackwell, Hopper, Ampere (W4A16) (see above) | Up to 4x faster throughput than BF16 (see above) |
| **BF16** (reference) | ~66GB, cache footprint ~63GB (see above) | 1x H100 80GB / A100 80GB, or 1x RTX 5090 via llama.cpp GGUF Q4_K_M (catalog.ngc.nvidia.com) | Hopper, Ampere native; RTX 5090 via quantized GGUF | Full precision reference implementation |

The practical read of Table 1 is that the NVFP4 checkpoint, not the BF16 reference weights, is the version most relevant to "local agent inference" as a category: it has published local deployment paths, but runtime configuration and usable context remain serving-dependent; the full-precision weights require data-center-class H100 or A100 accelerators. NVIDIA and its ecosystem partner **NeMo Switchyard** further extend this reach by allowing the same model to run "across PCs, workstations, data centers and the cloud" under one routing layer, rather than requiring a separate deployment recipe for each tier (blogs.nvidia.com), with NVIDIA specifically naming RTX PCs, DGX Spark, DGX Station, and Jetson as supported local systems (blogs.nvidia.com).

## Benchmark Performance: Vendor Claims and Independent Verification

NVIDIA's own model card reports a set of benchmark scores for both checkpoints. On **MMLU Pro**, a broad knowledge and reasoning test, NVIDIA reports **81.94** for BF16 and **81.62** for NVFP4; on **GPQA Diamond** (graduate-level science questions, no external tools), it reports **75.44** BF16 and **75.57** NVFP4 (huggingface.co). On coding and agentic benchmarks specifically, NVIDIA reports **51.56 / 52.80** on SWE-bench Verified, and **85.37 / 83.43** on **PinchBench**, an agentic task-completion benchmark (huggingface.co). NVIDIA's launch blog frames the PinchBench result as **86% accuracy while completing 10,000 agent tasks 30% faster than Qwen3.6 35B** at comparable accuracy, an explicit vendor claim against another open-weight model rather than a proprietary cloud

model (developer.nvidia.com). NVIDIA's documentation adds an important methodological caveat: figures reported by different vendors on the same public benchmarks routinely diverge because each organization runs its own evaluation harness, a distinction this report treats as material rather than boilerplate.

Independent evidence complicates parts of this picture. The inference marketplace **OpenRouter**, which tracks live throughput and accuracy across multiple third-party hosting providers, measured **GPQA Diamond accuracy of 68.8% via its auto-routing pool, 68.5% on CoreWeave, and 63.0 to 63.1% on Venice and DeepInfra**, all materially below NVIDIA's self-reported 75.44 to 75.57 (openrouter.ai). OpenRouter's live provider metrics are operational data that can change over time.

Consulting firm **Thoughtworks**, an NVIDIA early-access testing partner, independently measured the model on **tau-bench**, a multi-turn tool-calling customer-service agent benchmark, at **0.640**, and found the model's built-in speculative decoding head delivered **1.46 to 1.96 times the throughput** of unaccelerated decoding across 2,091 measurements spanning two inference engines and two GPU generations (H200 and B200) (thoughtworks.com) (thoughtworks.com). In their worked cost example, this native acceleration lowered self-hosted inference cost from **$0.477 to $0.250 per million output tokens** (thoughtworks.com). Notably, Thoughtworks also reported that a custom EAGLE-3 speculative draft head they trained specifically for the model's hybrid Mamba-2 architecture could only match, not beat, the vendor-supplied native prediction head (thoughtworks.com), independent evidence that NVIDIA's own acceleration engineering is difficult to improve on with off-the-shelf techniques.

