



NVIDIA NeMo

Unleash the power of enterprise-ready, customized generative AI models.

Quickly Overcome Challenges and Unlock the Power of Custom Generative AI

Enterprises are increasingly seeking to leverage generative AI and large language models (LLMs) to gain competitive advantage and drive innovation. However, the adoption of generative AI poses significant challenges, which impedes its integration into existing business processes.

- Leakage of intellectual property (IP): There's a risk of AI inadvertently exposing sensitive company information and IP.
- Brand alignment: Companies need to tailor AI behavior to align with their unique ethics and regulatory standards.
- Secure data access: There's a potential for AI systems to compromise client confidentiality, personally identifiable information (PII), and data security.
- Control and ownership: Companies need to maintain control over AI development and deployment, no matter where they're run.
- Proof of concept (POC) to production: Developers need enterprise-grade software that allows them to go from POC to production.

NVIDIA NeMo™ microservices are specifically designed to overcome the challenges of developing and deploying generative AI.

Get to Market Faster With Ready-to-Deploy Generative AI

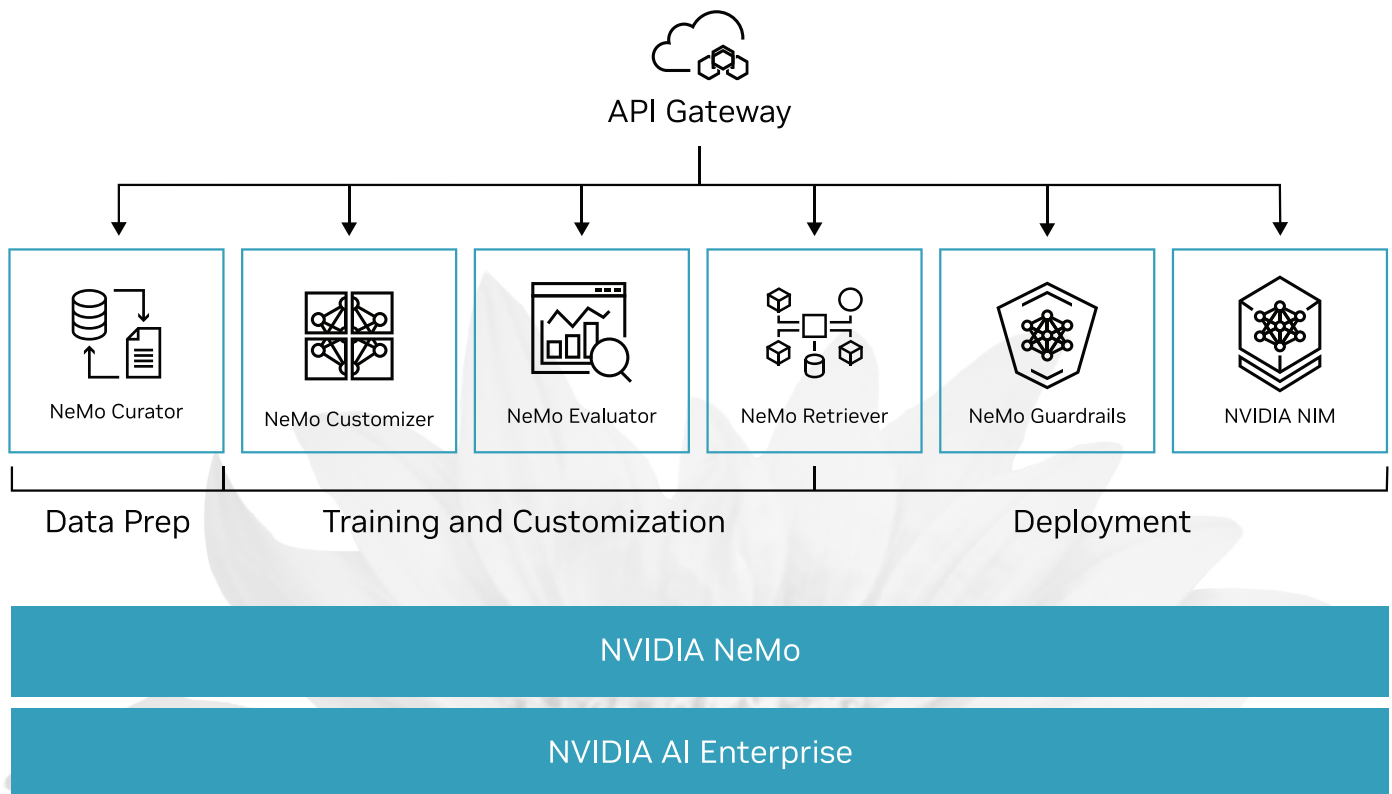
NVIDIA NeMo, a part of NVIDIA AI Enterprise, is an end-to-end platform for the development of custom generative AI. Deliver enterprise-ready models with precise data curation, cutting-edge customization, retrieval-augmented generation (RAG), and accelerated performance.