*Table 2* below consolidates the vendor-reported and independently measured figures side by side.

| Benchmark | NVIDIA BF16 | NVIDIA NVFP4 | Independent measurement | Source |
|---|---|---|---|---|
| MMLU Pro | 81.94 | 81.62 | Not independently reproduced in this research | (see above) |
| GPQA Diamond (no tools) | 75.44 | 75.57 | 63.0 to 68.8% across OpenRouter hosting providers | (see above) |
| SWE-bench Verified | 51.56 | 52.80 | Not independently reproduced in this research | (see above) |
| PinchBench (agentic tasks) | 85.37 | 83.43 | 86% accuracy, 30% faster than Qwen3.6 35B (vendor claim) | (see above) |
| tau-bench (customer service agent) | Not published by NVIDIA | Not published by NVIDIA | 0.640 (Thoughtworks) | (see above) |

The interpretive point of Table 2 is not that NVIDIA's figures are wrong, but that they are self-measured on NVIDIA's own harness, and the one broadly comparable independent measurement available (OpenRouter's GPQA Diamond tracking) sits **roughly 7 to 12 points lower** than the vendor number. Teams evaluating this model for production agent workloads should treat NVIDIA's benchmark table as an upper bound reference point and validate on their own task distribution, a standard caution that applies to nearly every vendor-published LLM benchmark, not to this model uniquely. No public **Berkeley Function Calling Leaderboard (BFCL)** score for Nemotron 3.5 Lightning was located during this research as of September 2026, despite the model's training data including function-calling-oriented datasets, which readers should treat as an open evidence gap rather than an implied weakness.

## Local vs. Cloud: When a Small Model Suffices and When to Escalate

The architectural bet behind Nemotron 3.5 Lightning rests on a specific claim about how agent workloads are structured. In a 2025 position paper, NVIDIA's own research group argued that **language models perform a small number of specialized tasks repetitively and with little variation**, and concluded that small language models (SLMs) are "sufficiently powerful, inherently more suitable, and necessarily more economical" than large ones for many agentic invocations (arxiv.org) (arxiv.org). NeMo Switchyard operationalizes that thesis: it decides per-request whether Nemotron 3.5 Lightning or a larger model handles the step, and NVIDIA reports internal benchmarks showing this routing **reduces task completion cost to nearly one-third of using Opus 4.8 alone** while maintaining comparable accuracy, a vendor claim not independently reproduced in this research (blogs.nvidia.com).

Independent cost measurements outside NVIDIA's own materials give a more mixed picture of when local inference actually saves money. A third-party measured comparison of local GPU electricity costs against cloud "Flash-class" APIs found that a smaller local model is not automatically cheaper: in one such test, a **32.8 billion parameter model (DeepSeek-R1-Distill) was the most expensive to run locally, at €1.526 per million tokens on an RTX 3090**, specifically because its lower throughput on that GPU outweighed its smaller size (towardsdatascience.com); the same methodology found other configurations reaching **about $0.60 per million output tokens** (towardsdatascience.com). The lesson generalizes directly to Nemotron 3.5 Lightning's own case: Thoughtworks' independent testing found the model's **cost per million output tokens fell from $0.477 to $0.250** only once its native speculative decoding was correctly configured (see above), meaning the economics of "local agent inference" depend heavily on serving configuration, not solely on parameter count. Separately, a secondary summary of NVIDIA's SLM position paper puts the general order of magnitude at **roughly 10 to 30 times lower cost per token** for small models relative to large ones in agentic settings, a figure attributed to that paper's analysis rather than independently re-derived here (arize.com).

For an on-device or on-premises deployment specifically, latency has a second-order benefit beyond raw token throughput: when the model is configured for local GPU deployment, an agent loop avoids the network round trip to a cloud endpoint, and NVIDIA markets this explicitly as enabling agent workloads to run "fully private" on an RTX-powered PC without cloud dependence (nvidia.com). The practical decision framework this evidence supports is: route narrow, repetitive, well-specified agent steps (structured extraction, routine tool calls, format validation) to a local model like Nemotron 3.5 Lightning, and escalate to a larger cloud-hosted model only for steps requiring broader world knowledge or multi-step planning where the independent GPQA gap documented above becomes material. Advisory teams helping life-sciences and other regulated organizations evaluate this kind of architecture, including consultancies such as IntuitionLabs, generally frame the decision as one of workflow governance and evidence tracking as much as raw model selection (intuitionlabs.ai), since a routing layer that silently escalates or fails to escalate has direct implications for auditability in validated environments.

## Competitive Landscape: Alternative Local and Edge Inference Hardware

Nemotron 3.5 Lightning is a model, not a chip, but its viability as a local agent inference option is inseparable from the hardware it runs on, and NVIDIA is not the only company selling specialized silicon for AI inference. IntuitionLabs' earlier report compared three alternative accelerator vendors, Cerebras, SambaNova, and Groq, each pursuing an architecture distinct from general-purpose GPUs (intuitionlabs.ai). None of these three vendors

targets the single-GPU, desktop-scale local deployment segment Nemotron 3.5 Lightning is built for; they instead compete for data-center-scale training and high-throughput inference, which makes them a useful contrast rather than a direct substitute.

**Cerebras'** third-generation Wafer-Scale Engine (WSE-3) packs **900,000 AI-optimized cores** and delivers **125 petaflops of peak AI performance** on a single wafer-scale die, with systems supporting up to **1.2 petabytes of external memory** (cerebras.ai) (cerebras.ai), and Cerebras claims its software stack requires **97% less code than GPUs** for training large language models (cerebras.ai). **Groq's** Language Processing Unit (LPU) takes the opposite design approach, a single deterministic streaming core rather than a wafer-scale array, with on-chip SRAM bandwidth Groq states is **roughly 10 times a GPU's off-chip HBM bandwidth (about 80 terabytes/second versus about 8 terabytes/second)**, and Groq claims this yields inference that is **up to 10 times more energy efficient** than GPU-based serving (groq.com) (groq.com). **SambaNova**, whose Reconfigurable Dataflow Unit (RDU) architecture targets both training and inference, completed the first close of a **$1 billion Series F round at an $11 billion valuation** in July 2026, and disclosed that JPMorganChase selected its **SN40 and SN50** systems as an on-premises inference infrastructure partner (sambanova.ai) (sambanova.ai), evidence that dedicated accelerators retain a real enterprise on-premises niche distinct from single-GPU local agent deployment.

*Table 3* situates NVIDIA's local inference hardware against these alternative accelerators on the dimensions most relevant to running an agent locally: available memory, headline compute figure, and target deployment scale.

| Platform | Type | Key spec | Typical deployment scale |
|---|---|---|---|
| NVIDIA DGX Spark (GB10) | Desktop AI system, Blackwell | 128GB unified memory, up to 1 PFLOP FP4 (nvidia.com) | Single desk / single developer |
| NVIDIA GeForce RTX 5090 | Consumer GPU | 32GB GDDR7, 3,352 AI TOPS (nvidia.com) | Single workstation |
| NVIDIA Jetson Thor | Embedded module | 128GB LPDDR5X, up to 2,070 FP4 TFLOPS sparse (nvidia.com) | Robotics / embedded edge |
| Cerebras WSE-3 / CS-3 | Wafer-scale training accelerator | 900,000 cores, 125 PFLOPs peak, up to 1.2PB system memory (cerebras.ai) | Data-center-scale training |
| SambaNova SN40 / SN50 (RDU) | Reconfigurable dataflow inference/training | Enterprise on-premises deployment (JPMorganChase) (sambanova.ai) | Enterprise data center |
| Groq LPU | Deterministic streaming inference chip | ~80TB/s on-chip SRAM bandwidth (groq.com) | Cloud-hosted low-latency inference |

The interpretive takeaway from Table 3 is that Nemotron 3.5 Lightning and its NVIDIA hardware targets occupy a distinct tier of the market, single-user or single-team local deployment, from the enterprise data-center accelerators offered by Cerebras, SambaNova, and Groq. A team choosing between these options is not really choosing a competing product for the same job; it is choosing between running a 30B-class agent model on hardware the team already owns and renting or purchasing a specialized accelerator for large-scale training or high-concurrency inference, which are adjacent but different procurement decisions.