NeMo Curator is a GPU-accelerated data-curation tool that prepares large-scale, high-quality datasets for pretraining generative AI models. Curator lets enterprises scale to thousands of compute cores and prepare large pretraining datasets in realistic time scales, as well as with state-of-the-art accuracy, by training LLMs with NeMo Curator-prepared tokens.

NVIDIA Cosmos™ tokenizers are open models designed to simplify the development and customization of VLMs and video AI models. They offer high-quality compression and fast, excellent visual reconstruction, lowering TCO during model development and deployments.

NVIDIA NeMo Use Cases

- Generative AI copilot: Enterprises can train LLMs on domain- and brand-specific data to assist humans by automating tedious tasks. Use cases include IT copilots, human resources copilots, and more.
- Generative AI virtual assistant: AI applications, powered by LLMs with access to company-specific data through RAG, can engage in interactive conversations with internal and external users.
- Summarization: Generative AI models can analyze documents, recordings, and videos and generate concise summaries in real time, breaking down complex content into plain language.
- Content generation: Generative AI can create content from text, images, 3D designs, code, and more. By training LLMs on proprietary content, enterprises can align generated content with their brand standards.
- Code generation: Generative AI empowers developers to speed up and optimize code generation.



NeMo Customizer, a high-performance, scalable microservice, simplifies fine-tuning and alignment of LLMs for domain-specific use cases. Customizer provides fine-tuning and alignment techniques, accelerates training performance with model parallelism, and supports fine-tuning of larger models by scaling to multiple GPUs and multiple nodes.

NeMo Evaluator is a cutting-edge automated evaluation microservice that empowers enterprises to assess custom LLMs and RAG. The comprehensive evaluation covers diverse academic and custom benchmarks with predefined metrics, human evaluations on custom datasets through crowdsourcing job templates, and LLM-as-a-judge techniques. Multiple evaluation jobs can be launched via API calls simultaneously on preferred cloud platforms or datacenters, with aggregated results provided in a comprehensive evaluation report.

NeMo Retriever offers a world-class information retrieval with the lowest latency, highest throughput, and maximum data privacy, enabling organizations to use their data better and generate business insights in real time. NeMo Retriever enhances generative AI applications with enterprise-grade RAG capabilities, which can be connected to business data wherever it resides.

NeMo Guardrails orchestrates dialog management, ensuring accuracy, appropriateness, and security in smart applications with LLMs. It safeguards organizations by overseeing generative AI systems.

NVIDIA NIM is a set of easy-to-use microservices designed to accelerate deployment of generative AI across an enterprise. This versatile runtime supports a broad spectrum of AI models—from open-source community models to NVIDIA AI Foundation models and custom AI models. Leveraging industry-standard APIs, developers can quickly build enterprise-grade AI applications with just a few lines of code. Built on robust foundations, including inference engines like NVIDIA Triton™ Inference Server, NVIDIA TensorRT™-LLM, and PyTorch, NIM is engineered to facilitate seamless AI inferencing at scale, ensuring that AI applications can be deployed anywhere with confidence. Whether on premises or in the cloud, NIM is the fastest way to achieve accelerated generative AI inference at scale.

Ready to get started?

Wildflower International Ltd. | 1516 Pacheco St, Santa Fe, NM 87505
+1 (505) 466-9111 | www.wildflowerintl.com