## Data Analysis and Evidence

Beyond the model-specific benchmarks above, several third-party market research figures contextualize how significant local and edge inference has become as a deployment category by late 2026. **Grand View Research** estimates the global edge AI market grew to **$30.0 billion in 2026**, up from **$24.9 billion in 2025**, projecting growth to **$118.7 billion by 2033** at a **21.7% compound annual growth rate (CAGR)**, with hardware holding the largest market share at **51.8%** in 2025 (grandviewresearch.com). **ABI Research** separately forecasts the global edge AI chipset market (inference and training combined) will grow from **$34.4 billion in 2026 to $96 billion by 2031**, and specifically projects edge AI GPU chipset shipments will grow at a **31% CAGR** over that period, with the revenue gap between GPUs and second-place ASIC accelerators widening from **$3.2 billion to $19.2 billion** (abiresearch.com) (abiresearch.com), evidence that GPU-based local inference, the category Nemotron 3.5 Lightning is built for, is expected to grow faster than dedicated ASIC alternatives over the same period.

On the inference market specifically, **Fortune Business Insights** values the global AI inference market at **$117.80 billion in 2026**, growing to **$312.64 billion by 2034** at a **12.98% CAGR**, and projects that **edge inference will account for 70.76% of the global market in 2026**, versus centralized cloud inference, with GPUs holding **35.32%** of the inference hardware segment (fortunebusinessinsights.com) (fortunebusinessinsights.com) (fortunebusinessinsights.com). At the broader semiconductor level, **Gartner** forecasts worldwide semiconductor revenue will reach **$1.6 trillion in 2026**, with the AI data-center ecosystem's share of that revenue rising from **36.5% in 2026 to more than 53% by 2030** (gartner.com). NVIDIA's own financial disclosures show the scale it brings to this market: the company reported record **Data Center revenue of $62.3 billion in its fourth fiscal quarter of 2026**, up 75% year over year, and full fiscal year 2026 revenue of **$215.9 billion**, up 65% (nvidianews.nvidia.com).

On quantization efficiency specifically, a directly relevant NVIDIA engineering disclosure for a related, larger Nemotron model shows how much a low-precision format like NVFP4 can compress an already-trained model: NVIDIA reported compressing its **Nemotron 3 Ultra** checkpoint from **1,121GB in BF16 down to 352.3GB**, a **3.2 times reduction**, while an optimized NVFP4 calibration approach retained **98.5% of BF16 accuracy**, compared with 96.8% under a simpler calibration method (developer.nvidia.com) (developer.nvidia.com). NVIDIA further reported that NVFP4-quantized decode-heavy inference on that larger model reached **up to 5.9 times higher throughput** than a comparably sized competing model run in FP4 (developer.nvidia.com). These figures come from a different, larger model in the same Nemotron family, not from Nemotron 3.5 Lightning itself, but they document NVIDIA's general NVFP4 methodology and accuracy-retention track record, which underlies the compression ratio (66GB to 22GB, a 3x reduction) observed directly on Nemotron 3.5 Lightning.

## Implications and Future Directions

The combination of a sub-25GB quantized checkpoint, single-GPU deployment on hardware ranging from a $4,699 desktop unit to a 32GB consumer graphics card, and an open routing layer in NeMo Switchyard signals a broader shift already visible in the market data above: **Fortune Business Insights'** projection that edge inference will represent over 70% of the AI inference market in 2026 suggests that models purpose-built for local execution, rather than only scaled-down versions of cloud models, are becoming a distinct product category rather than a compromise (fortunebusinessinsights.com). For organizations building agentic systems, particularly in regulated industries where data residency, auditability, and latency guarantees matter, a locally hosted model with a

documented routing policy addresses concerns that a purely cloud-based agent stack cannot: every tool call and intermediate decision can, in principle, be logged and inspected on infrastructure the organization physically controls.

At the same time, the evidence gaps identified in this report, the absence of an independently reproduced BFCL score, the roughly 7 to 12 point gap between NVIDIA's self-reported GPQA Diamond score and OpenRouter's measured figures, and the lack of any published head-to-head comparison against a proprietary cloud model on an agentic benchmark, mean organizations should treat Nemotron 3.5 Lightning as promising but not yet fully independently validated for high-stakes agentic use as of September 2026. The **ABI Research** projection that GPU-based edge AI chipsets will outgrow ASIC alternatives at a 31% CAGR through 2031 also suggests NVIDIA's general-purpose GPU and unified-memory approach, rather than a fixed-function accelerator, is likely to remain the dominant substrate for this category of local agent model in the near term (abiresearch.com). Whether NeMo Switchyard's internal claim of reducing task cost to roughly one-third of a frontier cloud model holds up under independent, published measurement is the single most consequential open question for teams evaluating this stack, and is the area most in need of third-party benchmarking as the ecosystem matures.

## Frequently Asked Questions (FAQs)

### What GPU is needed to run Nemotron 3.5 Lightning locally?

As detailed in Table 1 above, the NVFP4 checkpoint runs on a single DGX Spark (GB10) or a single H100, and is broadly compatible with NVIDIA's Blackwell, Hopper, and Ampere architectures. The full precision BF16 checkpoint requires an 80GB H100 or A100, or an RTX 5090 with GGUF quantization via llama.cpp (catalog.ngc.nvidia.com).

### How does Nemotron 3.5 Lightning compare to cloud LLM inference?

As discussed above, NVIDIA claims its NeMo Switchyard routing reduces task completion cost to nearly one-third of using a frontier cloud model alone, though this figure has not been independently reproduced in the sources reviewed for this report, and independent OpenRouter measurements show the model achieving 63 to 69% on GPQA Diamond depending on hosting provider, below NVIDIA's self-reported 75+ score, suggesting cloud frontier models retain an accuracy edge for knowledge-intensive steps even as local models close the gap on routine agent tasks.

### Can Nemotron 3.5 Lightning run on a laptop or edge device?

As covered above, NVIDIA states the model runs on RTX-powered PCs, DGX Spark, DGX Station, and Jetson devices, and it can be served through common local tools including LM Studio, llama.cpp, Ollama, and Unsloth. A dedicated embedded module, Jetson Thor, targets robotics-class edge deployment specifically (nvidia.com).

### Is Nemotron 3.5 Lightning open source?

As noted above, it is released under the OpenMDW License Agreement, version 1.1, an open weights license, with checkpoints published on Hugging Face and NGC.

### How does this compare to specialized inference chips like Cerebras, SambaNova, or Groq?

These accelerators target data-center-scale training or high-concurrency inference rather than single-GPU local deployment, making them adjacent rather than directly competing options; see Table 3 above and IntuitionLabs' dedicated comparison of those three vendors for further detail (intuitionlabs.ai).

## Conclusion

Nemotron 3.5 Lightning is a real, shipped, open-weight product as of its August 11, 2026 release, not an announced-but-unavailable model: it is downloadable in two precision formats, documented for deployment on named single-GPU configurations, and currently listed by OpenRouter as served by two providers, DeepInfra and CoreWeave. Its central proposition, that a 30B-parameter model with only 3B active parameters can handle the bulk of a long-running agent's tool-calling and validation steps locally, while a routing layer escalates harder steps to larger models, is architecturally coherent and grounded in NVIDIA's own published research on small language models for agentic systems. The hardware story includes published local deployment paths, but usable context and runtime memory depend on the serving configuration.

The evidence record is nonetheless mixed rather than uniformly favorable. NVIDIA's own benchmark numbers are strong, but the one broadly available independent measurement, OpenRouter's tracked GPQA Diamond scores, runs meaningfully below the vendor's self-reported figures, and no independent, published comparison against a frontier cloud model on an agentic benchmark yet exists. Organizations evaluating local agent inference on this model should use NVIDIA's published figures as a starting hypothesis, validate independently on their own workload, and pay particular attention to serving configuration, since Thoughtworks' independent testing showed that cost per token nearly halved once native speculative decoding was correctly enabled. As the broader edge and local inference market grows toward the roughly $118.7 billion by 2033 that Grand View Research projects, Nemotron 3.5 Lightning represents one of the first concrete, benchmarked entries in what is likely to become a much more crowded category of purpose-built local agent models.

---